Towards Trustworthy Modeling of Patient Trajectory with Longitudinal Electronic Health Records

Sihang Zeng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Ruth Etzioni, Chair

Meliha Yetişgen, Chair

Matthew Thompson

Hoifung Poon

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Towards Trustworthy Modeling of Patient Trajectory with Longitudinal Electronic Health Records

Sihang Zeng

Co-Chairs of the Supervisory Committee:
Ruth Etzioni
Fred Hutch Cancer Center

Meliha Yetişgen
University of Washington

Patient trajectory modeling, which predicts future clinical events using data from longitudinal electronic health records (EHRs), is expected to be of value for personalizing disease management. Yet the adoption of powerful deep learning (DL) models is often hindered by their "black-box" nature, creating a barrier to clinical trust. This challenge is compounded when modeling complex temporal dependencies between lab test results, treatments, and clinical events in the EHR, and is further exacerbated by the persistent difficulty these models face in generalizing across diverse patient populations, varying data elements, and different disease states.

This dissertation develops novel interpretable and generalizable frameworks for patient trajectory modeling, motivated by the hypothesis that models that account for the full dynamics of a patient's history will produce more reliable predictions than simpler models, and that these predictions can also be made transparent and robust across diverse clinical contexts. The work is structured around four complementary aims that collectively affirm this hypothesis while innovating in terms of both deep learning methods and interpretable learning tools. It begins by developing an interpretable and generalizable deep learning framework for predicting survival in metastatic prostate cancer from pre-metastasis serial PSA values and

treatments, establishing the value of trajectory-based modeling in a focused clinical setting. Building on this foundation, the work then advances from discrete-time modeling to a more precise continuous-time framework by introducing a model that learns continuous latent trajectories and uses a divide-and-conquer interpretation to explain how clinical changes drive outcomes. To broaden generalizability beyond training task-specific models, the dissertation next develops a multi-agent system that leverages a chain of large language model (LLM) agents with a long-term memory to reason over long, noisy, and heterogeneous EHR data for zero-shot cancer early detection. Finally, the work enables the self-evolving capability of this multi-agent system through an evolving experience pool and multi-agent reinforcement learning for lung cancer early detection, allowing the system to continuously adapt to new patient cohorts.

Through these complementary aims, this research traces an arc from task-specific prediction to generalizable and self-evolving reasoning, powered by DL-based sequential models and LLM-based systems. It shows that faithfully modeling the temporal information in patient histories can make the predictions accurate, robust, and interpretable, with interpretability and generalizability advancing together rather than in tension. In doing so, this dissertation seeks to contribute to the development of more trustworthy AI tools that can support personalized clinical decision-making across a spectrum of complex medical domains.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT): An encoder-only, transformer-based language model to encode texts into embeddings for downstream tasks.

CLINICAL DECISION SUPPORT: Systems designed to assist healthcare professionals by providing relevant, evidence-based clinical insights or recommendations.

ELECTRONIC HEALTH RECORD (EHR): A computer system for storing and managing important patient information documented during clinical care, including but not limited to structured data like diagnoses, laboratory tests, medications, procedures, and vital signs, and unstructured data like clinical notes and radiology reports.

FINE-TUNING: A supervised learning process that trains a pretrained language model by updating its parameters using labeled data.

IN-CONTEXT LEARNING (ICL): A technique where LLMs learn to perform tasks given a few examples provided within the input prompt, without additional parameter updates or fine-tuning.

LANGUAGE MODEL (LM): A computational model that estimates the probability distribution of a sequence of tokens. Learning the statistical patterns of language, it enables advanced natural language processing tasks such as text generation, prediction, translation, and comprehension.

LARGE LANGUAGE MODEL (LLM): An advanced LM characterized by its massive scale up to billions of parameters and trained on internet-scale datasets. This vast scaling unlocks emergent capabilities, allowing the model to perform complex reasoning, generation, and comprehension tasks with minimal or no task-specific training (zero-shot or few-shot learning).

LARGE LANGUAGE MODEL AGENTS (LLM AGENTS): LLMs enhanced with advanced capabilities such as memory construction and retrieval, coding, strategic planning, collaboration with other models, and tool calling.

LATENT TRAJECTORIES: Unobserved, continuous underlying health states or disease progressions inferred by a model from observed, often noisy and irregularly sampled, clinical data.

LLM-BASED MULTI-AGENT SYSTEM: A computational framework composed of multiple interacting LLM agents that collaborate, debate, or divide tasks to solve complex problems, such as reasoning over long and highly heterogeneous EHR data.

LONG-SHORT TERM MEMORY (LSTM): A recurrent neural network (RNN) architecture designed to model temporal sequences and long-range dependencies. By employing specialized gating mechanisms, LSTMs mitigate the vanishing gradient problem inherent in standard RNNs, enabling effective processing of long sequential data.

MULTI-AGENT REINFORCEMENT LEARNING (MARL): A subfield of RL where multiple interacting agents learn to make decisions in a shared environment, used to optimize complex, long-horizon tasks such as step-by-step clinical reasoning.

PATIENT TRAJECTORY MODELING: The computational process of analyzing longitudinal patient histories for downstream tasks such as predicting future clinical events, tracking disease progression, and supporting personalized disease management.

PRETRAINING: The training phase where an LM learns from large-scale unlabeled text corpora using self-supervised objectives to capture language patterns.

PROSTATE-SPECIFIC ANTIGEN (PSA): A protein biomarker often measured via serial blood tests; continuous tracking of PSA values is a crucial component in prostate cancer care.

REINFORCEMENT LEARNING (RL): A machine learning paradigm where an agent learns to make optimal sequential decisions by interacting with an environment. Through a process of trial and error, the agent receives feedback in the form of rewards or penalties and continuously updates its policy to maximize cumulative rewards over time.

REJECTION SAMPLING IN LLM: A technique where a model generates multiple candidate responses, evaluates them (often using a reward model), and retains only the highest-scoring outputs for final selection or further training.

RETRIEVAL-AUGMENTED GENERATION (RAG): A method that enhances LLM generation by incorporating relevant information retrieved from source data (e.g., external knowledge base, raw long-context data) into the input context.

SCALING LAWS: Empirical relationships demonstrating that LLM performance improves predictably in tandem with increases in computational resources, model size, and training data, thereby allowing models to internalize increasingly complex linguistic and semantic representations.

SELF-EVOLVING CAPABILITY: The ability of an AI system to continuously adapt and improve its capabilities over time through techniques like learning from experience or reflection without requiring complete retraining from scratch.

TRANSFORMER: A deep learning architecture that relies entirely on self-attention mechanisms to process sequential data in parallel. By dynamically weighing the contextual importance of all tokens in an input, it serves as the foundational architecture for modern large language models.

WORLD MODEL: A model architecture that builds an internal, abstract representation of the environment to simulate physics and predict the consequences of actions.

ACKNOWLEDGMENTS

The PhD journey has been unforgettable and rewarding, and I owe its completion to many people whose guidance and support I will always treasure.

I would like to express my deepest thanks to my advisors, Dr. Ruth Etzioni and Dr. Meliha Yetisgen. Their steady support and detailed guidance on how to be a good researcher have motivated me throughout this journey and will continue to shape my future career. Their mentorship helped me get through the hardest moments of my doctoral study, and for that I am deeply grateful. I would also like to thank Dr. Steve Zeliadt, who advised my projects at the VA and first introduced me to Ruth, a connection that turned out to be life-changing, as she has become a mentor not only in my research but also in my life. I feel very fortunate to have worked with each of them.

Throughout this journey, I have had the chance to learn from collaborators at Tsinghua University, the University of Washington, Fred Hutchinson Cancer Center, the Department of Veterans Affairs, Harvard University, and Truveta. I am especially thankful to Dr. Kaiyan Zhang, Dr. Matthew Thompson, Dr. Lucas Liu, Lukas Owens, Dr. Scott Coggeshall, Dr. Jun Wen, and Dr. Wilson Lau. These collaborations have strengthened my work at the intersection of AI and healthcare and given me valuable experience from both industry and academia. I would also like to thank Dr. Hoifung Poon and Dr. Noemi Kreif for kindly making time to serve on my committee and for the helpful feedback they provided during my exams.

I am lucky to have shared this journey with wonderful fellow students and labmates, including Zhaoyi Sun, Feng Chen, Luna Li, Zixuan Yu, Sitong Liu, Dr. Sitong Zhou, Dr. Yujuan Velvin Fu, and Wenyu Zeng. Thank you for the mutual support, the shared growth,

and all the time we spent together in Seattle. I am also grateful to my friends, including but not limited to Peiqi Gao, Yiming Lu, Yichen Zhang, Haokun Li, Jialu Gao, and Chuqi Zhang, for the chats, research discussions, and encouragements during my studies.

I feel especially lucky to have met Zihan Xu, whose company and support carried me through the final stretch of my PhD study.

Finally, I would like to thank my family for their unconditional love and support throughout my studies and in every life decision I have made. I could not have completed this work without you.

Chapter 1

INTRODUCTION

1.1 Context and Motivation

Longitudinal electronic health records (EHRs) provide a temporal view of a patient's health, capturing the complex relationships of lab results, treatments, and clinical events over time. A promise of this data is the ability to move beyond static snapshots of health and instead model the full dynamics of disease progression [26]. This is the core goal of patient trajectory modeling: to leverage these rich, temporal dynamics to predict clinical outcomes, personalize treatments, and ultimately improve patient care. In recent years, deep learning models have shown potential for this task, as they are uniquely capable of capturing the complex, non-linear patterns within this data [37, 124, 171, 91]. However, despite their predictive power, the widespread adoption of these advanced models in routine clinical practice is slow, hindered by concerns about their trustworthiness [205, 79]. In this dissertation, we focus on two significant barriers to building this trust: a lack of dynamic interpretability and a failure of generalizability [101].

The first barrier, interpretability, presents a unique challenge in the context of patient trajectories. Clinical decisions are rarely based on a single data point; rather, they rely on dynamic patterns. For example, acute kidney injury is defined by the changes in creatinine [190], and variations in blood biomarkers may indicate risk of diseases [54, 187]. Current explanatory tools are often not built for this [116, 186]. They may identify key features at a single moment but fail to explain how the evolution of a patient's trajectory—-including the features' trends, rates of change, and complex interactions over time—-drives a model's prediction. This is an important gap. We lack the tools to interpret these models dynamically. Furthermore, a truly interpretable model must faithfully represent the process it is model-

ing. Patient trajectories are inherently continuous-time processes, yet they are frequently approximated using summary features or discrete-time models, which may fail to capture the underlying biological dynamics accurately. These unique aspects create a significant challenge: developing models that faithfully emulate a patient's continuous-time biological dynamics and provide clinically meaningful, dynamic interpretations.

The second critical barrier is generalizability. Although the value of dynamic patient information is recognized, their practical superiority over simpler models using static summary features remains unproven and even debated in many settings [79]. The generalizable value of this information across different healthcare systems needs to be shown before their wide adoption. However, due to data heterogeneity across healthcare systems [59]—which differ in protocols, EHRs, and demographics—and temporal drifts, such as new imaging technologies [69], coding changes [128] or changes in patient population, models trained on one cohort often fail to adapt to new populations without complex re-engineering. This fragility often renders traditional external validation infeasible or unreliable [215], highlighting an urgent need for more flexible modeling paradigms and alternative methods to evaluate their generalizable value.

1.2 Research Problems and Contributions

Building on the two barriers identified above, this dissertation frames trustworthy patient trajectory modeling as a set of four concrete research problems. We state each problem explicitly, then present the contribution that addresses it. These problems trace a progression from interpretability in a focused clinical setting, to faithful continuous-time modeling, to generalization across tasks and data elements, and finally to systems that continue to improve as new patients are seen.

Research Problem 1 In a focused clinical setting where dynamic patient information has long been collected but is typically reduced to summary features (e.g., PSA doubling time or nadir PSA in metastatic prostate cancer), it remains unclear whether the full longitudinal

trajectory carries added prognostic value beyond these summaries, and whether that added value is robust across cohorts that differ in protocols, populations, and available data elements.

Contribution 1 We develop TrajSurv-mPC, an interpretable and generalizable deep learning framework for survival prediction in metastatic prostate cancer (mPC) patients using serial prostate-specific antigen (PSA) values and treatment histories. We demonstrate the generalizable added value of dynamic information over summary features on two different cohorts and two sequential model backbones, through both internal and external validation. To make external validation feasible, we created a harmonized subcohort from one institute compatible with the cohort from the other institute. This generalizability of added value is more relevant as available clinical data evolves and may vary across cohorts. We further conduct three interpretation analyses to verify how TrajSurv-mPC dynamically aggregates the temporal dependencies between serial PSA and treatments, thereby leveraging the added value of trajectory information.

Research Problem 2 Patient trajectories are inherently continuous-time biological processes, but most existing trajectory models discretize time and offer only static feature interpretations. This raises the question of how to learn latent representations that evolve continuously in a way that aligns with a patient's clinical progression, and how to interpret such representations so that clinicians can see how dynamic changes of features jointly drive predictions over time.

Contribution 2 To advance from discrete-time modeling to a more precise continuous-time formulation, we develop TrajSurv-Cont, which learns continuous latent trajectories that align with patients' evolving clinical status. We evaluate TrajSurv-Cont on two real-world public EHR datasets through internal and external validation. We also develop a novel end-to-end interpretation approach for TrajSurv-Cont that respects the dynamic and continuous-time nature of clinical progression, explaining how clinical feature changes drive the evolution of

the latent trajectory and link to the outcome. We compare this interpretation approach with existing methods to show its complementary value in explaining patient trajectories.

Research Problem 3 Existing patient trajectory models are typically tailored to a specific task, requiring labor-intensive feature engineering and model training using long, noisy, and heterogeneous EHR data. There is therefore a need for a more flexible modeling paradigm that can reason over raw longitudinal EHRs and generalize to multiple tasks, populations, and data elements, without complex feature engineering or task-specific retraining, while remaining clinically interpretable.

Contribution 3 Moving beyond task-specific models, we develop Traj-CoA, a generalizable multi-agent system for patient trajectory modeling on long and noisy EHR data. We design a simple yet effective preprocessing pipeline to convert EHR data into an LLM-friendly format, and use a chain-of-agents architecture with a long-term memory to model longitudinal records. We evaluate the generalizability and interpretability of Traj-CoA on cancer early detection tasks across 15 cancer types, using data from UWMC and Truveta. Traj-CoA demonstrates superior performance and temporal interpretability consistent with clinical knowledge.

Research Problem 4 Static predictive models, including LLM-based multi-agent systems, do not leverage the experience accumulated as new patients are seen, and therefore may be suboptimal. The open question is how to design a patient trajectory model that continually refines its temporal reasoning from its own experience, mirroring how clinicians refine judgment through accumulated cases, so that interpretability and performance improve together over time.

Contribution 4 We develop Traj-Evolve, a self-evolving multi-agent framework for patient trajectory modeling that learns from its own experience as it processes more patients, continually refining its temporal reasoning and learning from "patients-like-me". This approach enhances performance by actively learning from accumulated experience from evolving

populations. Interpretability is improved as Traj-Evolve mimics clinical practice, where diagnostic judgment is further refined by retrieving similar patients. We evaluate Traj-Evolve on UWMC lung cancer data and analyze its self-evolving dynamics.

1.3 Guide for the Reader

The remainder of this dissertation is organized as follows. Chapter 2 reviews background literature and motivation across the four research problems. Chapter 3 presents TrajSurv-mPC and the added value of full PSA trajectories for metastatic prostate cancer survival prediction. Chapter 4 introduces TrajSurv-Cont and its continuous-time interpretation framework. Chapter 5 develops the Traj-CoA multi-agent system and evaluates it for cancer early detection across UWMC and Truveta. Chapter 6 presents Traj-Evolve and its self-evolving dynamics on lung cancer early detection. Chapter 7 summarizes the key contributions, discusses limitations, and outlines directions for future work.

Chapter 2

BACKGROUND AND MOTIVATIONS

In this chapter, we review background literature that motivated our contributions in trustworthy patient trajectory modeling, including learning from PSA trajectory for progressed metastatic prostate cancer survival prediction in Chapter 3, continuous-time patient trajectory modeling for survival prediction in Chapter 4, a generalizable multi-agent framework for patient trajectory modeling in Chapter 5, and a self-evolving multi-agent system for patient trajectory modeling in Chapter 6.

2.1 Learning from PSA Trajectory for Metastatic Prostate Cancer Survival Prediction

In patients with localized prostate cancer, up to 46.8 per 10,000 person-years will eventually progress to metastatic prostate cancer (mPC) [143]. mPC exhibits varied disease progression and survival [2, 65, 53]. Decision-making about optimal mPC management must account for drivers of this variability, including cancer clinical features, the pace and pattern of recurrence, progression after the primary treatment, and patient characteristics, such as comorbid conditions. Individualizing treatment for mPC is therefore a considerable challenge [2, 139]. Novel metrics, including imaging modalities like PSMA-PET/CT and genomic tests, such as somatic DNA damage-repair assays [119, 60, 69], may be useful in this regard for a subset of patients. Yet, their high cost and limited accessibility present significant barriers to their widespread use [132, 24]. This raises the question of whether the wealth of routinely collected information in patients' clinical records might be leveraged instead of, or in addition to, these metrics, to more accurately and accessibly personalize prognosis.

The potential for serial clinical data to guide disease management has long been under-

stood in clinical practice. Practical algorithms to utilize this data for prognostication and treatment decision-making have been developed across a spectrum of conditions and care settings, such as mortality prediction in intensive care [124] and heart failure prediction [37]. While algorithms and methods continue to evolve, key challenges of generalizability and interpretability persist. Model performance, typically assessed in independent cohorts, often degrades due to identifiable differences in patient populations and data quality [129, 149]. This challenge is compounded when models rely on hand-curated features, whose calculation methods can vary across sites [187] and may limit cross-cohort generalization [129]. While modern deep learning architectures can mitigate these issues by learning high-level features directly from the longitudinal data [163], their inherent complexity often creates a "black box" problem [149]. Without sufficient model interpretation, this lack of transparency remains a significant barrier to clinical trust and adoption, particularly when models must interpret complex and dynamic patient trajectories.

In prostate cancer, longitudinal measurements like serial PSA are used in existing prognostic models, but typically through hand-curated summaries like PSA doubling time (PSADT) or nadir PSA [187, 183, 197]. Whether the full PSA trajectory adds prognostic value remains unclear. In this study, we hypothesized that the full, dynamic trajectory of serial PSA values, when integrated with baseline characteristics and treatment history, offers a richer source of prognostic information over summary features. Notably, our goal is not to develop a single, fixed model that can generalize across cohorts; rather, we demonstrate the generalizable added value derived from incorporating a patient's entire longitudinal PSA history over hand-curated summary features. This type of generalizability is more relevant as available clinical data evolves and may vary across cohorts; for example, some more contemporary cohorts may also include information from novel imaging or genomic data.

To facilitate the understanding of this added value, we present TrajSurv-mPC in Chapter 3. Specifically, we developed a flexible deep learning framework to leverage the temporal dependencies in pre-metastasis patient trajectory, including baseline clinical features, serial PSA values, and treatment histories to predict post-metastasis overall survival (OS). The deep

learning framework can be adapted to different sequential models including long-short term memory (LSTM) and Transformer. We evaluated the added prognostic value through internal and external validation, benchmarking TrajSurv-mPC on two fundamentally different cohorts against models that utilize summary features. To verify how TrajSurv-mPC leverages the temporal information within the trajectory, we designed complementary analyses including a post-hoc hidden state clustering analysis, a proactive simulation study, and a case study.

2.2 Learning Continuous Latent Trajectory for Trustworthy Survival Prediction

Accurate and transparent survival prediction is crucial for trustworthy clinical decision making [11]. Longitudinal EHRs, which capture patients' evolving clinical status through irregularly sampled clinical features, provide rich temporal information for survival prediction [98]. Although deep learning models have been developed to leverage this information and enhance accuracy, two key challenges remain unresolved: accurately modeling continuous clinical progression and transparently linking that progression to survival outcomes [205].

To accurately model the continuous clinical progression, it is important to aggregate irregularly sampled clinical features within a continuous-time framework in a clinically aligned way. Recurrent Deep Survival Machines (RDSM) [130] and Dynamic-DeepHit [98] leveraged recurrent neural networks (RNNs), which aggregate clinical features at discrete times, potentially discarding the underlying continuous-time patterns. Recent continuous-time models, including neural ordinary differential equations (NODEs) [31], ODE-RNN [161], and neural controlled differential equations (NCDEs) [86], map irregularly sampled time series into the continuously evolved latent states in the latent space through differential equations. SurvLatentODE [124] leveraged ODE-RNN to model time-varying clinical features, aggregating them into continuous-time latent states. However, without high-quality supervision signals to guide the latent states at earlier time points, it may learn latent states' trajectories that do not consistently align with the patient's actual clinical progression, even if the ultimate latent state is shown to be clinically relevant. These may lead to suboptimal modeling of clinical progression and suboptimal prediction accuracy.

To transparently link the clinical progression to the outcome, it is essential to provide an end-to-end interpretation of how changes in clinical features, i.e., feature velocities, lead to the outcome. However, existing models mainly interpret the contribution of clinical features' absolute values at observed time points or limit the interpretation to the last latent state without an end-to-end framework. For example, Dynamic-DeepHit offered interpretability through partial dependence-based feature importance, which measures feature importance by calculating the changes in predicted risk when varying the value of a target feature from its minimum to its maximum, temporal attention weights, and predicted risks over time [98]. SurvLatentODE interpreted the last latent state through clustering [124]. Other interpretation tools include SHAP [116] and permutation feature importance [21], which are frequently used for interpreting machine learning models with static features. Despite these efforts, there is still a lack of methods that provide an end-to-end explanation of how feature velocities drive the evolution of the latent trajectory and finally lead to the outcome. This is a critical gap, as clinicians recognize the predictive value of feature velocities in longitudinal data besides their absolute values. For example, creatinine kinetics are used to define acute kidney injury [190].

Among the various model families, the family of neural differential equations (NDEs) offers unique opportunity for continuous-time modeling [31, 86]. NDEs have been widely used in bioinformatics to model the continuous-time evolution of cell states and NDEs' vector field has been used to explain the cellular dynamics through RNA velocity [155, 194]. We hypothesize that this type of model can be used to learn continuous latent trajectories from longitudinal EHR to model patient trajectories, facilitating trustworthy survival prediction. Specifically, we present TrajSurv-Cont in Chapter 4. TrajSurv-Cont is an NDE-based model that learns continuous latent trajectories aligned with the patients' actual clinical progression for more accurate and transparent survival prediction. We evaluated the model through internal and external validation on two large-scale real-world EHR datasets. To facilitate end-to-end interpretation of such a continuous-time model, we designed a divide-and-conquer approach to split the prediction process into feature-to-trajectory and trajectory-to-outcome

processes, thereby explaning how clinical feature changes drive the evolution of the latent trajectory and finally lead to the outcome.

2.3 Generalizable Multi-Agent Framework for Patient Trajectory Modeling

Longitudinal EHRs provide rich, temporal data for modeling patient trajectories and predicting clinical outcomes [77]. Effective temporal reasoning is critical to unlocking this potential [40]. For instance, tracking a lung nodule's evolution is key to diagnosing cancer [134]. Traditional approaches required complex feature engineering and task-specific models, including recurrent neural networks (RNNs) [172, 25, 37, 130, 98, 157], NDEs [124, 10], and encoder-based transformers [107, 159, 94, 145]. Recently, EHR foundation models [160, 200, 171, 229] have demonstrated zero-shot generalizability by pre-training on large structured datasets. However, they are often constrained by limited code sets and short context windows (<16k tokens), preventing them from fully leveraging the rich information in unstructured text or modeling very long patient histories. As a contrast, modern large language models (LLMs) promise a more generalizable, zero-shot paradigm for clinical prediction [166, 174, 226]. Yet, the promise of LLMs is hindered by the unique challenges of EHR data: extremely long patient histories and inherent noisiness of the recorded clinical data [40, 91].

Patient trajectories accumulate multimodal data over years, creating records that often exceed the context windows of LLMs [91]. Even models with large context windows are hampered by the "lost-in-the-middle" problem, where performance degrades on long inputs as they struggle to attend to the middle of a long input sequence [112, 40]. Recent efforts to adapt LLMs for longitudinal EHR [40, 91] have shown that LLMs tend to fail on temporal reasoning over long EHR. While methods like temporal instruction tuning were proposed [40], these efforts have been largely confined to short EHR data or intensive care unit (ICU) data (<16k tokens). Temporal reasoning on very long EHR data over 32k or even 128k tokens remains an unclear challenge.

EHR data are inherently heterogeneous and primarily designed to support clinical care rather than research [87]. Consequently, they often contain noise arising from inconsistent

formats, typographical errors, missing data, and irregular sampling. For many predictive tasks, only a small portion of a patient's record is informative, while abundant irrelevant data can obscure key predictive signals. Early methods sought to address these challenges by enforcing standardized EHR formats [157] or applying extensive feature engineering during preprocessing [172, 25], which are difficult to generalize across diverse healthcare systems. Recently, LLMs offer a more flexible and broadly applicable framework for processing such data [242]. Yet, existing applications of LLMs in the EHR domain remain limited: many are restricted to single data modalities [241], while others struggle to capture temporal dependencies in long and complex patient histories [40, 91]. These gaps highlight the need for a mechanism to isolate relevant events scattered throughout patient histories.

Several strategies exist to manage long-context inputs for LLMs. These include memory-based methods, which utilize an external memory to store and dynamically update the context [34, 232]; retrieval-augmented generation (RAG), which retrieves relevant information based on the query to augment the model's input [58]; and agent-based approaches. Agent-based systems leverage memory, reflection, planning, and inter-agent communication to reason over extended contexts [111]. For instance, frameworks like LongAgent and chain-of-agents (CoA) segment the context into manageable chunks and employ multi-agent collaboration to process them [234, 231], forming a multi-agent system (MAS). Despite their success in the general domain, the application of these long-context modeling methods in healthcare remains limited, primarily confined to question-answering tasks [225, 102, 9]. Consequently, patient trajectory modeling with LLMs, typically prediction tasks, over long and noisy EHR remains an underexplored practical challenge.

We hypothesize that LLMs, with their vast pretrained knowledge and generalizability, hold the promise for a more generalizable and powerful paradigm for patient trajectory modeling, thereby facilitating predictive tasks. In Chapter 5, we present Traj-CoA, a generalizable multi-agent framework consisting of chain-of-agents and a long-term memory for patient trajectory modeling over long and noisy EHR data. We used cancer early detection task as the use case, and evaluated Traj-CoA on data from two organizations, UWMC and Truveta,

with different data elements. Specifically using Truveta data, we evaluated Traj-CoA's multi-cancer early detection ability on 15 cancer types. Traj-CoA achieved comparable performance as trained ML and DL models. We demonstrated Traj-CoA's interpretability and clinical relevance in temporal reasoning through topic modeling, case studies, statistical analyses, and LLM-as-a-judge evaluation.

2.4 Self-Evolving Multi-Agent System for Patient Trajectory Modeling

Extracting and reasoning over clinically relevant signals from long and noisy patient trajectories is challenging, and a substantial body of work has been developed to learn from longitudinal EHR data [36, 37, 107, 159, 89, 141, 221], including LLM-based approaches for generalizable modeling from heterogeneous data [173, 210, 241, 40, 91, 220, 118, 240, 222].

Despite these advances, existing LLM-based longitudinal EHR systems share a fundamental limitation: they are *static*. Every patient is processed in isolation, relying solely on the LLM's frozen parametric knowledge and a fixed prompt. This stands in sharp contrast to expert clinical practice, where diagnostic judgement is continually refined by accumulated experience with similar patients, a process that is central to how clinicians recognise atypical presentations, such as early lung cancer in a never-smoker with an otherwise unremarkable history [51, 140]. For example, for lung cancer early detection, where cases are rare, clinically heterogeneous, and often subtly distinguished from controls by patterns distributed across years of records, the inability of a system to learn from past verified cases can be a material bottleneck to performance and robustness, particularly in minority subgroups such as never-smokers.

Emerging research on self-evolving LLM agents directly addresses this gap. Rather than treating the model as immutable, self-evolving agents continually update their behaviour through interaction and feedback, evolving their memory, prompts, tools, or parameters as new experience accumulates [57, 232]. Memory-based approaches save the problem-solving trajectories as experience into an external database, which could guide future decisions through retrieval-augmented generation (RAG) [170, 233, 203, 238, 180]. In parallel, reinforcement-learning (RL) based approaches such as reward-ranked fine-tuning (RAFT) [45, 206] and

multi-agent RL variants [117, 109, 227] enable collaborative agent systems to internalise successful reasoning patterns directly into their parameters. In healthcare, self-evolving agents have been explored in synthetic or simulated patient interactions [102, 13] and medical question answering [30]. Designing self-evolving systems for patient trajectory modeling in predictive tasks like lung cancer early detection poses a distinct challenge: it requires complex temporal reasoning over years of noisy, multimodal data and the ability to draw actionable insights from rare clinical cases. Existing techniques may not readily applicable for this scenario and performance remains unclear. To our knowledge, no prior work has designed self-evolving agents to enhance longitudinal EHR modeling for high-stakes, real-world clinical predictive tasks.

We hypothesize that incorporating self-evolving designs into patient trajectory models can improve the system's performance as it has seen more patients. In Chapter 6, we present Traj-Evolve, a self-evolving multi-agent framework for patient trajectory modelling that extends our prior Traj-CoA architecture (Chapter 5) with two complementary evolutionary mechanisms, an evolving experience pool (ExPool) and multi-agent reinforcement learning (MARL). Collectively, these mechanisms enable Traj-Evolve to learn from its own experience as it processes more patients, continually refining its temporal reasoning and learning from "patients-like-me", eventually improving performance over time. These two mechanisms transform patient trajectory modelling from a static, isolated prediction task into a continuously improving clinical learning system. To our knowledge, this is the first self-evolving multi-agent framework for longitudinal EHR modeling applied to a real-world clinical prediction task. We demonstrate that Traj-Evolve achieves state-of-the-art discrimination for one-year lung cancer prediction in the overall population and never-smoker population in UWMC data. We further provide detailed analyses of the self-evolving dynamics, supporting the vision of a continuously improving clinical decision-support system.

Chapter 3

TRAJSURV-MPC: LEARNING FROM PSA TRAJECTORIES FOR INTERPRETABLE AND GENERALIZABLE SURVIVAL PREDICTION IN METASTATIC PROSTATE CANCER

This chapter is adapted from a paper titled "TrajSurv-mPC: Learning from PSA Trajectories for Interpretable and Generalizable Survival Prediction in Metastatic Prostate Cancer" authored by Sihang Zeng, Lukas Owens, Lucas J. Liu, Sigrid Carlsson, Jonathan S. Fainberg, Vincent P. Laudone, Martin W. Schoen, Isla P. Garraway, and Ruth Etzioni, under revision at JCO Clinical Cancer Informatics.

In this chapter, we develop TrajSurv-mPC, a deep learning framework that utilizes pre-metastasis patient trajectories to predict post-metastasis overall survival (OS) for metastatic prostate cancer patients progressing from localized disease. To demonstrate its architectural robustness, the framework's sequential modeling was implemented with both LSTM [68] and Transformer [186] backbones. The models utilized baseline characteristics alongside longitudinal PSA and treatment data to predict OS from the time of metastasis detection. To demonstrate the generalizable added value of pre-metastasis PSA trajectory over summary features, we conducted separate internal validations on two largely distinct cohorts (MSK and VA) and performed a cross-cohort external validation by training on a harmonized VA sub-cohort and testing on the MSK cohort. We explicitly addressed model interpretation through three complementary analyses, including post-hoc clustering, a proactive simulation study, and a case study, to confirm the clinical plausibility of TrajSurv-mPC's predictions. Our findings make the case that prognostic algorithms for mPC may, in general, benefit from incorporating longitudinal PSA information via a TrajSurv-mPC-like approach.

3.1 Methods

We were provided access to patient data from two cohorts: Memorial Sloan Kettering (MSK) and Veterans Affairs (VA). The distinct clinical environments permitted rigorous evaluation of TrajSurv-mPC's generalizability. The Central VA Institutional Review Board and Research and Development and the Fred Hutch Cancer Center Institutional Review Board both approved this retrospective cohort study. The Strengthening and Reporting of Observation Studies in Epidemiology (STROBE) reporting guideline was followed.

3.1.1 Study Population

We utilized data from two independent cohorts, MSK (n=403) and VA (n=9,984). The MSK cohort consisted of patients treated with radical prostatectomy (RP) between 2000 and 2022 who subsequently progressed to metastatic disease. Metastasis was ascertained via manual chart review. The VA cohort comprised veterans diagnosed with localized prostate cancer between 2005 and 2019 who developed metastasis, identified via a validated natural language processing algorithm [12]. To ensure sufficient data for trajectory modeling, we excluded patients in either cohort with fewer than three PSA measurements or those developing metastasis within one year of the primary event (RP for MSK; diagnosis for VA). Detailed inclusion criteria and variable definitions are provided in the Appendix A.1.

3.1.2 Data Preparation and Feature Extraction

For TrajSurv-mPC, we extracted baseline clinical features (e.g., age, Gleason score, stage), longitudinal PSA measurements, and treatment histories (radiation therapy (RT) and hormone therapy (HT) for MSK; RP, RT, anti-androgen therapy (AAT), and androgen deprivation therapy (ADT) for VA) prior to metastasis detection. PSA values were log-transformed and interpolated to monthly intervals. logPSA and treatments were used to create sequential patient trajectories. The primary outcome was OS measured from the date of metastasis detection, and the prediction task was to predict post-metastasis OS using pre-metastasis

information. To evaluate the added value of trajectory modeling, we compared TrajSurv-mPC against standard machine learning models trained on baseline features supplemented with summary features derived from the patient trajectory (e.g., PSA doubling time, time to metastasis). Specific preprocessing steps, feature lists, and imputation methods are detailed in the Appendix A.1.

3.1.3 TrajSurv-mPC Framework

TrajSurv-mPC is a flexible deep learning framework designed to process sequential patient trajectories (Figure 3.1). To demonstrate its architectural robustness, the framework's core sequential modeling was implemented using both LSTM [68] and Transformer encoder [186] backbones. In both architectures, a patient's baseline clinical features are first encoded to initialize the sequence, serving as the initial hidden and cell states for the LSTM, or as the initial prefix token for the Transformer. The model then processes the sequence of logPSA and treatment data at monthly intervals. We employed unidirectional LSTM and causal attention masking in the Transformer to emulate real-world patient progression and prevent information leakage from future time steps. To ensure the models capture relevant temporal disease dynamics, we utilized a multi-task learning strategy [230] with two shared auxiliary objectives: predicting the subsequent month's PSA value and the probabilities of treatment initiation. Finally, the sequence representation at the time of metastasis is fed into a non-linear Cox proportional hazards component [38, 84] to predict survival. Detailed model architecture and training procedures are provided in the Appendix A.2.1.

3.1.4 Internal Validation

To evaluate the added value of full trajectory over summary features on distinct clinical settings, we compared TrajSurv-mPC with LSTM and Transformer backbones against four established models: Cox proportional hazard model (CoxPH) [38], random survival forest (RSF) [74], gradient boosting [56], and DeepSurv [84]. Models were developed separately for each cohort using a 70% training, 10% validation (for hyperparameter tuning via grid search

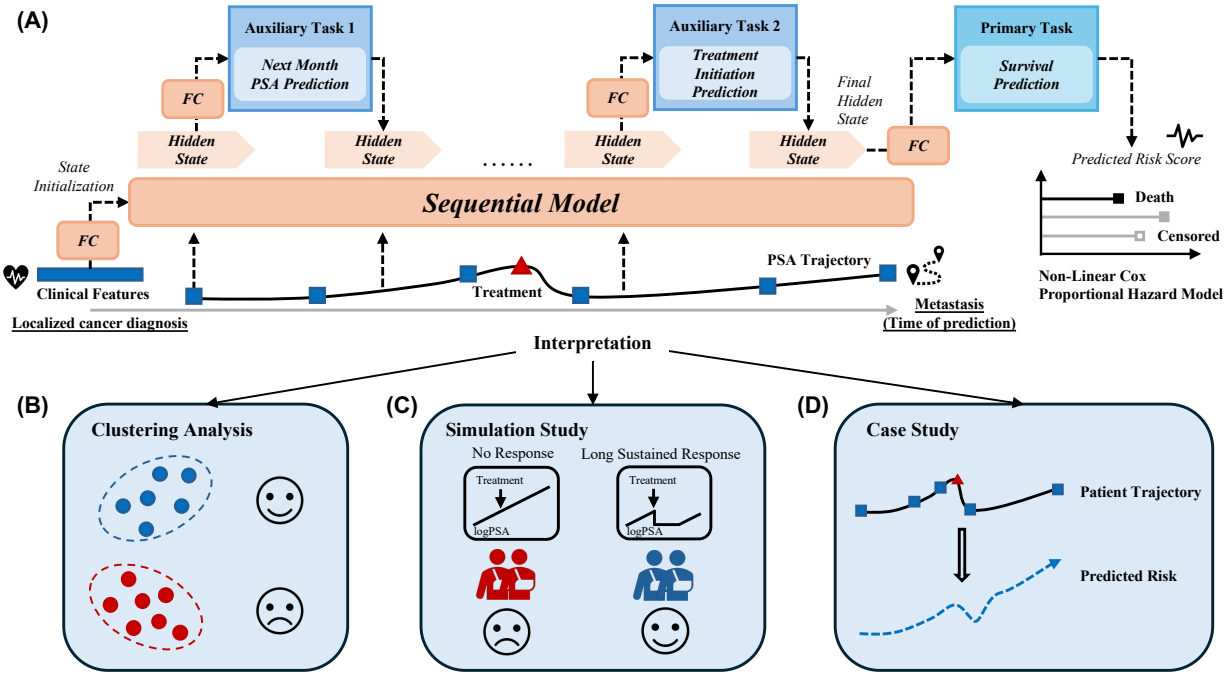

Figure 3.1: An illustration of TrajSurv-mPC framework and its complementary interpretation analyses. (A) TrajSurv-mPC uses a sequential model (e.g., LSTM, Transformer) and multi-task learning strategy to learn clinically relevant survival prediction from baseline clinical features and the patient trajectory; (B) The clustering analysis probes whether the final hidden states represent clinically meaningful patterns and corresponding outcomes; (C) The simulation study validates the learned temporal patterns through predictions across simulated patient groups with different synthetic PSA trajectories; and (D) The case study analyzes the correspondence between changes in the patient trajectory and trends in the predicted risk.

[110]), and 20% testing split. The survival prediction performance was evaluated using the concordance index (C-index), reported as the mean and standard deviation across 10 random splits. To isolate the specific contributions of the auxiliary tasks and baseline features, we performed an ablation study, which is detailed in the Appendix A.3.2.

3.1.5 Cross-Cohort External Validation

To demonstrate the generalizability of TrajSurv-mPC, we conducted a cross-cohort external validation by training on a harmonized VA sub-cohort and validating on the full MSK cohort. We first identified a sub-cohort of patients within the VA cohort (n=587) who had RP within 3 months of diagnosis, matching the post-RP condition of the MSK cohort. Feature sets were harmonized to ensure compatibility: granular VA treatment variables (ADT and AAT) were aggregated into a unified HT variable, and AJCC overall stages were mapped to approximate pathologic T stages (Stages 1-2 $\rightarrow$ T0-T2; Stage 3 $\rightarrow$ T3; Stage 4 $\rightarrow$ T4). Furthermore, PSA trajectories were temporally aligned by defining the date of RP as time zero. Models utilized a common intersection of baseline features (CCI, age, Gleason score, derived T stage, and pre-RP PSA). We created patient trajectories using logPSA, RT, and HT for TrajSurv-mPC, while comparison methods used summary features including time to metastasis, the type of salvage therapy, time to salvage therapy, and PSA values at metastasis and the highest recorded level. This harmonization strategy enabled a direct evaluation of model generalizability across these distinct clinical settings.

3.1.6 Model Interpretation

To verify the clinical relevance of TrajSurv-mPC's predictions, we designed a post-hoc clustering analysis and a proactive simulation study for interpretation. We conducted the interpretation analyses on the MSK data using the best-performing external validation model.

Clustering Analysis

The final hidden state summarizes the patient's pre-metastasis status. To analyze the clinical relevance of this representation, we conducted unsupervised clustering of final hidden states through k-means clustering [66], with the optimal cluster number determined by silhouette analysis. We generated Kaplan-Meier survival curves for each cluster to demonstrate the prognostic stratification capacity of the learned latent representations. We also compared the

clinical feature differences between clusters.

Simulation Study

While deep learning models can achieve high predictive performance, they operate as "black boxes". They are prone to learning spurious correlations or relying on simplistic shortcuts [153, 76], such as baseline features or highest PSA value, while ignoring complex temporal dynamics. Without a dedicated interpretation analysis, it remains unclear whether TrajSurv-mPC has learned the underlying mechanisms of the temporal dependencies between PSA and treatments, or if it is merely relying on these shortcuts. Therefore, we designed a proactive mechanistic validation analysis. We hypothesized that if TrajSurv-mPC's temporal aggregation was functioning correctly, it should assign different risk scores to patients with distinct treatment response patterns, such as sustained response versus non-response, aligned with established clinical knowledge.

We generated five synthetic patient groups (n=1,000 each) emulating distinct post-RP scenarios: (1) No Treatment; (2) No Response (salvage treatment administered but PSA trajectory unchanged); (3) Response and Recurrence (immediate PSA drop after treatment followed by immediate rise); (4) Short-Sustained Response; and (5) Long-Sustained Response. Synthetic patients were created by combining bootstrapped baseline features and parameters from the MSK cohort with mathematically defined PSA and treatment trajectories specific to each scenario (Figure 3.4A). For each group, we obtained the final predicted risk scores of the patients. We then compared the predicted risks across groups to assess whether the differences are clinically relevant, thereby validating our hypothesis. Detailed generation protocols and trajectory equations are provided in the Appendix A.2.2.

3.1.7 Case Study

To analyze how TrajSurv-mPC dynamically aggregates patient trajectories, we selected two representative patients with different PSA trajectories from each cohort. We then computed predicted risk scores at monthly intervals throughout their pre-metastasis period, using the

hidden states at certain times. Finally, we compared the trends of these dynamic predictions against the observed patient trajectories to assess the model's ability to integrate temporal information and update prognostic estimates in clinically relevant ways. TrajSurv-mPC with LSTM trained on each cohort was used in this analysis.

3.1.8 Statistical Analysis

Experiments were performed using Python 3.10. We assessed TrajSurv-mPC's performance improvement over the best comparison model in external validation using bootstrapping with 1,000 iterations. In the clustering analysis, risk stratification was evaluated via log-rank tests, and feature differences were compared using Wilcoxon Rank Sum (continuous), Chi-squared (categorical) tests, and median differences (MD). Simulation group risks were compared using Kruskal-Wallis one-way ANOVA and Wilcoxon Rank Sum test.

3.2 Results

3.2.1 Cohort Description

The MSK (n=403) and VA (n=9,984) cohorts represented clinically distinct populations (Table 3.1). The VA cohort was older, had a higher burden of comorbidities, and experienced higher overall mortality compared to the MSK cohort. Treatment landscapes also differed substantially: the MSK cohort consisted exclusively of post-RP patients receiving salvage therapies, whereas the VA cohort included a diverse mix of primary therapies, with only 15.7% of patients having undergone RP.

3.2.2 Internal Validation Performance

TrajSurv-mPC demonstrated better performance in predicting post-metastasis OS compared to models utilizing summary features across both sequential architectures (Figure 3.2A). In MSK cohort, TrajSurv-mPC achieved C-indices of 0.723 (LSTM) and 0.747 (Transformer), higher than comparison models (C-indices 0.650–0.677). In VA cohort, TrajSurv-mPC yielded

Table 3.1: Patient Characteristics for the MSK and VA Cohorts

Characteristic	**Category**	**MSK (n=403)**	**VA (n=9,984)**
Age (Mean, SD)	–	62.05 (7.42)	66.77 (8.09)
Gleason Grade	6	8 (2.0%)	2,645 (26.5%)
	7	206 (51.1%)	4,044 (40.5%)
	8	59 (14.6%)	1,421 (14.2%)
	9–10	130 (32.3%)	1,661 (16.6%)
	Unknown	0 (0.0%)	213 (2.1%)
T Stage	T0–T2	62 (15.4%)	–
	T3	301 (74.7%)	–
	T4	40 (9.9%)	–
	Unknown	0 (0.0%)	–
AJCC Overall Stage	I–II	–	9,205 (92.2%)
	III	–	317 (3.2%)
	IV	–	48 (0.5%)
	Unknown	–	414 (4.1%)
CCI	0	257 (63.8%)	4,575 (45.8%)
	1	67 (16.6%)	2,396 (24.0%)
	2+	79 (19.6%)	3,013 (30.2%)
Therapy Type*	RP	403 (100.0%)	1,565 (15.7%)
	RT	129 (32.0%)	6,306 (63.2%)
	HT	292 (72.5%)	–
	ADT	–	6,521 (65.3%)
	AAT	–	5,839 (58.5%)
Therapy Time in Years** (Mean, SD)	RP	0.00 (0.00)	1.03 (2.01)
	RT	3.00 (2.53)	1.53 (2.50)
	HT	3.65 (3.58)	–
	ADT	–	2.62 (3.48)
	AAT	–	3.27 (3.66)
Time to Metastasis in Years** (Mean, SD)	–	5.84 (3.97)	6.73 (3.87)
log PSA at Diagnosis (Mean, SD)	–	1.94 (0.72)	1.79 (1.14)
Post-metastasis Follow-up in Years (Mean, SD)	–	5.03 (3.58)	3.12 (2.99)
Death events	–	187 (46.4%)	5,960 (59.7%)

*Patients may have received multiple therapy types.

**Therapy time and time to metastasis measured from diagnosis.

C-indices of 0.711 (LSTM) and 0.754 (Transformer), with comparison models' C-indices ranging from 0.657 to 0.673.

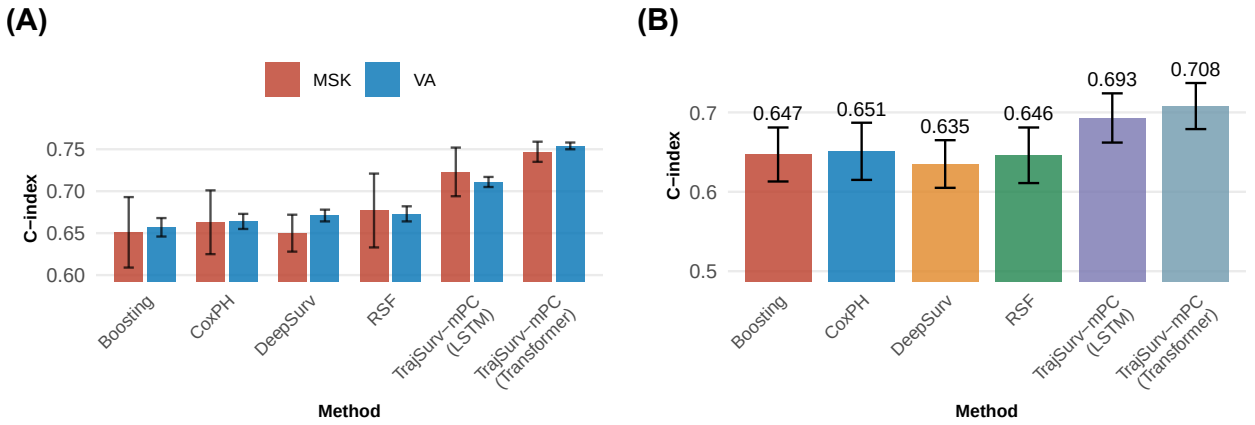

Figure 3.2: Internal and external performance comparison. (A) TrajSurv-mPC achieves a higher C-index compared to models using summary features in separate internal validations on MSK and VA cohort across two different backbones, LSTM and Transformer; (B) TrajSurv-mPC with both backbones achieve higher C-index than comparison methods in the cross-cohort external validation (train on VA post-RP sub-cohort, test on MSK). Error bar shows the standard errors of the performance estimates.

3.2.3 External Validation Performance

In the cross-cohort external validation, TrajSurv-mPC demonstrated improved performance over comparison methods utilizing summary features. TrajSurv-mPC achieved C-indices of 0.693 (LSTM, bootstrapping p=0.02) and 0.708 (Transformer, bootstrapping p=0.008), compared to baseline models (0.635–0.651) (Figure 3.2B), indicating that the granular prognostic signal captured by the full PSA trajectory is robust to cross-institutional heterogeneity and modeling architectures, offering generalizable added value beyond static summary features.

3.2.4 Clustering Analysis

The clustering analysis identified two distinct patient clusters (Figure 3.3A) with significantly different prognoses ($p<0.001$). As shown in Figure 3.3B, Cluster 2 (n=121) was associated with a significantly poorer post-metastasis OS outcome compared to Cluster 1 (n=282). Patients in the high-risk Cluster 2 had a significantly higher logPSA ($p<0.001$, MD=1.99, 1.07 vs 3.06), shorter PSADT ($p=0.002$, MD=-3.17 months, 4.37 vs 1.20 months), larger proportion of Gleason score 9-10 ($p<0.001$, 26% vs 46%) and higher utilization of hormone therapy ($p<0.001$, 57% vs 78%) (Figure 3.3C).

3.2.5 Simulation Study

As shown in Figure 3.4B, the No Response group received the highest median predicted risk score, while the Long-Sustained Response group received the lowest ($p<0.001$). Notably, despite identical PSA progression, the No Response group showed a significantly higher risk than the No Treatment group. These validate our hypothesis that the model correctly learned to associate different treatment response patterns with varying levels of mortality risk, leveraging the temporal dependencies in the patient trajectory rather than purely relying on baseline or trivial features. The fact that it confirmed our expectations increases our confidence that TrajSurv-mPC is learning the correct mechanisms.

3.2.6 Case Study

Figure 3.5 shows that the predicted risk score evolves in response to changes in the PSA trajectory, including increases in PSA and responses to therapy. For example, patient P1 had no response to the therapy, reflected by a sudden increase in predicted risk thereafter. In contrast, patient P2's predicted risk drops after treatment initiation, reflecting the patient's PSA response to treatment. From the VA cohort, patient P3 demonstrates an increasing predicted risk for a persistent PSA despite AAT and ADT, while the predicted risk drops along with the response to RT. Patient P4 has a sudden increase in predicted risk for the limited

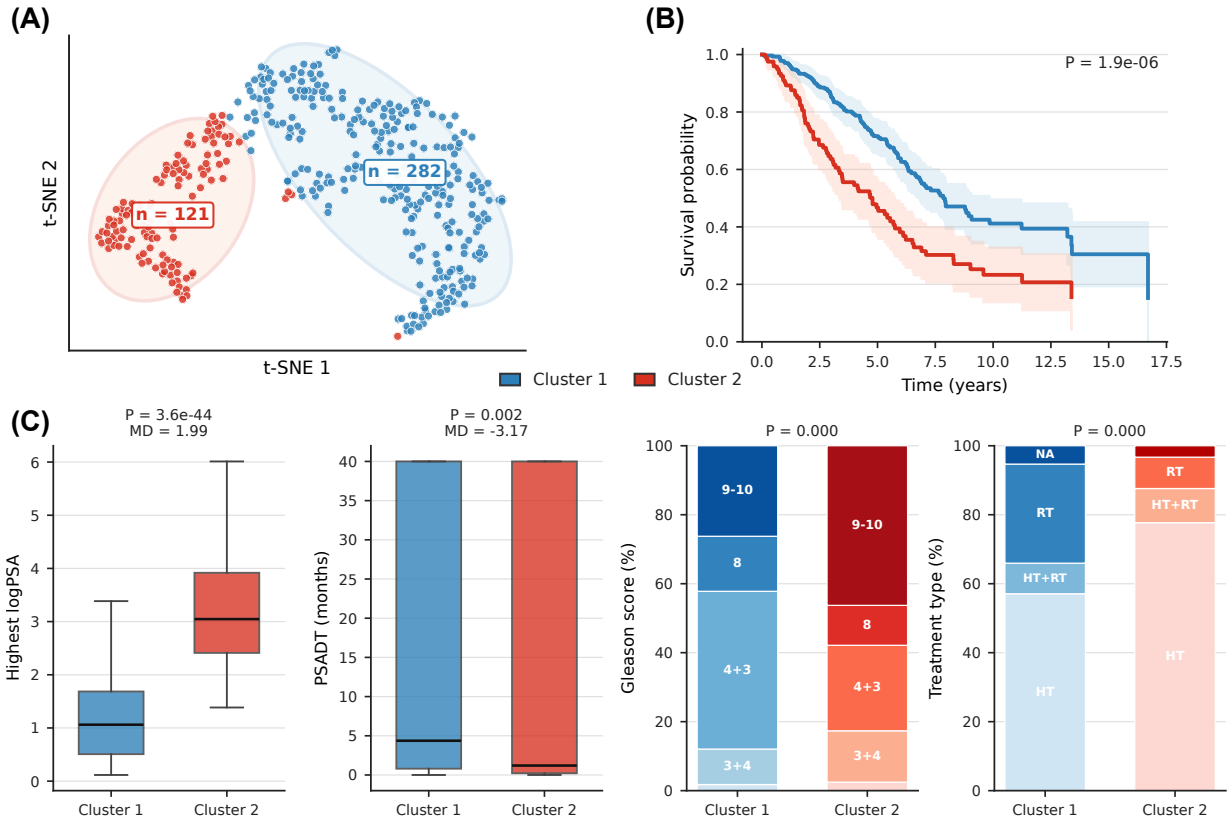

Figure 3.3: Clustering Analysis on the MSK cohort using an external validation TrajSurv-mPC (Transformer) model. (A) The t-SNE plot visualizes the distribution of final hidden states of the two clusters; (B) The KM curve shows statistically significant differences in post-metastasis OS in the two clusters; and (C) Comparisons of clinical summaries between the two clusters with p-values and median differences. PSADT is capped at 40 months for infinite values.

response to AAT and RT. This demonstrates the model's ability to integrate longitudinal data and dynamically update a patient's prognostic assessment in a clinically relevant way.

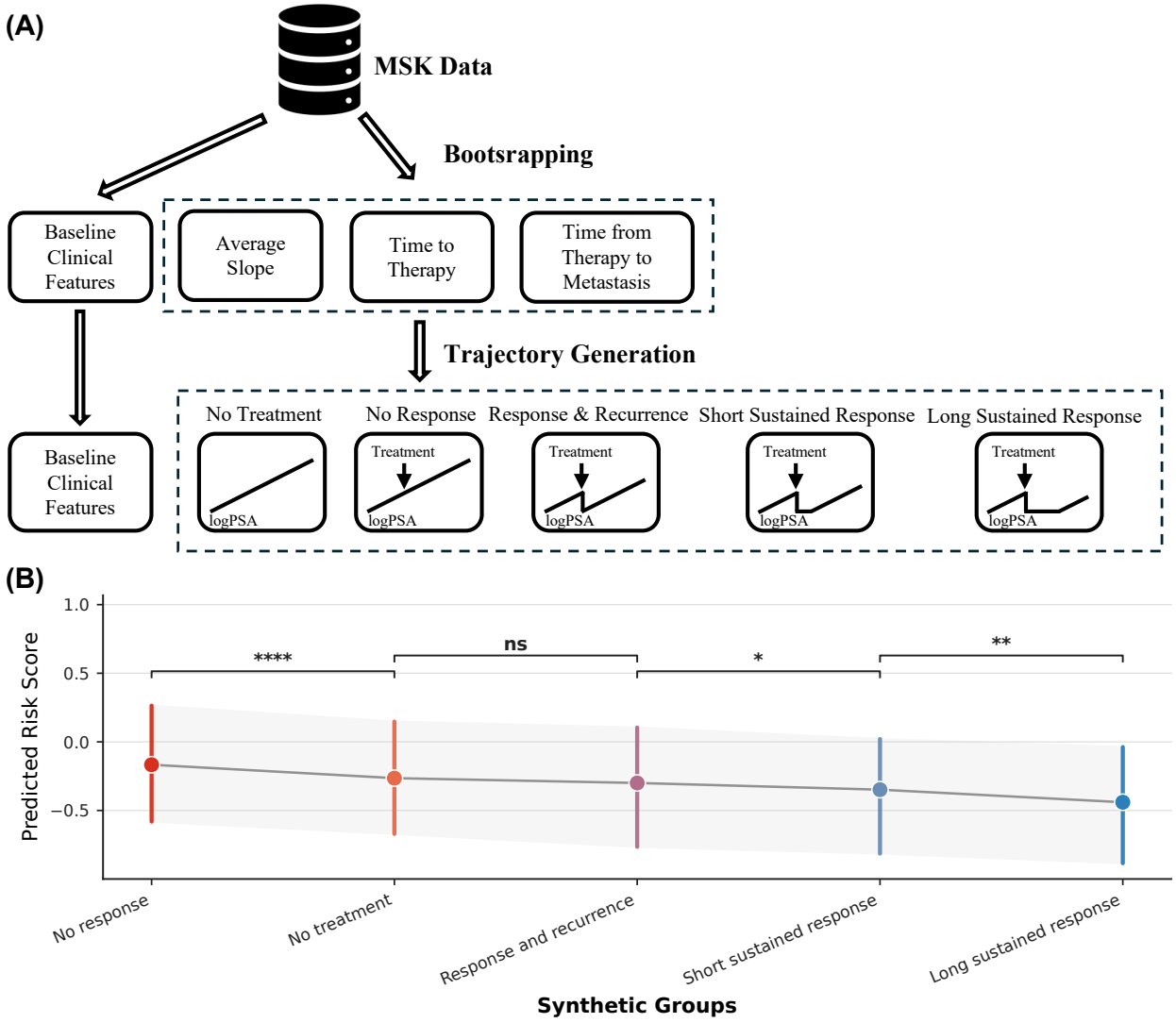

Figure 3.4: Simulation Study to proactively probe whether TrajSurv-mPC's temporal aggregation is clinically relevant using TrajSurv-mPC (Transformer). (A) The sampling process to generate synthetic patients with 5 different treatment response patterns in the study; and (B) The distributions of predicted risk score for OS among different groups, where No Response group has the highest risk and Long-Sustained Response group has the lowest risk. (ns: non-significant, *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$, ****: $p<0.0001$)

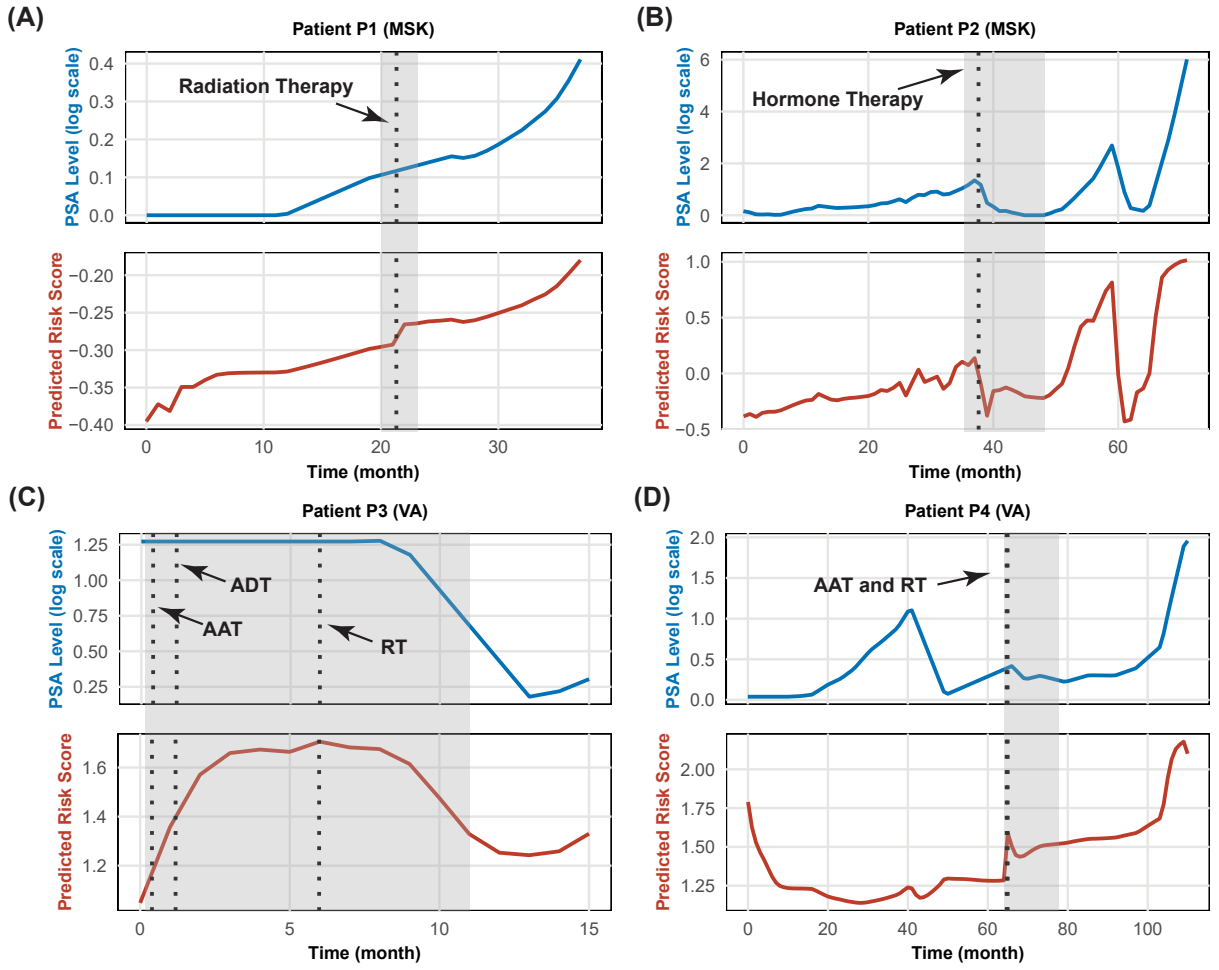

Figure 3.5: Case study displaying the logPSA trajectory with treatments recorded in the data and their corresponding predicted risk score over time from TrajSurv-mPC (LSTM) model. Treatments that are not recorded in the data are unknown and thereby omitted. The predicted risk score at each time is obtained by the hidden state at that time point passing through the FC for survival prediction, representing an estimated risk profile for post-metastasis OS at that time point. The relative changes of the risk score reflect how TrajSurv-mPC dynamically integrates longitudinal patient trajectory over time.

3.3 Conclusion

We developed TrajSurv-mPC, a deep learning framework that predicts post-metastasis OS using pre-metastasis patient trajectories, demonstrating improved prognostic accuracy

over summary-feature models in internal and external validations across distinct cohorts. Complementary analyses revealed its ability to capture clinically relevant temporal patterns.

To assess the generalizability of the PSA trajectory signal, we evaluated TrajSurv-mPC across two fundamentally different healthcare settings: a specialized tertiary referral center (MSK) and a national integrated health system (VA). These cohorts exhibited substantial heterogeneity in demographics, available clinical features, treatment paradigms (post-RP vs. mixed), and metastasis ascertainment methods. By demonstrating improved performance in both separate internal validations and a cross-cohort external validation, we established that the prognostic signal embedded in pre-metastasis PSA dynamics is robust to institutional variability and generalizable across distinct clinical populations.

Our primary objective was to evaluate the added prognostic value of the full PSA trajectory, rather than to establish a singular state-of-the-art architecture. To ensure our findings were not artifacts of a specific architectural choice, we evaluated the TrajSurv-mPC framework using both LSTM and Transformer backbones. Both models improved over comparison methods in internal and external validation, confirming that the prognostic signal is robust and intrinsic to the longitudinal trajectory data itself, regardless of the specific modeling backbone. TrajSurv-mPC with Transformer yielded better performance than with LSTM, consistent with recent studies on its advantages in modeling sequential data [186, 178]. The interpretation analyses further verified that TrajSurv-mPC could leverage the added value of PSA trajectory by learning temporal dependencies between PSA and treatments.

Existing OS prediction models [63, 175, 184, 55, 177, 64] have focused mostly on either de novo metastatic prostate cancer at initial diagnosis or metastatic castration-resistant prostate cancer (mCRPC). These represent fundamentally different disease states in the continuum of care compared to our focus: patients progressing to metastasis from localized disease. Consequently, we were not able to identify existing published models that serve as direct empirical comparators for this specific clinical setting. Instead, we used insights from the broader literature to inform the summary features, such as PSADT [187] and time to metastasis [55], that we included in our comparison models.

Looking forward, we anticipate that integrating PSA trajectories with emerging modalities, such as PSMA-PET/CT [69] and genomic features [23] will further refine prognosis. Our finding that the pre-metastasis PSA trajectory adds value in predicting outcomes in two cohorts with different patient populations and data elements offers some confidence that this will be the case in future settings. In conclusion, TrajSurv-mPC represents a step toward personalized prognostic modeling for metastatic prostate cancer, demonstrating the potential for the added value using data routinely recorded in the longitudinal EHR of prostate cancer patients.

3.4 Limitations

First, while we demonstrated relative improvements of TrajSurv-mPC over comparison models, the study was restricted to a core set of widely accessible clinical features. The moderate performance (C index 0.70-0.75) may likely be improved by incorporating additional laboratory variables with known prognostic value, such as hemoglobin [63, 64], alkaline phosphatase [175, 64], and novel biomarkers. As new prognostic metrics and tools become available [69, 50, 176, 41], we expect that PSA trajectory would provide similar added value from its dynamic information complementary to these static metrics, but this would be indeterminate without rigorous experiments. Second, the model's survival predictions are marginalized over treatments received after metastasis. Although this provides a useful average prognosis under common practice, it limits the ability to predict outcomes for specific treatments. Third, potential inaccuracies in ascertaining metastatic progression may impact the results. Cancer registries primarily document metastatic disease present at diagnosis, but the metastasis detection from subsequent progression is not standardized across healthcare systems and can lead to variability in endpoint capture. While the VA cohort used an NLP approach to ascertain metastasis, the algorithm had high validity (PPV 0.969, NPV 0.944) [12]. While a $\sim$3% misclassification rate could theoretically bias towards longer survival times and attenuate sensitivity, this uncertainty is unlikely to materially affect TrajSurv-mPC's overall prognostic performance. Finally, our model predicts OS using predominantly prostate

cancer-related features. This is an inherently challenging task, as non-cancer-related deaths are common in an elderly population with prostate cancer recurrence. A future model focused on prostate cancer-specific survival might yield predictions more directly attributable to the disease course, but the cause of death was not reliably available in our datasets.

Chapter 4

TRAJSURV-CONT: LEARNING CONTINUOUS LATENT TRAJECTORY FOR TRUSTWORTHY SURVIVAL PREDICTION

This chapter is adapted from a paper titled "TrajSurv: Learning Continuous Latent Trajectories from Electronic Health Records for Trustworthy Survival Prediction" authored by Sihang Zeng, Lucas Jing Liu, Jun Wen, Meliha Yetisgen, Ruth Etzioni, and Gang Luo, published at the Machine Learning for Healthcare Conference 2025. [221]

We introduce TrajSurv-Cont, which learns continuous latent trajectories from longitudinal EHR for trustworthy survival prediction. To ensure the accurate modeling of continuous clinical progression, TrajSurv-Cont uses a neural controlled differential equation (NCDE) [86] to aggregate clinical feature changes over time into continuous latent trajectories, and designs a time-aware contrastive learning (TACL) objective to explicitly align the latent trajectories with actual clinical progression. TACL not only improves the prediction accuracy, but also learns clinically aligned latent trajectories that split the model into feature-to-trajectory and trajectory-to-outcome processes. This enables a divide-and-conquer approach for end-to-end transparency, leveraging a learned vector field and latent trajectory clustering. The implementation of TrajSurv-Cont is available at `https://github.com/zengsihang/TrajSurv`. The contributions of TrajSurv-Cont include:

- **Accurate modeling of continuous clinical progression improves trustworthiness for longitudinal EHR analysis.** With NCDE and TACL, TrajSurv-Cont accurately models the continuous clinical progression, improving prediction accuracy and enabling transparency.

- **Divide-and-conquer approach improves end-to-end model transparency.**

TrajSurv-Cont achieves its end-to-end transparency by splitting the model into feature-to-trajectory and trajectory-to-outcome steps for interpretation.

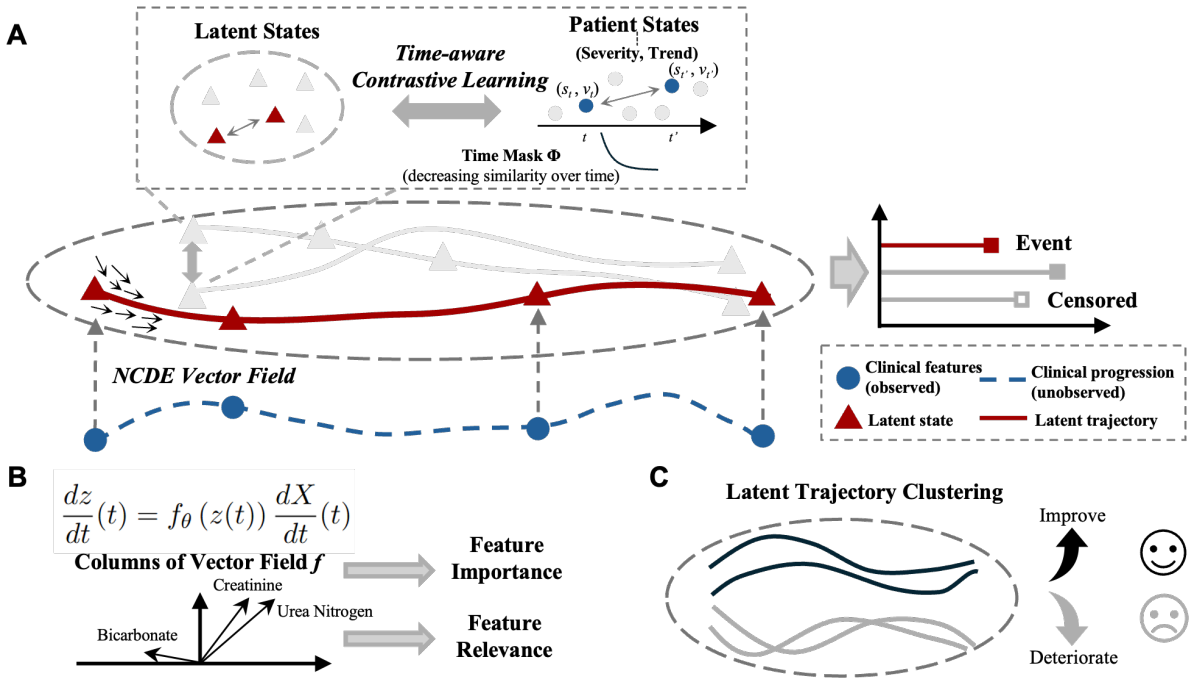

Figure 4.1: An illustration of TrajSurv-Cont and its two-step interpretation. (A) Model architecture. (B) TrajSurv-Cont's vector field interpretation. (C) TrajSurv-Cont's latent trajectory clustering.

4.1 Methods

4.1.1 Problem Formulation

Survival Prediction

We consider a right-censored survival dataset $\mathcal{D} = \{(\mathcal{X}^i, T^i, \delta^i)\}_{i=1}^N$, consisting of N patients with longitudinal EHR data $\mathcal{X}^i$, time-to-event T^i, and event indicator δ^i. For each patient i, the longitudinal EHR data $\mathcal{X}^i = \{(t_j^i, x_j^i)\}_{j=1}^{n_i}$ represents an irregularly sampled time series

of clinical features, where $x_j^i \in \mathbb{R}^d$ is a d-dimensional clinical feature vector at time t_j^i, and $t_0^i = 0$ is defined the same for all patients as a starting time point (e.g., time of admission). $\mathcal{X}^i$ is irregularly sampled, which means that the total number of observations n_i and intervals between observations vary across patients, with clinical features containing missing values. The time-to-event T^i is defined by the time from the last observation $t_{n_i}^i$ to the event of interest (e.g., in-hospital death) or censoring (e.g., discharge). The event indicator $\delta^i = 1$ if the event occurs and $\delta^i = 0$ if censored.

The task of survival prediction is to model the survival distribution given $\mathcal{X}^i$ up to the last observation, predicting the hazard function:

$$\lambda(t|\mathcal{X}^i) = \lim_{\Delta t \to 0} \frac{P(t \leq T^i \leq t + \Delta t | T^i \geq t, \mathcal{X}^i)}{\Delta t} \tag{4.1}$$

or equivalently, the survival function $S(t|\mathcal{X}^i) = P(T^i > t|\mathcal{X}^i)$.

In this study, we aggregate longitudinal EHR data $\mathcal{X}^i$ in a shared latent space $\mathcal{Z}$, where the patient status at time t is encoded as a latent state $z_t^i \in \mathcal{Z}$. The latent state z_t^i summarizes the patient's historical information up to t without knowing the clinical features after t, and the continuous latent trajectory $\tau^i = \{z_t^i, t \in [0, t_{n_i}^i]\}$ summarizes how latent states evolve in $\mathcal{Z}$. We seek to train a clinically aligned latent space $\mathcal{Z}$ for accurate and transparent survival prediction, where the continuous latent trajectory τ^i represents patients' clinical progression. For simplicity, we omit the superscript i in the following sections, except where differentiation between patients is required.

Clinical Alignment

To make the continuous latent trajectory clinically meaningful, we have to align the entire latent space with the clinical meaning of patient states, where a patient state captures the patient's current condition and the trend in the current condition. We define clinical alignment for latent states z_t at a given time point t as the property whereby close latent states correspond to similar patient conditions and their ongoing trends, while distant latent

states correspond to distinct conditions or trends. This definition allows us to assess the local clinical relevance of the latent space at any specific time. However, directly comparing latent states at different time points does not provide a reliable measure of clinical alignment across these points. This is because the distance between z_t and $z_{t'}$ (where $t \neq t'$) could arise from genuine differences in patient states or simply from drift over time. Therefore, we consider entire latent trajectories τ, which inherently encode both time and the evolution of patient states. We define clinical alignment for latent trajectories as the property whereby similar trajectories represent similar evolutions of patient conditions, while dissimilar trajectories represent distinct evolutions. This definition focuses on the overall pattern of change over time, rather than point-to-point comparisons of individual latent states. As the trend information is inherently included in the evolution of patient conditions, it is not used as a separate metric to define clinical alignment for latent trajectories.

To practically capture clinical alignment, we utilize existing clinical assessments of severity as proxies for the patient's condition. These assessments represent established clinical knowledge about patient conditions. Examples of such assessments include the Sequential Organ Failure Assessment (SOFA) score and the Acute Physiology And Chronic Health Evaluation (APACHE) score for intensive care unit (ICU) settings, and the Model for End-Stage Liver Disease (MELD) score for chronic liver disease. We denote the severity at time t as s_t and define the ongoing trend of severity as $v_t = (s_{t+\Delta t} - s_{t-\Delta t})/2\Delta t$, representing the changing rate of severity at t. The choice of Δt is tailored to the specific clinical scenario, ensuring that the severity trend captures meaningful changes in the patient's condition over time, rather than merely reflecting noise from small time intervals. Both severity and its trend can be high-dimensional if the assessment involves multiple components (e.g., different components in the SOFA score).

4.1.2 TrajSurv-Cont: Model Architecture

In this section, we introduce the architecture of TrajSurv-Cont. TrajSurv-Cont uses an NCDE to map irregularly sampled clinical features in longitudinal EHR into continuous latent

trajectories, where the last latent states are linked to the survival outcomes. We further design a time-aware contrastive learning objective (TACL) to align latent trajectories with actual clinical progression. Figure 4.1A shows an illustration of TrajSurv-Cont.

Mapping Longitudinal EHR to Continuous Latent Trajectories

TrajSurv-Cont uses NCDE as an encoder to aggregate temporal information from longitudinal EHR data. NCDE is a latent state-based method that learns continuous-time latent states driven by an irregularly sampled input process [86]. Similar to previous work [165], the use of NCDE in TrajSurv-Cont is motivated by its ability to capture the underlying continuous process from irregularly sampled clinical features, rather than discretely modeling the data (e.g., RNNs and transformers) or modeling through latent trajectories that only depend on initial values (e.g., NODEs).

Specifically, TrajSurv-Cont maps longitudinal EHR $\mathcal{X} = \{(t_j, x_j)\}_{j=0}^{n}$ into a continuous latent trajectory $\tau = \{z_t, t \in [0, t_n]\}$. Following previous practices [86, 125], the longitudinal EHR is interpolated into a continuous control signal $\{X_t, t \in [0, t_n]\}$ through the cubic Hermite splines with backward differences scheme such that $X_{t_j} = (x_j, t_j) \in \mathbb{R}^{d+1}$. The initial observation X_0 is mapped into a d_z-dimensional latent space using a feed forward network (FFN) $g_\phi : \mathbb{R}^{d+1} \to \mathbb{R}^{d_z}$. Subsequent continuous-time latent states are the solutions to an NCDE parameterized by a vector field $f_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z \times (d+1)}$:

$$z_t = z_0 + \int_0^t f_\theta(z_s) dX_s \tag{4.2}$$

where $z_0 = g_\phi(X_0)$, $t \in [0, t_n]$, and the integral is a Riemann-Stieltjes integral. Therefore, the latent trajectory τ is the solution to an NCDE controlled by the underlying process of irregularly sampled clinical features. This transformation is realized through the vector field f_θ, which maps the changes of clinical features dX_t into the evolution of latent states.

Linking Latent Trajectory to Survival Outcome

Because TrajSurv-Cont transforms irregularly sampled clinical features into evolved latent states, the last latent state z_{t_n} accumulates the temporal information up to t_n. Therefore, z_{t_n} is used as an aggregated predictor for the survival outcome. Similar to DeepSurv [84], we use a nonlinear Cox proportional hazard model leveraging an FFN for more accurate prediction. In particular, the hazard function $\lambda(t|\mathcal{X})$ is decomposed into a baseline hazard $\lambda_0(t)$ and the exponential of a risk score r, where r is derived from z_{t_n} through an FFN $G_\eta : \mathbb{R}^{d_z} \to \mathbb{R}$:

$$\lambda(t|\mathcal{X}) = \lambda(t|z_{t_n}) = \lambda_0(t) \cdot r = \lambda_0(t) \cdot \exp\left(G_\eta(z_{t_n})\right) \tag{4.3}$$

We optimize TrajSurv-Cont's survival prediction using two loss functions. The first is the negative partial likelihood loss, commonly applied in Cox proportional hazard models [38]. For any $t \geq 0$, the risk set $R(t)$ is defined as the index set of patients still at risk (i.e., surviving equal or longer than t): $R(t) = \{i|T_i \geq t\}$. The negative log partial likelihood loss $\mathcal{L}_{PL}$ is then formulated as:

$$\mathcal{L}_{PL} = -\frac{1}{N_{\delta=1}} \sum_{i=1}^{N} \mathbb{I}(\delta^i = 1) \log \frac{\exp(r^i)}{\sum_{k \in R(T^i)} \exp(r^k)} \tag{4.4}$$

where $N_{\delta=1}$ is the number of patients who experienced the event and $\mathbb{I}(.)$ is the indicator function. Minimizing $\mathcal{L}_{PL}$ corresponds to maximizing the risk score of patients who had events, relative to those who survived longer. To further improve the concordance of predicted risk scores between patients, similar to previous work [192], we introduce a smoothed pairwise ranking loss $\mathcal{L}_{PR}$ to encourage patients with shorter survival times to have higher risk scores through pairwise ranking, defined as:

$$\mathcal{L}_{PR} = \frac{1}{N_{\delta=1}} \sum_{i=1}^{N} \sum_{k \in R(T^i)} \mathbb{I}(\delta^i = 1) \sigma(r^k - r^i) \tag{4.5}$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function, providing a smoothed and differentiable approximation of the indicator function. The final survival loss $\mathcal{L}_{SURV}$ is the sum of the two

loss functions.

$$\mathcal{L}_{SURV} = \mathcal{L}_{PL} + \mathcal{L}_{PR} \tag{4.6}$$

Clinical Alignment of Latent Trajectory

Previous studies have shown that linking the last latent state z_{t_n} to the survival outcome could learn clinically meaningful last states, where closer states represent similar outcomes [124]. However, because there is no explicit supervision on prior latent states, TrajSurv-Cont's latent trajectory τ is not guaranteed to be clinically aligned, i.e., the proximity of latent trajectories may not correspond to similar clinical progression, though they ultimately arrive at a clinically meaningful distribution. This may affect trustworthiness because the clinical progression information is not accurately and transparently captured. To ensure TrajSurv-Cont's latent trajectories reflect clinical progression, we design a TACL objective.

While aligning two continuous-time processes, the latent trajectory and clinical progression, is challenging, the TACL objective simplifies this alignment by only aligning anchor latent states with high-quality labels while accounting for the time intervals between them. It is inspired by the anchor-based methods that select representative anchor points to enhance computational efficiency and reduce noise [209], facilitating robust alignment. Although the alignment is performed at discrete time points for a latent trajectory, we expect it to propagate across the timeline due to the time-aware design and the continuous-time nature of the latent space, allowing for consistent alignment throughout the entire trajectory.

We use latent states at observation time points as anchor latent states. For simplicity, we denote the set of anchor latent states as $Z = \{z_k\} = \bigcup_{i=1}^{N} \{z_{t_j}^i\}_{j=0}^{n_i}$. To avoid confusion, in the remainder of this section, $|Z|$ will refer to the number of anchor latent states in a single batch, and the subscripts will refer to distinct anchor latent states and their properties.

Specifically, we align anchor latent states z_i across all patients in a batch with their corresponding severity s_i and severity trend v_i through contrastive learning [85]. The contrastive learning guides anchor latent states with similar labels, which are severity and trend in our scenario, to be closer in the latent space. Because both severity and its trend

are continuous variables, we employ a contrastive learning for regression framework called Rank-N-Contrast [223]. This framework learns the latent distance based on the rank of label distance. If we directly adopt Rank-N-Contrast, we have the negative log-likelihood loss:

$$\mathcal{L}_{RNC} = -\frac{1}{|Z|^2}\sum_{i=1}^{|Z|}\sum_{j=1}^{|Z|}\log\frac{\exp(\text{sim}(z_i, z_j)/\kappa_1)}{\sum_{z_k\in\mathcal{S}_{i,j}}\exp(\text{sim}(z_i, z_k)/\kappa_1)} \tag{4.7}$$

where $\mathcal{S}_{i,j} = \{z_k | k \neq i, |s_i - s_k| + \delta|v_i - v_k| > |s_i - s_j| + \delta|v_i - v_j|\}$ is the set of latent states with larger label distance with z_i than the label distance between z_i and z_j, δ is a hyperparameter that balances the severity and its trend, κ_1 is a hyperparameter that controls the sensitivity of contrast, and $\text{sim}(\cdot, \cdot)$ is the similarity between two latent states.

However, this direct adoption does not account for the time intervals between latent states, which may inadequately capture the time-dependent nature of clinical progression. Therefore, we introduce a time mask $\Phi(t_i, t_j; \kappa_2) = \exp(-\frac{|t_i - t_j|}{\kappa_2})$ applied to the log-likelihood term in $\mathcal{L}_{RNC}$, forming our TACL objective. The time mask penalizes the contrasts between latent states separated by larger time intervals. This allows latent states with a large time interval to separate even if they share similar patient states. The hyperparameter κ_2 controls the strength of this penalty and can be adjusted for different clinical scenarios to reflect the time sensitivity of clinical progression. Formally, the TACL objective $\mathcal{L}_{TACL}$ is defined as:

$$\mathcal{L}_{TACL} = -\frac{1}{|Z|^2}\sum_{i=1}^{|Z|}\sum_{j=1}^{|Z|}\Phi(t_i, t_j; \kappa_2)\log\frac{\exp(\text{sim}(z_i, z_j)/\kappa_1)}{\sum_{z_k\in\mathcal{S}_{i,j}}\exp(\text{sim}(z_i, z_k)/\kappa_1)} \tag{4.8}$$

Overall, TrajSurv-Cont's loss function is:

$$\mathcal{L} = \mathcal{L}_{SURV} + \alpha\mathcal{L}_{TACL} \tag{4.9}$$

where α adjusts the relative strength of clinical alignment to survival prediction.

4.1.3 TrajSurv-Cont: Model Interpretation

While the relationship between changes in clinical features and survival outcomes is complex, TrajSurv-Cont's latent trajectory divides and conquers this process into two steps: mapping

longitudinal EHR into continuous latent trajectories and linking latent trajectories to survival outcomes. The transparency can then be achieved by analyzing NCDE's vector field to understand how changes in clinical features lead to the evolution of latent trajectories, and by clustering latent trajectories to identify key clinical progression patterns linked to different survival outcomes. An illustration of model interpretation is shown in Figure 4.1B and C.

Step 1: Vector Field Interpretation

The vector field f_θ maps the changes of clinical features into continuous latent trajectories. To further understand this transformation, we first transform the integral form of the NCDE 4.2 to the derivative form:

$$\frac{dz}{dt}(t) = f_\theta\left(z(t)\right)\frac{dX}{dt}(t) \tag{4.10}$$

For a specific time point t, the vector of latent velocity $\frac{dz}{dt}(t)$ is computed as the product of the matrix $f_\theta(z(t))$ and the vector of clinical feature velocity $\frac{dX}{dt}(t)$. Expanding the matrix f_θ into column vectors, the latent velocity can be represented as the linear combination of the columns of f_θ weighed by the clinical feature velocity. Therefore, each column of f_θ represents the influence of a unit change in a specific clinical feature on the magnitude and direction of the latent state's moving velocity. We analyze the columns of f_θ as follows:

- **Feature Importance:** The magnitude of each column indicates the importance of the corresponding clinical feature in driving latent state evolution. Features with larger column magnitudes have a greater impact on the evolution of the latent state.
- **Feature Relevance:** The cosine similarity between two columns captures the relevance of the corresponding features in driving latent states evolution. If the cosine similarity is high, it suggests that similar changes in these features tend to move the latent state in similar directions. Conversely, if the cosine similarity is low, it indicates that similar changes in these features tend to move the latent state in opposing directions.

This approach provides an elegant and interpretable framework for understanding how changes in clinical features drive latent state evolution and produce the latent trajectory,

through the feature importance and relevance. Practically, instead of analyzing the vector field at specific times for specific patients, we estimate the average vector field $\bar{f}_\theta$ by sampling latent states $z(t)$ across patients and time points and averaging their vector field $f_\theta(z(t))$. This average field provides an overall perspective on feature importance and relevance. We note that these average patterns may not generalize to individual cases and should be interpreted cautiously.

Step 2: Latent Trajectory Clustering

Trained with the TACL objective, TrajSurv-Cont's continuous latent trajectory is expected to be clinically aligned. This means that similar latent trajectories, reflecting the evolution of underlying patient states, can correlate with similar clinical progression and survival outcomes. To investigate this, we employ dynamic time warping (DTW) [127], a time series clustering method that finds optimal alignments between sequences of different lengths and outputs distances. Using DTW, we cluster latent trajectories $(\tau_i, i = 1, ..., N)$ into C distinct groups. We visualize the cluster centroids—derived using the DTW barycenter averaging algorithm—in the latent space. We compute the average severity trajectory over time for each cluster, aiming to reveal key clinical progression patterns. We further assess the association between the clusters and survival outcomes using Kaplan-Meier (KM) curves, hypothesizing that latent trajectory patterns may serve as prognostic indicators.

4.2 Experiments

In this section, we briefly introduce the experiment settings and refer readers to the Appendix for more details.

4.2.1 Dataset and Setup

We evaluated TrajSurv-Cont's survival prediction performance on two real-world EHR datasets in the ICU setting, MIMIC-III [78] and eICU [146]. Similar to prior work [124], the prediction

task is to estimate time to in-hospital mortality based on longitudinal EHR data in the first 36 hours of admission. This prediction task is crucial for timely care and optimal resource allocation in ICU [124].

In ICU, the SOFA score assesses the performance of six organ systems (respiration, coagulation, liver, cardiovascular, neurologic, and renal), with each component ranging from 0 to 4 and higher scores indicating more severe organ dysfunction. The overall SOFA score is the sum of these 6 components and higher scores indicate a higher risk of ICU mortality [168]. In the experiments, we used the SOFA score and its six components as s_t, representing patient conditions over time. SOFA scores s_t were processed hourly from forward-imputed data, and SOFA trends v_t were computed by the SOFA's average changing rate over 4-hour intervals ($\Delta t = 2$) to smooth fluctuations and capture meaningful trends.

4.2.2 MIMIC-III and eICU Cohort

The MIMIC-III cohort included 20,258 hospital admissions and 53 clinical features in 1-hour resolution. The eICU cohort included 116,503 hospital admissions and 53 clinical features in 1-hour resolution. Both cohorts featured irregularly sampled time series data. To ensure computational efficiency, the time series data was pre-processed to only include hourly time points with more than 20 observed features (See Appendix B.1.5 and B.2.1). For both cohorts, t_0 was defined as the time of hospital admission, and the time-to-event T was the time from the last record $(t_n, x_n); t_n \leq 36$ to in-hospital mortality. Patients who survived were right censored at discharge. Clinical features were standardized. Each dataset was randomly split into training (70%), validation (10%), and test sets (20%).

4.2.3 Comparison Methods

We compared the survival prediction performance of TrajSurv-Cont with three machine learning models, one clinical model, and four deep learning models. The machine learning models included the Cox proportional hazards model with elastic net penalty (CoxPH) [38], gradient-boosted Cox proportional hazards model (Boosting) [38, 56], and random survival

forest (RSF) [74]. For the clinical model, we computed the longitudinal overall SOFA scores every 4 hour to create an aggregated 9-dimensional SOFA feature, and applied the Cox proportional hazards model (SOFA) for survival prediction. In the deep learning category, we evaluated models designed for survival prediction using longitudinal EHR data, including RDSM [130], SurvLatent ODE (SLODE) [124], and Dynamic-DeepHit (DDH) [98].

4.2.4 Evaluation Metrics

We evaluated survival prediction performance using three commonly used time-dependent metrics: the concordance index (C-index), the time-dependent Brier score, and the dynamic area under the curve (AUC). The C-index and dynamic AUC measure the model's discrimination ability to distinguish patient risks. The Brier score assesses both the model's discrimination ability and its calibration. Specifically, we implemented the metrics using the *concordance_index_ipcw*, *brier_score*, and *cumulative_dynamic_auc* functions from scikit-survival module [147]. We computed the average across quartiles of follow-up times in the dataset for these metrics. We reported the average performance across 5 runs with different random seeds. Hyperparameters were tuned on the validation set.

For TrajSurv-Cont, we further evaluated the clinical alignment between latent states and patient states by computing Spearman's correlation between latent distance and SOFA or SOFA trend distance at each time point.

4.3 Results

4.3.1 Survival Prediction Performance

Table 4.1 presents the survival prediction performance across comparison models and TrajSurv-Cont. Compared to existing models, TrajSurv-Cont achieved a higher C-index and dynamic AUC and comparable Brier score, indicating improved discrimination and competitive calibration performance. (For more details, see Appendix B.1)

Table 4.1: Survival Prediction Performance Comparison, mean ± std. **Bold** values indicate the best performance among all models or deep learning models for a given metric.

		MIMIC-III			eICU		
	Model	C-index	Brier	AUC	C-index	Brier	AUC
ML	CoxPH	0.706 ± 0.012	0.058 ± 0.002	0.698 ± 0.012	0.751 ± 0.007	0.044 ± 0.001	0.733 ± 0.007
	Boosting	0.766 ± 0.014	0.054 ± 0.002	0.754 ± 0.014	0.781 ± 0.005	0.043 ± 0.001	0.761 ± 0.005
	RSF	0.784 ± 0.013	**0.053 ± 0.002**	0.774 ± 0.014	0.804 ± 0.004	**0.042 ± 0.001**	0.783 ± 0.004
Clinical	SOFA	0.754 ± 0.012	0.053 ± 0.003	0.729 ± 0.013	0.755 ± 0.006	0.044 ± 0.001	0.726 ± 0.006
DL	RDSM	0.772 ± 0.012	0.061 ± 0.002	0.757 ± 0.012	0.796 ± 0.004	0.048 ± 0.001	0.779 ± 0.004
	SLODE	0.762 ± 0.018	0.064 ± 0.004	0.745 ± 0.020	0.780 ± 0.007	0.045 ± 0.000	0.756 ± 0.007
	DDH	0.768 ± 0.014	0.072 ± 0.005	0.748 ± 0.014	0.818 ± 0.005	0.051 ± 0.002	0.796 ± 0.005
	TrajSurv-Cont	**0.803 ± 0.011**	**0.056 ± 0.002**	**0.790 ± 0.013**	**0.823 ± 0.006**	**0.044 ± 0.001**	**0.803 ± 0.006**

Ablation Study To examine how our additional objectives improve the performance over vanilla NCDE, we conducted ablation studies on MIMIC-III by removing the time mask in $\mathcal{L}_{TACL}$ (A1), removing the entire $\mathcal{L}_{TACL}$ (A2), and removing both $\mathcal{L}_{TACL}$ and $\mathcal{L}_{PR}$ (A3). As shown in Figure 4.2A and B, comparing A2 and A3, we found that adding $\mathcal{L}_{PR}$ improves the performance over vanilla NCDE. Comparing A1 and A2, we found that contrastive learning slightly improves the discrimination ability, potentially due to additional supervision for the latent states. TrajSurv-Cont outperforms both A1 and A2, implying the validity of our model design involving the time-dependent nature of clinical progression.

Cross-Cohort Generalization To assess out-of-distribution generalization, we performed a cross-cohort evaluation by training TrajSurv-Cont on the full MIMIC-III training set and testing it on a random sample of 4,000 patients from the eICU dataset. The model achieved a C-index of 0.760 and a Brier score of 0.038, demonstrating robust performance on unseen data from different hospital systems.

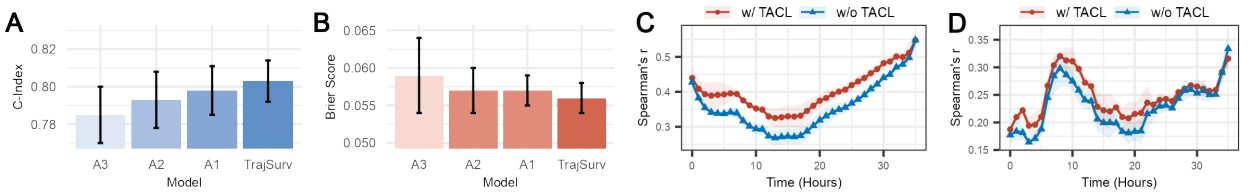

Figure 4.2: Ablation study showing improved (A) C-index and (B) Brier score. Spearman's correlation between latent distance and label distance, (C) SOFA distances, and (D) SOFA trend distances, with or without the TACL objective.

4.3.2 Clinical Alignment Performance

Figure 4.2C and D shows how the correlations change over time with and without TACL on MIMIC-III data. Without TACL, the model has relatively high correlations with SOFA and SOFA trend at the end of the 36-hour period, while they drop significantly in the middle. This suggests that despite the clinical alignment of the last latent state, the evolution of latent states does not align with the clinical progression. Trained with the TACL objective, TrajSurv-Cont retains the correlations at the end, and has significantly higher correlations with SOFA and slightly higher correlations with SOFA trend in the middle of the 36 hours. This suggests that latent states across the time horizon are more clinically aligned with patient states. Therefore, the evolution of latent states better reflects the clinical progression.

4.3.3 Interpretation

In this section, we validate TrajSurv-Cont's two-step interpretation process with known clinical knowledge on the test set of MIMIC-III data at the population level.

Feature Importance and Relevance from TrajSurv-Cont's Vector Field

From the columns of $\bar{f}_\theta$ (Figure 4.3A), we obtained feature importance and feature relevance following the procedure in section 4.1.3. As shown in Figure 4.3B, the top 15 influential

features are ranked by the magnitude of the columns in the average vector field. Among the most important are blood urea nitrogen, total bilirubin, white blood cell count, creatinine, and hematocrit. These findings are consistent with previous ICU studies identifying these as significant indicators of patient states [35, 75]. For instance, blood urea nitrogen, creatinine, hematocrit, lactate, and PH were identified among the top 6 significant clinical features in a previous study [35]. The feature importance derived from the vector field provides a dynamical perspective, suggesting that changes in these features drive latent state shifts more rapidly than others.

Additionally, we plotted a heatmap of cosine similarities between columns of the vector field to reveal specific relationships among features (Figure 4.3C), providing an understanding of feature relevance in driving latent state transitions. For instance, creatinine and urea nitrogen display high positive cosine similarity, suggesting that their similar change patterns result in latent state shifts in nearly identical directions. This alignment is clinically consistent, as both are indicators of kidney function, where elevated levels may indicate renal impairment [71]. In contrast, urea nitrogen and bicarbonate show a high negative cosine similarity, meaning their parallel increases or decreases drive latent states in opposing directions. This finding aligns with evidence that urea nitrogen and bicarbonate are often inversely related in certain conditions [137, 17]. For example, bicarbonate supplementation is associated with reduced blood urea nitrogen levels in chronic renal failure patients [137].

These relationships derived from TrajSurv-Cont's vector field confirm known clinical associations while illustrating how TrajSurv-Cont captures the dynamical interdependencies between features that influence latent state evolution. We found these interpretations to be robust across different random data splits. For instance, blood urea nitrogen, total bilirubin, and white blood cell count were consistently ranked as top-5 important features, and the cosine similarity between creatinine and urea nitrogen remained stable at 0.60 ± 0.06 across three different random seeds. Note that the feature importance and relevance were data-driven and should be interpreted with caution.

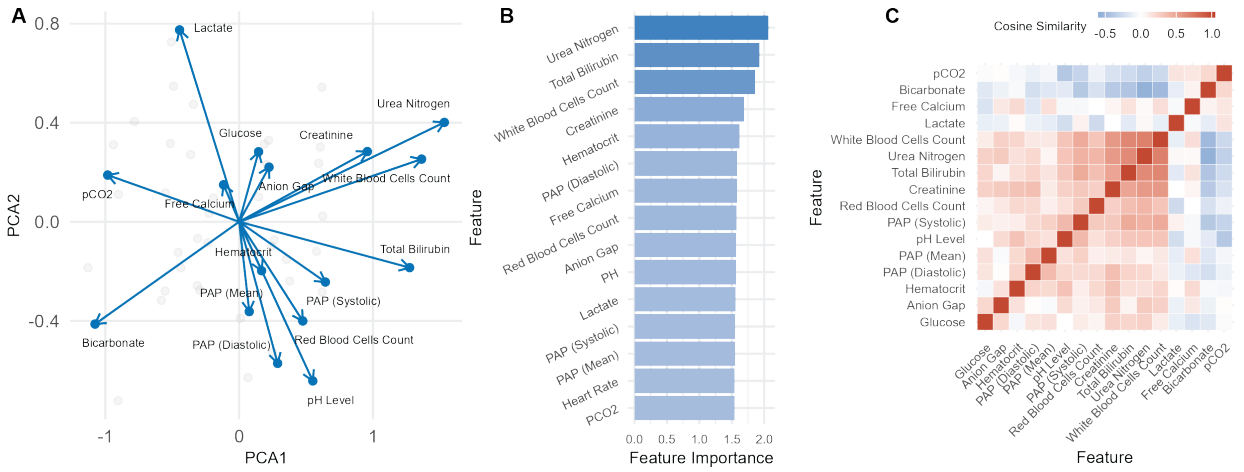

Figure 4.3: Feature importance and relevance derived from TrajSurv-Cont's vector field. (A) Columns of the average vector field; (B) Features with top 15 importance; (C) Cosine similarities between selected features.

Latent Trajectory Clustering

Clustering latent trajectories using DTW, we obtained four clusters. Figure 4.4A shows the centroids of four clusters in the latent space, dividing the space into subdomains. Examining the average SOFA score trajectories (Figure 4.4B) of these clusters reveals four distinct disease progression patterns. ISQI cluster starts with high SOFA scores that decrease rapidly (initially severe, quickly improving - ISQI). ISK cluster maintains a consistently high SOFA score with a slight increase early on, indicating an initially severe condition that gradually worsens (initially severe, keep - ISK). IMK cluster and IMSI cluster exhibit lower SOFA scores, with scores remaining stable over time for the former (initially mild, keep - IMK) and slowly improving for the latter (initially mild, slowly improving - IMSI). These clusters suggest that TrajSurv-Cont's continuous latent trajectories represent different clinical progression patterns in the ICU.

The KM survival curve for each cluster reveals distinct survival patterns aligned with the clinical progression identified in the latent trajectories, hence stratifying risks. The

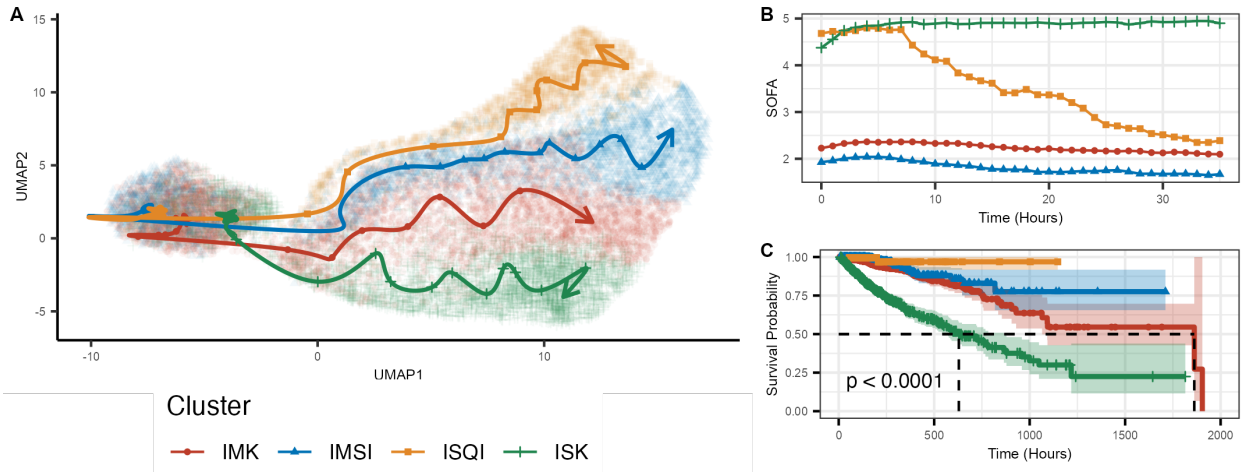

Figure 4.4: Clustering of latent trajectories with (A) the centroids of clusters; (B) the average SOFA trajectories of clusters; and (C) survival curves of clusters.

ISQI subgroup, characterized by severe initial conditions that rapidly improve, shows the highest survival probability over time, suggesting that patients with early improvement after critical onset are less likely to experience adverse outcomes. In contrast, the ISK subgroup, with consistently high severity and a gradually worsening trend, displays the lowest survival probability, indicating a sustained high risk. The IMSI and IMK subgroups, representing milder initial conditions with slow improvement or stability, show intermediate survival outcomes. The statistically significant separation among the KM curves highlights the strength of TrajSurv-Cont's continuous latent trajectories in capturing meaningful clinical progression patterns, effectively stratifying risk based on the evolved patient conditions.

4.3.4 Case Study

A case study of two patients in MIMIC-III demonstrates TrajSurv-Cont's patient-level interpretation of how latent trajectories lead to predictions consistent with actual clinical progression.

For patient P1, TrajSurv-Cont predicts a favorable survival outcome after the last obser-

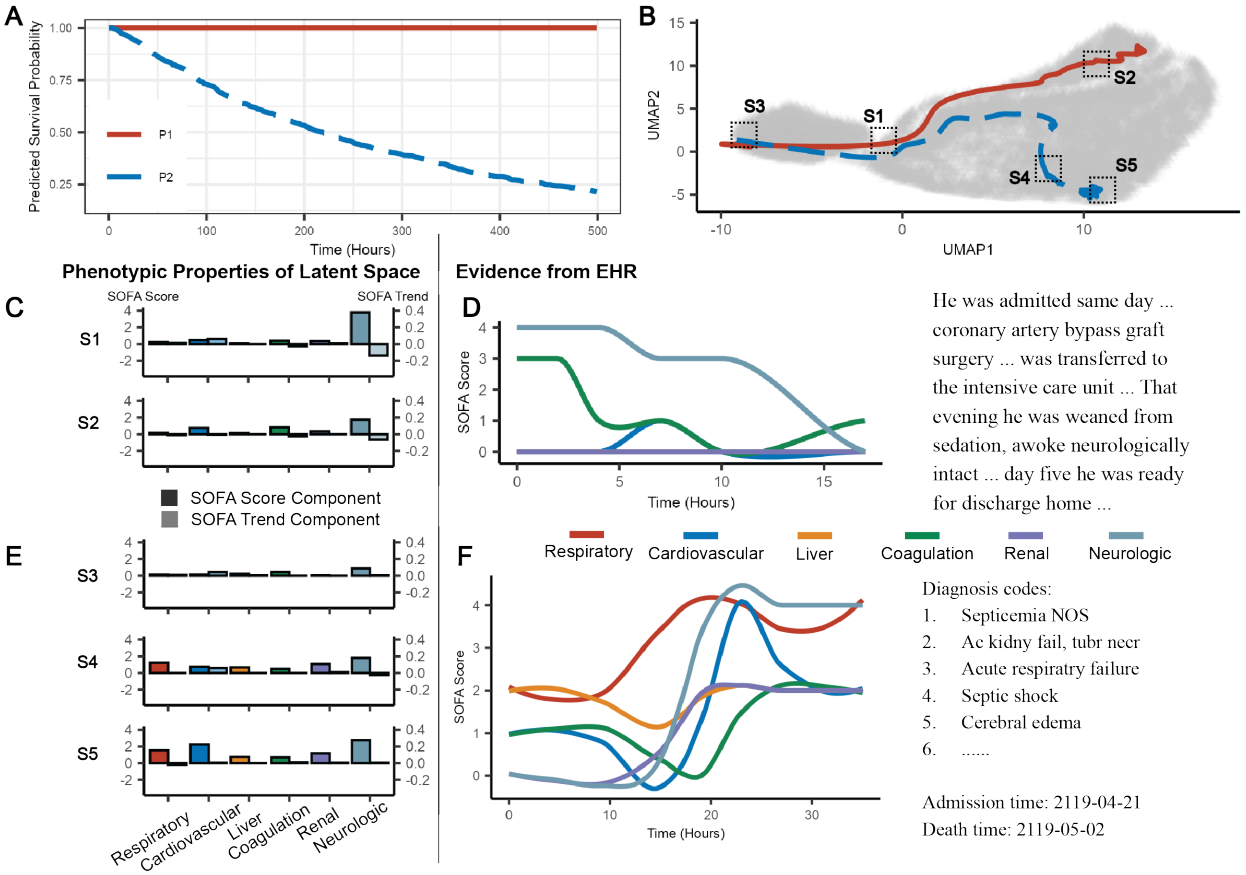

Figure 4.5: Case studies of two patients (P1 and P2). (A) Predicted survival probability after the last observation; (B) Continuous latent trajectories until the last observation; (C) Average phenotypic properties of S1 and S2 of the latent space; (D) Evidence from EHR for P1, including the ground truth SOFA trajectory and the discharge note; (E) Average phenotypic properties of S3, S4, and S5; (F) Evidence from EHR for P2, including the ground truth SOFA trajectory and the Diagnosis codes.

vation (Figure 4.5A). This prediction is supported by several aspects of P1's latent trajectory. First, as shown in Figure 4.5B, P1's latent trajectory falls within the ISQI subgroup, indicating a generally positive prognosis. Additionally, P1's latent trajectory sequentially goes through regions S1 and S2 on the latent space, where the phenotypic properties of latent space (see

Appendix for the calculation) show severe but improving neurologic symptoms at S1 and mild symptoms at S2 (Figure 4.5C). This is consistent with P1's actual SOFA trajectory and discharge note, confirming neurologic improvement post-coronary artery bypass surgery on the first admission day (Figure 4.5D).

For patient P2, TrajSurv-Cont predicts a poor survival outcome following the last observation (Figure 4.5A). This result is supported by P2's latent trajectory (Figure 4.5B). P2's trajectory shows a sudden shift from the IMSI subdomain to ISK subdomain, suggesting an abrupt deterioration mid-visit. P2's trajectory passes through S3, S4, and S5 regions, which indicate initial presentation with mild symptoms (S3) that progress to multiple organ dysfunction (S4 and S5). In the absence of discharge notes, we compared this interpretation to P2's actual SOFA trajectory, diagnosis codes, and actual survival time, showing consistency with a sepsis diagnosis and rapid decline leading to death within 300 hours.

This case study suggests TrajSurv-Cont's potential to provide clinicians with interpretable insights consistent with patient outcomes and clinical progression.

4.4 Discussion

In this study, we propose TrajSurv-Cont, a method for survival prediction that learns continuous latent trajectories from longitudinal EHR. TrajSurv-Cont leverages NCDE to capture the continuous process of evolving clinical features, mapping this evolution to continuous latent trajectories. We further incorporate a TACL objective to ensure the clinical alignment of these trajectories, improving both accuracy and transparency. Paired with a two-step divide-and-conquer interpretation process, TrajSurv-Cont enables trustworthy survival prediction, achieving competitive prediction performance and superior transparency.

With increasing emphasis on the need for reliable and transparent AI systems in healthcare, TrajSurv-Cont provides a viable direction for the transparent modeling of longitudinal EHR by learning clinically aligned continuous latent trajectories and interpreting the model through a divide-and-conquer approach. The divide-and-conquer interpretation process provides a dynamical perspective on how the temporal changes of clinical features link to the outcome,

enriching existing interpretation methods and fostering trust in the model's outputs. By providing a clear and visual representation of clinical progression, TrajSurv-Cont contributes to the development of trustworthy AI tools for clinical use.

A key innovation in TrajSurv-Cont lies in its TACL objective. This objective works as a soft regularization to the latent space by explicitly aligning the latent trajectory with actual clinical progression while respecting its time-dependent nature. By contrasting latent representations of patient states at different time points, TACL encourages the model to learn trajectories that are not only predictive of survival but also clinically meaningful. This enables clinically meaningful vector field analysis and cluster patterns in our two-step interpretation process. Through our ablation, we found that TACL improved both performance and clinical alignment.

The patient's clinical progression is a complex and dynamic process, and TrajSurv-Cont's TACL offers insights into how latent trajectories align with this process. Our analysis reveals that TACL learns latent states that correspond more closely to severity levels than to severity trends (Figure 4.2C and D). This finding is consistent with prior research showing that mean and peak SOFA scores are stronger ICU prognosticators than changes in SOFA (Δ-SOFA) [52], suggesting that TACL may prioritize the most salient dimensions of multidimensional patient states. Future research may further refine this alignment by incorporating more comprehensive patient state definitions or advancing alignment techniques. As one of the early efforts, TrajSurv-Cont establishes a framework for transparent longitudinal EHR modeling by quantifying patient states, aligning latent trajectories with clinical progression, and interpreting with a divide-and-conquer approach.

Our work builds upon and extends previous research utilizing NCDEs. While studies like TE-CDE [165] and CoxSig [20] have employed NCDEs for tasks like continuous-time counterfactual prediction and dynamical survival prediction, they have largely underexplored the rich information contained within the latent trajectory and vector field. TrajSurv-Cont specifically leverages NCDE's capacity to represent the continuous nature of clinical progression transparently and quantitatively, linking clinical feature evolution to the latent

trajectory. Crucially, the feature importance and relevance derived from the NCDE vector field provide a novel and elegant way to interpret how changes in different clinical features jointly influence changes in the latent space. For example, similar changes in creatinine and urea nitrogen lead to the similar transition of latent states.

Finally, TrajSurv-Cont holds considerable potential for clinical translation. Its accurate survival prediction and risk stratification can facilitate clinical resource allocation and decision-making. Critically, as shown in the case study, the interpretable outputs provided by TrajSurv-Cont can empower clinicians by providing insights into the prediction process, fostering trust and acceptance in clinical settings. While comprehensive clinical validation across diverse cohorts remains necessary, we believe TrajSurv-Cont represents a significant step toward trustworthy AI tools in clinical practice and can inspire the development of future explainable AI models for longitudinal EHR data.

4.5 Limitations

Our model design relies on an existing clinical assessment like SOFA as the supervision signal. While these assessments offer a practical way to quantify patient states, they do not fully represent patient states and reduce the model flexibility in various clinical contexts, as not all clinical contexts provide such assessments. Future research may explore purely data-driven methods to ensure clinical alignment or define patient states more comprehensively using more adaptable labels, such as knowledge graphs or key clinical variables, rather than pre-existing assessments. Furthermore, future research is needed to understand the clinical alignment between latent trajectories and clinical progression from a more holistic view. Finally, population-level properties may not fit individual cases, and we leave individual-level vector field analysis for future studies.

Chapter 5

TRAJ-COA: A GENERALIZABLE MULTI-AGENT FRAMEWORK FOR PATIENT TRAJECTORY MODELING FOR CANCER EARLY DETECTION

This chapter is adapted from two papers published at NeurIPS 2025 GenAI4Health Workshop and under review at npj Digital Medicine, titled "Traj-CoA: Patient Trajectory Modeling via Chain-of-Agents for Lung Cancer Risk Prediction " authored by Sihang Zeng, Yujuan Fu, Sitong Zhou, Zixuan Yu, Lucas Jing liu, Jun Wen, Matthew Thompson, Ruth Etzioni, and Meliha Yetisgen, and titled "TrajOnco: a multi-agent framework for temporal reasoning over longitudinal EHR for multi-cancer early detection" authored by Sihang Zeng, Young Won Kim, Wilson Lau, Ehsan Alipour, Ruth Etzioni, Meliha Yetisgen, and Anand Oka, respectively [220, 222]. Results in this chapter were also submitted as abstracts and published at ISPOR 2026 and ASCO 2026 as a poster and oral presentation, respectively.

In this chapter, we introduce Traj-CoA, a generalizable multi-agent framework for patient trajectory modeling. Traj-CoA employs a chain-of-agents architecture [231] with an external long-term memory system to perform complex temporal reasoning on long patient histories. Our framework accepts a unified XML input that minimizes feature engineering and decomposes the longitudinal EHR into manageable time-aware chunks for more effective reasoning and summarization. These chunks are processed through a multi-agent workflow comprising a chain of specialized worker agents and a manager agent, which interact with a long-term memory module (LTM). We evaluate Traj-CoA on two separate clinical tasks on cancer early detection, predicting lung cancer diagnosis within 1 year and multi-cancer early detection at 1 year, using data from University of Washington Medical Center (UWMC) and Truveta, respectively. The main contributions of this work are:

- We propose Traj-CoA, a generalizable multi-agent framework for temporal reasoning over long and noisy EHRs.

- To handle data noisiness, worker agents process unified XML inputs in sequential time-aware chunks, extracting salient signals for the task while removing localized noise.

- To manage long contexts, Traj-CoA leverages multi-agent communication and LTM for effective temporal reasoning with global context.

- We demonstrate Traj-CoA's strong zero-shot performance and interpretability on two clinical tasks with data from two organizations.

5.1 Method

In this section, we introduce Traj-CoA, a generalizable framework for patient trajectory modeling with longitudinal EHR. We outline the problem formulation and Traj-CoA's core components: a unified data preprocessing pipeline that represents heterogeneous EHR in an XML format with time-aware chunking, a chain-of-agents (CoA) workflow that sequentially processes the long and noisy EHR, and a long-term EHR memory system that memorizes detailed clinical events. The framework is described in a general context before being applied to specific use cases of cancer early detection. An illustration of Traj-CoA is shown in Figure 5.1a.

5.1.1 Problem Formulation

We consider a patient trajectory as a longitudinal, multimodal sequence of n observations from the EHR, denoted as $\mathcal{X} = \{x_i, m_i, t_i\}_{i=1}^{n}$. Each tuple consists of a timestamp t_i, a data modality m_i (e.g., diagnosis, lab result, clinical note), and the corresponding event data x_i within the modality. A prediction task $\mathcal{T}$ is defined as a mapping $f : \mathcal{X} \to y$, where y is the label of a task-specific outcome that occurs after the final observation time t_n. We seek a

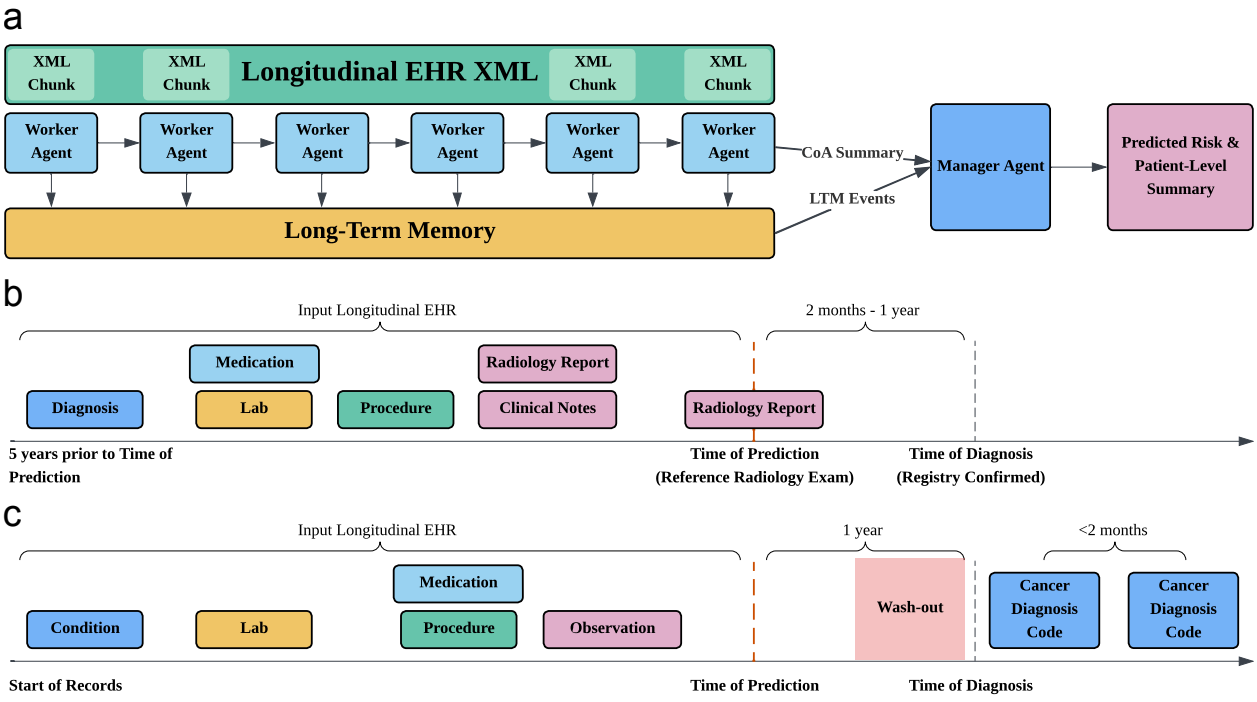

Figure 5.1: **a.** Traj-CoA architecture consisting of a chain of worker agents, a manager agent, and LTM. **b.** Study design on UWMC lung cancer data. **c.** Study design on Truveta data.

generalized framework that learns task $\mathcal{T}$ with minimum task-specific data preprocessing and model training.

The task is challenging due to the inherent properties of EHR data. The observation window (t_1 to t_n) may span over years. The data itself is heterogeneous, potentially from different EHR vendors and with inconsistent coding systems (e.g., ICD-9 vs. ICD-10) across patients; irregularly sampled, with non-uniform time gaps between observations; and noisy, containing missing data and irrelevant events to $\mathcal{T}$. An effective solution must therefore identify task-related events from this long, noisy data and model their temporal patterns in a generalizable manner. For instance, while lung nodule growth indicates cancer risk [134], this temporal pattern is often scattered in EHRs, requiring algorithms to extract and interpret these critical dynamics.

In this study, we consider two clinical tasks on cancer early detection to evaluate the general Traj-CoA framework: **(1) Lung cancer early detection using UWMC data.** Given the full patient trajectory $\mathcal{X}$ spanning up to 5 years before a reference date t_n, the task is to predict whether the patient will be diagnosed with lung cancer within the following year. **(2) Multi-cancer early detection using Truveta data.** Given the full structured patient trajectory $\mathcal{X}$ before 1 year prior to the diagnosis or index date, the task is to predict whether the patient will be diagnosed with one of the 15 cancer types at 1 year. For both tasks, we have $y = 1$ for cases and $y = 0$ for controls.

5.1.2 Data Preprocessing

To handle the noisy, long, and heterogeneous EHR data $\mathcal{X}$, we design a simple yet effective preprocessing pipeline to create a unified input representation X. Instead of relying on the complex task-specific feature engineering and cleaning common in prior work [25, 193, 195], our approach preserves the heterogeneity in the data while transforming it into an LLM-friendly format for reasoning. This consists of two steps: data unification into XML format and time-aware chunking.

Data Unification We convert a patient's entire multimodal history $\mathcal{X}$ into a single, unified XML format X. This strategy is motivated by the proven ability of LLMs to effectively comprehend structured, tag-based data [15] and inspired by similar practice on EHR data [40]. As illustrated in Figure 5.1a, we organize the longitudinal data chronologically within a nested XML structure. The root contains patient demographics, followed by a sequence of timestamped records, each encapsulating all data modalities and events observed at that time. This approach yields a clean, well-structured representation of the patient's timeline for Traj-CoA, preserving the data heterogeneity in text format.

Time-Aware Chunking Long XML inputs pose computational and reasoning challenges for LLMs. For example, in the UWMC lung cancer dataset, the 75th percentile token count reaches

120k (Table 5.1). Despite large context windows (>128k tokens) in recent LLMs, performance degrades significantly with longer contexts due to the “lost-in-the-middle" phenomenon, where models fail to process information from the middle sections effectively [112].

To address this limitation, we follow the CoA approach [231] which splits the context into chunks and uses multi-agent communication to ensure sequential information aggregation and seamless reasoning (see Section 5.1.3). Instead of hard chunking based on a fixed chunk size, we design a time-aware chunking strategy that avoids missing timestamp information caused by chunking. Specifically, we partition the XML EHR input into chunks of maximum k tokens while preserving temporal ordering and timestamp completeness. Since our data structure is organized by timestamps, we split XML input into segments by timestamp, which are aggregated into chunks under the token limit k. When a single timestamp's records exceed k tokens, we split them further while maintaining the original timestamp for each resulting chunk. This dynamic process converts the full XML input X into a temporally coherent series of C chunks $\{c_1, c_2..., c_C\}$, guaranteeing that all information within any given chunk is closely related in time. Note that the number of chunks C may vary by patient.

5.1.3 Chain-of-Agent

We adapt the CoA algorithm [231] to reason over the chunked longitudinal EHR data. The vanilla CoA consists of two stages involving a chain of worker agents to conduct temporal reasoning chunk-by-chunk and a manager agent for the final prediction. Instead of relying on task-specific feature engineering or model training, CoA operates via task-specific instructions provided to its LLM agents, an approach that offers greater flexibility and efficiency for complex reasoning tasks. [231, 162]

In stage one, a series of worker agents $\mathbf{W}_i$ sequentially process each chunk c_i. Each worker agent takes the current chunk c_i, a task-specific instruction I_W, and the summary message S_{i-1} from the preceding agent as input. Its function is to extract salient task-related information from c_i, analyze temporal patterns in relation to the aggregated summary, and produce an updated summary message S_i. This sequential process allows for the progressive

task-related information aggregation across the entire longitudinal EHR. The operation of each worker agent is defined as:

$$S_i = \mathbf{W}_i \left(I_W, S_{i-1}, c_i \right) \tag{5.1}$$

In stage two, a manager agent $\mathbf{M}$ receives the final summary message S_C from the last worker agent $\mathbf{W}_C$ along with a task-specific instruction I_M. The manager agent's role is to synthesize the comprehensive information contained in S_C to produce the final output O. This is formulated as:

$$O = \mathbf{M} \left(I_M, S_C \right) \tag{5.2}$$

The CoA framework transforms a long-context reasoning problem into a structured agent communication chain, with each agent assigned a shorter context, thereby improving the reasoning quality and mitigating the LLM's "lost-in-the-middle" phenomenon common in long-context reasoning. [231]

5.1.4 Long-Term Memory

In our experiments, we found that a direct application of the vanilla CoA framework to longitudinal EHR data can lead to the progressive abstraction and loss of critical task-related information over long sequences. In other words, early clinical events may be vital for accurate prediction but may be "forgotten" by the final summary message S_C. To mitigate this, we introduce a structured long-term memory (LTM) module storing task-related events and timestamps, denoted by $\mathcal{M}$.

LTM is populated during stage one, where each worker agent extracts new clinical events or risk factors that are potentially task-related, and stores their contents and timestamps as entries E_i in $\mathcal{M}$. To prevent overwhelmingly redundant entries caused by EHR "copy-forwarding", [200] we employ a deduplication mechanism: each agent's prompt is augmented with the last k events from $\mathcal{M}$ and is instructed to only store new, unrecorded information. In stage two, the manager agent's decision-making is augmented by the global context in $\mathcal{M}$,

conditioning its output on both the final summary message S_C and the entire memory $\mathcal{M}$. The Traj-CoA's operation is thus redefined as:

$$S_i, E_i = \mathbf{W}_i \left(I_W, S_{i-1}, c_i, \mathcal{M}[-k:] \right) \tag{5.3}$$

$$\mathcal{M} \leftarrow \mathcal{M} \oplus E_i \tag{5.4}$$

$$O = \mathbf{M}(I_M, S_C, \mathcal{M}) \tag{5.5}$$

where $\mathcal{M}[-k:]$ denotes the last k events in $\mathcal{M}$, and $\oplus$ means concatenation. The LTM module serves two primary functions: (1) it constructs a distilled clinical timeline, effectively reducing the noise inherent in raw EHR, and (2) it provides the manager agent with a structured global context complementary to the unstructured worker agents' summary, enabling more robust reasoning across the entire patient history.

Crucially, the extraction heuristic for populating LTM is intentionally inclusive. Worker agents identify a slightly broader set of events potentially relevant to $\mathcal{T}$, rather than strictly filtering for those with an immediate, obvious connection to the task. This design choice acknowledges that local worker agents lack the global context to definitively assess an event's long-term significance. By preserving a richer, slightly redundant set of events in $\mathcal{M}$, we delegate the final synthesis and attribution of importance to the manager agent, which can leverage the complete temporal context for a more informed judgment.

5.2 Results on UWMC Data

5.2.1 Clinical Context

Lung cancer remains one of the most lethal malignancies, yet early detection provides a critical window of opportunity to significantly improve patient prognosis and survival rates [95, 43]. Accurate estimation of lung cancer risk from longitudinal EHRs is important for optimizing screening protocols and facilitating earlier clinical intervention. However, modeling the complex patient trajectories that precede a diagnosis presents a significant clinical challenge [135]. Over time, patients accumulate diverse, multimodal data encompassing

subtle symptoms, lifestyle factors, and chronic comorbidities. The early predictive signals of malignancy are often deeply embedded within these extensive and inherently noisy clinical histories, making it difficult to accurately estimate a patient's lung cancer risk.

The complexity is further amplified in real-world clinical pathways where risk assessment is often anchored to a specific radiological event, such as chest X-ray or CT. In this context, the patient has already undergone radiological imaging due to prior clinical indications. Consequently, the predictive rask requires distinguishing radiological findings associated with primary lung cancer from those attributable to benign conditions or existing comorbidities, such as cardiovascular disease [105]. Successfully differentiating these nuanced, overlapping clinical presentations requires a holistic integration of the patient's long-term history rather than relying on isolated snapshots of their health. Traj-CoA offers a solution for temporal reasoning over longitudinal multimodal EHRs for accurate risk estimation, summarizing dynamics such as enlarging nodules or the gradual exacerbation of respiratory symptoms which might otherwise be obscured by the sheer volume of data.

5.2.2 Dataset

We experimented on a proprietary case-control dataset from UWMC on lung cancer risk assessment. Each instance in the dataset is anchored to a specific chest-related radiology exam (chest X-ray, chest CT, or abdomen CT) capable of visualizing the lungs. The prediction task is to determine, at the time of this index exam (t_n), whether a patient will receive a primary lung cancer diagnosis within 1 year. Cases are defined as subjects diagnosed with primary lung cancer within one year of the index exam, with diagnoses cross-validated against a state cancer registry. Controls are subjects with no cancer diagnosis. Cases and controls were matched at a 1:10 ratio on the time and type of the index exam. For each instance, up to five years of the patient's longitudinal EHR history prior to the index exam were recorded. This rich, multi-modal data encompasses both structured records (e.g., ICD codes, lab results, vitals) and unstructured text (e.g., clinical notes, radiology reports). The study design is illustrated in Figure 5.1b.

Table 5.1: Dataset Statistics by Case-Control Status. Values shown as median (IQR) unless otherwise specified.

Variable	**Cases (n=28)**	**Controls (n=272)**
Sex (Female/Male)	14/14	115/157
Year of last record	2015.5 (2014.0–2018.0)	2016.5 (2013.0–2018.0)
Year span	4.0 (2.0–5.0)	4.0 (1.0–5.0)
XML tokens	61,270 (28,675–121,722)	51,610 (19,442–132,777)
# Timestamps	42.5 (24.0–79.8)	38.5 (14.0–96.0)
# EHR System 1 notes	7.5 (1.0–29.3)	8.0 (0.0–29.0)
# EHR System 2 notes	3.5 (0.0–12.8)	4.0 (0.0–14.0)
# Radiology reports	10.0 (4.8–17.3)	8.0 (3.0–15.3)
# Diagnosis codes	156.0 (92.8–271.3)	128.5 (42.5–349.3)
# Medication codes	42.0 (17.8–135.3)	64.0 (9.0–184.5)
# Procedure codes	98.0 (51.5–197.8)	106.5 (38.0–231.5)
# Lab codes	186.0 (26.8–439.3)	201.0 (49.8–523.0)
# Vital sign codes	28.0 (0.0–91.0)	20.0 (0.0–79.5)

We randomly sampled 13,629 instances (1,239 cases and 12,390 controls) for model development, which were further randomly split into a training (12,266 samples) and validation (1,363 samples) set. These data were used for fine-tuning and **not** used under a zero-shot setting. From the rest, we randomly sampled 300 instances (28 cases and 272 controls) to construct a test set for feasible and fair evaluation. We highlight that the token count of XML input is substantial, with an interquartile range (IQR) of 28k–121k for cases and 19k–132k for controls in the test set (Table 5.1).

5.2.3 Experimental Settings

We used MedGemma-27B [166] as the base model of Traj-CoA, with a default chunk size of 8k tokens and a maximum of 15 chunks to accommodate contexts up to 120k tokens. We benchmarked Traj-CoA against four baseline categories: (1) ML: logistic regression (LR) and XGBoost [32] trained on summary EHR features. (2) DL: The RNN-based trajectory models RETAIN [37] and PatientTM [172]. (3) BERT-based: Clinical ModernBERT (C-MBERT) [99], fine-tuned with LoRA [72] using an 8k context window. (4) LLM: Zero-shot MedGemma using two strategies: direct prompting (Vanilla, up to 64k context) and RAG with a bge-m3 [29] retriever on the time-aware chunks.

For LR, XGBoost, and RETAIN, we used diagnosis codes as input features. For PatientTM, we used the text descriptions of medical codes and unstructured notes as input. For BERT and LLM methods, we used the same XML input and employed a middle truncation strategy for context beyond the window. Specifically, we alternately selected texts from the beginning and end of the patient record until the context window limit is met, maintaining their relative chronological order. Compared to left truncation, where the most recent records are used, this method retains a longer and more challenging temporal span for evaluating temporal reasoning. We further report the performance using left truncation for BERT and vanilla baselines in Appendix C.1.1, a case study in Appendix C.1.3, and error analysis in Appendix C.1.4.

We evaluated all methods on AUROC and the best F1 score among all thresholds, with its corresponding precision and recall. We report AUPRC and discuss the differences in the Appendix C.1.1. For BERT, the risk score was derived from the output logit. For all LLM-based methods, we prompted the model to output a risk score between 1 and 10. Further experimental details are in the Appendix C.1.5.

Table 5.2: Performance comparison of different models. **Bold** indicates the best AUROC and F1 among all models or zero-shot models. SFT means supervised fine-tuning. The average performance (mean±std) for Traj-CoA across 5 runs with different random seeds is reported.

Model Family	Model	Prediction Method	Context Window	AUROC	Precision	Recall	F1
ML	LR	SFT	—	0.741	0.306	0.393	0.344
	XGBoost	SFT	—	0.763	0.367	0.393	0.379
DL	RETAIN	SFT	—	0.757	0.346	0.321	0.333
	PatientTM	SFT	—	0.730	0.361	0.464	**0.406**
BERT	C-MBERT	SFT	8k	0.749	0.367	0.393	0.379
LLM (MedGemma)	Vanilla	Zero-shot	32k	0.743	0.345	0.357	0.351
	RAG	Zero-shot	1k × 32	0.753	0.221	0.607	0.324
	Traj-CoA (w/o LTM)	Zero-shot	8k × 15	0.748	0.183	0.821	0.299
	Traj-CoA	Zero-shot	8k × 15	**0.766** ± 0.019	0.358 ± 0.057	0.436 ± 0.105	**0.380** ± 0.018

5.2.4 Results

Table 5.2 reports the performance of each method. Vanilla MedGemma with a 32k context window and RAG with top 32 chunks of maximum 1k tokens each achieves an AUROC of 0.74–0.75. However, expanding vanilla MedGemma's context to 64k degrades its performance to an AUROC of 0.714, suggesting difficulty in utilizing longer input sequences effectively (Table C.1). In contrast, Traj-CoA, with its 120k context and dedicated design, substantially outperforms all zero-shot baselines, achieving an AUROC of 0.766 and an F1 score of 0.380, outperforming most SFT baselines and comparable to the best. These results highlight Traj-CoA's superior capability to conduct temporal reasoning over very long EHRs, overcoming the limitations observed in standard long-context models.

Ablation Study To analyze the contribution of the LTM component, we performed an ablation study by removing it from our architecture. As detailed in Table 5.2, this modification

significantly impairs model performance, causing an absolute drop of 1.8% in AUROC and 8.1% in F1 score. This result underscores the importance of maintaining a detailed long-term memory of clinical events, as it provides predictive signals that are not fully captured by the evolving summary alone.

To further probe the mechanisms by which Traj-CoA synthesizes temporal information from longitudinal EHR, we conducted additional analyses to answer five research questions. We fixed a random seed for the following analyses.

5.2.5 Sensitivity Analysis

Q1: How does the chunk size affect the performance?

Motivated by recent findings that LLM performance can degrade in very long contexts [112], we analyzed Traj-CoA's sensitivity to chunk size. To isolate this effect, we fixed the total context length at 80k tokens and varied the chunk size from 2k to 16k, which correspondingly adjusted the number of chunks from 40 down to 5. Figure 5.2A shows that performance peaks with a moderate chunk size of 8k, while both smaller (2k) and larger (16k) chunks result in lower AUROC scores.

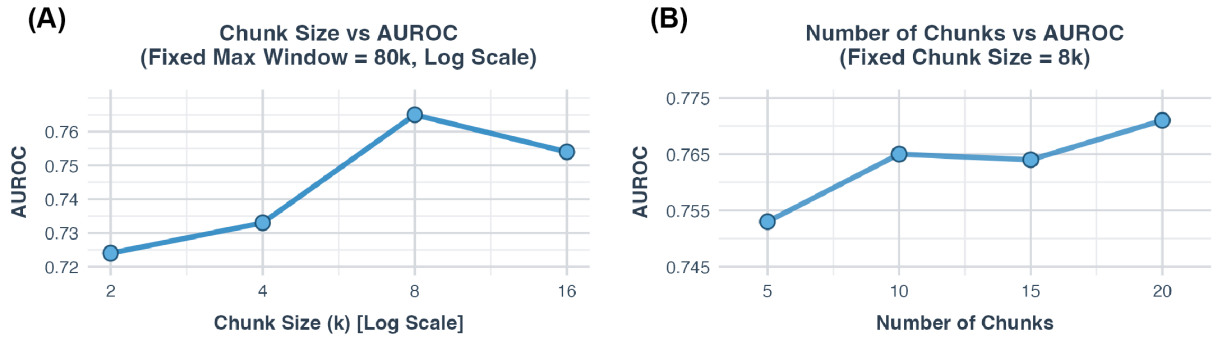

Figure 5.2: Sensitivity analysis on (A) chunk size and (B) number of chunks.

This reveals a fundamental trade-off. Small chunks force a long chain of iterative summarizations, risking catastrophic forgetting [88] where early, critical details are abstracted away.

Conversely, large chunks shorten the chain but are susceptible to the "lost-in-the-middle" issue [112], where each worker agent fails to identify fine-grained signals within a vast context. The 8k chunk size appears to provide an optimal balance, preserving local detail while enabling effective global aggregation.

Q2: How does the maximum context window affect the performance?

We next analyzed how performance scales with the total context window to determine if more temporal information is beneficial. We fixed the chunk size at 8k and increased the maximum number of chunks from 5 to 20, thereby expanding the context window from 40k to 160k tokens.

As shown in Figure 5.2B, AUROC consistently improves as the context window expands. While the vanilla LLM baseline also improves when scaling from an 8k to a 32k context, its performance degrades significantly at 64k (Table C.1). In contrast, Traj-CoA maintains its positive trend up to 160k tokens. This demonstrates that EHRs contain rich, albeit noisy, predictive signals and that Traj-CoA's architecture is uniquely capable of leveraging these ultra-long sequences where standard methods fail.

5.2.6 Temporal Reasoning Analysis

To investigate Traj-CoA's temporal reasoning, we performed a topic modeling analysis on the clinical events it identified as salient for lung cancer prediction. We used an LLM-based pipeline inspired by TopicGPT [142] to systematically categorize these events. First, for all positive cases, we extracted the relevant events from the model's final output O. Next, using Qwen2.5-7B-Instruct [156], we conducted a three-step analysis: (1) each event was classified into one of seven predefined categories (diagnosis, procedure, lab tests, vital, medication, health behaviors or symptoms, and demographics); (2) for each category, the top three common themes were generated based on all events under it; and (3) each event was mapped to its most relevant theme, if any, or recorded as "Others". This structured thematic analysis of the model's output forms the basis to answering the subsequent Q3–Q5.

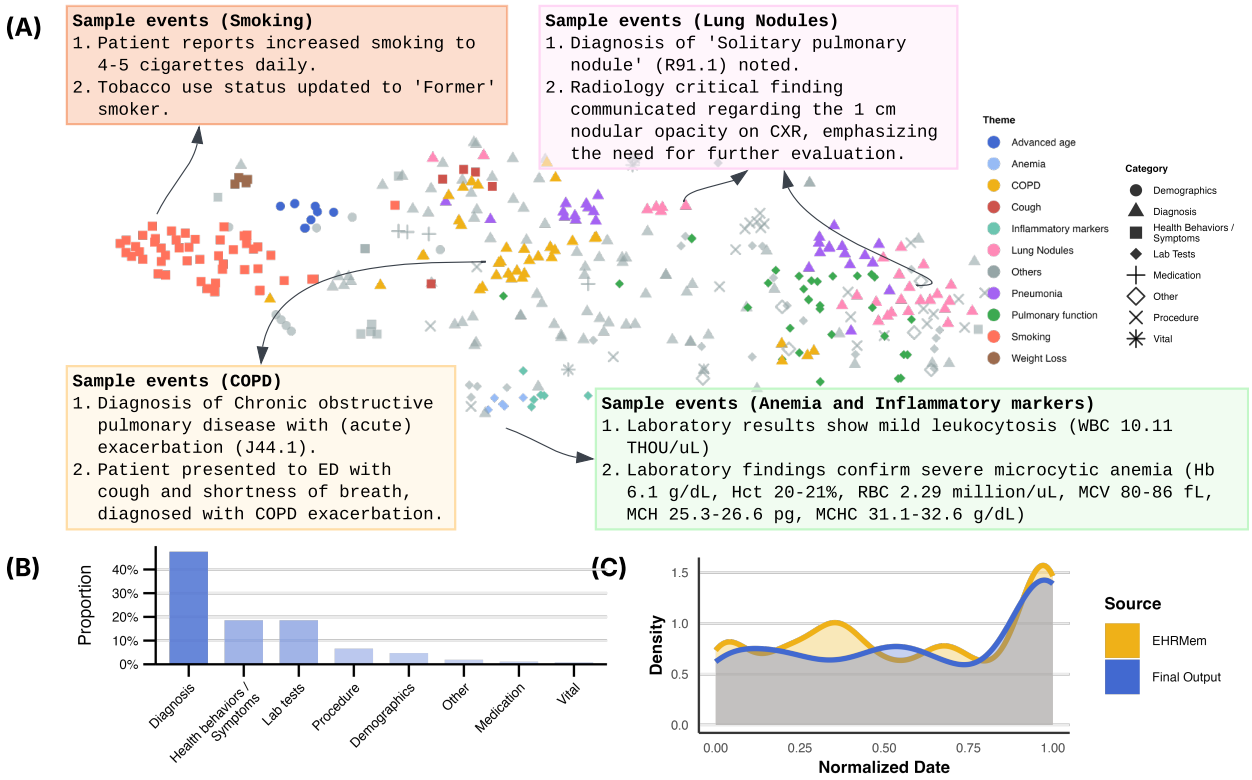

Figure 5.3: Analysis of Traj-CoA's behavior on UWMC lung cancer early detection task. (A) t-SNE plot visualizing the distribution of lung cancer related events in all cases' output O and sample events (timestamps in sample events are omitted for de-identification purposes). The events were embedded through nomic-embed-text-v1.5 [133]; (B) Distribution of categories in the lung cancer related events; and (C) Normalized date distribution of the events.

Q3: Does Traj-CoA reason across diverse event categories?

We sought to verify that Traj-CoA's reasoning extends beyond trivial heuristics (e.g., identifying smoking status) to encompass a broad range of clinical events. Our analysis confirms that Traj-CoA identifies salient events from all seven predefined categories. As shown in the distribution in Figure 5.3B, the most frequently utilized categories are diagnosis, health behaviors or symptoms, and lab tests.

The diversity of these identified events is further underscored by the t-SNE visualization in Figure 5.3A. The event embeddings are scattered across the semantic space rather than forming a single cluster, indicating that the model draws upon a wide variety of clinical concepts. The qualitative examples presented in the figure corroborate this ability. This demonstrates that Traj-CoA performs multifaceted reasoning by integrating signals from a medically diverse set of events for its predictions.

Q4: Can Traj-CoA reason over the entire time horizon?

To determine if Traj-CoA utilizes the full patient trajectory or suffers from the "lost-in-the-middle" phenomenon, we analyzed the temporal distribution of the events it identified as salient. In Figure 5.3C, we plot the normalized date distribution of events from the LTM ($\mathcal{M}$) and the final output (O).

Both distributions show a concentration of events in the final year before the prediction date, which aligns with the clinical intuition that recent events are often most critical for diagnosis. Crucially, however, the model still identifies events from earlier periods. The presence of historical events in the final output demonstrates that Traj-CoA effectively synthesizes information over long time horizons, appropriately weighing recent events more heavily without discarding valuable historical context.

Q5: Is Traj-CoA's reasoning clinically relevant?

Finally, we assessed the clinical relevance of Traj-CoA's reasoning by analyzing the themes of the salient events. The t-SNE visualization of the event embeddings (Figure 5.3A) reveals that the most frequently identified themes form distinct semantic clusters, indicating thematic consistence.

Importantly, these data-driven themes align directly with well-established clinical knowledge. The top themes identified include advanced age, anemia, COPD, cough, inflammatory markers, lung nodules, pneumonia, pulmonary function, smoking, and weight loss. These are either widely recognized by clinical practice and screening guidelines for assessing lung cancer

risk [198, 164], or consistent with existing evidence [82, 49, 148]. This validates the model's interpretability and demonstrates that its risk predictions are based on clinically meaningful patterns within the EHR data.

5.3 Results on Truveta Data

5.3.1 Clinical Context

Early detection is an aspirational approach for significantly reducing the burden of cancer morbidity and mortality [196, 39]. There is an emerging interest in harnessing large-scale longitudinal electronic health record (EHR) data to model multi-cancer risk [81, 235, 171]. Modeling this risk generally serves two primary purposes: tailoring screening and surveillance for high-risk individuals, or advancing a future cancer diagnosis for those already exhibiting subtle clinical signals. EHR data contain detailed records of patient trajectories capturing these clinical signals, such as diagnoses, laboratory tests, and medications over time. For example, gradually worsening anemia and repeated imaging for unexplained abdominal pain may precede the diagnosis of certain cancers [70, 123]. Identifying these evolving clinical patterns offers a potential modality built directly into routine clinical care to advance the time of clinical diagnosis.

Depending on the time horizon before a cancer diagnosis and the true disease status of the patient, EHR-based prediction problems can generally be categorized into three distinct tasks. First, *near-term diagnosis prediction* addresses the very short-term horizon (weeks to months) where a patient is already exhibiting obvious clinical symptoms, aiming to predict an imminent diagnosis [103]. Second, *early detection* (or advanced clinical diagnosis) utilizes a moderate time horizon (1-3 years) to predict short-term cancer risk based on early, subtle, or evolving clinical signals before full symptomatic disease is recognized. Finally, *long-term risk prediction* targets a longer time horizon (5+ years) to predict future cancer risk in patients who currently do not have the disease, guiding tailored surveillance. These settings represent increasing levels of difficulty as signals become weaker and more temporally diffuse.

Specifically for early detection, capturing these subtle, accumulating signals across a patient's trajectory requires robust temporal reasoning over longitudinal EHR data [196, 39].

Prior studies detecting early cancer from EHRs have predominantly relied on data from large national healthcare systems [81, 171]. Multi-cancer early detection with EHRs across diverse, multi-system environments remains challenging. Existing approaches typically rely on supervised machine learning (ML) models that require extensive feature engineering and training data [138, 235, 81]. In addition, their outputs and interpretations, such as hazard ratios or SHAP values [116], offer limited insight into the complex temporal patterns within patient histories that jointly shape the evolution of cancer risk.

LLMs have emerged as a flexible and generalizable paradigm, offering evidence-driven reasoning with improved interpretability. In oncology, prior applications have focused on short-context tasks such as cancer progression prediction, phenotype extraction, and diagnosis generation from individual notes or reports [103, 240, 96]. However, extending LLMs to early detection requires modeling temporal patterns and interactions among diverse clinical factors over long patient trajectories which may span several years. Conventional single-agent LLMs struggle with such long-context reasoning due to limitations such as the "lost-in-the-middle" effect [112], which may lead to the omission of evolving EHR signals. Agent-based LLM systems have recently been proposed to decompose complex reasoning into coordinated steps across multiple agents [104], offering a promising approach for temporal reasoning over long and heterogeneous EHR data.

In this section, we apply Traj-CoA on the task of multi-cancer early detection using Truveta data. We define multi-cancer early detection as prediction of short-term risk across multiple distinct cancer types within a shared modeling framework. Traj-CoA addresses this setting in a zero-shot manner using the same base model and architecture across cancers, with only minimal prompt adaptation to specify the target cancer type.

Using de-identified structured EHR data encompassing multiple healthcare systems from Truveta, we evaluated Traj-CoA to predict cancer risk across 15 cancer types. To focus on early detection and avoid overlap with near-term diagnosis prediction, we imposed a 1-year

gap between the time of prediction and the diagnosis or index date (Figure 5.1c), leveraging all available EHR data prior to the time of prediction. In zero-shot settings, Traj-CoA achieved AUROCs ranging from 0.64 to 0.80 across cancers and performed comparably to supervised ML models in a lung cancer benchmark. Notably, its multi-agent design enables effective temporal reasoning without relying on very large models. In addition to individual-level predictions, Traj-CoA identifies key clinical events associated with cancer risk, informing our understanding of the most common factors that portend a future cancer diagnosis.

5.3.2 Data Source and Cohort Construction

Truveta data comprise de-identified EHRs from over 30 US health systems, covering more than 120 million patients across 900 hospitals and 20,000 clinics. Using these data, we identified cohorts for 15 cancer types, including liver, lung, prostate, breast, bladder, cervical, esophageal, endometrial, gastric, ovarian, pancreatic, colorectal cancers, leukemia, multiple myeloma, and lymphoma, based on clinician-curated diagnosis codes. These represent major solid and hematologic malignancies commonly studied in early detection research [81, 138]. We used Truveta Studio to identify the case and control population of 15 cancer types using clinician curated diagnosis codes, including SNOMED CT, ICD-9-CM, and ICD-10-CM. We used an existing phenotyping algorithm to define new-onset cancer, which was based on observation of two diagnosis codes associated with the cancer of interest within two months of each other, with a cancer-free washout period of 6 months [67, 167]. The first record in the first qualifying pair was used as the index date for the cases. For controls, we randomly sampled an index date among all available visit dates, and age/sex matched to the cases in a 1:1 manner, with the ratio aligned with previous studies [22, 179]. The prediction task in this study was predicting the risk of cancer at 1 year, therefore we used all longitudinal EHR from the start of the records to 1 year before the index date as the input (Figure 5.1c). Patients with fewer than 2 visits in the input EHR were excluded during the sampling procedure. The EHR was structured data, including timestamped conditions, observations, lab results, medications, and procedures, as well as demographics including birth date, sex, ethnicity,

and race.

5.3.3 Zero-Shot Multi-Cancer Early Detection

To evaluate Traj-CoA's performance across cancer types, we conducted three analyses: (i) cancer-specific case-control prediction for each cancer, (ii) benchmarking on a lung cancer cohort against established machine learning models, and (iii) sensitivity analyses on efficiency and scalability. Unless otherwise specified, GPT-4.1-mini was used as the base model for both worker and manager agents due to its favorable balance of performance, latency, and cost. The context length of each XML chunk was set as 16k tokens. We also evaluated a composite any-cancer prediction task following prior work [235] (Appendix C.2.2).

Cancer-Specific Prediction across 15 Cancers

For each cancer type, we constructed a case-control test set comprising 500 randomly sampled cases and 500 controls. This moderate sample size was determined to balance the variance in performance estimates and the experimental costs. Controls were randomly sampled and matched 1:1 to cases by sex and 10-year age group to reduce demographic confounding. For Traj-CoA, worker and manager agent prompts were kept consistent across cancer types, with only the cancer type modified, ensuring comparable model behavior (Appendix C.2.7).

Table 5.3 summarizes baseline characteristics for the 15 cancer-specific case-control cohorts, with 500 cases and 500 matched controls per cancer type. Median age varied by cancer, from younger cohorts such as cervical cancer (60.3 years among cases) and ovarian cancer (63.8 years) to older cohorts such as bladder cancer (74.1 years), leukemia (72.6 years), multiple myeloma (72.3 years), and prostate cancer (72.3 years). Sex distributions were consistent with disease epidemiology, including predominantly female cohorts for breast, all-female cohorts for cervical, endometrial, and ovarian cancers and an all-male prostate cohort, while other cancers had more balanced or male-predominant distributions. Across most cancer types, the majority of patients were non-Hispanic/Latino and White, though the degree of racial diversity differed by cohort. Cases generally had longer longitudinal EHR history than

Table 5.3: Baseline characteristics by cancer type and case/control status. Age, EHR span, and EHR XML token count are median [IQR]; sex, ethnicity, and race are n (%). EHR XML token count was estimated with `tiktoken` using `cl100k_base`.

	Bladder		Breast		Cervical		Colorectal		Endometrial	
Characteristic	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl
N	500	500	500	500	500	500	500	500	500	500
Age, y	74.1 [67.1, 80.7]	73.2 [65.0, 80.8]	68.1 [60.0, 76.3]	68.6 [60.0, 76.4]	60.3 [46.7, 68.9]	60.0 [48.0, 68.8]	70.3 [60.5, 79.7]	70.2 [60.4, 79.8]	67.1 [58.3, 74.0]	66.4 [57.0, 74.0]
Sex										
Female	135 (27.0)	135 (27.0)	498 (99.6)	498 (99.6)	500 (100.0)	500 (100.0)	257 (51.4)	257 (51.4)	500 (100.0)	500 (100.0)
Male	365 (73.0)	365 (73.0)	2 (0.4)	2 (0.4)	0 (0.0)	0 (0.0)	243 (48.6)	243 (48.6)	0 (0.0)	0 (0.0)
Ethnicity										
Hisp./Latino	18 (3.6)	31 (6.2)	17 (3.4)	47 (9.4)	38 (7.6)	53 (10.6)	27 (5.4)	44 (8.8)	40 (8.0)	45 (9.0)
Not Hisp./Latino	482 (96.4)	469 (93.8)	483 (96.6)	453 (90.6)	462 (92.4)	447 (89.4)	473 (94.6)	456 (91.2)	460 (92.0)	455 (91.0)
Race										
White	435 (87.0)	369 (73.8)	399 (79.8)	382 (76.4)	406 (81.2)	356 (71.2)	399 (79.8)	392 (78.4)	398 (79.6)	371 (74.2)
Black	24 (4.8)	65 (13.0)	51 (10.2)	43 (8.6)	50 (10.0)	59 (11.8)	56 (11.2)	55 (11.0)	49 (9.8)	65 (13.0)
Asian	9 (1.8)	12 (2.4)	19 (3.8)	18 (3.6)	8 (1.6)	25 (5.0)	11 (2.2)	8 (1.6)	12 (2.4)	22 (4.4)
AI/AN	1 (0.2)	2 (0.4)	0 (0.0)	5 (1.0)	5 (1.0)	4 (0.8)	0 (0.0)	3 (0.6)	4 (0.8)	5 (1.0)
NH/PI	0 (0.0)	2 (0.4)	0 (0.0)	2 (0.4)	1 (0.2)	3 (0.6)	1 (0.2)	3 (0.6)	2 (0.4)	1 (0.2)
Other	12 (2.4)	30 (6.0)	13 (2.6)	28 (5.6)	15 (3.0)	27 (5.4)	12 (2.4)	22 (4.4)	12 (2.4)	19 (3.8)
Unknown	19 (3.8)	20 (4.0)	18 (3.6)	22 (4.4)	15 (3.0)	26 (5.2)	21 (4.2)	17 (3.4)	23 (4.6)	17 (3.4)
EHR span, years	5.0 [2.0, 9.3]	2.9 [0.6, 6.5]	4.9 [1.9, 8.5]	2.5 [0.8, 6.1]	4.9 [1.6, 8.9]	2.3 [0.8, 5.9]	4.0 [1.4, 7.5]	2.5 [0.7, 6.4]	5.5 [2.2, 9.3]	2.7 [0.8, 6.1]
XML tokens	25.1k [7.2k, 65.3k]	11.5k [3.2k, 32.7k]	12.8k [4.5k, 32.8k]	9.2k [2.9k, 29.0k]	15.4k [4.0k, 47.2k]	7.5k [2.3k, 22.8k]	5.1k [1.4k, 14.5k]	4.1k [1.3k, 11.3k]	16.5k [5.3k, 47.1k]	10.4k [2.6k, 34.1k]

	Esophageal		Gastric		Leukemia		Liver		Lung	
Characteristic	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl
N	500	500	500	500	500	500	500	500	500	500
Age, y	71.1 [63.9, 78.2]	71.2 [63.3, 77.7]	69.0 [58.5, 76.9]	68.4 [58.2, 76.4]	72.6 [62.6, 78.6]	72.0 [61.3, 78.7]	68.9 [63.3, 75.8]	69.4 [61.6, 76.0]	71.6 [65.6, 78.2]	71.2 [64.7, 77.4]
Sex										
Female	120 (24.0)	120 (24.0)	231 (46.2)	231 (46.2)	205 (41.0)	205 (41.0)	174 (34.8)	174 (34.8)	292 (58.4)	292 (58.4)
Male	380 (76.0)	380 (76.0)	269 (53.8)	269 (53.8)	295 (59.0)	295 (59.0)	326 (65.2)	326 (65.2)	208 (41.6)	208 (41.6)
Ethnicity										
Hisp./Latino	19 (3.8)	46 (9.2)	47 (9.4)	42 (8.4)	19 (3.8)	35 (7.0)	36 (7.2)	35 (7.0)	13 (2.6)	31 (6.2)
Not Hisp./Latino	481 (96.2)	454 (90.8)	453 (90.6)	458 (91.6)	481 (96.2)	465 (93.0)	464 (92.8)	465 (93.0)	487 (97.4)	469 (93.8)
Race										
White	426 (85.2)	396 (79.2)	363 (72.6)	373 (74.6)	420 (84.0)	396 (79.2)	375 (75.0)	383 (76.6)	421 (84.2)	385 (77.0)
Black	28 (5.6)	52 (10.4)	55 (11.0)	52 (10.4)	34 (6.8)	53 (10.6)	50 (10.0)	50 (10.0)	38 (7.6)	59 (11.8)
Asian	3 (0.6)	12 (2.4)	34 (6.8)	18 (3.6)	10 (2.0)	18 (3.6)	16 (3.2)	20 (4.0)	10 (2.0)	27 (5.4)
AI/AN	1 (0.2)	1 (0.2)	5 (1.0)	3 (0.6)	2 (0.4)	0 (0.0)	2 (0.4)	2 (0.4)	0 (0.0)	1 (0.2)
NH/PI	2 (0.4)	1 (0.2)	1 (0.2)	5 (1.0)	2 (0.4)	0 (0.0)	2 (0.4)	1 (0.2)	0 (0.0)	1 (0.2)
Other	12 (2.4)	18 (3.6)	19 (3.8)	26 (5.2)	10 (2.0)	14 (2.8)	18 (3.6)	20 (4.0)	8 (1.6)	11 (2.2)
Unknown	28 (5.6)	20 (4.0)	23 (4.6)	23 (4.6)	22 (4.4)	19 (3.8)	37 (7.4)	24 (4.8)	23 (4.6)	16 (3.2)
EHR span, years	4.9 [1.8, 8.9]	2.8 [0.9, 6.3]	5.0 [2.0, 9.4]	2.7 [0.7, 6.4]	4.4 [1.7, 7.7]	2.7 [0.7, 5.8]	5.0 [1.9, 8.3]	2.6 [0.6, 6.3]	4.7 [1.9, 8.0]	2.6 [0.8, 6.6]
XML tokens	26.2k [7.7k, 76.8k]	10.0k [3.0k, 39.2k]	27.1k [7.0k, 72.3k]	9.6k [3.0k, 25.9k]	20.3k [6.4k, 53.1k]	10.2k [2.8k, 29.6k]	35.7k [10.3k, 89.1k]	9.8k [3.9k, 30.4k]	26.4k [8.7k, 73.2k]	11.5k [3.5k, 32.6k]

	Lymphoma		Mult. Myeloma		Ovarian		Pancreatic		Prostate	
Characteristic	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl
N	500	500	500	500	500	500	500	500	500	500
Age, y	70.3 [59.7, 78.2]	70.3 [59.4, 78.9]	72.3 [64.6, 78.5]	72.3 [64.3, 78.9]	63.8 [52.9, 73.2]	64.6 [52.6, 73.0]	71.3 [64.2, 79.1]	71.3 [63.8, 79.3]	72.3 [66.3, 78.7]	71.9 [64.8, 77.8]
Sex										
Female	229 (45.8)	229 (45.8)	238 (47.6)	238 (47.6)	500 (100.0)	500 (100.0)	251 (50.2)	251 (50.2)	0 (0.0)	0 (0.0)
Male	271 (54.2)	271 (54.2)	262 (52.4)	262 (52.4)	0 (0.0)	0 (0.0)	249 (49.8)	249 (49.8)	500 (100.0)	500 (100.0)
Ethnicity										
Hisp./Latino	29 (5.8)	41 (8.2)	28 (5.6)	37 (7.4)	32 (6.4)	48 (9.6)	17 (3.4)	33 (6.6)	36 (7.2)	26 (5.2)
Not Hisp./Latino	471 (94.2)	459 (91.8)	472 (94.4)	463 (92.6)	468 (93.6)	452 (90.4)	483 (96.6)	467 (93.4)	464 (92.8)	474 (94.8)
Race										
White	401 (80.2)	394 (78.8)	351 (70.2)	367 (73.4)	387 (77.4)	367 (73.4)	394 (78.8)	394 (78.8)	363 (72.6)	400 (80.0)
Black	46 (9.2)	44 (8.8)	91 (18.2)	63 (12.6)	53 (10.6)	63 (12.6)	48 (9.6)	55 (11.0)	88 (17.6)	43 (8.6)
Asian	7 (1.4)	12 (2.4)	12 (2.4)	17 (3.4)	21 (4.2)	20 (4.0)	13 (2.6)	17 (3.4)	9 (1.8)	14 (2.8)
AI/AN	3 (0.6)	2 (0.4)	1 (0.2)	0 (0.0)	1 (0.2)	3 (0.6)	2 (0.4)	3 (0.6)	4 (0.8)	2 (0.4)
NH/PI	0 (0.0)	2 (0.4)	2 (0.4)	1 (0.2)	3 (0.6)	0 (0.0)	0 (0.0)	1 (0.2)	2 (0.4)	1 (0.2)
Other	17 (3.4)	23 (4.6)	10 (2.0)	21 (4.2)	15 (3.0)	24 (4.8)	6 (1.2)	12 (2.4)	19 (3.8)	10 (2.0)
Unknown	26 (5.2)	23 (4.6)	33 (6.6)	31 (6.2)	20 (4.0)	23 (4.6)	37 (7.4)	18 (3.6)	15 (3.0)	30 (6.0)
EHR span, years	4.4 [1.7, 8.1]	3.0 [0.7, 6.9]	4.6 [1.7, 8.1]	2.6 [0.7, 6.3]	4.2 [1.6, 8.5]	2.6 [0.7, 6.0]	5.3 [2.1, 9.0]	2.5 [0.6, 5.9]	4.6 [1.7, 8.2]	2.9 [0.7, 6.0]
XML tokens	14.0k [4.9k, 45.6k]	8.3k [2.4k, 31.2k]	21.9k [5.6k, 62.2k]	10.0k [2.6k, 28.8k]	5.2k [1.5k, 15.6k]	4.0k [1.0k, 10.6k]	21.7k [7.7k, 60.6k]	10.1k [3.2k, 31.1k]	15.3k [4.5k, 42.8k]	13.3k [4.2k, 33.6k]

controls, with median EHR span for cases typically around 4-5.5 years versus 2.3-3.0 years for controls. Similarly, EHR XML token counts were usually higher among cases than controls, indicating greater documentation volume; this difference was especially pronounced for liver, gastric, esophageal, lung, bladder, pancreatic, leukemia, and multiple myeloma cohorts.

Model discrimination varied across the 15 cancer types (AUROC 0.64-0.80; Figure 5.4a). Performance was highest for liver and lung cancers and lowest for colorectal cancer, while most other solid and hematologic malignancies demonstrated comparable discrimination with AUROCs above 0.70. This indicates that the framework generalizes across diverse cancer types, albeit with cancer-specific variations in predictive accuracy.

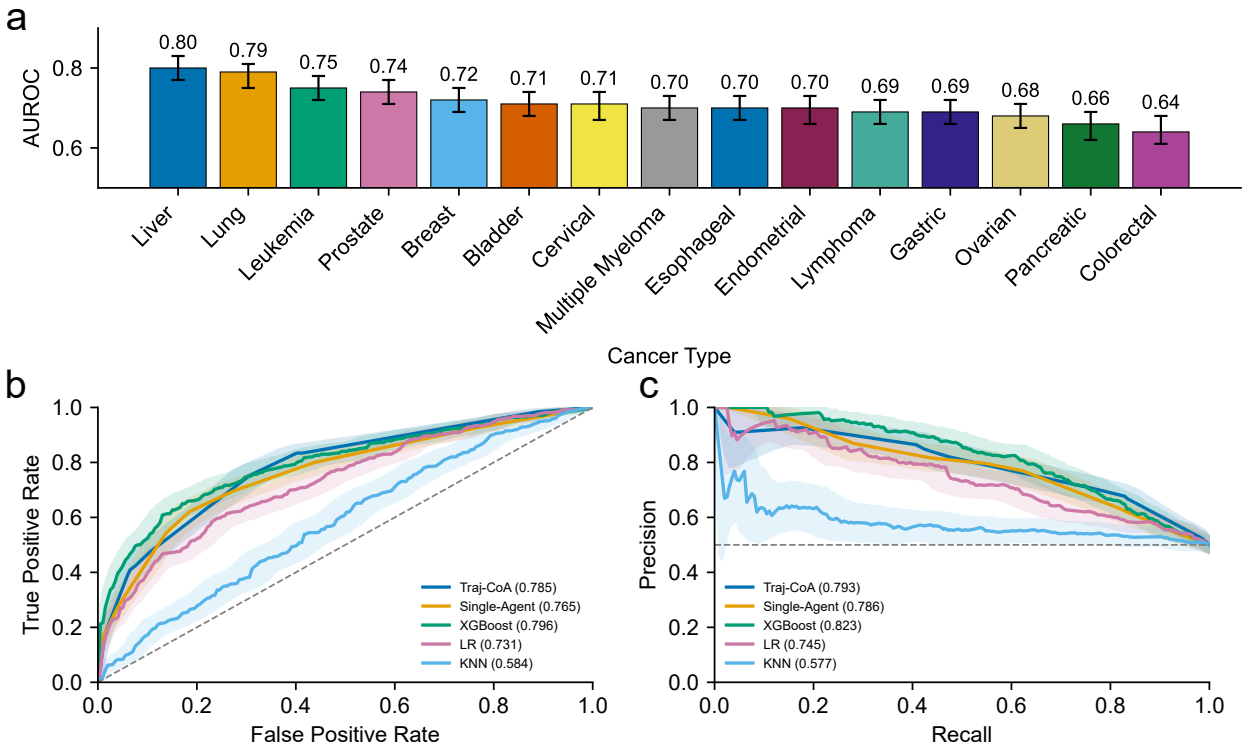

Figure 5.4: **Zero-shot prediction performance of Traj-CoA on Truveta data. a.** Area under receiver operating characteristic (AUROC) of cancer-specific prediction on each cancer type; **b.** ROC curves and **c.** precision-recall curves for different methods on the lung cancer benchmark.

Benchmarking on Lung Cancer Cohort

To benchmark zero-shot performance, we compared Traj-CoA with logistic regression (LR), k-nearest neighbors (KNN), XGBoost [32], and a single-agent baseline on an age- and sex-matched lung cancer cohort of 500 random cases and 500 controls. For supervised models, we derived summary feature vectors from longitudinal EHR data, retaining features with at least 1% frequency (1,253 features) following prior work [106]. The single-agent baseline used the same base LLM, the same XML input representation, and the same final risk output format as Traj-CoA, but processed the entire available input in one pass. Supervised models were trained and validated on a separate dataset of 170,464 cases and 166,903 controls using an 8:2 split, while LLM-based methods remained zero-shot prediction.

On this benchmarking cohort, Traj-CoA achieved an AUROC of 0.785, outperforming the single-agent baseline (AUROC = 0.765) and substantially exceeding the performance of logistic regression and KNN, while performing slightly below XGBoost (AUROC = 0.796) (Figure 5.4b,c). These results indicate that Traj-CoA, despite not being trained specifically for cancer prediction, achieves performance comparable to a trained gradient boosting model and markedly surpasses conventional linear and distance-based approaches for lung cancer early detection.

Sensitivity Analyses

To evaluate the scalability and efficiency of Traj-CoA, we conducted sensitivity analyses examining (1) base model selection, (2) patient trajectory length, and (3) context detailedness.

Base model selection We replaced the default base model (GPT-4.1-mini) with alternative GPT models, including GPT-4o, GPT-4.1-nano, GPT-4.1, GPT-5-nano, GPT-5-mini, and GPT-5. For reasoning models, we used a medium reasoning effort setting. All models were evaluated within the identical Traj-CoA multi-agent framework on the lung cancer benchmark.

We observed that more advanced reasoning models (e.g., GPT-5) did not substantially

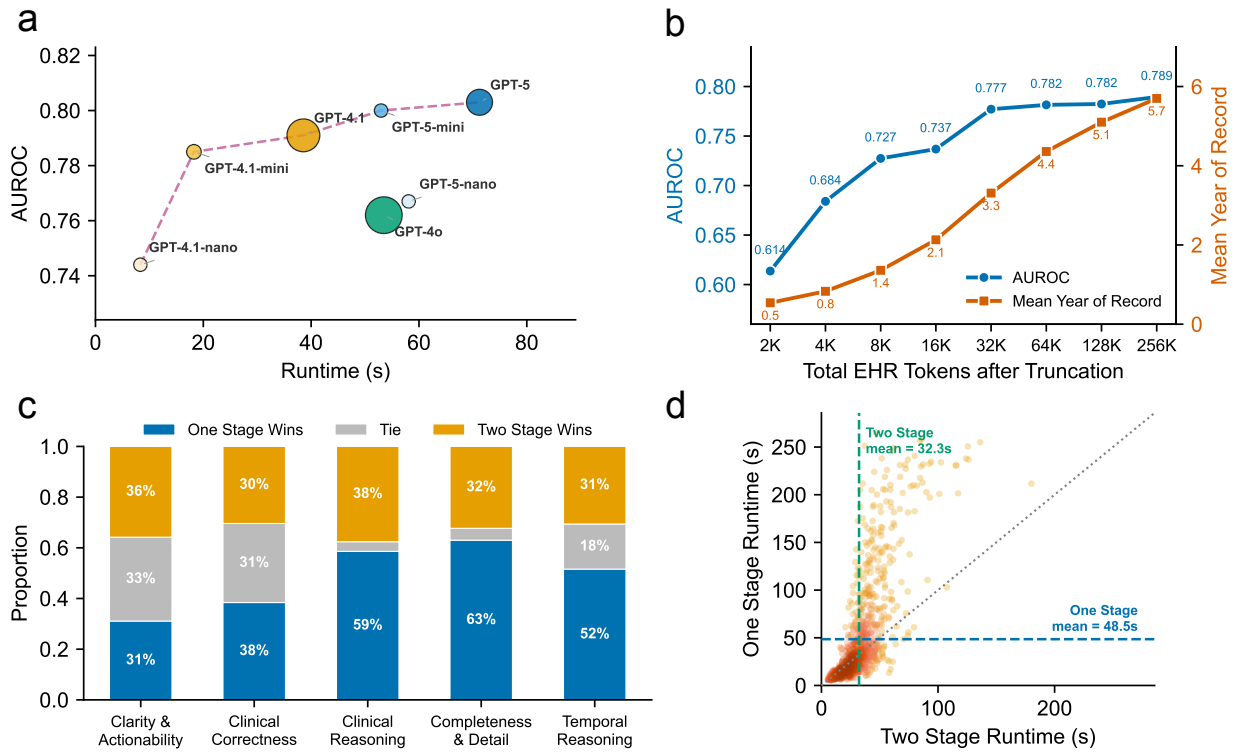

Figure 5.5: **Sensitivity analyses on the scalability and efficiency of Traj-CoA on Truveta data. a.** AUROC and median runtime of Traj-CoA on the lung cancer benchmark using different base models. The radius of each point is proportional to the API price of the input token of the model. The pink dashed line represents the Pareto front. Among these base models, GPT-4.1-mini achieved a balance between performance, latency, and cost. **b.** Improved predictive performance when expanding the maximum length of left-truncated patient trajectory. **c.** LLM-as-a-judge comparison between one-stage Traj-CoA and two-stage Traj-CoA. The one-stage model had better clinical reasoning, completeness of detail, and temporal reasoning than the two-stage model. **d.** The two-stage model achieved lower latency especially for very long EHR sequences.

outperform smaller models such as GPT-4.1-mini (Figure 5.5a). We hypothesize that this is because the Traj-CoA architecture decomposes complex EHR understanding into smaller,

well-defined subtasks with structured inter-agent communication and summarization. This design reduces the need for complex global reasoning within a single model invocation, thereby diminishing the performance advantage typically associated with larger reasoning models on complex tasks.

Within each model family, larger models generally achieved improved predictive performance, but at the expense of higher computational cost and increased latency (Figure 5.5a). Overall, GPT-4.1-mini provided the most favorable balance between runtime, cost, and predictive performance, and was therefore selected as the default base model.

Patient trajectory length To assess the impact of patient trajectory length, i.e., how far back in a patient's history Traj-CoA uses for prediction, we conducted experiments on the lung cancer benchmark using left-truncated EHRs with varying maximum token limits, simulating scenarios where only partial historical data are available. Specifically, we truncated the input EHR XML to retain only the most recent 2k to 256k tokens before prediction.

Increasing the trajectory length from 2k to 256k tokens extended the average history from 0.5 to 5.7 years and improved AUROC from 0.614 to 0.789 on the same benchmark (Figure 5.5b). These results indicate that longer patient trajectories consistently improve predictive performance, suggesting that richer longitudinal clinical information provides stronger signals for cancer risk assessment. Overall, the findings demonstrate that Traj-CoA benefits from extended patient records when available, supporting its scalability and applicability to comprehensive real-world EHR data.

Context detailedness For patients with very long EHRs, the number of sequential worker agents increases proportionally, resulting in high latency due to the inherently sequential design of the framework. To mitigate this limitation, we evaluated a two-stage architecture (Figure C.1). In this variant, each EHR chunk was first processed independently by a preprocessor agent in parallel. The preprocessor extracted informative events, which were then concatenated into a shorter, structured XML representation for downstream consumption

by the Traj-CoA pipeline. This design reduces the number of sequential worker agents and converts part of the workflow from sequential to parallel execution.

The two-stage approach reduced average latency by 25% relative to the one-stage model (32.3 vs. 48.5 s), with larger gains in patients with very long EHRs (Figure 5.5d). Theoretical complexity analysis was provided in Appendix C.2.1. Predictive performance was largely preserved (AUROC, 0.78 vs. 0.79). However, compared with the original Traj-CoA framework, the two-stage model produced less detailed summaries and weaker clinical reasoning (Figure 5.5c), likely because parallel preprocessors operated without global context, resulting in less complete information extraction and a more abstract XML representation of the EHR. These findings indicate that parallel preprocessing in the two-stage model improves efficiency for very long EHRs, but may reduce summary fidelity and reasoning quality, highlighting a trade-off between latency and interpretability.

5.3.4 Traj-CoA Generates Clinically Meaningful Patient-level Summaries

Beyond producing a final risk score, Traj-CoA's sequential worker agents maintain an evolving summary over time, while the manager agent generates a final structured patient-level summary for each individual. These summaries provide a condensed view of the patient trajectory, enhancing interpretability. We evaluated the quality and clinical relevance of these summaries through: (i) comparison with a single-agent baseline using an LLM-as-a-judge framework [237] (Section 5.3.4), (ii) human evaluation assessing the fidelity of the summary (Section 5.3.4), and (iii) a case study illustrating how the worker agents dynamically aggregate clinical events and update the predicted risk (Section 5.3.4).

Comparison with a Single-Agent Baseline using LLM-as-a-judge Evaluation

Traj-CoA was designed to enhance temporal reasoning over longitudinal EHR through its CoA and LTM architecture. To assess whether this design improves summary quality, we conducted a pairwise LLM-as-a-judge evaluation [237] comparing Traj-CoA with a single-agent LLM baseline on the lung cancer case cohort. Evaluations were performed using GPT-5 configured

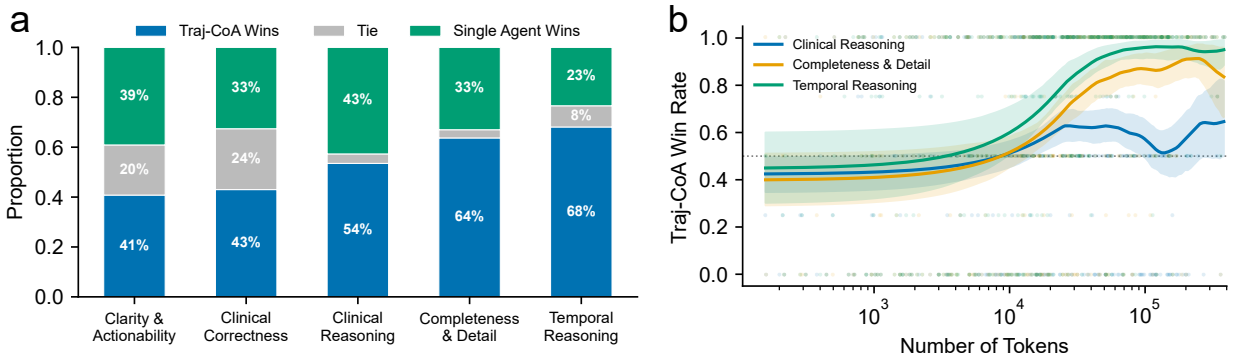

Figure 5.6: **LLM-as-a-judge evaluation of Traj-CoA versus a single-agent baseline. a,** Pairwise comparison across five dimensions using GPT-5 with high reasoning effort as the judge. Bars show the proportions of Traj-CoA wins (blue), ties (gray), and single-agent wins (green). Traj-CoA outperforms the baseline across all dimensions, with the largest advantage in Temporal Reasoning (68% wins vs. 23% losses). **b,** Traj-CoA win rate by EHR length (input tokens; log scale) for selected dimensions. Solid lines indicate smoothed win rates and shaded areas indicate confidence intervals. Each scattered point represents the average win rate over two runs with different candidate orders. For each run, "win" assigns a score of 1.0, "tie" assigns a score of 0.5, and "loss" assigns a score of 0.0.

with high reasoning effort. We defined five prespecified evaluation dimensions: (i) Clarity & Actionability, (ii) Clinical Correctness, (iii) Clinical Reasoning, (iv) Completeness & Detail, and (v) Temporal Reasoning. To mitigate the position bias from the order of two candidates [169], we report the average win/tie/loss rates over two runs with different candidate orders.

Across all five dimensions, Traj-CoA achieved more wins than losses compared to the single-agent model (Figure 5.6a), with the largest advantage in temporal reasoning (68% wins vs. 23% losses). It also outperformed in completeness and detail (64% vs. 33%) and clinical reasoning (54% vs. 43%), indicating more comprehensive and better-integrated summaries. Notably, this advantage increased with longer EHRs (Figure 5.6b), suggesting that the multi-agent design and memory mechanisms are particularly effective for long and

complex patient trajectories.

Human Evaluation

To assess the fidelity of Traj-CoA-generated summaries, we conducted a human evaluation with two independent annotators (a PhD candidate in biomedical informatics and a medical doctor). Annotators assessed whether each identified cancer-related event accurately reflected the corresponding EHR evidence, labeling events as *correct* or *inaccurate*. We randomly sampled 30 patients from 15 cancer case-control test sets, yielding 113 events. Error modes were then summarized from the inaccurate events. Disagreements between annotators were resolved through discussion.

Cohen's kappa was 0.72, indicating substantial inter-annotator agreement. Of the 113 evaluated events, 102 (90.27%) were annotated as correct. The remaining errors were primarily attributable to temporal inaccuracies (n = 5) and interpretive inaccuracies (n = 6). Temporal errors typically occurred when Traj-CoA identified an event as "new" even though it had been documented previously. Interpretive errors were mainly due to minor hallucinations or overly assertive inferences, such as adding unsupported qualifiers (e.g., referring to a "*fasting* glucose test" when fasting status was not documented in the original EHR) or failing to preserve clinical uncertainty. In some cases, the system described findings such as "suggestive of disease progression" or "significantly elevating liver cancer risk" although the available evidence could also have been consistent with alternative explanations, such as nonspecific abnormalities or a benign lesion. Overall, these inaccuracies were infrequent and did not materially compromise the clinical validity or usefulness of the patient summaries.

Case Study

To illustrate the dynamic nature of Traj-CoA's risk assessment, we present a longitudinal case study tracking a single patient's clinical trajectory from 2012 to 2020 (Figure 5.7). As worker agents process temporal EHR chunks, they extract key events into LTM and continuously update risk. For example, active smoking in 2012 was recognized and increased the risk to

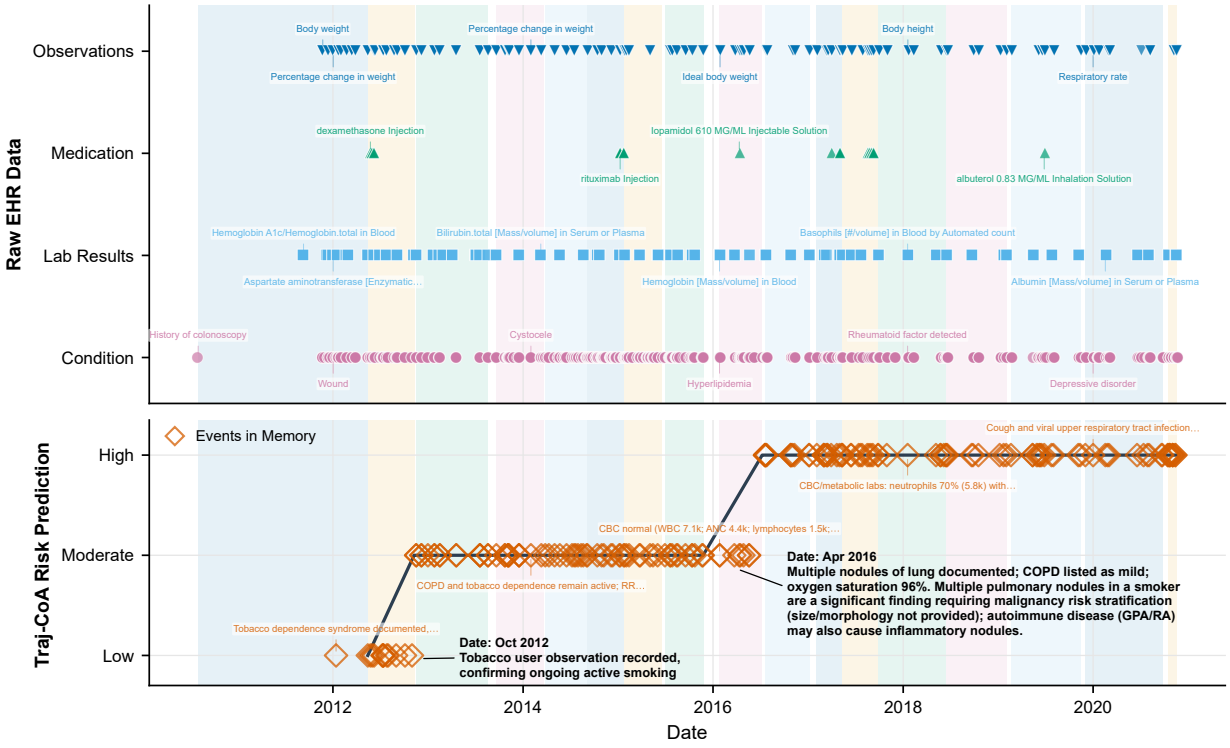

Figure 5.7: **Dynamic short-term risk prediction and event aggregation by Traj-CoA worker agents.** The top panel displays the patient's raw EHR data, encompassing observations, medications, lab results, and conditions from 2012 to 2020. Different temporal chunks processed sequentially by the worker agents are distinguished by alternating background colors. The bottom panel shows Traj-CoA's continuously updated short-term risk prediction corresponding to the timeline. Key clinical insights synthesized by the agents and stored as events in LTM are explicitly marked with diamond symbols along the risk trajectory. Two critical updates are highlighted: (1) In October 2012, an observation confirming ongoing active smoking shifts the baseline risk. (2) In April 2016, the discovery of multiple pulmonary nodules and mild COPD triggers a risk escalation to high, with the system's memory specifically noting the need for malignancy risk stratification in a smoker, while remaining robust to alternative inflammatory etiologies such as autoimmune disease.

moderate, while the emergence of multiple lung nodules in 2016, combined with prior smoking history and COPD, escalated the risk to high. This demonstrates how Traj-CoA's sequential agents conduct clinical reasoning by synthesizing historical context with emerging events to dynamically update patient-level risk.

5.3.5 Traj-CoA Generates Population-level Insights in Cancer Risk

Each patient-level summary contains the cancer-related events identified by Traj-CoA, enabling aggregation into population-level insights for the cancer. In this section, (i) we use lung cancer as a case study to characterize the thematic EHR signals underlying Traj-CoA's risk estimates for this cancer type (Section 5.3.5), and (ii) we summarize the top five themes identified across all 15 cancer types together with the cross-cancer patterns derived from these summaries (Section 5.3.5). In the Appendix C.2.3, we further analyze shared temporal patterns in predicted risk trajectories and the reasons for risk transitions derived from worker agents' outputs. These analyses illustrate Traj-CoA's utility not only as a predictive model, but also as a framework for generating interpretable population-level insights from longitudinal EHR data.

Thematic EHR Signals Underlying Lung Cancer Risk Estimation

We use lung cancer as an example to illustrate how the cancer-related events in the final summaries can be aggregated for population-level interpretation. Specifically, we applied a topic modeling approach similar to TopicGPT [142] to uncover common themes within these cancer-related events. In lung cancer, the events identified by Traj-CoA form distinct, clinically interpretable clusters (Figure 5.8a), with the dominant themes (Figure 5.8b) reflecting established risk factors, comorbidities, and early symptoms across various event categories, including advanced age [108], smoking exposure, COPD [191], respiratory symptoms like shortness of breath and chronic cough [148], lung nodules [115], and laboratory abnormalities like anemia [144]. Traj-CoA also captured signals associated with benign or stable trajectories, such as normal imaging and serially normal laboratory findings, indicating that its predictions

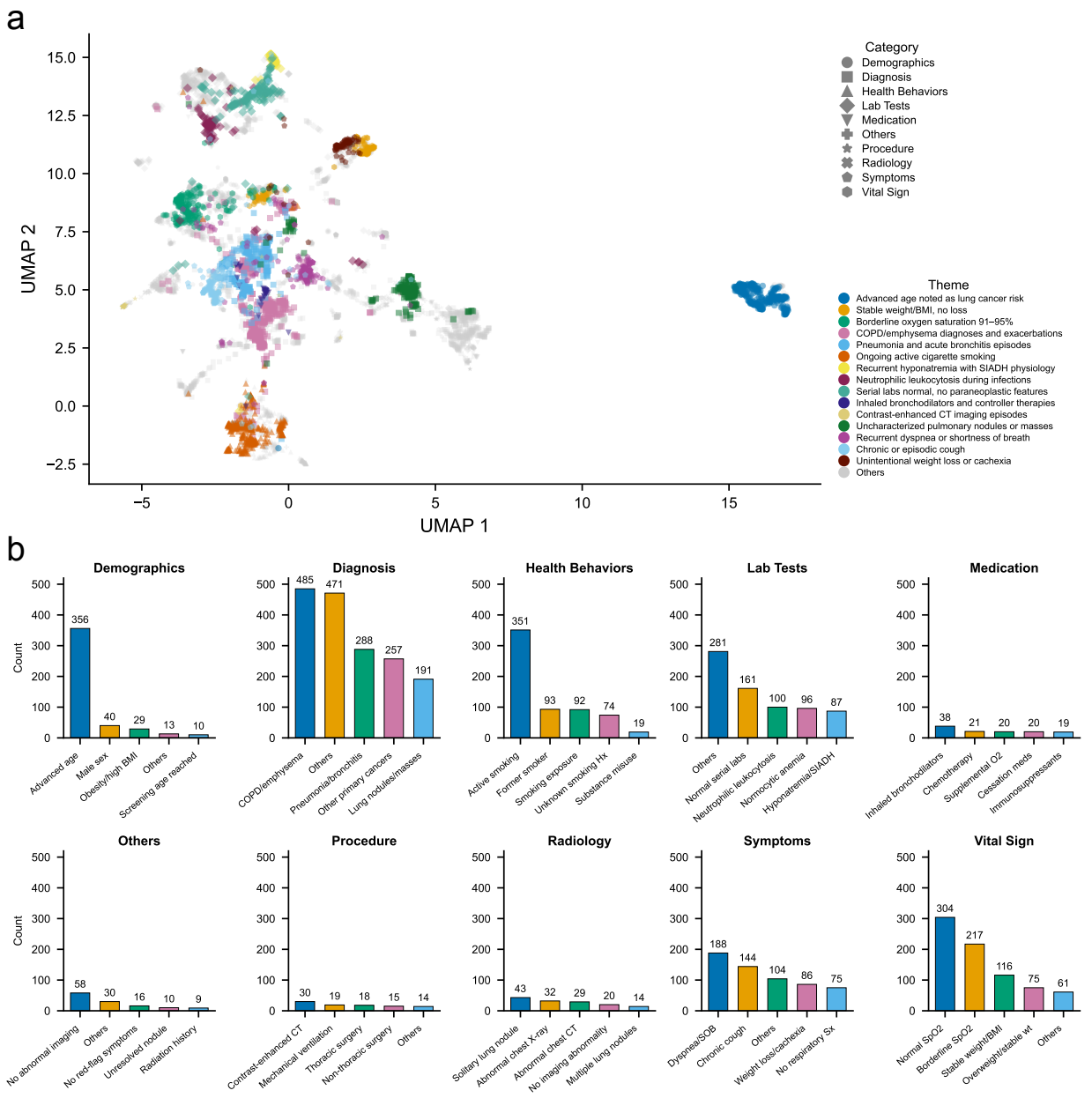

Figure 5.8: **Thematic patterns of lung cancer-related events identified by Traj-CoA among case patients. a.** UMAP projection of cancer-related events output by Traj-CoA, demonstrating coherent clustering of events into clinically interpretable themes across demographic factors, comorbid diagnoses, health behaviors, laboratory findings, medications, procedures, radiology, symptoms, and vital signs. **b.** Distribution of the most frequently identified events within each category, highlighting dominant population-level signals such as smoking exposure. Note that the counts were the mentions of all identified events within each category, which can be more than one for each patient.

integrate both positive and negative evidence. These patterns not only illustrate that the aggregation of Traj-CoA's output summaries recovers known clinical associations, but also show how Traj-CoA integrates these multidimensional clinical signals into interpretable estimates of lung cancer risk.

Cross-Cancer Relationships

Extending the topic modeling analysis across all 15 cancer types showed that Traj-CoA can generate interpretable, data-driven insights across diseases (Figure 5.9). Using each patient's full set of identified cancer-related events as an individual summary, the summary-level UMAP analysis [121] revealed that patients with the same cancer type tend to occupy nearby regions of the latent space, indicating that Traj-CoA captures meaningful within-cancer signal structure rather than diffuse, nonspecific comorbidity patterns (Figure 5.9a). Consistent with this, the similarity matrix showed greater thematic resemblance within major cancer groupings, including gastrointestinal, hematologic, and genitourinary cancers, while still preserving disease-specific profiles (Figure 5.9b). The prevalence analysis further clarified the dominant signals underlying each cancer risk and cross-cancer relationships, consistent with clinical knowledge. For instance, colorectal cancer was characterized by metabolic factors [216, 122] and iron-deficiency anemia [28] (Figure 5.9c). From these prevalent themes, common themes among cancer types can be derived. Examples of these derived cross-cancer relationships include that gastrointestinal cancers were linked by metabolic and nutrition-related themes [27], hematologic malignancies were linked by cytopenia- and anemia-related symptoms [80], and lung and bladder cancer shared smoking-associated signals [185]. Obesity-related and broader metabolic themes recurred across multiple cancers. These results highlight Traj-CoA's utility not only for prediction, but also for generating population-level insight into shared and cancer-specific themes from longitudinal EHR data.

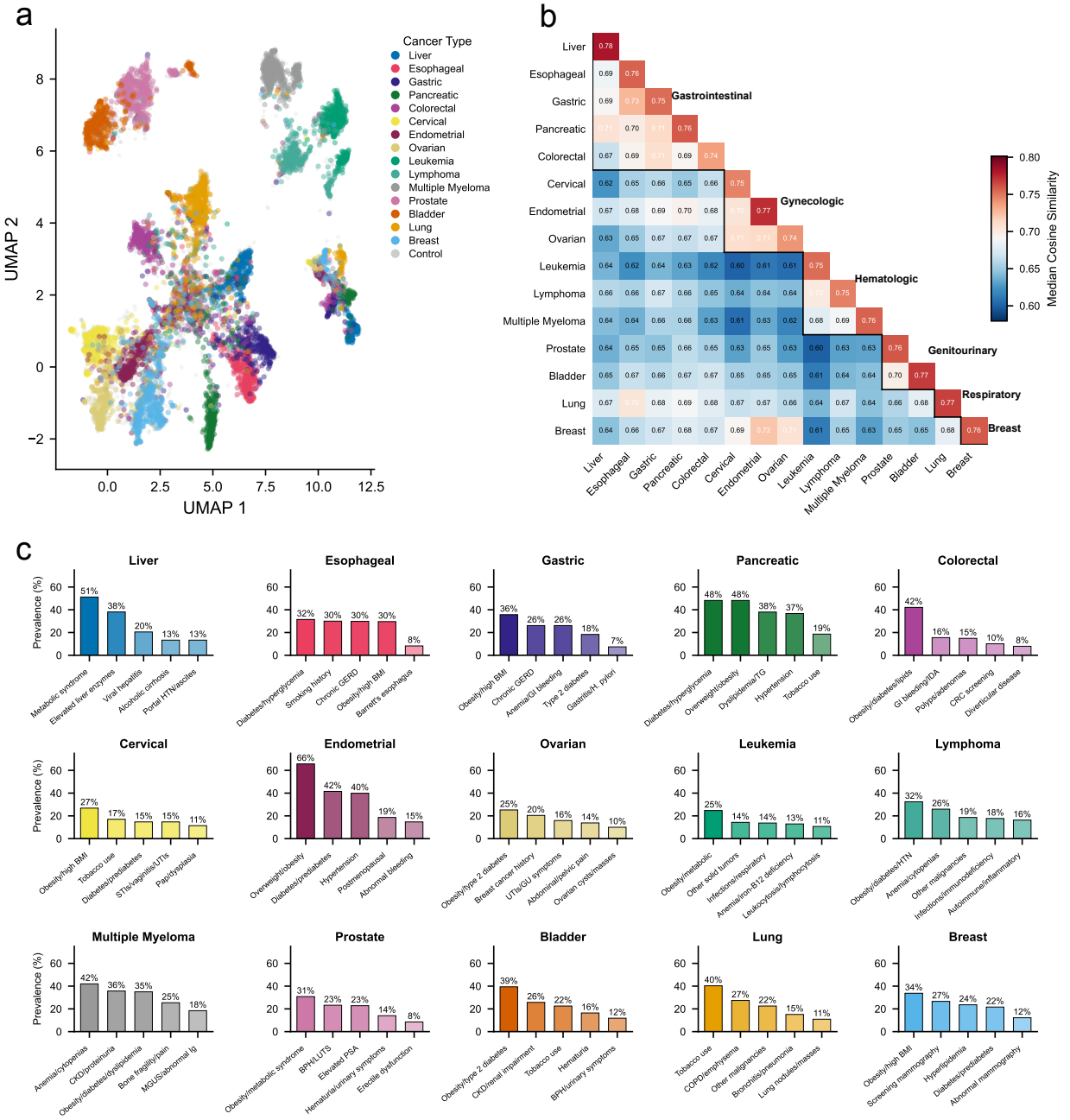

Figure 5.9: **Cancer-specific themes and relationships derived from patient summaries.** **a.** UMAP of patient summaries colored by cancer type, showing that patients with the same cancer tend to cluster together. **b.** Heatmap of median cosine similarity between patients from pairwise cancer types, highlighting higher within-cancer similarity and relationships among related cancers. **c.** Top prevalent themes identified for each cancer type from patient summaries.

5.3.6 Discussion

In this study, we applied Traj-CoA for 1-year multi-cancer early detection from longitudinal structured EHR data. Across 15 cancer types, Traj-CoA showed variable discrimination in a zero-shot setting, with AUROCs ranging from 0.64 to 0.80, and the majority of them exceeding 0.70. It achieved performance on the lung cancer benchmark comparable to that of a supervised XGBoost model, despite requiring no task-specific training. Beyond risk estimation, the framework generated patient-level summaries linked to temporally ordered clinical evidence and recovered clinically coherent population-level themes across cancers, suggesting Traj-CoA's potential in both prediction and interpretation from routine EHR data.

From a clinical perspective, this study lies in the area of multi-cancer early detection, where longitudinal history, risk factors, early symptoms, and physiologic changes may jointly indicate short-term cancer risk. This setting is clinically relevant because it could support earlier diagnosis and more timely screening [39]. Prior EHR-based approaches have shown promise in multi-cancer early detection [81, 138], but have often required supervised training and offered limited interpretability. In contrast, Traj-CoA harnesses LLMs to demonstrate consistent *zero-shot* performance on Truveta data spanning more than 30 US healthcare systems and produces rich thematic interpretation from the summaries consistent with clinical knowledge.

Traj-CoA demonstrated variable performance across the evaluated cancer types, with the strongest performance observed in liver cancer and more challenging results in colorectal cancer, aligning with prior studies [235, 81]. It remains unclear exactly why this variability exists. It is possible that some cancers inherently lack detectable clinical signals in structured EHR data more than 1 year prior to diagnosis, often presenting instead with nonspecific symptoms closer to diagnosis [47]. To analyze whether this was the case, we performed sensitivity analyses using time gaps of 0.5, 1, 2, 3, 4, and 5 years (Appendix C.2.4). Performance generally improved as the time gap decreased, suggesting the presence of stronger predictive signals

nearer to diagnosis (Figure C.4). However, these gains were not uniform across cancer types, supporting variable cancer signal evolutions. For example, performance for liver and lung cancer appeared to saturate at a 1-year gap, whereas colorectal cancer showed substantially better performance at 0.5 years than at 1 year.

5.4 Conclusion

In this chapter, we developed Traj-CoA, a training-free multi-agent framework designed to perform temporal reasoning on long and noisy multimodal longitudinal EHR data for patient trajectory modeling. Results of Traj-CoA on distinct data from UWMC and Truveta and on different clinical tasks have shown its generalizability over different populations, clinical tasks, and data elements, and its interpretability in identifying cancer-related events in cancer early detection tasks.

The main technical contribution of Traj-CoA is a zero-shot, multi-agent framework that combines temporal decomposition with explicit memory to enable interpretable and generalizable modeling of long patient trajectories. By converting multimodal EHR data into a unified, LLM-friendly XML representation and organizing it into temporally coherent chunks, Traj-CoA avoids extensive feature engineering, which is costly and difficult to generalize. Its CoA design, coupled with LTM, separates local signal extraction from global synthesis, allowing the model to capture complex temporal dependencies that are difficult for single-agent LLMs due to long-context limitations such as the "lost-in-the-middle" effect [112].

Compared to conventional EHR ML models that require task-specific training [106, 221], EHR foundation models with limited interpretability [229, 160, 171], and prior single-agent LLM approaches that struggle with long-context reasoning [40, 91, 200], Traj-CoA explicitly structures temporal reasoning and embeds interpretability directly into the inference process through evidence-linked summaries and extracted clinical events. Its zero-shot formulation further enables generalization across populations, clinical tasks, and data elements through minimal prompt adaptation.

Traj-CoA demonstrates favorable scaling properties in base model size and patient trajec-

tory length. First, the CoA with LTM design decomposes a complex longitudinal reasoning task into simpler sequential subtasks, enabling effective temporal reasoning even with smaller models such as GPT-4.1-mini and MedGemma, with only a modest performance gap between GPT-4.1-mini and larger and more expensive models such as GPT-5 in the Truveta study. Second, performance improved as more historical patient trajectory was retained in both cohorts, suggesting that Traj-CoA can effectively leverage richer longitudinal EHR context. In addition, Traj-CoA's output quality advantage over a single-agent baseline increased with longer patient trajectories in the Truveta study, further supporting its strength in long-horizon temporal reasoning.

Broadly, our findings suggest that CoA reasoning with LTM constitutes an interpretable and generalizable framework for modeling long-term patient trajectories. Integrating tool use or agent skills [208] and specialized subagents for distinct modalities such as lab values may enable more flexible and complete reasoning across heterogeneous data streams. Emerging medical agent frameworks [73, 236, 104] suggest a direction in which multiple agents with complementary expertise jointly deliberate over complex cases. Traj-CoA's temporal reasoning could provide longitudinal enrichment to such often cross-sectional multi-agent designs, thereby paving the way for more dynamic and adaptive clinical problem-solving across various long-horizon tasks.

5.5 Limitations

Although Traj-CoA is a task-agnostic framework, it requires carefully crafted prompts for specific tasks to optimize the performance. Research into prompt optimization [158] or data-driven hypothesis generation [151, 152] could reduce this dependency and allow for more general modeling of patient trajectories. From a study design perspective, we used 1:10 and 1:1 case-control matching in the UWMC and Truveta studies, respectively. Although these do not reflect true cancer incidence, AUROC provides an unbiased metric for the performance. In a sensitivity analysis using an unmatched lung cancer cohort with a 1:250 case-to-control ratio using Truveta data, which approximated the cumulative incidence rate in the database,

Traj-CoA achieved a higher AUROC of 0.871 (95% CI, 0.855–0.885) likely due to age as a major predictor. Finally, while we have validated the fidelity of LLM-generated summaries and themes through human evaluation and known clinical knowledge, inaccuracies exist and they should be interpreted or used with caution.

Chapter 6

TRAJ-EVOLVE: A SELF-EVOLVING MULTI-AGENT SYSTEM FOR PATIENT TRAJECTORY MODELING IN LUNG CANCER EARLY DETECTION

This chapter is adapted from a paper titled "Traj-Evolve: A Self-Evolving Multi-Agent System for Patient Trajectory Modeling in Lung Cancer Early Detection" authored by Sihang Zeng, Matthew Thompson, Ruth Etzioni, and Meliha Yetisgen, under review at EMNLP 2026.

In this chapter, we present Traj-Evolve, a self-evolving multi-agent system built on top of Traj-CoA for early lung cancer detection. Traj-Evolve has two self-evolving mechanisms, an experience pool (ExPool) and multi-agent reinforcement learning (MARL). ExPool contains past verified predictions and their ground truth diagnostic status, explicitly providing "patients-like-me" when predicting for a new patient. In contrast, MARL internalizes experience into the parameters of the worker and manager agents, facilitating better inter-agent and agent-memory collaboration.

We evaluate Traj-Evolve on a large longitudinal cohort from the University of Washington Medical Center (UWMC), matched at a 1:10 case-to-control ratio, using five years of multimodal EHR history to predict incident lung cancer within the subsequent year. We benchmark against a comprehensive suite of baselines spanning clinical risk models, classical machine learning, sequential deep learning, clinical BERT-based models, and state-of-the-art LLM-based systems, on an overall population and a never-smoker population. We further demonstrate the self-evolving dynamics of the two mechanisms.

The main contributions of this work are:

- We introduce Traj-Evolve, to our knowledge, the first self-evolving multi-agent framework

for longitudinal EHR modelling applied to a real-world clinical prediction task.

- We design two complementary evolutionary mechanisms: an evolving ExPool that provides non-parametric, few-shot "patients-like-me" retrieval, and a MARL procedure that parametrically optimises inter-agent and agent-memory collaboration using rejection-sampled high-reward trajectories.

- We demonstrate that Traj-Evolve achieves state-of-the-art discrimination for one-year lung cancer prediction in the overall population and the never-smoker population.

- We provide detailed analyses of the self-evolving dynamics, supporting the vision of a continuously improving clinical decision-support system.

6.1 Methods

6.1.1 Study Design and Participants

This retrospective case-control study utilized a UWMC longitudinal dataset to predict first primary lung cancer diagnoses within a one-year prediction period [239]. The index date for prediction of both cases and controls was defined as the completion time of a qualifying chest-related radiology exam (chest CT, abdomen CT, or chest X-ray). To model patient trajectories, up to five years of EHR history prior to the index date was utilized. This multimodal EHR data encompasses both structured records (diagnosis, medication, lab, vital, and procedure) and unstructured text (clinical notes and radiology reports).

Eligible patients were over 40 years old. Cases were defined as individuals with a valid primary lung cancer diagnosis, excluding those with prior cancers. To ensure predictions informed early detection and excluded active diagnostic workups, the index exam for cases was required to be completed between two months and one year prior to diagnosis. Controls were defined as individuals with no cancer registry records of any cancer type; their index exams were selected to guarantee at least three years of cancer-free follow-up. To ensure

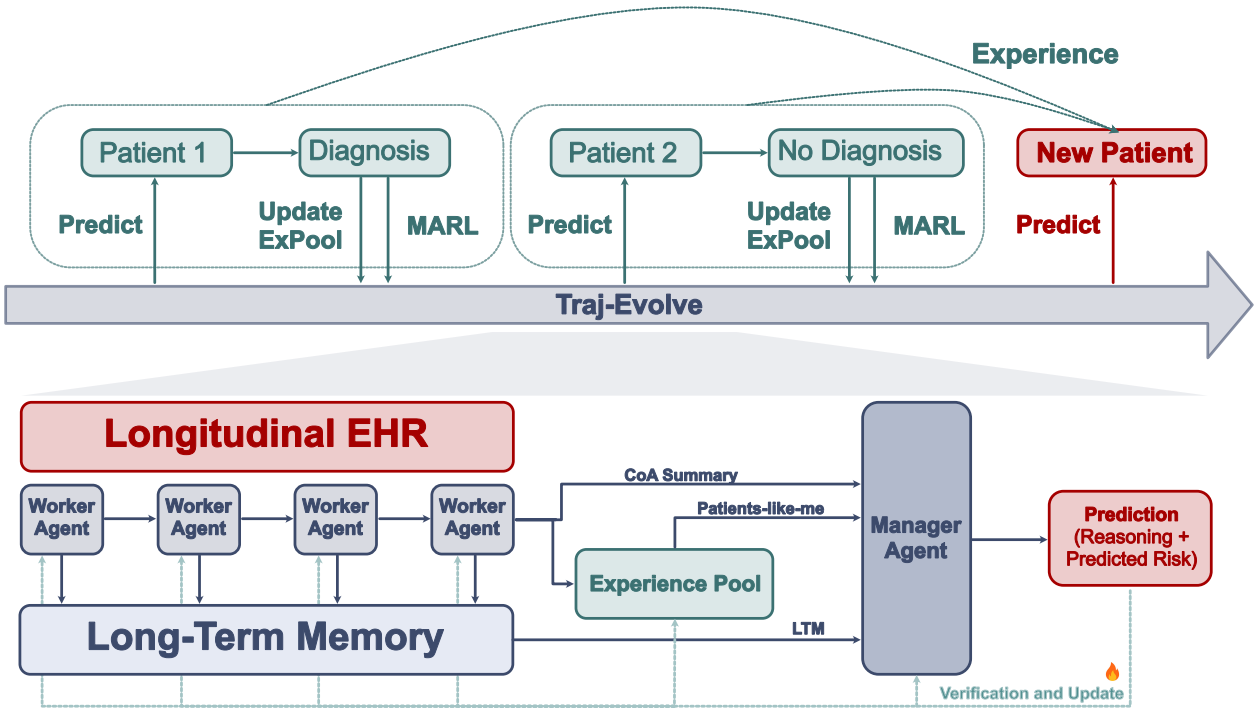

Figure 6.1: **Overview of the Traj-Evolve architecture and self-evolving workflow.** The top panel illustrates the self-evolving process, wherein the system accumulates experience from prior verified patients to iteratively update the ExPool and optimize model parameters via MARL. The bottom panel details the pipeline. Longitudinal EHRs are processed sequentially by worker agents to summarize the patient journey and distill relevant temporal events into an LTM module. A manager agent subsequently synthesizes the final worker summary, memory contents, and retrieved historical cases ("patients-like-me") from the ExPool to generate a final lung cancer risk prediction alongside clinical reasoning. Verified predictions are then used to update the ExPool and the agents for self-evolution.

longitudinal data, included patients should have an active clinical encounter history exceeding 120 days within the prior five years. Controls were matched to cases based on the index exam type and date (overlapping within a three-month window) using a 1:10 case-to-control matching schema.

The matched cohort was randomly partitioned into mutually exclusive training (n=13,629) and validation (n=300) sets, with the remainder assigned to a held-out test set. To assess model performance and generalizability, we constructed two evaluation cohorts via independent random sampling from this test set: an overall cohort of 1000 patients (90 cases, 910 controls) encompassing all smoking statuses, and a never-smoker cohort of 835 patients (27 cases, 808 controls).

6.1.2 Data Preprocessing and Temporal Reasoning

To prepare the longitudinal EHR for comprehensive temporal reasoning by Traj-Evolve, we followed the data preprocessing methodology established in our previous work [220, 222]. Specifically, multimodal longitudinal EHR were unified into a LLM-friendly XML format and segmented into chronologically ordered chunks to avoid LLMs' performance degradation on long-context data.

Similar to our previous work, Traj-Evolve equips its temporal reasoning capability through chain-of-agents (CoA) and a long-term episodic memory (LTM). Within this architecture (Figure 6.1), a sequence of worker agents iteratively processes the time-aware XML chunks. As each agent reviews a specific temporal window, it updates an accumulated patient summary, effectively capturing the progression of the patient's health status over time. Ultimately, the final worker agent produces a comprehensive unstructured summary that encapsulates the entire five-year observation period. Simultaneously, these agents identify and extract salient lung cancer-related clinical events, such as emerging symptoms, relevant diagnoses, and medical interventions, distilling them into LTM. This episodic memory module filters out the extensive noise in raw EHR irrelevant to lung cancer, yielding a condensed, structured timeline of the patient's longitudinal history. A manager agent then synthesizes the last worker's summary alongside the distilled temporal events within the LTM to formulate a final reasoning trace and risk estimation (an integer between 1 and 10) for lung cancer.

However, while this multi-agent mechanism successfully models the temporal dynamics of an individual patient's trajectory, its zero-shot nature treats each patient in isolation, relying

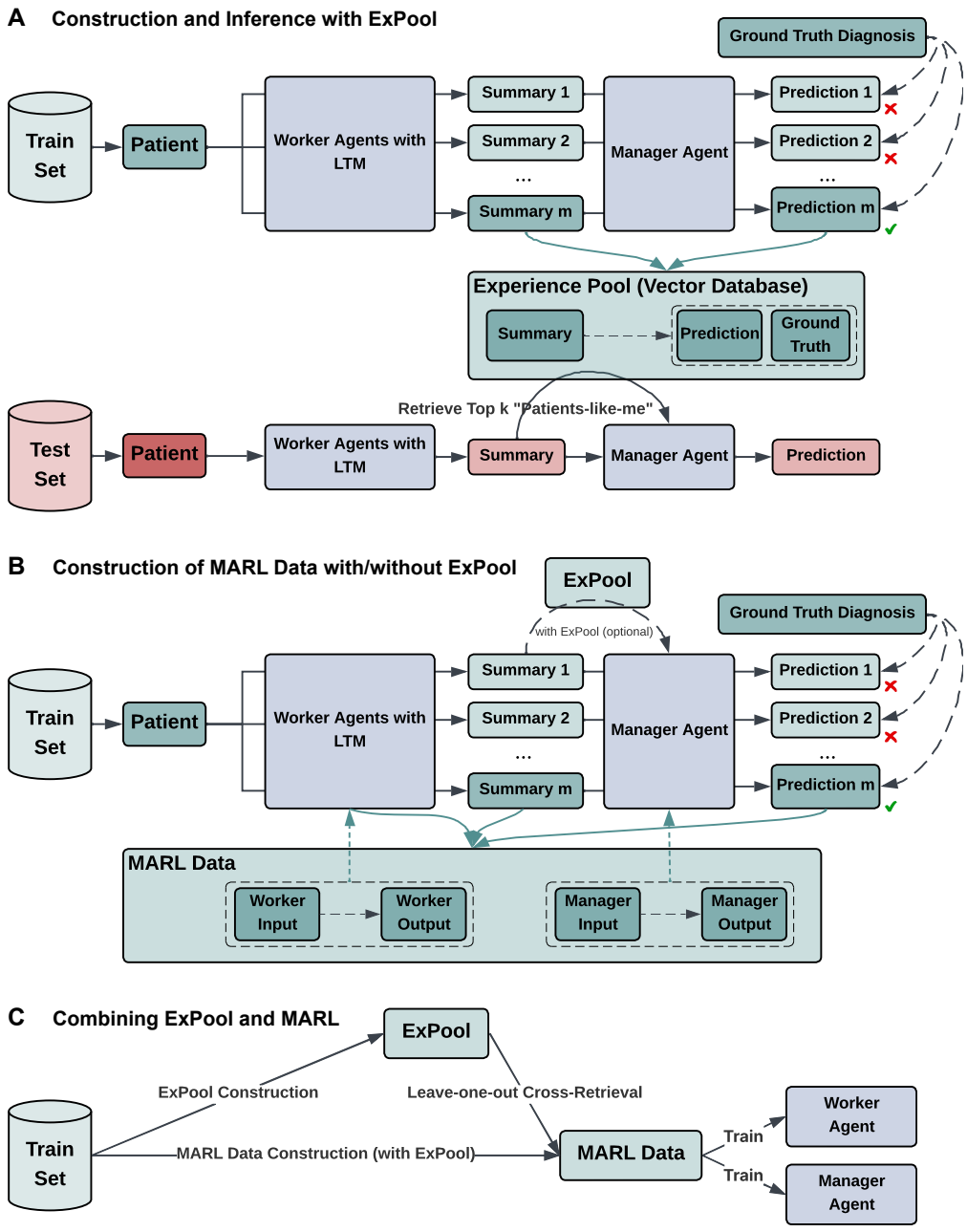

Figure 6.2: **Methodological design of the self-evolving mechanisms.** (A) Construction and inference pipeline for the ExPool. (B) Construction of MARL training data. (C) Integration of ExPool and MARL. The training set is utilized to construct an initial ExPool. A leave-one-out cross-retrieval strategy is then applied to generate context-augmented MARL training data, which then optimizes the worker and manager agents.

solely on the LLM's internal knowledge. This diverges from expert clinical practice, which also relies on past experience on similar patients and reasoning processes (Figure 6.1). To bridge this gap and emulate experiential learning, we introduce two self-evolving mechanisms for Traj-Evolve: an evolving experience pool (ExPool) and multi-agent reinforcement learning (MARL).

6.1.3 Evolving Experience Pool (ExPool)

To emulate the accumulation of clinical experience over time, we developed an evolving procedural memory module termed ExPool (Figure 6.2A). As Traj-Evolve generates predictions that can be subsequently verified against ground-truth diagnostic status (e.g., confirmed cancer diagnosis or confirmed benign status), these verified reasoning traces can serve as experience for future cases. Correct predictions may offer reusable diagnostic patterns, while incorrect predictions may expose model vulnerabilities and trigger self-reflection. These data-driven patterns evolve the system beyond its current boundary, forming the self-evolving capability. To capture the current system's boundary, i.e., the optimal diagnostic ability, we constructed ExPool using a rejection sampling approach that selected best-of-N from a set of roll-out reasoning traces. For each patient in the training set, we generated m independent reasoning roll-outs utilizing an elevated LLM temperature to induce trajectory diversity. The optimal reasoning trace for each patient was selected based on prediction accuracy: the trace yielding the highest predicted risk was chosen for confirmed cases, and the lowest predicted risk for controls. Notably, these optimal reasoning traces may mix correct and incorrect predictions, which provides different signals for ExPool. These selected traces populate ExPool, a vector database where the retrieval keys are the final worker agent summaries (encoded into embeddings via `nomic-embed-text-v1.5` [133]), and the values comprise the manager agent's reasoning, predicted risks, and actual diagnostic status. ExPool expands dynamically as new verified patient trajectories are incorporated.

During inference, ExPool functions using a retrieval-augmented generation (RAG) [100] mechanism (Figure 6.2A). For each new patient, the final worker summary is encoded using

the identical embedding model to conduct a similarity search within ExPool. The top k most semantically similar historical cases ("patients-like-me") are retrieved. The reasoning traces and verified diagnostic status of these retrieved patients are then integrated as in-context, few-shot examples for the new patient's manager agent. This retrieval process triggers in-context learning [46], fostering pattern recognition and critical reflection. As ExPool continuously scales with newly diverse verified patients, the retrieval of highly specific clinical neighbors becomes increasingly precise, shifting the system from standalone prediction to a progressively self-evolving predictive framework.

6.1.4 Multi-Agent Reinforcement Learning (MARL)

While ExPool enables retrieval-augmented in-context learning, it does not update the underlying model parameters nor directly optimize the intermediate temporal reasoning steps. To intrinsically enhance the system's reasoning and multi-agent collaboration, including the agent-memory interactions, we implemented a MARL framework (Figure 6.2B). Specifically, we adopted reward-ranked fine-tuning (RAFT) [45], a reinforcement learning approach that optimizes models using the reasoning trace with the highest reward among all self-generated roll-outs.

Parallel to the construction of ExPool, we generated diverse reasoning traces via elevated LLM temperature roll-outs. Similarly, ground-truth diagnostic labels were utilized as the reward signal for rejection sampling. However, in MARL, we restricted our training dataset strictly to correct reasoning traces, defined by a predicted risk score of ≥ 6 for confirmed cases and ≤ 4 for controls. By filtering for these optimal roll-outs, we ensure the models learn directly from their most logically sound and clinically accurate reasoning pathways.

To preserve the specialized roles within the framework, the optimization was decoupled for worker and manager agents. Worker agents were trained using worker's input-output pairs. To maintain longitudinal balance, we sampled interactions from the first, the last, and two randomly selected intermediate worker agents per reasoning trace. This training step explicitly optimizes sequential inter-agent temporal reasoning and agent-memory collaboration, refining

how effectively lung cancer-related clinical events are progressively summarized and distilled into the LTM. Conversely, the manager agent was trained using only manager-specific traces. This refines its capacity for risk prediction by optimizing how worker summary and the distilled evidence in LTM are synthesized and aggregated into a final risk estimation.

As the verified patient cohort expands, this MARL approach benefits from larger training data and drives continuous self-evolution through reinforcing successful temporal reasoning and collaborative memory distillation to progressively enhance the system's intrinsic diagnostic capabilities.

6.1.5 Combining ExPool and MARL

The final Traj-Evolve framework integrates the two complementary self-evolving mechanisms, ExPool and MARL. While ExPool facilitates few-shot prediction by retrieving patients-like-me, it does not update underlying model parameters. Conversely, MARL parametrically optimizes temporal reasoning and agent-memory interactions based on prior experience, yet it lacks an explicit mechanism to incorporate similar patient cohorts during final risk estimation. By unifying these approaches, the worker agents and the manager agent are explicitly trained to temporally reason over the patient trajectory while synthesizing more accurate predictions derived from retrieved patients-like-me within the ExPool (Figure 6.2C).

To generate the training reasoning traces that incorporate patients-like-me while strictly preventing data leakage, we implemented a leave-one-out cross-retrieval strategy. For each patient in the training cohort, the remainder of the training set served as a dynamic ExPool from which the top k most semantically similar patients were retrieved during the MARL roll-out phase. The resultant reasoning traces were subsequently filtered using the established MARL rejection sampling approach. This methodological synthesis yields optimal traces that benefit from the augmented context provided by patients-like-me from the dynamic ExPool, thereby generating potentially better training data compared to MARL in isolation. The worker and manager agents were subsequently trained using the standard MARL protocol, while inference was conducted utilizing the standard ExPool RAG methodology.

6.1.6 Evaluation and Analysis

We benchmarked Traj-Evolve against a comprehensive suite of baseline models, categorized into: a traditional clinical risk model (LCRAT) [83]; standard machine learning (ML) models (logistic regression and XGBoost [32]); sequential deep learning models (RETAIN [37] and PatientTM [172]); a specialized clinical BERT-based model (Clinical ModernBERT [99]); and static LLM-based systems utilizing GPT-OSS-20B [136] with or without long-context modeling strategies (zero-shot vanilla LLM, RAG, and Traj-CoA [220]).

To ensure feasible and fair comparisons, input data modalities were tailored to each model's specific architectural requirements. Standard ML models utilized structured medical codes, whereas PatientTM incorporated both codes and unstructured clinical texts (i.e., clinical notes and radiology reports). For BERT- and LLM-based architectures, we employed the unified XML-formatted EHR representation that preserved multimodal information, including codes, free text, and numerical laboratory values. For long sequences exceeding model context limits, left-truncation was applied to prioritize the most recent clinical records.

Supervised baselines, Clinical ModernBERT, and the MARL-optimized Traj-Evolve were trained on the same training set. In contrast, the ExPool variant of Traj-Evolve operated in a few-shot manner, dynamically retrieving "patients-like-me" from the ExPool constructed using the training set, while zero-shot baselines operated without access to the training set. Optimal hyperparameter configurations for all trained baselines were determined using the validation set. During MARL, Traj-Evolve's worker and manager agents were fine-tuned using QLoRA [44] for one epoch.

Our primary evaluation focused on binary risk classification. We assessed predictive performance on the hold-out test sets primarily using the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and F1-score, with other metrics including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) reported in the Appendix. To quantify uncertainty, we reported the bootstrap average and standard error for all metrics across 1,000 bootstrap

iterations.

In secondary analyses, we analyzed the reasoning quality and self-evolving dynamics of Traj-Evolve. Because Traj-Evolve generates both a quantitative risk score (ranging from 1 to 10) and an explicit natural language rationale, we employed a pairwise LLM-as-a-judge framework [62] to compare the clinical validity and reasoning quality of Traj-Evolve's rationales against those generated by Traj-CoA and vanilla LLM. Furthermore, to analyze the system's self-evolving properties, we tracked predictive performance continuously as the verified patient cohort dynamically expanded, simulating a real-world, streaming patient population.

6.2 Results

6.2.1 Lung Cancer Dataset

Baseline demographic, clinical, and lifestyle characteristics of the two test cohorts are summarized in Table 6.1. For overall population, patient trajectories spanned a median of 4.5 years for cases and 3.7 years for controls, and contained tens to hundreds of dated entries drawn from diagnoses, procedures, laboratories, medications, vital signs, and free-text notes pulled from UWMC clinical research repository. The per-patient input was also exceptionally long, with median XML token counts exceeding 60,000 in the overall sample and 80,000 in the never-smoker cases. Time-aware chunking reduced the mean per-chunk token count to approximately 16,000 (Figure D.1), allowing Traj-Evolve to process the full trajectory effectively while preserving chronology.

The cohort was clinically heterogeneous across subpopulations. As shown in Figure D.2, cases had higher frequencies of lung mass, cough, and COPD relative to controls. These clinical distributions differed significantly between ever-smoker and never-smoker cases: never-smoker cases exhibited markedly higher rates of lung mass (15.0% vs. 5.8%) and specific radiographic findings such as ground glass opacity (15.0% vs. 6%). Conversely, ever-smoker cases presented substantially higher rates of tobacco-related comorbidities, notably COPD

Table 6.1: Baseline characteristics of cases and controls in the overall and nonsmoker cohorts.

	Overall sample		**Never-smoker sample**	
Characteristic	Cases (n=90)	Controls (n=910)	Cases (n=27)	Controls (n=808)
Demographics				
Age at index date, years, median (IQR)	68 (60–76)	68 (59–78)	74 (64–78)	68 (58–80)
Sex, n (%)				
Female	49 (54.4)	423 (46.5)	24 (88.9)	463 (57.3)
Male	41 (45.6)	487 (53.5)	3 (11.1)	345 (42.7)
Race, n (%)				
White	74 (82.2)	657 (72.2)	21 (77.8)	572 (70.8)
Black or African American	8 (8.9)	102 (11.2)	0 (0.0)	80 (9.9)
Asian	5 (5.6)	77 (8.5)	6 (22.2)	99 (12.3)
Other[a]	1 (1.1)	28 (3.1)	0 (0.0)	22 (2.7)
Unknown[b]	2 (2.2)	46 (5.1)	0 (0.0)	35 (4.3)
Ethnicity, n (%)				
Hispanic or Latino	4 (4.4)	48 (5.3)	4 (14.8)	47 (5.8)
Not Hispanic or Latino	80 (88.9)	758 (83.3)	23 (85.2)	722 (89.4)
Unknown[b]	6 (6.7)	104 (11.4)	0 (0.0)	39 (4.8)
Clinical index and EHR history				
Index year, median (IQR)	2016 (2014–2018)	2016 (2014–2018)	2018 (2016–2018)	2017 (2015–2018)
EHR look-back duration, years, median (IQR)	4.5 (2.0–4.9)	3.7 (1.1–4.9)	4.2 (2.4–4.8)	4.5 (2.8–4.9)
Time to event, years, median (IQR)	0.5 (0.4–0.7)	—	0.5 (0.3–0.8)	—
XML token count, median (IQR)	61 678 (26 535–151 282)	51 904 (18 429–125 871)	84 406 (46 292–131 768)	75 070 (35 060–153 634)
Distinct dated EHR entries, median (IQR)	44 (19–113)	39 (12–95)	74 (30–102)	58 (28–114)
Index exam modality, n (%)				
Computed radiography (CR)	45 (50.0)	516 (56.7)	15 (55.6)	493 (61.0)
Computed tomography (CT)	45 (50.0)	394 (43.3)	12 (44.4)	315 (39.0)
Index exam body site, n (%)				
Chest	59 (65.6)	548 (60.2)	18 (66.7)	550 (68.1)
Abdomen/pelvis[f]	8 (8.9)	135 (14.8)	5 (18.5)	140 (17.3)
Extremity[g]	5 (5.6)	95 (10.4)	2 (7.4)	49 (6.1)
Other[h]	0 (0.0)	4 (0.4)	0 (0.0)	4 (0.5)
Missing	18 (20.0)	128 (14.1)	2 (7.4)	65 (8.0)
Low-dose CT lung cancer screening, n (%)	1 (1.1)	18 (2.0)	0 (0.0)	0 (0.0)
Clinical text and structured data volume, median (IQR)				
Epic notes	10 (0–50)	6 (0–28)	32 (7–58)	20 (8–45)
ORCA notes	2 (0–12)	4 (0–13)	0 (0–10)	2 (0–12)
Radiology reports	10 (5–19)	8 (3–14)	9 (6–18)	8 (3–15)
Diagnosis entries	144 (53–340)	118 (34–346)	283 (115–377)	202 (86–451)
Medication entries	52 (9–200)	59 (10–164)	77 (37–173)	72 (23–184)
Procedure entries	118 (37–220)	98 (34–221)	100 (72–170)	124 (52–272)
Laboratory entries	167 (54–538)	177 (49–484)	293 (160–627)	230 (93–577)
Vital-sign entries	28 (0–118)	18 (0–83)	81 (22–172)	61 (24–135)
Lifestyle factors				
Smoking status, n (%)[c]				
Current	16 (17.8)	82 (9.0)	0 (0.0)	0 (0.0)
Former	29 (32.2)	180 (19.8)	0 (0.0)	0 (0.0)
Never	11 (12.2)	289 (31.8)	27 (100.0)	808 (100.0)
Passive exposure only	0 (0.0)	2 (0.2)	0 (0.0)	0 (0.0)
Unknown[d]	34 (37.8)	357 (39.2)	0 (0.0)	0 (0.0)
Alcohol use, n (%)				
Current use	16 (17.8)	219 (24.1)	10 (37.0)	285 (35.3)
No current use[e]	30 (33.3)	260 (28.6)	11 (40.7)	426 (52.7)
Unknown[d]	44 (48.9)	431 (47.4)	6 (22.2)	97 (12.0)

(42.0% vs. 13.3%) and emphysema (28.4% vs. 5.3%). The never-smoker group was also enriched for female and Asian patients (Table 6.1). Therefore, this heterogeneity requires robust model, and Traj-Evolve addresses this by learning from experience as verified cohort accumulates.

6.2.2 Performance Comparison

In the overall study population, the self-evolving strategies of the Traj-Evolve framework demonstrated superior predictive performance compared with baseline models, highlighting the value of integrating past experiential learning for complex clinical temporal reasoning (Figure 6.3A). The combined approach, Traj-Evolve (ExPool+MARL), achieved the highest overall discrimination (AUROC 0.86; AUPRC 0.32; F1 0.42), surpassing the individual dynamic ExPool variant (AUROC 0.84; AUPRC 0.32; F1 0.39) and the MARL-optimized variant (AUROC 0.84; AUPRC 0.31; F1 0.41). Traditional approaches struggled to process the noisy longitudinal data; the standard clinical LCRAT model and the XGBoost machine learning baseline yielded notably lower AUROCs (0.69 and 0.76, respectively). Furthermore, the combined Traj-Evolve model outperformed standard LLM pipelines, including zero-shot generation (AUROC 0.79) and RAG (AUROC 0.81), suggesting that the dual self-evolving architecture is better equipped to summarize and predict from complex patient trajectories.

Accurate risk stratification among never-smokers represents a critical clinical challenge due to the absence of overt traditional risk factors. While baseline models exhibited significant performance degradation in this subgroup (AUROC LCRAT:0.61; XGBoost:0.71; vanilla LLM:0.76), Traj-Evolve maintained robust discriminatory power (Figure 6.3B). The integrated Traj-Evolve (ExPool+MARL) strategy achieved the highest performance in this cohort (AUROC 0.83; AUPRC 0.28; F1 0.31), while the standalone ExPool strategy (AUROC 0.82; AUPRC 0.27; F1 0.29) and MARL strategy (AUROC 0.82; AUPRC 0.26; F1 0.26) also performed strongly. The complementary strengths of these combined mechanisms successfully avoided the severe degradation seen in traditional models. This sustained accuracy indicates that the framework effectively identifies and synthesizes nuanced, temporally distant clinical

events rather than over-relying on a history of smoking.

Beyond quantitative risk stratification, pairwise LLM-as-a-judge evaluations confirmed the qualitative reasoning superiority of the combined Traj-Evolve (ExPool+MARL) framework over both the vanilla LLM and the static Traj-CoA architecture. When benchmarked against Traj-CoA, Traj-Evolve achieved an overall preference rate of 69% (Figure 6.3C), demonstrating notable strengths in executing complex clinical reasoning (73% win rate), capturing fine-grained details (69% win rate), and maintaining comprehensive temporal reasoning (68% win rate). These advantages were similarly robust against the vanilla LLM (Figure 6.3D). Traj-Evolve's comparative advantage over vanilla LLM scaled positively with sequence length; its overall win rate increased consistently as patient records expanded toward 1.5 million tokens (Figure 6.3E). This scaling behavior underscores the self-evolving framework's distinct capacity to distill and synthesize exceptionally long, noisy EHR trajectories.

6.2.3 Evolving Properties of ExPool

To understand the self-evolving capability of the ExPool, we analyzed its retrieval properties as the pool of verified patient predictions progressively expanded. Sensitivity analyses demonstrated that a larger ExPool systematically retrieved more clinically similar historical cases (Figure 6.4A, B, C). As the pool size increased, the average semantic distance between index patients and their retrieved neighbors consistently decreased. Furthermore, the retrieved cohorts exhibited increasing Spearman correlations with the anchor patients in terms of age and predicted risk scores. The clinical purity of the retrieved neighbors, measured by concordance in case status, sex, and never-smoking status, also improved as the pool evolved, remaining well above the expected purity of random retrieval.

This continuous expansion revealed a dynamic evolutionary tradeoff between retrieval diversity and specificity. By evaluating predictive performance across varying retrieval sizes ($k = 5, 10, 15$) and ExPool size, we observed a shift in the optimal number of retrieved "patients-like-me" (Figure 6.4D). In the early stages of evolution, when ExPool was relatively small, larger retrieved cohorts ($k = 15$) were favored, suggesting that early performance may

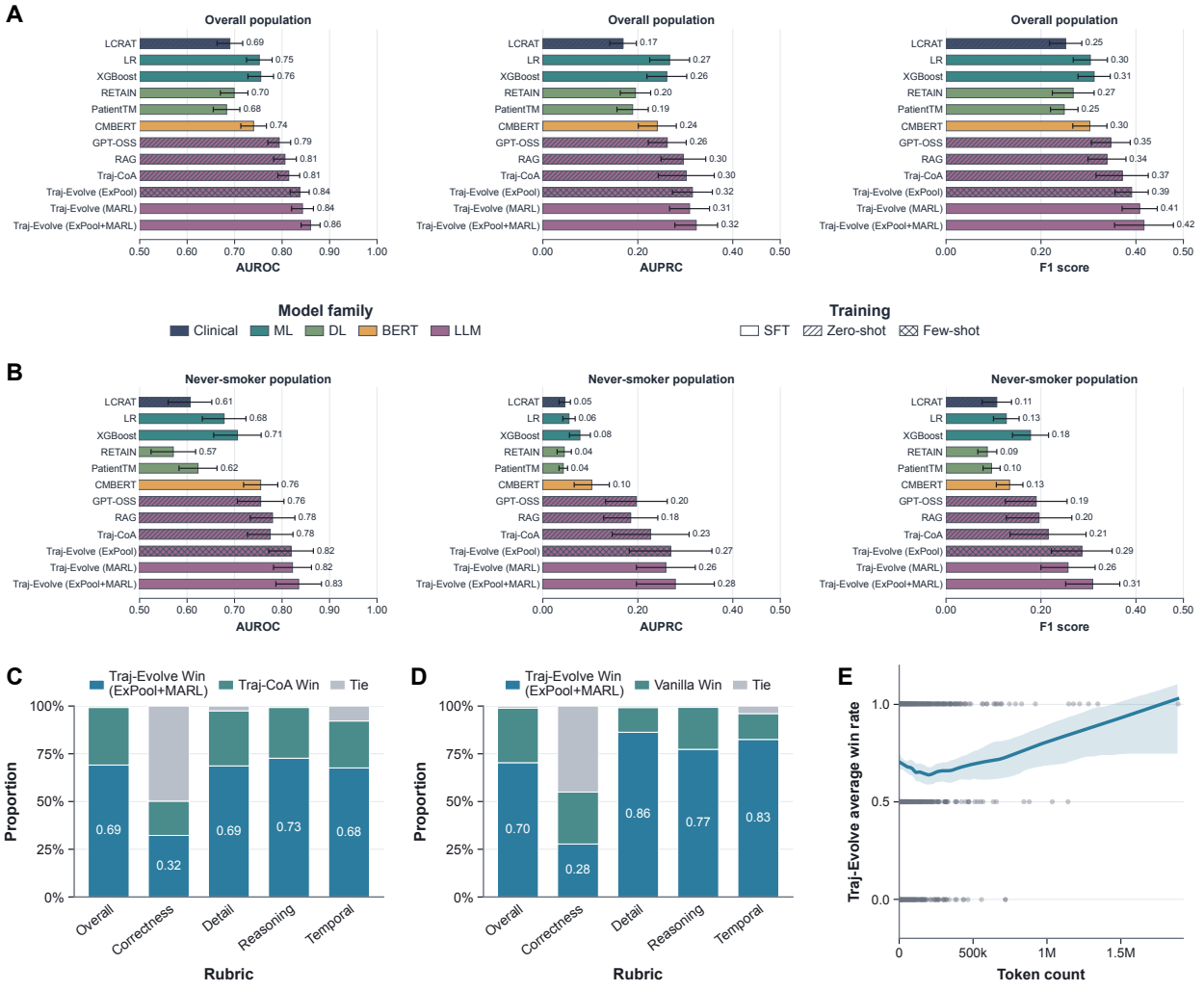

Figure 6.3: **Model performance comparison for 1-year patient outcomes.** **(A)** Performance metrics including AUROC, AUPRC, and F1 score across clinical, machine learning, deep learning, BERT, and LLM model families for the overall population. **(B)** Robustness evaluation showing performance metrics for the targeted nonsmoker population. **(C)** Pairwise LLM-as-a-judge evaluation displaying the proportion of Traj-Evolve wins, ties, and Traj-CoA wins across multiple rubrics: Overall, Correctness, Detail, Reasoning, and Temporal. **(D)** Pairwise LLM-as-a-judge evaluation comparing Traj-Evolve and vanilla LLM. **(E)** Traj-Evolve average win rate against vanilla LLM scaled by token count.

rely more on broad pool diversity to capture sufficient context. Conversely, as ExPool matured and densely expanded, smaller, highly specific cohorts ($k = 5$) yielded the highest AUROC. This may indicate that a mature ExPool can prioritizes specificity over diversity, leveraging highly similar neighbors to minimize noise and maximize few-shot accuracy. Throughout this progression, the few-shot framework consistently maintained an AUROC above the zero-shot baseline of 0.814.

6.2.4 Evolving Properties of MARL

To evaluate the evolving properties of MARL, we tracked the training dynamics and predictive performance of Traj-Evolve optimized with MARL as it processed an expanding cohort of verified patient trajectories. By filtering for optimal roll-outs through rejection sampling, the models learned directly from their most logically sound and clinically accurate reasoning pathways.

Figure 6.5A illustrates the training loss curves for both the worker and manager agents across the training iterations. Following the decoupled optimization strategy, both agents exhibited a sharp initial decline in loss and stabilized thereafter, indicating effective internalization of successful experience. However, the loss curve of the manager agent quickly converged, while the worker agent's loss slowly but continuously decreased with more training samples. This may indicate that the manager agent's summarization and prediction ability can be easier to learn, while the worker agents' temporal reasoning and collaborative memory distillation consistently benefitting from more verified patients.

This parameter optimization directly translated to progressive improvements in the overall risk estimation capability as the system accumulated experience. As depicted in Figure 6.5B, the framework's discriminative performance continuously evolved as the number of training samples increased up to 5,000. Aligning this upward trend with the observed loss dynamics (Figure 6.5A), we hypothesize that these continuous performance gains are driven by two synergistic mechanisms: the enhanced predictive accuracy of the manager agent and the progressively refined temporal reasoning of the worker agents.

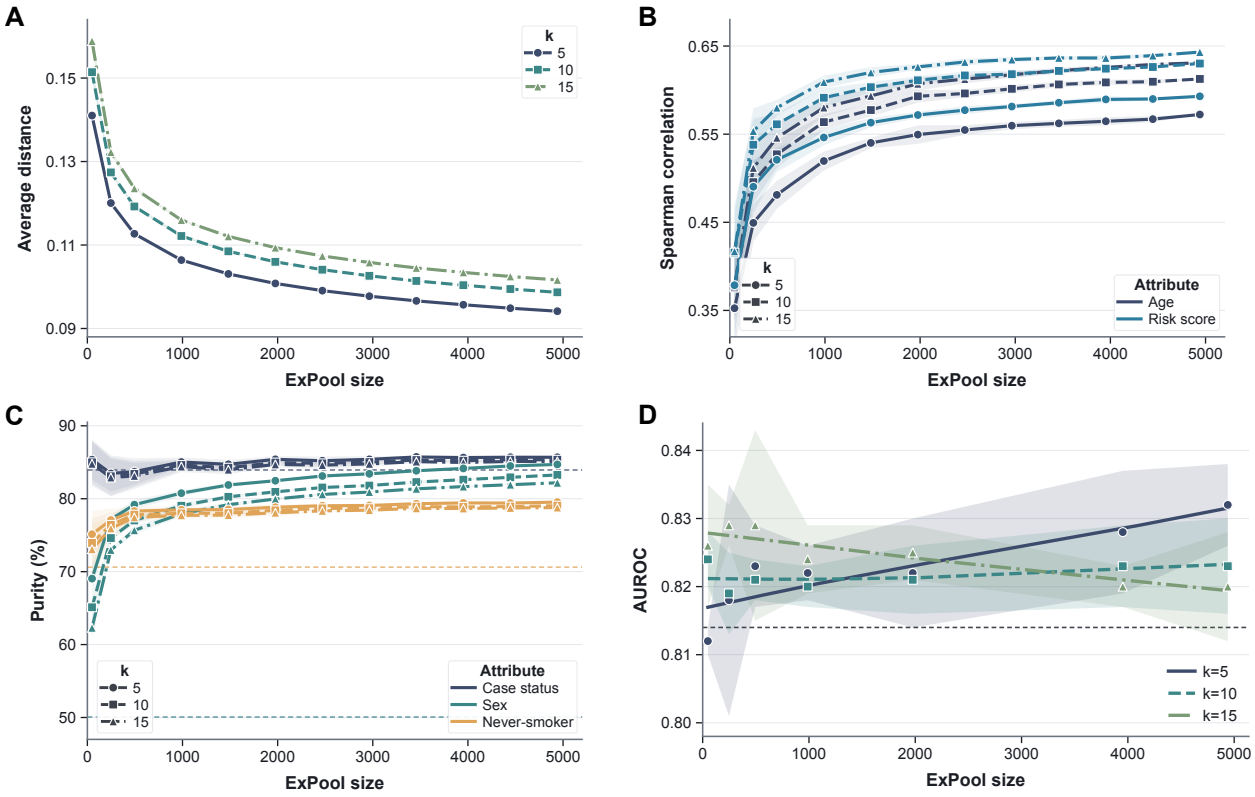

Figure 6.4: **Evolution of the ExPool. A,** As ExPool size increased, average distance from each index patient to the retrieved neighbors decreased; **B,** Spearman correlations in age and risk score between index patients and retrieved neighbors increased; and **C,** purity by case status, sex, and never-smoking status increased. Shaded bands indicate variability across runs; dashed horizontal lines in the purity panel indicate expected purity under random retrieval. **D,** Predictive performance during ExPool evolution, measured by AUROC for retrieval with $k = 5$, $k = 10$, and $k = 15$ patients. Each point indicates average performance across 3 runs with different random seeds. Shaded bands indicate uncertainty across runs, and the horizontal dashed line indicates performance without ExPool (AUROC=0.814).

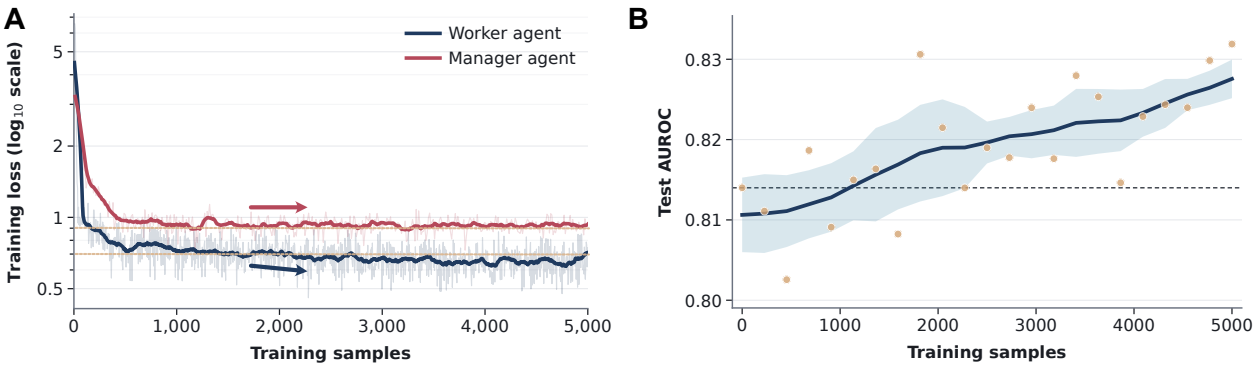

Figure 6.5: **Evolving MARL performance during agent training. A,** Loss curves for the worker and manager agents across training iterations. **B,** Model performance as the number of training samples increases, reflecting accumulated experience during training.

6.2.5 Optimization Properties of Traj-Evolve

To elucidate how the individual self-evolving mechanisms contribute to the model's enhanced discriminative ability, we compared the predicted risk score distributions of the Traj-Evolve variants against the static Traj-CoA baseline (Figure 6.6).

The ExPool mechanism improves predictive performance primarily through conservative calibration; it systematically lowers the predicted risk scores of control patients, shifting their distribution downward relative to the Traj-CoA baseline. However, this broader downward shift also collaterally lowers the predicted risk for some true cases. Conversely, the MARL optimization uniquely benefits the system by acting on intermediate predictions. It effectively lifts the predicted risk scores of cases that initially received ambiguous or middle-range scores under the static model, pushing them toward a higher, more accurate risk estimation.

By integrating both mechanisms, the combined Traj-Evolve (ExPool + MARL) framework strikes a critical evolutionary balance. It successfully synthesizes the specificity gained from ExPool's suppression of control scores with the heightened sensitivity of MARL's upward adjustment for mid-range cases, ultimately yielding a distinctly better separation between

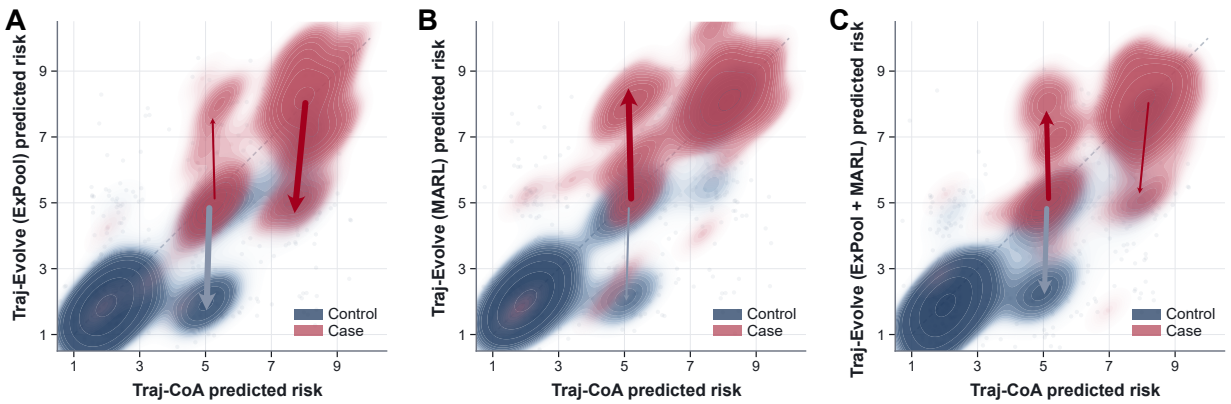

Figure 6.6: **Optimization properties of Traj-Evolve's self-evolving mechanisms.** Density scatter plots comparing the predicted risk scores of Traj-Evolve variants against the static Traj-CoA baseline. **A,** The ExPool mechanism improves discrimination primarily by lowering the predicted risk scores of controls, though it also slightly depresses scores for some cases. **B,** The MARL variant predominantly enhances performance by elevating the risk scores of cases that previously clustered in the middle range. **C,** The combined ExPool and MARL framework achieves a complementary balance. Arrows illustrate the strength of how Traj-Evolve changes the scores over Traj-CoA (wider means stronger). Scattered points are presented in a jittered way to facilitate visualization.

cases and controls for optimal predictive performance.

6.2.6 Case Study

To qualitatively demonstrate the nuanced clinical reasoning capabilities enabled by the ExPool, we examined the diagnostic trajectory of a 50-year-old Asian female never-smoker with a history of occupational asbestos exposure. The index patient presented with a normal pulmonary function test (PFT) and cough in 2015, formally documented asbestos exposure in 2016, and the discovery of a 16 mm non-calcified left upper lobe (LUL) nodule in 2018, with normal PFT and absence of pulmonary symptoms. She was subsequently diagnosed with lung cancer.

Traj-Evolve used information from two of the retrieved 10 patients from ExPool. Patient 8 was a 55-year-old woman who shared the documented asbestos exposure and progressive pulmonary fibrosis, but lacked a discrete mass and remained cancer-free after 3 years. Patient 10 was a 57-year-old Asian female never-smoker who presented with an 18 mm LUL mass and was confirmed to have lung cancer within 6 months.

In its generated reasoning rationale, Traj-Evolve explicitly estimated the index patient's risk by comparing her longitudinal trajectory to these retrieved cohorts. The manager agent identified that the index patient "sits biologically between these extremes", correctly weighing the mitigating factor of normal pulmonary function against the high-risk modifiers of progressive environmental carcinogen exposure and a suspicious nodule. The system assigned a risk score of 7/10. This case exemplifies how the incorporation of verified clinical experience enables Traj-Evolve to execute robust, comparative clinical judgment in atypical and highly challenging clinical presentations.

6.3 Conclusion

In this chapter, we presented Traj-Evolve, a self-evolving multi-agent system for longitudinal EHR modeling, and applied it to the challenging task of one-year lung cancer risk prediction. Building upon the temporal reasoning foundation of Traj-CoA, Traj-Evolve introduces two

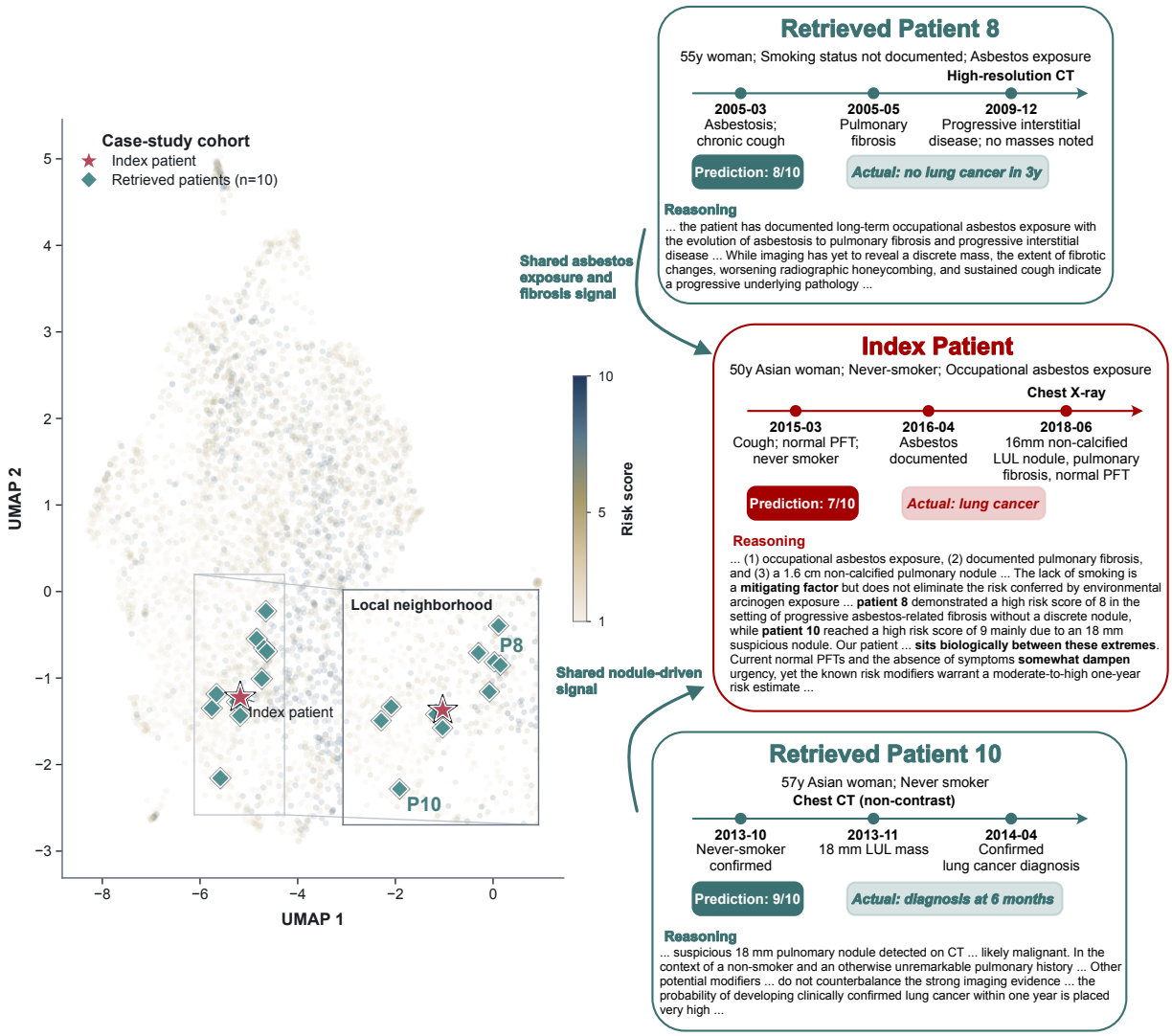

Figure 6.7: Case study demonstrating Traj-Evolve (ExPool+MARL)'s reasoning for a never-smoker patient. A UMAP projection maps the local semantic neighborhood of the index patient (star) alongside retrieved historical cases (diamonds). The right panel demonstrates how Traj-Evolve balances the shared signals from retrieved patients and the index patient's own characteristics.

complementary self-evolving mechanisms that emulate the experiential learning intrinsic to expert clinical practice. The non-parametric ExPool functions as a procedural memory module, enabling the manager agent to retrieve verified "patients-like-me" and ground its reasoning in concrete historical precedents. In parallel, the parametric MARL framework leverages rejection-sampled high-reward trajectories to refine the intrinsic temporal reasoning of worker agents and the synthesis capability of the manager agent. When unified through a leave-one-out cross-retrieval training strategy, these mechanisms operate synergistically rather than redundantly.

Detailed analyses of Traj-Evolve's self-evolving dynamics revealed that ExPool progressively retrieves more clinically homogeneous neighbors and shifts from a diversity-favoring to a specificity-favoring retrieval regime as it matures, while MARL exhibits decoupled convergence behavior in which manager-level synthesis is learned rapidly and worker-level temporal reasoning continues to benefit from accumulating experience. Risk-score-distribution analyses clarified the complementary mechanistic roles of the two components: ExPool primarily improves specificity by suppressing inflated control scores, whereas MARL primarily improves sensitivity by elevating ambiguous mid-range case scores.

Taken together, these results support the broader vision of clinical decision-support systems that do not remain static after deployment but instead continuously refine their reasoning as verified patient experience accumulates. By coupling explicit case-based retrieval with parametric reinforcement of successful reasoning pathways, Traj-Evolve offers a concrete instantiation of how multi-agent LLM systems can move toward continuously evolving, experientially grounded clinical temporal reasoning.

6.4 Limitations

Several limitations of this work warrant consideration and motivate future research. First, our evaluation was conducted within a single-institution retrospective case-control design at UWMC. Prospective and multi-institutional validation will be essential to characterize calibration, generalizability across health systems and patient populations, and performance

under genuine deployment prevalence. Second, the self-evolving mechanisms depend on ground-truth diagnostic labels for both ExPool population and MARL rejection sampling. In clinical practice, definitive labels for incident lung cancer may take months to years to materialize and are subject to verification noise, loss to follow-up, and ascertainment bias. The current framework also assumes that label-conditioned rejection sampling reliably identifies the most clinically sound reasoning trace, yet a roll-out with the correct final prediction does not guarantee correct intermediate reasoning. Incorporating process-level reward signals [228] may mitigate these issues. Furthermore, although Traj-Evolve generates explicit natural-language rationales and grounds its predictions in retrieved historical cases, we did not formally evaluate clinician trust, downstream decision-making impact, or potential harms arising from over-reliance on retrieved analogues. Rigorous human-factors studies, fairness audits across demographic subgroups, and prospective evaluations of clinical utility are needed before such systems can be responsibly deployed. Finally, the present work focused on one-year incident lung cancer prediction with a five-year look-back window. Extending Traj-Evolve to other outcomes, longer or shorter prediction horizons, and continuously updating patient trajectories in true streaming settings will be necessary to establish the framework as a general-purpose paradigm for self-evolving longitudinal EHR modeling.

Chapter 7

CONCLUSION AND FUTURE WORK

This chapter concludes the dissertation by summarizing its key contributions, discussing limitations, and proposing directions for future work.

7.1 Key Contributions

This dissertation set out to develop interpretable and generalizable frameworks for patient trajectory modeling with longitudinal EHRs. The four aims collectively verified the hypothesis that models built on the full dynamics of a patient's history can produce more accurate predictions than simpler models, and that these predictions and the temporal reasoning process can be made transparent and robust across diverse clinical contexts. Progressively, the four aims established the added value of trajectory information, advanced from discrete-time modeling to continuous-time modeling, and generalized patient trajectory modeling beyond task-specific models with self-evolving multi-agent systems.

Establishing the added value of trajectory information in a focused clinical setting. Chapter 3 introduced TrajSurv-mPC, a deep learning framework that predicts post-metastasis overall survival in progressed metastatic prostate cancer patients, using pre-metastasis serial PSA values and treatment histories. The central contribution is not a single fixed model, but evidence that the performance improvement over summary-feature-models can generalize across cohorts, and that the information embedded in the trajectories can be interpret through our three complementary analyses, consistent with clinical knowledge. These give us confidence that the patient trajectory contains robust and useful added value for prediction in this specific clinical task.

Advancing from discrete-time modeling to more faithful continuous-time modeling. Chapter 4 introduced TrajSurv-Cont, which learns continuous latent trajectories from irregularly sampled longitudinal EHRs using an NCDE, paired with a TACL objective that aligns the latent trajectory with the patient's actual clinical progression. This alignment improved both accuracy and transparency, and enabled a divide-and-conquer interpretation that splits the model into feature-to-trajectory and trajectory-to-outcome processes. The contribution is a viable direction for transparent longitudinal EHR modeling: rather than interpreting feature values at isolated time points, TrajSurv-Cont explains how the velocities of clinical features drive the evolution of a latent state and ultimately the outcome, offering a dynamical perspective that complements existing interpretation tools. Furthermore, TrajSurv-Cont represents the patient trajectory as a latent trajectory, more expressive than latent states at single time points.

Generalizing beyond task-specific models with a multi-agent framework. Chapter 5 introduced Traj-CoA, a training-free multi-agent framework that performs temporal reasoning over long, noisy, and multimodal EHRs. By converting heterogeneous records into a unified, LLM-friendly representation and decomposing them into time-aware chunks processed by a chain of worker agents, a manager agent, and a long-term memory module, Traj-CoA separates local signal extraction from global synthesis and mitigates the long-context limitations of single-agent LLMs. Traj-CoA's temporal reasoning ability was shown using data from two organizations and 15 cancer types. The contribution is a generalizable and interpretable patient trajectory modeling paradigm that reasons over raw trajectories without task-specific feature engineering or retraining.

Enabling self-evolution through experience accumulation. Chapter 6 introduced Traj-Evolve, which extends Traj-CoA with two complementary self-evolving mechanisms: a non-parametric evolving ExPool that retrieves verified "patients-like-me," and a MARL procedure that internalizes successful reasoning patterns into the parameters of the worker

and manager agents. To our knowledge, this is the first self-evolving multi-agent framework for longitudinal EHR modeling applied to a real-world clinical prediction task. Traj-Evolve demonstrates that a self-evolving design can be combined synergistically with temporal reasoning and prediction in patient trajectory modeling. Analyses characterized how ExPool and MARL play complementary mechanistic roles in the model's evolution. The contribution is a methodological proof-of-concept of a clinical decision-support system utilizing long patient trajectory data that does not remain static after deployment but continually refines its reasoning as verified patient experience accumulates.

Across these four contributions, we found that faithfully modeling the temporal information in patient histories makes predictions more accurate, more robust, and more interpretable, and interpretability and generalizability advance together rather than in tension.

7.2 Limitations

Each chapter discusses limitations specific to its methods and datasets. Beyond these, several limitations cut across the dissertation as a whole and motivate the future work below.

First, the evaluations in this dissertation are predominantly retrospective. TrajSurv-mPC and TrajSurv-Cont were validated on retrospective cohorts, and Traj-CoA and Traj-Evolve were evaluated on retrospective case-control designs. Although cross-cohort and cross-organization evaluation provides meaningful evidence of generalizability, retrospective designs cannot fully characterize calibration, performance under genuine deployment prevalence, or behavior under prospective temporal drift. Prospective and multi-institutional validation remains necessary before any of these frameworks could be considered for clinical use.

Second, the frameworks model patient trajectories at a fixed prediction point with a bounded look-back window, rather than as continuously updating streams. TrajSurv-mPC and TrajSurv-Cont produce predictions over time but are trained and evaluated on fixed observation windows; Traj-CoA and Traj-Evolve reason over a defined history up to an index date. The online prediction capabilities of these frameworks need to be evaluated in real-world streaming context.

Third, the dissertation spans two methodological families, DL-based sequential models (TrajSurv-mPC and TrajSurv-Cont) and LLM-based multi-agent systems (Traj-CoA and Traj-Evolve), each with distinct trade-offs. The DL models offer faithful continuous-time dynamics and quantitative, mechanistic interpretation, but require task-specific training. The multi-agent systems offer generalizability, multimodality, and natural-language interpretability, but are difficult to provide probabilistic predictions and often generate interpretations that are descriptive rather than mechanistic. The dissertation demonstrates each family's value but does not unify them.

Finally, interpretability is assessed primarily through validation with known clinical knowledge, case studies, statistical analyses, and LLM-as-a-judge evaluation. While these approaches successfully interpret the otherwise black-box models, they do not measure whether the interpretations actually translate to improved clinician decision-making and trust. Rigorous human-factors studies and prospective evaluations of clinical utility remain outstanding.

7.3 Future Work

The four aims of this dissertation suggest a natural next question: what would a general-purpose framework for patient trajectory modeling look like, one that retains the faithful temporal dynamics and mechanistic interpretability of the DL-based models while also achieving the multimodality, generalizability, and experiential learning of the LLM-based systems? Below we outline two complementary research directions toward this goal, agentic systems and world models, and position the contributions of this dissertation as preliminary work along each.

7.3.1 Agentic System for Patient Trajectory Modeling

One promising direction is to build agentic systems that more fully leverage the capabilities of foundation models for patient trajectory modeling. The motivation is that foundation models have demonstrated strong general reasoning, instruction-following, and tool-use abilities, and

that agentic harness including decomposition, memory, planning, reflection, and inter-agent communication can extend these abilities to long-horizon tasks that exceed what a single forward pass can handle [212, 114]. Patient trajectory modeling is a natural target for this paradigm, because a patient's record is long and heterogeneous, requiring a collection of capabilities to fully understand and leverage. Traj-CoA and Traj-Evolve are preliminary work toward this agentic direction. We envision a future system equipped with at least three properties.

Multimodal longitudinal reasoning To achieve comprehensive modeling of patient trajectory, the system should jointly reason over multimodal data, including structured data, texts (e.g., clinical notes and imaging reports), images (e.g., pathology images, radiology images), time series data, genomics data, and other modalities, integrating them along timeline. Traj-CoA represents an initial attempt by converting heterogeneous data into a unified XML format for temporal reasoning. However, a fuller system may choose the best approach for each modality while aggregate multimodal data in a synergistic way. For example, specific tools, subagents, or encoder modules may be introduced to incorporate time series data. Recently, EHR foundation models has been developed to learn from patient trajectories [171, 229, 224]. Incorporating EHR foundation models into the agentic system may improve its temporal modeling of trajectory data while retaining agentic reasoning ability.

Self-evolution In a streaming clinical setting, there are two evolving aspects. First, with new observations and interventions, the modeling of each patient should be updated. Second, with more patients seen and verified, experience accumulates and should further improve the system. Traj-CoA and Traj-Evolve provides preliminary realization of these ideas, but does not unify them in a streaming setting. From a methodological perspective, future work could extend self-evolution of the agentic system along several axes including evolving the agents' tools and skills over time, incorporating process-level reward signals, and supporting

continual adaptation in true streaming settings with delayed and noisy labels.

Factual reliability In high-stake clinical settings, the system should be grounded by evidence and uncertainties without hallucinations. Existing work has designed SFT and RL training paradigms for medical LLMs that incorporate factual validation using knowledge graphs or web search [201, 181, 182]. However, how to ensure the factual reliability when dealing with long and noisy patient trajectory data is unknown. In our experiments, we found Traj-CoA may be overly assertive in some cases (Section 5.3.4). Future work may design novel methods to ensure faithful modeling.

7.3.2 World Models for Patient Trajectory Modeling

A second, complementary direction is to learn world models of patients: models that learn good representations of a patient's state and of how that state evolves over time. This is motivated by a line of work showing that learning predictive representations in an abstract latent space can reflect physical and causal dynamics of the world [97, 150, 16, 18]. This is particularly appealing for patient trajectory modeling because disease progression and patient state evolution naturally or under interventions can be regarded as a world modeling problem, learning representations that simulate patient states over time. TrajSurv-Cont is a preliminary realization of this idea, showing that patient trajectories can be represented as latent trajectories from which interpretation of the patient dynamics can be achieved. However, TrajSurv-Cont requires intermediate supervision signals for the TACL objective, which deviates from the unsupervised or self-supervised principle of the world models. We envision three properties for world models for patient trajectory modeling.

Self-supervised learning A patient world model should learn the dynamics of patient states from longitudinal EHRs without relying on task-specific labels, in the same self-supervised spirit that underlies JEPA and related architectures [97, 16]. This is both a conceptual and a practical motivation. Conceptually, learning to predict how a patient's

representation evolves, rather than learning a mapping to one labeled outcome, encourages the model to capture the intrinsic structure of disease progression itself, which can then transfer across outcomes, prediction horizons, and populations. Practically, labels for many clinical outcomes are scarce, delayed, or noisily ascertained, whereas unlabeled longitudinal trajectories are abundant; self-supervised pretraining is what has allowed EHR foundation models to improve sample efficiency and robustness to distribution shift. Therefore, a next step for TrajSurv-Cont is to replace the external supervision of survival outcome and SOFA labels with self-supervised predictive objectives, for example, predicting the latent representation of a patient's future state from the representation of observed history, following the JEPA blueprint [97], so that the alignment between latent dynamics and clinical progression emerges from the data rather than from a hand-chosen label. While there were emerging research on JEPA for patient trajectory modeling [211, 8], they are mainly based on EHR foundation models rather than continuous-time models, which may not fully model the complex patient dynamics.

Interpretability through intrinsic dynamics Because a world model learns how a patient's latent state evolves rather than only a feature-to-outcome mapping, the latent space itself becomes an object of clinical interpretation. Its geometry, its phenotypic structure, and the trajectories patients trace through it can be examined and related to clinical knowledge, offering a window into the intrinsic dynamics of disease progression. This is a different and complementary form of interpretability from the feature-attribution methods commonly used for EHR models: instead of asking which features mattered at a single time point, it asks how a patient's state is moving and why. TrajSurv-Cont demonstrates that this latent trajectory view is feasible and interpretable for longitudinal EHRs, with its learned vector field and trajectory clustering explaining how changes in clinical features drive the evolution of the latent state and link to the outcome. We expect to see similar interpretability on a continuous-time patient world model.

Counterfactual prediction and interventions Finally, because a world model represents dynamics rather than a single fixed outcome, it can in principle be combined with interventions and used for counterfactual reasoning, asking how a patient's trajectory would differ under an alternative treatment. This capability is central to clinical decision support, since clinicians routinely weigh how outcomes might change under different management choices. The path toward this capability has been explored by recent works [165, 171, 126] through simulation of future trajectories under different intervention options. A patient world model that factors the effect of interventions on latent dynamics, learning not only how a patient's state evolves but how that evolution responds to treatment, would move patient trajectory modeling from prediction toward interpretable decision support. However, these preliminary efforts leave gap for future work. For example, existing patient world models do not account for the time-dependent confounding just like what TE-CDE [165] explicitly models. This may induce confounding bias from the observational EHR data.

7.3.3 Toward a Unified View

A world model learns the dynamics of patient state evolution, providing a simulator for future patient trajectories. The agentic model can leverage the strong capabilities of foundation models to plan, reason, act (e.g., tool-calling), and answer for queries on patient trajectories. A possible combination of the two is using the agentic system as the configurator module [97] and the world model as the simulator. For any queries from the user, it can be decomposed by the agentic model, generating an explicit plan for reasoning and simulating on the patient trajectories. Then the agentic system can call the world model as a tool to accurately simulate future patient states possibly under certain counterfactual scenarios. The results are then returned to the agentic system for further reasoning and answer generation. These steps can be done in a loop, under a model predictive control (MPC) style [1]. Recent advances have confirmed the feasibility of such a combination system [207, 202] but still fall short in previously mentioned properties.

7.4 Final Remarks

This dissertation began from a simple premise that a patient's history is a trajectory, not a snapshot, and that modeling its dynamics well can make clinical prediction models more trustworthy. The four aims pursued this premise from complementary angles, establishing the added value of trajectory information, modeling it faithfully in continuous time, generalizing across tasks and data through multi-agent reasoning, and enabling systems that improve as they see more patients. A recurring theme is that interpretability and generalizability need not be sacrificed for one another; in each chapter, modeling temporal information more faithfully improved accuracy, transparency, and robustness together.

The frameworks developed here are steps, not destinations. The trajectory of the field is encouraging. As foundation models, agentic systems, and world models mature, the tools available for modeling patient trajectories are becoming both more capable and more amenable to interpretation. It is my hope that this dissertation contributes, in a modest way, to a future in which AI tools for longitudinal EHRs are trusted not because they are accurate alone, but because clinicians can see how they reason about the unfolding story of a patient's health.

BIBLIOGRAPHY

[1] Fast Model Predictive Control Using Online Optimization. URL https://stanford.edu/~boyd/papers/fast_mpc.html.

[2] Metastatic prostate cancer - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/metastatic-prostate-cancer/symptoms-causes/syc-20377966.

[3] Cervical Cancer Screening - NCI, October 2022. URL https://www.cancer.gov/types/cervical/screening. Archive Location: nciglobal,ncienterprise.

[4] Screening for Breast Cancer - NCI, August 2025. URL https://www.cancer.gov/types/breast/screening. Archive Location: nciglobal,ncienterprise.

[5] Colorectal Cancer Screening - NCI, February 2026. URL https://www.cancer.gov/types/colorectal/patient/colorectal-screening-pdq. Archive Location: nciglobal,ncienterprise.

[6] Lung Cancer Screening - NCI, February 2026. URL https://www.cancer.gov/types/lung/patient/lung-screening-pdq. Archive Location: nciglobal,ncienterprise.

[7] Prostate Cancer Screening - NCI, February 2026. URL https://www.cancer.gov/types/prostate/patient/prostate-screening-pdq. Archive Location: nciglobal,ncienterprise.

[8] Irsyad Adam, Zekai Chen, David Laprade, Shaun Porwal, David Laub, Erik Reinertsen, Arda Pekis, and Kevin Brown. The Patient is not a Moving Document: A World Model Training Paradigm for Longitudinal EHR, January 2026. URL http://arxiv.org/abs/2601.22128. arXiv:2601.22128 [cs] version: 1.

[9] Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17, 2025.

[10] Asem Alaa, Erik Mayer, and Mauricio Barahona. ICE-NODE: Integration of Clinical Embeddings with Neural Ordinary Differential Equations. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pages 537–564. PMLR, December 2022.

[11] Abdallah Alabdallah. *Towards Trustworthy Survival Analysis with Machine Learning Models*. PhD thesis, Halmstad University Press, 2025. URL `https://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-55202`.

[12] Patrick R. Alba, Anthony Gao, Kyung Min Lee, Tori Anglin-Foote, Brian Robison, Evangelia Katsoulakis, Brent S. Rose, Olga Efimova, Jeffrey P. Ferraro, Olga V. Patterson, Jeremy B. Shelton, Scott L. Duvall, and Julie A. Lynch. Ascertainment of Veterans With Metastatic Prostate Cancer in Electronic Health Records: Demonstrating the Case for Natural Language Processing. *JCO Clinical Cancer Informatics*, (5):1005–1014, December 2021. doi: 10.1200/CCI.21.00030. URL `https://ascopubs.org/doi/full/10.1200/CCI.21.00030`.

[13] Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 362–372. Springer, 2025.

[14] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[15] Anthropic. Use xml tags to structure your prompts. `https://docs.anthropic.com/en/`

docs/build-with-claude/prompt-engineering/use-xml-tags, 2025. Accessed: 2025-08-16.

[16] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, June 2025. URL http://arxiv.org/abs/2506.09985. arXiv:2506.09985 [cs].

[17] M. Balakrishnan, R. Tucker, B. E. Stephens, and J. M. Bliss. Blood urea nitrogen and serum bicarbonate in extremely low birth weight infants receiving higher protein intake in the first week after birth. *Journal of Perinatology: Official Journal of the California Perinatal Association*, 31(8):535–539, August 2011. ISSN 1476-5543. doi: 10.1038/jp.2010.204.

[18] Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, November 2025. URL http://arxiv.org/abs/2511.08544. arXiv:2511.08544 [cs].

[19] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305, 2012.

[20] Linus Bleistein, Van-Tuan Nguyen, Adeline Fermanian, and Agathe Guilloux. Dynamical survival analysis with controlled latent states. *arXiv preprint arXiv:2401.17077*, 2024.

[21] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

[22] Adam R. Brentnall, Emma C. Atakpa, Harry Hill, Ruggiero Santeramo, Celeste Damiani, Jack Cuzick, Giovanni Montana, and Stephen W. Duffy. An optimization framework to guide the choice of thresholds for risk-based cancer screening. *npj Digital Medicine*, 6 (1):223, November 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00967-9.

[23] Nicole Brighi, Giuseppe Schepisi, Vincenza Conteduca, Emanuela Scarpi, Paola Caroli, Federica Matteucci, and Cristian Lolli. Validation of a Combined Prognostic Score Using Plasma Tumor DNA and Clinical Features in Metastatic Castration-Resistant Prostate Cancer Patients Treated with Taxanes. *OncoTargets and Therapy*, 18:667–677, May 2025. ISSN 1178-6930. doi: 10.2147/OTT.S509285. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12117575/.

[24] Matthew D. Bucknor, Daphne Y. Lichtensztajn, Tracy K. Lin, Hala T. Borno, Scarlett L. Gomez, and Thomas A. Hope. Disparities in PET Imaging for Prostate Cancer at a Tertiary Academic Medical Center. *Journal of Nuclear Medicine*, 62(5):695–699, May 2021. ISSN 0161-5505, 2159-662X. doi: 10.2967/jnumed.120.251751.

[25] Manuel Burger, Gunnar Rätsch, and Rita Kuznetsova. Multi-modal graph learning over umls knowledge graphs. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 52–81. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/burger23a.html.

[26] Lucía A Carrasco-Ribelles, José Llanes-Jurado, Carlos Gallego-Moll, Margarita Cabrera-Bean, Mònica Monteagudo-Zaragoza, Concepción Violán, and Edurne Zabaleta-del-Olmo. Prediction models using artificial intelligence and longitudinal data from electronic health records: A systematic methodological review. *Journal of the American Medical Informatics Association : JAMIA*, 30(12):2072–2082, September 2023. ISSN 1067-5027. doi: 10.1093/jamia/ocad168.

[27] Chiara Cencioni, Ilaria Trestini, Geny Piro, Emilio Bria, Giampaolo Tortora, Carmine Carbone, and Francesco Spallotta. Gastrointestinal Cancer Patient Nutritional Management: From Specific Needs to Novel Epigenetic Dietary Approaches. *Nutrients*, 14(8):1542, April 2022. ISSN 2072-6643. doi: 10.3390/nu14081542. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9024975/.

[28] Leonidas Chardalias, Ioannis Papaconstantinou, Antonios Gklavas, Marianna Politou, and Theodosios Theodosopoulos. Iron Deficiency Anemia in Colorectal Cancer Patients: Is Preoperative Intravenous Iron Infusion Indicated? A Narrative Review of the Literature. *Cancer Diagnosis & Prognosis*, 3(2):163, March 2023. doi: 10.21873/cdp.10196. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9949551/.

[29] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL https://arxiv.org/abs/2402.03216.

[30] Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*, 2025.

[31] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[32] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016. doi: 10.1145/2939672.2939785. URL http://arxiv.org/abs/1603.02754. arXiv:1603.02754 [cs].

[33] Elvin S Cheng, Marianne Weber, Julia Steinberg, and Xue Qin Yu. Lung cancer risk in

never-smokers: An overview of environmental and genetic factors. *Chinese Journal of Cancer Research*, 33(5):548, 2021.

[34] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

[35] Alvin Har Teck Chia, May Sze Khoo, Andy Zhengyi Lim, Kian Eng Ong, Yixuan Sun, Binh P. Nguyen, Matthew Chin Heng Chua, and Junxiong Pang. Explainable machine learning prediction of ICU mortality. *Informatics in Medicine Unlocked*, 25:100674, January 2021. ISSN 2352-9148. doi: 10.1016/j.imu.2021.100674.

[36] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[37] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[38] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 0035-9246.

[39] David Crosby, Sangeeta Bhatia, Kevin M. Brindle, Lisa M. Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C. Fitzgerald, Sanjiv S. Gambhir, Peter Kuhn, Timothy R. Rebbeck, and Shankar Balasubramanian. Early detection of cancer. *Science*, 375(6586):eaay9040, March 2022. ISSN 1095-9203. doi: 10.1126/science.aay9040.

[40] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H. Shah. TIMER: Temporal instruction modeling and evaluation for

longitudinal clinical records. *npj Digital Medicine*, 8(1):577, September 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01965-9.

[41] Deepansh Dalela, Björn Löppenberg, Akshay Sood, Jesse Sammon, and Firas Abdollah. Contemporary Role of the Decipher® Test in Prostate Cancer Management: Current Practice and Future Perspectives. *Reviews in Urology*, 18(1):1–9, 2016. ISSN 1523-6161. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4859922/.

[42] Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

[43] Tatiana V. Denisenko, Inna N. Budkevich, and Boris Zhivotovsky. Cell death-based treatment of lung adenocarcinoma. *Cell Death & Disease*, 9(2):117, January 2018. ISSN 2041-4889. doi: 10.1038/s41419-017-0063-y. URL https://www.nature.com/articles/s41419-017-0063-y.

[44] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115, 2023.

[45] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[46] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A Survey on In-context Learning, October 2024. URL http://arxiv.org/abs/2301.00234. arXiv:2301.00234 [cs].

[47] Baojun Duan, Yaning Zhao, Jun Bai, Jianhua Wang, Xianglong Duan, Xiaohui Luo, Rong Zhang, Yansong Pu, Mingqing Kou, Jianyuan Lei, and Shangzhen Yang. Colorectal

Cancer: An Overview. In Jose Andres Morgado-Diaz, editor, *Gastrointestinal Cancers*. Exon Publications, Brisbane (AU), 2022. ISBN 978-0-6453320-6-3. URL http://www.ncbi.nlm.nih.gov/books/NBK586003/.

[48] Daniel E. Ehrmann, Shalmali Joshi, Sebastian D. Goodfellow, Mjaye L. Mazwi, and Danny Eytan. Making machine learning matter to clinicians: model actionability in medical decision-making. *npj Digital Medicine*, 6(1):7, January 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00753-7. URL https://www.nature.com/articles/s41746-023-00753-7.

[49] Eric A Engels. Inflammation in the development of lung cancer: epidemiological evidence. *Expert review of anticancer therapy*, 8(4):605–615, 2008.

[50] Andre Esteva, Jean Feng, Douwe van der Wal, Shih-Cheng Huang, Jeffry P. Simko, Sandy DeVries, Emmalyn Chen, Edward M. Schaeffer, Todd M. Morgan, Yilun Sun, Amirata Ghorbani, Nikhil Naik, Dhruv Nathawani, Richard Socher, Jeff M. Michalski, Mack Roach, Thomas M. Pisansky, Jedidiah M. Monson, Farah Naz, James Wallace, Michelle J. Ferguson, Jean-Paul Bahary, James Zou, Matthew Lungren, Serena Yeung, Ashley E. Ross, Howard M. Sandler, Phuoc T. Tran, Daniel E. Spratt, Stephanie Pugh, Felix Y. Feng, and Osama Mohamad. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *npj Digital Medicine*, 5(1):1–8, June 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00613-w. URL https://www.nature.com/articles/s41746-022-00613-w. Number: 1.

[51] Kevin W Eva. What every teacher needs to know about clinical reasoning. *Medical education*, 39(1):98–106, 2005.

[52] Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758, 2001.

[53] Simone Ferretti, Chiara Mercinelli, Laura Marandino, Giulio Litterio, Michele Marchioni, and Luigi Schips. Metastatic castration-resistant prostate cancer: Insights on current therapy and promising experimental drugs. *Research and Reports in Urology*, 15: 243–259, 2023. ISSN 2253-2447. doi: 10.2147/RRU.S385257.

[54] Brody H. Foy, Rachel Petherbridge, Maxwell T. Roth, Cindy Zhang, Daniel C. De Souza, Christopher Mow, Hasmukh R. Patel, Chhaya H. Patel, Samantha N. Ho, Evie Lam, Camille E. Powe, Robert P. Hasserjian, Konrad J. Karczewski, Veronica Tozzo, and John M. Higgins. Haematological setpoints are a stable and patient-specific deep phenotype. *Nature*, 637(8045):430–438, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08264-5.

[55] Sebastian Frees, Shusuke Akamatsu, Samir Bidnur, Daniel Khalaf, Claudia Chavez-Munoz, Werner Struss, Bernhard J. Eigl, Martin Gleave, Kim N. Chi, and Alan So. The impact of time to metastasis on overall survival in patients with prostate cancer. *World Journal of Urology*, 36(7):1039–1046, July 2018. ISSN 1433-8726. doi: 10.1007/s00345-018-2236-4.

[56] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.

[57] Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025.

[58] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL `https://arxiv.org/abs/2312.10997`.

[59] Earl F Glynn and Mark A Hoffman. Heterogeneity introduced by EHR system imple-

mentation in a de-identified data resource from 100 non-affiliated organizations. *JAMIA Open*, 2(4):554–561, August 2019. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooz035.

[60] Amir Goldkorn, Catherine Tangen, Melissa Plets, Daniel Bsteh, Tong Xu, Jacek K. Pinski, Sue Ingles, Timothy Junius Triche, Gary R. MacVicar, Daniel A. Vaena, Anthony W. Crispino, David James McConkey, Primo N. Lara, Jr, Maha H. A. Hussain, David I. Quinn, Tanya B. Dorff, Seth Paul Lerner, Ian Thompson, Jr, and Neeraj Agarwal. Circulating Tumor Cell Count and Overall Survival in Patients With Metastatic Hormone-Sensitive Prostate Cancer. *JAMA Network Open*, 7(10):e2437871, October 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.37871.

[61] Zhichen Gong and Huanhuan Chen. Dynamic state warping, 2017. URL `https://arxiv.org/abs/1703.01141`.

[62] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A Survey on LLM-as-a-Judge, November 2024. URL `http://arxiv.org/abs/2411.15594`. arXiv:2411.15594 [cs] version: 1.

[63] Susan Halabi, Chen-Yen Lin, Eric J. Small, Andrew J. Armstrong, Ellen B. Kaplan, Daniel Petrylak, Cora N. Sternberg, Liji Shen, Stephane Oudard, Johann de Bono, and Oliver Sartor. Prognostic model predicting metastatic castration-resistant prostate cancer survival in men treated with second-line chemotherapy. *Journal of the National Cancer Institute*, 105(22):1729–1737, November 2013. ISSN 1460-2105. doi: 10.1093/jnci/djt280.

[64] Susan Halabi, Chen-Yen Lin, W. Kevin Kelly, Karim S. Fizazi, Judd W. Moul, Ellen B. Kaplan, Michael J. Morris, and Eric J. Small. Updated Prognostic Model for Predicting Overall Survival in First-Line Chemotherapy for Patients With Metastatic Castration-Resistant Prostate Cancer. *Journal of Clinical Oncology*, 32(7):671–677, March 2014.

ISSN 0732-183X. doi: 10.1200/JCO.2013.52.3696. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3927736/.

[65] Anis A. Hamid, Nicolas Sayegh, Bertrand Tombal, Maha Hussain, Christopher J. Sweeney, Julie N. Graff, and Neeraj Agarwal. Metastatic hormone-sensitive prostate cancer: Toward an era of adaptive and personalized treatment. *American Society of Clinical Oncology Educational Book*, (43):e390166, May 2023. ISSN 1548-8748. doi: 10.1200/EDBK_390166.

[66] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. ISSN 0035-9254. doi: 10.2307/2346830. URL https://www.jstor.org/stable/2346830.

[67] Nicole R Haug, Anita K Wagner, Katherine A McGlynn, Charles E Leonard, Michael D Nguyen, and Jacqueline M Major. New-onset cancer cases in FDA's Sentinel System: A large distributed system of US electronic healthcare data. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 31(10):1890–1895, October 2022. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-21-1451.

[68] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

[69] Michael S. Hofman, Nathan Lawrentschuk, Roslyn J. Francis, Colin Tang, Ian Vela, Paul Thomas, Natalie Rutherford, Jarad M. Martin, Mark Frydenberg, Ramdave Shakher, Lih-Ming Wong, Kim Taubman, Sze Ting Lee, Edward Hsiao, Paul Roach, Michelle Nottage, Ian Kirkwood, Dickon Hayne, Emma Link, Petra Marusic, Anetta Matera, Alan Herschtal, Amir Iravani, Rodney J. Hicks, Scott Williams, Declan G. Murphy, and proPSMA Study Group Collaborators. Prostate-specific membrane antigen PET-CT in patients with high-risk prostate cancer before curative-intent

surgery or radiotherapy (proPSMA): A prospective, randomised, multicentre study. *Lancet (London, England)*, 395(10231):1208–1216, April 2020. ISSN 1474-547X. doi: 10.1016/S0140-6736(20)30314-7.

[70] Knut Holtedahl, Peter Hjertholm, Lars Borgquist, Gé A Donker, Frank Buntinx, David Weller, Tonje Braaten, Jörgen Månsson, Eva Lena Strandberg, Christine Campbell, Joke C Korevaar, and Ranjan Parajuli. Abdominal symptoms and cancer in the abdomen: prospective cohort study in European primary care. *The British Journal of General Practice*, 68(670):e301–e310, May 2018. ISSN 0960-1643. doi: 10.3399/bjgp18X695777. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5916077/.

[71] Adrian O Hosten. Bun and creatinine. *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition*, 1990.

[72] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[73] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv: The Preprint Server for Biology*, page 2025.05.30.656746, June 2025. ISSN 2692-8205. doi: 10.1101/2025.05.30.656746.

[74] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS169.

[75] Shinya Iwase, Taka-aki Nakada, Tadanaga Shimada, Takehiko Oami, Takashi Shimazui, Nozomi Takahashi, Jun Yamabe, Yasuo Yamao, and Eiryo Kawakami. Prediction

algorithm for ICU mortality and length of stay using machine learning. *Scientific Reports*, 12:12912, July 2022. doi: 10.1038/s41598-022-17091-5.

[76] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G. Wilson. On Feature Learning in the Presence of Spurious Correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fb64a552feda3d981dbe43527a80a07e-Abstract-Conference.html.

[77] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6): 395–405, 2012.

[78] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

[79] Nari Johnson, Sonali Parbhoo, Andrew S Ross, and Finale Doshi-Velez. Learning Predictive and Interpretable Timeseries Summaries from ICU Data. *AMIA Annual Symposium Proceedings*, 2021:581–590, February 2022. ISSN 1942-597X.

[80] R. A. Johnson and G. D. Roodman. Hematologic manifestations of malignancy. *Disease-a-month: DM*, 35(11):721–768, November 1989. ISSN 0011-5029. doi: 10.1016/0011-5029(89)90011-4.

[81] Alexander W Jung, Peter C Holm, Kumar Gaurav, Jessica Xin Hjaltelin, Davide Placido, Laust Hvas Mortensen, Ewan Birney, Moritz Gerstung, et al. Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study. *The Lancet Digital Health*, 6(6):e396–e406, 2024.

[82] Hye Seon Kang, Ah Young Shin, Chang Dong Yeo, Chan Kwon Park, Ju Sang Kim, Jin Woo Kim, Seung Joon Kim, Sang Haak Lee, and Sung Kyoung Kim. Clinical significance of anemia as a prognostic factor in non-small cell lung cancer carcinoma with activating epidermal growth factor receptor mutations. *Journal of Thoracic Disease*, 12 (5):1895, 2020.

[83] Hormuzd A. Katki, Stephanie A. Kovalchik, Christine D. Berg, Li C. Cheung, and Anil K. Chaturvedi. Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. *JAMA*, 315(21):2300–2311, June 2016. ISSN 1538-3598. doi: 10.1001/jama.2016.6255.

[84] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1): 24, February 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1.

[85] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. URL `https://arxiv.org/abs/2004.11362`.

[86] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural Controlled Differential Equations for Irregular Time Series. In *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707. Curran Associates, Inc., 2020.

[87] Ellen Kim, Samuel M. Rubinstein, Kevin T. Nead, Andrzej P. Wojcieszynski, Peter E. Gabriel, and Jeremy L. Warner. The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*, 29(4):354–361, October 2019. ISSN 1532-9461. doi: 10.1016/j.semradonc.2019.05.010.

[88] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka

Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

[89] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6 (4):e281–e290, 2024.

[90] Kurt Kroenke, Kathryn J. Ruddy, Deirdre R. Pachman, Veronica Grzegorczyk, Jeph Herrin, Parvez A. Rahman, Kyle A. Tobin, Joan M. Griffin, Linda L. Chlan, Jessica D. Austin, Jennifer L. Ridgeway, Sandra A. Mitchell, Keith A. Marsolo, and Andrea L. Cheville. Using Electronic Health Records to Classify Cancer Site and Metastasis. *Applied Clinical Informatics*, 16(3):556–568, June 2025. ISSN 1869-0327. doi: 10.1055/a-2544-3117. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12176508/.

[91] Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, volume 2025, pages 20715–20735, 2025.

[92] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression, September 2019. URL http://arxiv.org/abs/1907.00825. arXiv:1907.00825 [stat].

[93] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for

large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[94] Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. DuETT: Dual event time transformer for electronic health records. In *Machine Learning for Healthcare Conference*, pages 403–422. PMLR, 2023.

[95] Harriet L. Lancaster, Marjolein A. Heuvelmans, and Matthijs Oudkerk. Low-dose computed tomography lung cancer screening: Clinical evidence and implementation research. *Journal of Internal Medicine*, 292(1):68–80, July 2022. ISSN 1365-2796. doi: 10.1111/joim.13480.

[96] Wilson Lau, Youngwon Kim, Sravanthi Parasa, Md Enamul Haque, Anand Oka, and Jay Nanduri. Predicting Early-Onset Colorectal Cancer with Large Language Models, June 2025. URL `http://arxiv.org/abs/2506.11410`. arXiv:2506.11410 [cs].

[97] Yann LeCun. A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27.

[98] Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, January 2020. ISSN 1558-2531. doi: 10.1109/TBME.2019.2909027.

[99] Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text, 2025. URL `https://arxiv.org/abs/2504.03964`.

[100] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive

NLP Tasks, April 2021. URL http://arxiv.org/abs/2005.11401. arXiv:2005.11401 [cs].

[101] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Comput. Surv.*, 55(9), January 2023. ISSN 0360-0300. doi: 10.1145/3555803. URL https://doi.org/10.1145/3555803.

[102] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

[103] Mingchen Li, Zaifu Zhan, Jiatan Huang, Jeremy Yeung, Kai Ding, Anne Blaes, Steven Johnson, Hongfang Liu, Hua Xu, and Rui Zhang. CancerLLM: a large language model in cancer domain. *npj Digital Medicine*, February 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02441-8. URL https://www.nature.com/articles/s41746-026-02441-8.

[104] Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Caread: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, August 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01940-4.

[105] Wang-Jia Li, Fa-Jin Lv, Yi-Wen Tan, Bin-Jie Fu, and Zhi-Gang Chu. Benign and malignant pulmonary part-solid nodules: differentiation via thin-section computed tomography. *Quantitative Imaging in Medicine and Surgery*, 12(1), 2021. ISSN 2223-4306. URL https://qims.amegroups.org/article/view/77781.

[106] Xiudi Li, Erin Y. Yuan, Stephen J. Kuperberg, Clara-Lea Bonzel, Mary I. Jeffway, Tianrun Cai, Katherine P. Liao, Raquel Aguiar-Ibáñez, Yu-Han Kao, Melissa L. Santorelli, David C. Christiani, Tianxi Cai, and Rui Duan. Early detection of non-small cell lung cancer: An electronic health record data-driven approach. *BMC Medicine*, 23: 551, October 2025. ISSN 1741-7015. doi: 10.1186/s12916-025-04289-3.

[107] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155, April 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-62922-y.

[108] Yuwei Li, Chengting Lin, Lei Cui, Chao Huang, Liting Shi, Shiyang Huang, Yue Yu, Xianglan Zhou, Qian Zhou, Kun Chen, and Lei Shi. Association between age and lung cancer risk: evidence from lung lobar radiomics. *BMC Medical Imaging*, 25: 204, June 2025. ISSN 1471-2342. doi: 10.1186/s12880-025-01747-5. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12143070/.

[109] Junwei Liao, Muning Wen, Jun Wang, and Weinan Zhang. Marft: Multi-agent reinforcement fine-tuning. *arXiv preprint arXiv:2504.16129*, 2025.

[110] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS, December 2019. URL http://arxiv.org/abs/1912.06059. arXiv:1912.06059 [cs].

[111] Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang, Sujian Li, and Zhaoxiang Zhang. A comprehensive survey on long context language modeling, 2025. URL https://arxiv.org/abs/2503.17407.

[112] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

[113] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[114] Xinghua Lou, Miguel Lázaro-Gredilla, Antoine Dedieu, Carter Wendelken, Wolfgang Lehrach, and Kevin P. Murphy. AutoHarness: improving LLM agents by automatically synthesizing a code harness, February 2026. URL https://arxiv.org/abs/2603.03329v1.

[115] Konstantinos Loverdos, Andreas Fotiadis, Chrysoula Kontogianni, Marianthi Iliopoulou, and Mina Gaga. Lung nodules: A comprehensive review on current approach and management. *Annals of Thoracic Medicine*, 14(4):226–238, 2019. ISSN 1817-1737. doi: 10.4103/atm.ATM_110_19. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6784443/.

[116] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[117] Hao Ma, Tianyi Hu, Zhiqiang Pu, Boyin Liu, Xiaolin Ai, Yanyan Liang, and Min Chen. Coevolving with the other you: Fine-tuning llm with sequential cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 37:15497–15525, 2024.

[118] Nikita Makarov, Maria Bordukova, Papichaya Quengdaeng, Daniel Garger, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P Menden. Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine*, 8(1): 588, 2025.

[119] Joaquin Mateo, Suzanne Carreira, Shahneen Sandhu, Susana Miranda, Helen Mossop, Raquel Perez-Lopez, Daniel Nava Rodrigues, Dan Robinson, Aurelius Omlin, Nina Tunariu, Gunther Boysen, Nuria Porta, Penny Flohr, Alexa Gillman, Ines Figueiredo,

Claire Paulding, George Seed, Suneil Jain, Christy Ralph, Andrew Protheroe, Syed Hussain, Robert Jones, Tony Elliott, Ursula McGovern, Diletta Bianchini, Jane Goodall, Zafeiris Zafeiriou, Chris T. Williamson, Roberta Ferraldeschi, Ruth Riisnaes, Bernardette Ebbs, Gemma Fowler, Desamparados Roda, Wei Yuan, Yi-Mi Wu, Xuhong Cao, Rachel Brough, Helen Pemberton, Roger A'Hern, Amanda Swain, Lakshmi P. Kunju, Rosalind Eeles, Gerhardt Attard, Christopher J. Lord, Alan Ashworth, Mark A. Rubin, Karen E. Knudsen, Felix Y. Feng, Arul M. Chinnaiyan, Emma Hall, and Johann S. de Bono. DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *New England Journal of Medicine*, 373(18):1697–1708, October 2015. ISSN 0028-4793. doi: 10.1056/NEJMoa1506859.

[120] Matthew B. McDermott, Haoran Zhang, Lasse Hyldig Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at auroc and auprc under class imbalance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, page 44102–44163. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4df3510ad02a86d69dc32388d91606f8-Paper-Conference.pdf.

[121] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [stat].

[122] Bárbara Cristina Jardim Miranda, Francisco Tustumi, Eric Toshiyuki Nakamura, Victor Haruo Shimanoe, Daniel Kikawa, and Jaques Waisberg. Obesity and Colorectal Cancer: A Narrative Review. *Medicina*, 60(8):1218, July 2024. ISSN 1010-660X. doi: 10.3390/medicina60081218. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11355959/.

[123] Lidia-Maria Mondoc, Alina-Camelia Catana, Liiana-Carmen Prodan, Madalina Valeanu,

and Romeo-Gabriel Mihaila. The impact of anemia on the survival of patients diagnosed with low-grade malignant b-cell lymphomas. *Cureus*, 16(7):e65441, 2024.

[124] Intae Moon, Stefan Groha, and Alexander Gusev. SurvLatent ODE : A Neural ODE based time-to-event model with competing risks for longitudinal data improves cancer-associated Venous Thromboembolism (VTE) prediction. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pages 800–827. PMLR, December 2022.

[125] James Morrill, Patrick Kidger, Lingyi Yang, and Terry Lyons. Neural controlled differential equations for online prediction tasks. *arXiv preprint arXiv:2106.11028*, 2021.

[126] Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. EHRWorld: A Patient-Centric Medical World Model for Long-Horizon Clinical Trajectories, February 2026. URL `http://arxiv.org/abs/2602.03569`. arXiv:2602.03569 [cs].

[127] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[128] Linda Mulvihill. The National Transition from ICD-9 to ICD-10. *Journal of registry management*, 38(2):100–109, 2011. ISSN 1945-6123.

[129] Behzad Naderalvojoud, Catherine M Curtin, Chen Yanover, Tal El-Hay, Byungjin Choi, Rae Woong Park, Javier Gracia Tabuenca, Mary Pat Reeve, Thomas Falconer, Keith Humphreys, Steven M Asch, and Tina Hernandez-Boussard. Towards global model generalizability: Independent cross-site feature evaluation for patient-level risk prediction models using the OHDSI network. *Journal of the American Medical Informatics Association : JAMIA*, 31(5):1051–1061, February 2024. ISSN 1067-5027. doi: 10.1093/jamia/ocae028.

[130] Chirag Nagpal, Vincent Jeanselme, and Artur Dubrawski. Deep Parametric Time-to-Event Regression with Time-Varying Covariates. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, pages 184–193. PMLR, May 2021.

[131] Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. *arXiv preprint arXiv:2204.07276*, 2022.

[132] Engineering National Academies of Sciences, Health and Medicine Division, Board on Health Sciences Policy, and Roundtable on Genomics and Precision Health. Exploring the Barriers to Accessing Genomic and Genetic Services. In *Understanding Disparities in Access to Genomic Medicine: Proceedings of a Workshop*. National Academies Press (US), November 2018.

[133] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2025. URL `https://arxiv.org/abs/2402.01613`.

[134] Beatriz Ocaña-Tienda, Alba Eroles-Simó, Julián Pérez-Beteta, Estanislao Arana, and Víctor M Pérez-García. Growth dynamics of lung nodules: implications for classification in lung cancer screening. *Cancer Imaging*, 24(1):113, 2024.

[135] Amy L. Olex and Bridget T. McInnes. Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be. *Journal of Biomedical Informatics*, 118:103784, June 2021. ISSN 1532-0480. doi: 10.1016/j.jbi.2021.103784.

[136] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien

Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b Model Card, August 2025. URL http://arxiv.org/abs/2508.10925. arXiv:2508.10925 [cs].

[137] N. J. Papadoyannakis, C. J. Stefanidis, and M. McGeown. The effect of the correction of metabolic acidosis on nitrogen and potassium balance of patients with chronic renal failure. *The American Journal of Clinical Nutrition*, 40(3):623–627, September 1984. ISSN 0002-9165. doi: 10.1093/ajcn/40.3.623.

[138] Jiheum Park, Chao Pang, Tristan Y. Lee, Jeong Yun Yang, Jacob Berkowitz, Alexander Z. Wei, and Nicholas Tatonetti. Toward Scalable Early Cancer Detection: Evaluating

EHR-Based Predictive Models Against Traditional Screening Criteria. *ArXiv*, page arXiv:2511.11293v2, January 2026. ISSN 2331-8422.

[139] C. Parker, E. Castro, K. Fizazi, A. Heidenreich, P. Ost, G. Procopio, B. Tombal, and S. Gillessen. Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Annals of Oncology*, 31(9):1119–1134, September 2020. ISSN 0923-7534. doi: 10.1016/j.annonc.2020.06.011.

[140] Vimla L Patel, José F Arocha, and Jiajie Zhang. Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning*, 14(727-750):1, 2005.

[141] Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Matthias Keicher, and Nassir Navab. From ehrs to patient pathways: Scalable modeling of longitudinal health trajectories with llms. *arXiv preprint arXiv:2506.04831*, 2025.

[142] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A Prompt-based Topic Modeling Framework. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.164. URL https://aclanthology.org/2024.naacl-long.164/.

[143] Paul F. Pinsky, Amanda Black, Sarah E. Daugherty, Robert Hoover, Howard Parnes, Zachary L. Smith, Scott Eggener, Gerald Andriole, and Sonja I. Berndt. Metastatic Prostate Cancer at Diagnosis and through Progression in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Cancer*, 125(17):2965–2974, September 2019. ISSN 0008-543X. doi: 10.1002/cncr.32176.

[144] Robert Pirker, Katrin Wiesenberger, Gudrun Pohl, and Wilma Minar. Anemia in Lung Cancer: Clinical Impact and Management. *Clinical Lung Cancer*, 5(2):90–97,

September 2003. ISSN 1525-7304, 1938-0690. doi: 10.3816/CLC.2003.n.022. URL https://www.clinical-lung-cancer.com/article/S1525-7304(11)70326-6/abstract.

[145] Davide Placido, Bo Yuan, Jessica X. Hjaltelin, Chunlei Zheng, Amalie D. Haue, Piotr J. Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, Alexander Chowdhury, Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S. Marks, Aviv Regev, Siamack Ayandeh, Mary T. Brophy, Nhan V. Do, Peter Kraft, Brian M. Wolpin, Michael H. Rosenthal, Nathanael R. Fillmore, Søren Brunak, and Chris Sander. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 29(5):1113–1122, May 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02332-5.

[146] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[147] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html.

[148] Maria G Prado, Larry G Kessler, Margaret A Au, Hannah A Burkhardt, Monica Zigman Suchsland, Lesleigh Kowalski, Kari A Stephens, Meliha Yetisgen, Fiona M Walter, Richard D Neal, et al. Symptoms and signs of lung cancer prior to diagnosis: case–control study using electronic health records from ambulatory care within a large us-based tertiary care centre. *BMJ open*, 13(4):e068832, 2023.

[149] W. Nicholson Price. Big Data and Black-Box Medical Algorithms. *Science translational medicine*, 10(471):eaao5333, December 2018. ISSN 1946-6234. doi: 10.1126/scitranslmed.aao5333.

[150] Mohammad Areeb Qazi, Maryam Nadeem, and Mohammad Yaqub. Beyond Generative

AI: World Models for Clinical Prediction, Counterfactuals, and Planning, November 2025. URL http://arxiv.org/abs/2511.16333. arXiv:2511.16333 [cs].

[151] Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. Large language models are zero shot hypothesis proposers, 2023. URL https://arxiv.org/abs/2311.05965.

[152] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation, 2024. URL https://arxiv.org/abs/2407.08940.

[153] GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity Matters: Dynamics of Feature Learning in the Presence of Spurious Correlations, August 2024. URL http://arxiv.org/abs/2403.03375. arXiv:2403.03375 [cs] version: 3.

[154] Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16(1): 9799, November 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64769-1. URL https://www.nature.com/articles/s41467-025-64769-1.

[155] Xiaojie Qiu, Yan Zhang, Jorge D. Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N. Pogson, Marco Y. Hein, Kyung Hoi (Joseph) Min, Li Wang, Emanuelle I. Grody, Matthew J. Shurtleff, Ruoshi Yuan, Song Xu, Yian Ma, Joseph M. Replogle, Eric S. Lander, Spyros Darmanis, Ivet Bahar, Vijay G. Sankaran, Jianhua Xing, and Jonathan S. Weissman. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711.e45, February 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2021.12.045.

[156] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang,

Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[157] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):1–10, May 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0029-1.

[158] Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, Panpan Xu, and Lin Lee Cheong. A systematic survey of automatic prompt optimization techniques, 2025. URL https://arxiv.org/abs/2502.16923.

[159] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, May 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00455-y.

[160] Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer.

npj Digital Medicine, 7(1):256, September 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01235-0. URL https://www.nature.com/articles/s41746-024-01235-0.

[161] Yulia Rubanova, Ricky T. Q. Chen, and David K Duvenaud. Latent Ordinary Differential Equations for Irregularly-Sampled Time Series. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[162] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025. URL https://arxiv.org/abs/2402.07927.

[163] Iqbal H. Sarker. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *Sn Computer Science*, 2(6):420, August 2021. doi: 10.1007/s42979-021-00815-1.

[164] Matthew B Schabath and Michele L Cote. Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, 28(10):1563–1579, 2019.

[165] Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

[166] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos,

Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

[167] Soko Setoguchi, Daniel H. Solomon, Robert J. Glynn, E. Francis Cook, Raisa Levin, and Sebastian Schneeweiss. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data. *Cancer Causes & Control*, 18(5):561–569, June 2007. ISSN 0957-5243, 1573-7225. doi: 10.1007/s10552-007-0131-1.

[168] Navid Shafigh, Morteza Hasheminik, Elnaz Shafigh, Haleh Alipour, Shahram Sayyadi, Neda Kazeminia, Batoul Khoundabi, and Sara Salarian. Prediction of mortality in ICU patients: A comparison between the SOFA score and other indicators. *Nursing in Critical Care*, July 2023. ISSN 1478-5153. doi: 10.1111/nicc.12944.

[169] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh, editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

[170] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

[171] Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, pages 1–9, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09529-3. URL https://www.nature.com/articles/s41586-025-09529-3.

[172] João Figueira Silva and Sérgio Matos. Modelling patient trajectories using multimodal information. *Journal of Biomedical Informatics*, 134:104195, October 2022. ISSN 1532-0464. doi: 10.1016/j.jbi.2022.104195. URL https://www.sciencedirect.com/science/article/pii/S1532046422002003.

[173] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[174] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[175] Oren Smaletz, Howard I. Scher, Eric J. Small, David A. Verbel, Alex McMillan, Kevin Regan, W. Kevin Kelly, and Michael W. Kattan. Nomogram for Overall Survival of Patients With Progressive Metastatic Prostate Cancer After Castration. *Journal of Clinical Oncology*, October 2002. doi: 10.1200/JCO.2002.11.021. URL https://ascopubs.org/doi/10.1200/JCO.2002.11.021.

[176] Daniel E Spratt, Siyi Tang, Yilun Sun, Huei-Chung Huang, Emmalyn Chen, Osama

Mohamad, Andrew J Armstrong, Jonathan D Tward, Paul L Nguyen, Joshua M Lang, Jingbin Zhang, Akinori Mitani, Jeffry P Simko, Sandy DeVries, Douwe Van Der Wal, Hans Pinckaers, Jedidiah M Monson, Holly A Campbell, James Wallace, Michelle J Ferguson, Jean-Paul Bahary, Edward M Schaeffer, Nrg Prostate Cancer Ai Consortium, Howard M Sandler, Phuoc T Tran, Joseph P Rodgers, Andre Esteva, Rikiya Yamashita, and Felix Y Feng. Artificial Intelligence Predictive Model for Hormone Therapy Use in Prostate Cancer. preprint, In Review, April 2023. URL https://www.researchsquare.com/article/rs-2790858/v1.

[177] Felix Steffens, Frederik Wessels, Svetlana Hetjens, Nicolas Carl, Katja Nitschke, Daniel Uysal, Nadim Moharam, Paul Patroi, Thomas Stefan Worst, Karl Friedrich Kowalewski, Maurice Stephan Michel, and Manuel Neuberger. Prognostic factors for overall survival in castration-resistant metastatic prostate cancer treated with docetaxel (MeProCSS): results from a German real-world cohort. *International Urology and Nephrology*, 57(7): 2063–2072, July 2025. ISSN 1573-2584. doi: 10.1007/s11255-025-04389-2.

[178] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

[179] Sunao Tanaka, Lynne R. Wilkens, Loïc Le Marchand, Gong Yang, Jian-Min Yuan, Woon-Puay Koh, Martha J. Shrubsole, Michael Luu, Ian Pagano, Jane C. Figueiredo, Hideki Furuya, and Charles J. Rosser. Developing a prediction model in a large case-control study for the early detection of bladder cancer. *Journal of Translational Medicine*, 24:49, December 2025. ISSN 1479-5876. doi: 10.1186/s12967-025-07511-1.

[180] Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, et al. Agent kb: Leveraging cross-domain experience for agentic problem solving. *arXiv preprint arXiv:2507.06229*, 2025.

[181] Baichuan-M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, Jinyang Tai, Jixiang Hong, Kai Lu, Linzhuang Sun, Peidong Guo, Qian Ma, Rihui Xin, Shihui Yang, Shusen Zhang, Yichuan Mo, Zheng Liang, Zhishou Zhang, Hengfu Cui, Zuyi Zhu, and Xiaochuan Wang. Baichuan-M2: Scaling Medical Capability with Large Verifier System, September 2025. URL http://arxiv.org/abs/2509.02208. arXiv:2509.02208 [cs].

[182] Baichuan-M3 Team, Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, Jinyang Tai, Linzhuang Sun, Peidong Guo, Yichuan Mo, Xiaochuan Wang, Hengfu Cui, and Zhishou Zhang. Baichuan-M3: Modeling Clinical Inquiry for Reliable Medical Decision-Making, February 2026. URL http://arxiv.org/abs/2602.06570. arXiv:2602.06570 [cs].

[183] Abhishek Tripathi, Yu-Hui Chen, David Frazier Jarrard, Jorge A. Garcia, Robert Dreicer, Glenn Liu, Maha H. A. Hussain, Daniel Shevrin, Matthew M. Cooney, Mario A. Eisenberger, Manish Kohli, Elizabeth R. Plimack, Nicholas J. Vogelzang, Joel Picus, Michael A Carducci, Robert S. DiPaola, and Christopher Sweeney. Impact of PSA nadir on long-term survival in metastatic hormone-sensitive prostate cancer: Final 10-year analysis of the ECOG-ACRIN E3805 CHAARTED trial. *Journal of Clinical Oncology*, 43(5_suppl):167–167, February 2025. ISSN 0732-183X. doi: 10.1200/JCO.2025.43.5_suppl.167.

[184] Koichi Uemura, Yasuhide Miyoshi, Takashi Kawahara, Shuko Yoneyama, Yusuke Hattori, Jun-ichi Teranishi, Keiichi Kondo, Masatoshi Moriyama, Shigeo Takebayashi, Yumiko Yokomizo, Masahiro Yao, Hiroji Uemura, and Kazumi Noguchi. Prognostic value of a computer-aided diagnosis system involving bone scans among men treated with docetaxel for metastatic castration-resistant prostate cancer. *BMC Cancer*, 16

(1):109, February 2016. ISSN 1471-2407. doi: 10.1186/s12885-016-2160-1. URL https://doi.org/10.1186/s12885-016-2160-1.

[185] Michael Vainrib and Ilan Leibovitch. Urological implications of concurrent bladder and lung cancer. *The Israel Medical Association journal: IMAJ*, 9(10):732–735, October 2007. ISSN 1565-1088.

[186] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[187] Andrew J. Vickers and Simon F. Brewster. Psa velocity and doubling time in diagnosis and prognosis of prostate cancer. *British journal of medical surgical urology*, 5(4): 162–168, 2012. ISSN 1875-9742. doi: 10.1016/j.bjmsu.2011.08.006.

[188] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[189] Ananta Wadhwa, Charlotte Roscoe, Elizabeth A. Duran, Lorna Kwan, Candace L. Haroldsen, Jeremy B. Shelton, Jennifer Cullen, Beatrice S. Knudsen, Mathew B. Rettig, Saiju Pyarajan, Nicholas G. Nickols, Kara N. Maxwell, Kosj Yamoah, Brent S. Rose, Timothy R. Rebbeck, Hari S. Iyer, and Isla P. Garraway. Neighborhood Deprivation, Race and Ethnicity, and Prostate Cancer Outcomes Across California Health Care Systems. *JAMA network open*, 7(3):e242852, March 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.2852.

[190] Sushrut S. Waikar and Joseph V. Bonventre. Creatinine Kinetics and the Definition of

Acute Kidney Injury. *Journal of the American Society of Nephrology : JASN*, 20(3): 672–679, March 2009. ISSN 1046-6673. doi: 10.1681/ASN.2008070669.

[191] Tonya Walser, Xiaoyan Cui, Jane Yanagawa, Jay M. Lee, Eileen Heinrich, Gina Lee, Sherven Sharma, and Steven M. Dubinett. Smoking and Lung Cancer. *Proceedings of the American Thoracic Society*, 5(8):811–815, December 2008. ISSN 1546-3222. doi: 10.1513/pats.200809-100TH. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4080902/.

[192] Lu Wang, Yan Li, and Mark Chignell. Combining ranking and point-wise losses for training deep survival analysis models. In *2021 IEEE international conference on data mining (ICDM)*, pages 689–698. IEEE, 2021.

[193] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, ACM CHIL '20, page 222–235. ACM, April 2020. doi: 10.1145/3368555.3384469. URL http://dx.doi.org/10.1145/3368555.3384469.

[194] Yajunzi Wang, Jing Li, Haoruo Zha, Shuhe Liu, Daiyun Huang, Lei Fu, and Xin Liu. Paradigms, innovations, and biological applications of RNA velocity: A comprehensive review. *Briefings in Bioinformatics*, 26(4):bbaf339, July 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf339.

[195] Jun Wen, Jue Hou, Clara-Lea Bonzel, Yihan Zhao, Victor M. Castro, Vivian S. Gainer, Dana Weisenfeld, Tianrun Cai, Yuk-Lam Ho, Vidul A. Panickan, Lauren Costa, Chuan Hong, J. Michael Gaziano, Katherine P. Liao, Junwei Lu, Kelly Cho, and Tianxi Cai. Latte: Label-efficient incident phenotyping from longitudinal electronic health records, 2023. URL https://arxiv.org/abs/2305.11407.

[196] Katriina Whitaker. Earlier diagnosis: the importance of cancer symptoms. *The Lancet Oncology*, 21(1):6–8, 2020.

[197] Julia Wilkerson, Kald Abdallah, Charles Hugh-Jones, Greg Curt, Mace Rothenberg, Ronit Simantov, Martin Murphy, Joseph Morrell, Joel Beetsch, Daniel J. Sargent, Howard I. Scher, Peter Lebowitz, Richard Simon, Wilfred D. Stein, Susan E. Bates, and Tito Fojo. Estimation of tumour regression and growth rates during treatment in patients with advanced prostate cancer: a retrospective analysis. *The Lancet Oncology*, 18(1):143–154, January 2017. ISSN 1470-2045, 1474-5488. doi: 10.1016/S1470-2045(16)30633-7.

[198] Andrew MD Wolf, Kevin C Oeffinger, Tina Ya-Chen Shih, Louise C Walter, Timothy R Church, Elizabeth TH Fontham, Elena B Elkin, Ruth D Etzioni, Carmen E Guerra, Rebecca B Perkins, et al. Screening for lung cancer: 2023 guideline update from the american cancer society. *CA: A Cancer Journal for Clinicians*, 74(1):50–81, 2024.

[199] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[200] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, 2024.

[201] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng,

Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs, April 2025. URL http://arxiv.org/abs/2504.00993. arXiv:2504.00993 [cs].

[202] Minghao Wu, Yuting Yan, Zhenyang Cai, Ke Ji, Chuangsen Fang, Ziying Sheng, Xidong Wang, Rongsheng Wang, Hejia Zhang, Shuang Li, Benyou Wang, and Hongyuan Zha. Agentifying Patient Dynamics within LLMs through Interacting with Clinical World Model, May 2026. URL http://arxiv.org/abs/2605.14723. arXiv:2605.14723 [cs].

[203] Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, et al. Evolver: Self-evolving llm agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*, 2025.

[204] Xifeng Wu, Huakang Tu, Qingfeng Hu, Shan Pou Tsai, David Ta-Wei Chu, and Chi-Pang Wen. Novel machine learning algorithm in risk prediction model for pan-cancer risk: application in a large prospective cohort. *BMJ Oncology*, 3(1):e000087, July 2024. ISSN 2752-7948. doi: 10.1136/bmjonc-2023-000087. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11261702/.

[205] Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of Biomedical Informatics*, 126:103980, February 2022. ISSN 1532-0480. doi: 10.1016/j.jbi.2021.103980.

[206] Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, et al. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*, 2025.

[207] Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. medDreamer: Model-Based Reinforcement Learning with Latent Imagination on Complex EHRs for Clinical Decision Support, December 2025. URL http://arxiv.org/abs/2505.19785. arXiv:2505.19785 [cs].

[208] Renjun Xu and Yang Yan. Agent Skills for Large Language Models: Architecture, Acquisition, Security, and the Path Forward, February 2026. URL http://arxiv.org/abs/2602.12430. arXiv:2602.12430 [cs].

[209] Beihua Yang, Peng Song, Yuanbo Cheng, Shixuan Zhou, and Zhaowei Liu. Enhanced tensor based embedding anchor learning for multi-view clustering. *Information Sciences*, 690:121532, 2025.

[210] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

[211] Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, Alasdair Edward Gent, Victor Moas, and Rishikesan Kamaleswaran. Clin-JEPA: A Multi-Phase Co-Training Framework for Joint-Embedding Predictive Pretraining on EHR Patient Trajectories, May 2026. URL http://arxiv.org/abs/2605.10840. arXiv:2605.10840 [cs].

[212] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, March 2023. URL http://arxiv.org/abs/2210.03629. arXiv:2210.03629 [cs].

[213] Huaiyuan Ying, Zhengyun Zhao, Yang Zhao, Sihang Zeng, and Sheng Yu. CoRTEx: contrastive learning for representing terms via explanations with applications on construct-

ing biomedical knowledge graphs. *Journal of the American Medical Informatics Association*, 31(9):1912–1920, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae115. URL https://doi.org/10.1093/jamia/ocae115.

[214] Huaiyuan Ying, Hongyi Yuan, Jinsen Lu, Zitian Qu, Yang Zhao, Zhengyun Zhao, Isaac Kohane, Tianxi Cai, and Sheng Yu. GENIE: Generative Note Information Extraction model for structuring EHR data, January 2025.

[215] Alex Youssef, Michael Pencina, Anshul Thakur, Tingting Zhu, David Clifton, and Nigam H. Shah. All models are local: Time to replace external validation with recurrent local validation, May 2023.

[216] Guan-Hua Yu, Shuo-Feng Li, Ran Wei, and Zheng Jiang. Diabetes and Colorectal Cancer Risk: Clinical and Therapeutic Implications. *Journal of Diabetes Research*, 2022:1747326, March 2022. ISSN 2314-6745. doi: 10.1155/2022/1747326. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8920658/.

[217] Sheng Yu, Zheng Yuan, Jun Xia, Shengxuan Luo, Huaiyuan Ying, Sihang Zeng, Jingyi Ren, Hongyi Yuan, Zhengyun Zhao, Yucong Lin, Keming Lu, Jing Wang, Yutao Xie, and Heung-Yeung Shum. BIOS: An Algorithmically Generated Biomedical Knowledge Graph, April 2022.

[218] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.01825.

[219] Sihang Zeng, Zheng Yuan, and Sheng Yu. Automatic Biomedical Term Clustering by Learning Fine-grained Term Representations. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 91–96, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.8.

[220] Sihang Zeng, Yujuan Fu, Sitong Zhou, Zixuan Yu, Lucas Jing Liu, Jun Wen, Matthew Thompson, Ruth Etzioni, and Meliha Yetisgen. Traj-coa: Patient trajectory modeling via chain-of-agents for lung cancer risk prediction. *arXiv preprint arXiv:2510.10454*, 2025.

[221] Sihang Zeng, Lucas Jing Liu, Jun Wen, Meliha Yetisgen, Ruth Etzioni, and Gang Luo. TrajSurv: Learning Continuous Latent Trajectories from Electronic Health Records for Trustworthy Survival Prediction. In *Proceedings of the 10th Machine Learning for Healthcare Conference*. PMLR, October 2025.

[222] Sihang Zeng, Young Won Kim, Wilson Lau, Ehsan Alipour, Ruth Etzioni, Meliha Yetisgen, and Anand Oka. Trajonco: a multi-agent framework for temporal reasoning over longitudinal ehr for multi-cancer early detection. *arXiv preprint arXiv:2604.10386*, 2026.

[223] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-N-Contrast: Learning Continuous Representations for Regression. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[224] Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale, April 2026. URL `https://arxiv.org/abs/2604.18570v2`.

[225] Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, et al. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*, 8(1): 239, 2025.

[226] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma,

Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. Ultramedical: Building specialized generalists in biomedicine, 2024.

[227] Kaiyan Zhang, Kai Tian, Runze Liu, Sihang Zeng, Xuekai Zhu, Guoli Jia, Yuchen Fan, Xingtai Lv, Yuxin Zuo, Che Jiang, et al. Marti: A framework for multi-agent llm systems reinforced training and inference. In *The Fourteenth International Conference on Learning Representations*, 2025.

[228] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A Survey of Reinforcement Learning for Large Reasoning Models, October 2025. URL `http://arxiv.org/abs/2509.08827`. arXiv:2509.08827 [cs].

[229] Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for ehr foundation models. *arXiv preprint arXiv:2505.22964*, 2025.

[230] Yu Zhang and Qiang Yang. A Survey on Multi-Task Learning, March 2021. URL `http://arxiv.org/abs/1707.08114`. arXiv:1707.08114 [cs].

[231] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.

[232] Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large

language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025.

[233] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642, 2024.

[234] Jun Zhao, Can Zu, Hao Xu, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. Longagent: Scaling language models to 128k context through multi-agent collaboration, 2024. URL https://arxiv.org/abs/2402.11550.

[235] Renjia Zhao, Huangbo Yuan, Yanfeng Jiang, Zhenqiu Liu, Ruilin Chen, Shuo Wang, Linyao Lu, Ziyu Yuan, Zhixi Su, Qiye He, et al. Development and validation of an integrative 54 biomarker-based risk identification model for multi-cancer in 42,666 individuals: a population-based prospective study to guide advanced screening strategies. *Biomarker Research*, 13(1):101, 2025.

[236] Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, Xin Sun, Ya Zhang, Yongguo Yu, Kun Sun, and Weidi Xie. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, pages 1–10, February 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-10097-9. URL https://www.nature.com/articles/s41586-025-10097-9.

[237] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. URL http://arxiv.org/abs/2306.05685. arXiv:2306.05685 [cs].

[238] Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, et al. Memento: Fine-tuning llm agents without fine-tuning llms. *arXiv preprint arXiv:2508.16153*, 2025.

[239] Sitong Zhou, Meliha Yetisgen, and Mari Ostendorf. RadTimeline: Timeline Summarization for Longitudinal Radiological Lung Findings, March 2026. URL https://arxiv.org/abs/2603.22820v1.

[240] Menglei Zhu, Hui Lin, Jue Jiang, Abbas J. Jinia, Justin Jee, Karl Pichotta, Michele Waters, Doori Rose, Nikolaus Schultz, Sulov Chalise, Lohit Valleru, Olivier Morin, Jean Moran, Joseph O. Deasy, Shirin Pilai, Chelsea Nichols, Gregory Riely, Lior Z. Braunstein, and Anyi Li. Large language model trained on clinical oncology data predicts cancer progression. *npj Digital Medicine*, 8(1):397, July 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01780-2. URL https://www.nature.com/articles/s41746-025-01780-2.

[241] Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*, 2024.

[242] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey, 2024. URL https://arxiv.org/abs/2308.07107.

Appendix A

TRAJSURV-MPC

A.1 Datasets

A.1.1 Data Curation

All variables in the original MSK data were manually extracted via chart review.

Variables in the original VA data were curated in a mixed way. The detection of metastasis and Gleason scores was extracted through natural language processing (NLP) algorithms. Treatments and demographics were existing variables in the patient's electronic medical records. The AJCC stage was obtained from the cancer registry. As for the treatments, ADT groups the use of orchiectomy, leuprolide, goserelin, triptorelin, histrelin, buserelin, degarelix, or relugolix, and AAT groups the use of flutamide, bicalutamide, nilutamide, enzalutamide, apalutamide, darolutamide, or abiraterone.

A.1.2 MSK Cohort

The MSK cohort was derived from a larger cohort of 16,025 patients from Memorial Sloan Kettering Cancer Center with localized prostate cancer treated with radical prostatectomy (RP) between 2000 and 2022. Time of administration of the first salvage therapy after biochemical recurrence (BCR, defined as two consecutive PSA measures not lower than 0.1 ng/mL at least 3 months after RP) and time of clinical diagnosis with mPC were recorded, as were time of death and last follow-up. The metastasis was identified manually from the patient's medical records.

We identified 403 progressed metastatic patients for the MSK cohort in this study. Patients were excluded if they had fewer than three PSA measurements or developed metastasis

within one year of RP. The baseline clinical features included pre-RP PSA, age, pathologic Gleason score, pathologic T-stage, surgical margins, lymph node involvement, and Charlson Comorbidity Index (CCI). The PSA trajectory involved PSA measurements following RP and before metastasis. The treatment information included the time of administration of the first salvage therapy, including radiation therapy (RT), hormone therapy (HT), or both.

A.1.3 VA Cohort

The VA cohort was derived from a larger cohort of male veterans at risk of prostate cancer who received VA care on at least two or more visits between 2005 and 2020 and had a minimum of five years of follow-up. Localized prostate cancer cases were identified, and clinical/demographic data were obtained from the VA Multi-OMICS Analysis Platform for Prostate Cancer (VA-MAPP) database housed on the VA Informatics and Computing Infrastructure (VINCI) server [189]. Serial PSA measurements and the first initiation of each prostate cancer therapy, including RP, RT, anti-androgen therapy (AAT), and androgen deprivation therapy (ADT), were recorded. The metastasis was identified and recorded through a natural language processing (NLP) algorithm deployed on oncology, urology, and radiology clinical notes [12].

We identified 9,984 progressed metastatic patients for the VA cohort in this study. Patients were included if they were in the VA Central Cancer Registry (VACCR), diagnosed with localized prostate cancer between 01/01/2005 and 01/01/2019, and developed metastasis before 07/30/2024. Similar to the MSK cohort, patients with fewer than 3 PSA measurements between diagnosis and metastasis or a time from localized disease diagnosis to metastasis of less than 1 year were excluded.

The baseline clinical features were measured at localized disease diagnosis, including CCI, Gleason score, body mass index (BMI), AJCC overall stage, age, and race and ethnicity. All PSA measurements following localized disease diagnosis and before metastasis were extracted. The treatment information included the time of initiation of RP, RT, AAT, and ADT.

A.1.4 Data Preprocessing and Feature Extraction

The model input of TrajSurv-mPC consisted of the pre-metastasis patient data, which included: baseline clinical features, longitudinal PSA trajectories, and treatment histories before the detection of metastasis. To prepare these data and account for the inherent irregularities of real-world EHRs, several preprocessing steps were performed. To handle the irregularly sampled PSA trajectories (Table A.2, A.4), PSA values were log-transformed (log(PSA+1)) and linearly interpolated to regular monthly intervals.

For treatment histories, we utilized the first documented initiation of each therapy. Treatment trajectories were encoded for each therapy as binary monthly indicators (0=pre-initiation, 1=post-initiation). If a specific treatment was not recorded in the EHR, we assumed the patient did not receive it. Because the data were extracted via manual chart review or from comprehensive, integrated networks in VA, where records are linked across network facilities, we assume that patients with complete diagnosis and metastasis records were actively receiving care, hence minimizing the impact of incomplete follow-up windows.

The combination of PSA and treatment trajectory formed a sequential multivariate time series data, which we called the patient trajectory. Specifically, the patient trajectory in the MSK cohort had 3 dimensions (logPSA, RT, and HT), while the patient trajectory in the VA cohort had 5 dimensions (logPSA, RP, RT, AAT, and ADT). Missing baseline clinical features were imputed using the mean for continuous variables or the mode for categorical variables, after which all baseline features were standardized. The missingness burden for these baseline clinical features was exceptionally low. The MSK cohort had 0.0% missingness across all baseline features (Table A.1). The VA cohort exhibited minimal missingness, limited to AJCC overall stage (4.1%), BMI (4.1%), and Gleason score (2.1%) (Table A.3). The primary outcome was OS from the detection of metastasis, with data censored at the last follow-up in MSK data or 07/30/2024 in VA data.

To establish a performance benchmark and highlight the added value of using the full patient trajectory, we developed several standard machine learning models for comparison.

These comparison models utilized the same set of baseline clinical features as TrajSurv-mPC, but were supplemented with a set of summary features derived from the longitudinal data. For the MSK cohort, the summary features included: time to BCR, time to metastasis, the type of salvage therapy administered, time to salvage therapy, the inverse of the PSA doubling time, and PSA values measured at BCR, at metastasis, and the highest recorded level. For the VA cohort, the summary features included: time to metastasis, indicator variables for the utilization of each therapy, time to each therapy, and PSA values at diagnosis, at metastasis, and the highest and lowest observed levels. Following extraction, all features were standardized before model training.

Table A.1: Missingness of baseline clinical features in the MSK cohort.

MSK Baseline Clinical Features	**Missingness (n, %)**
Age at RP	0, 0.0%
CCI	0, 0.0%
Lymph node involvement	0, 0.0%
Pathologic Gleason score	0, 0.0%
Pathologic T-stage	0, 0.0%
Pre-RP PSA	0, 0.0%
Surgical margins	0, 0.0%

Table A.2: PSA trajectory characteristics in the MSK cohort.

MSK PSA Trajectory Characteristics	**Value**
PSA measurements per patient, Median (IQR)	18.0 (10.5 – 28.0)
PSA trajectory duration in months, Median (IQR)	54.9 (32.6 – 92.9)
Time Gap between consecutive PSA measurements in months, Median (IQR)	2.7 (1.2 – 3.7)

Table A.3: Missingness of baseline clinical features in the VA cohort.

VA Baseline Clinical Features	**Missingness (n, %)**
AJCC overall stage	414, 4.1%
Age	0, 0.0%
BMI	412, 4.1%
CCI	0, 0.0%
Gleason score	213, 2.1%
Race/ethnicity	0, 0.0%

Table A.4: PSA trajectory characteristics in the VA cohort.

VA PSA Trajectory Characteristics	**Value**
PSA measurements per patient, Median (IQR)	14.0 (9.0 – 23.0)
PSA trajectory duration in months, Median (IQR)	70.1 (40.3 – 107.6)
Time Gap between consecutive PSA measurements in months, Median (IQR)	3.4 (2.1 – 6.0)

A.2 Method

A.2.1 TrajSurv-mPC Framework

TrajSurv-mPC employs a sequential-model-agnostic deep learning framework to effectively capture temporal dependencies within clinical features and patient trajectories. We implemented this framework using both unidirectional long short-term memory (LSTM) networks [68] and causal Transformer encoder [186] architectures.

A key feature of both implementations is the unified integration of static baseline data and dynamic trajectory data. A patient's baseline clinical features are first passed through fully connected layers (FCs) to create an initial state representation. For the LSTM, this representation is used to initialize the hidden and cell states. For the Transformer, it

serves as the prefix (the first token) of the sequence. Subsequently, the sequential patient trajectory (logPSA and treatments) is processed at monthly intervals. For the Transformer implementation, we applied Rotary Position Embedding (RoPE) [178] and causal attention masking to the sequence. The causal masking ensures the model strictly adheres to the temporal flow of the patient's history, allowing predictions at any given month to depend only on the prefix and preceding trajectory data, effectively preventing future information leakage.

The primary objective for both architectures is survival prediction. The final sequence representation, which serves as a comprehensive summary of the patient's status at the time of metastasis, is fed into a non-linear Cox proportional hazards model [38, 84]. This is trained using a partial likelihood loss [38].

To enhance the ability to learn from temporal dynamics, both TrajSurv-mPC implementations incorporate a multi-task learning strategy [230] with two auxiliary objectives. Specifically, FCs are applied at each time step to predict the subsequent month's PSA value and the probabilities of treatment initiation. These auxiliary tasks are trained with a mean square error loss for the continuous PSA prediction and a binary cross-entropy loss for the probabilistic treatment initiation prediction. By explicitly learning these auxiliary tasks, TrajSurv-mPC is forced to better capture the temporal dependencies underlying the trajectory, improving the clinical relevance and accuracy of the primary survival prediction.

A.2.2 Synthetic Patient Generation in Simulation Study

We created five simulated post-RP patient groups that each emulate a different clinically important treatment pattern and PSA history and evaluated TrajSurv-mPC's predictions in each group. The five groups were as follows: patients with no salvage treatment exposure (No Treatment), patients receiving treatment with no PSA response (No Response), patients receiving treatment with response followed by instant recurrence (Response and Recurrence), patients receiving treatment with short-term sustained response (Short-Sustained Response), and patients receiving treatment with long-term sustained response (Long-Sustained Response). We hypothesize that TrajSurv-mPC will predict different average levels of risk for

each of these groups, with No Response having the greatest risk and Long-Sustained Response having the lowest risk.

For each group, we generated 1,000 patients. The baseline clinical features of each synthetic patient were randomly sampled with replacement from the MSK cohort. To generate the monthly PSA trajectory of each synthetic patient, we first randomly sampled the following parameters from the MSK cohort for each patient: the average slope of pre-metastasis logPSA (denoted by v), initiation time of salvage therapy (denoted by t), and the time from salvage therapy to the detection of metastasis (denoted by d). The time of metastasis was then derived $T = t + d$. The treatment trajectory of the No Treatment group was a constant zero trajectory, and we generated the treatment trajectories for other groups using the initiation time of salvage therapy t following the same data preprocessing procedure of TrajSurv-mPC. We then mathematically defined the monthly PSA trajectories (logPSA y against month m) of the five groups to simulate a post-RP scenario similar to the MSK cohort:

$$y = \begin{cases} v \cdot m, \quad m < T & \text{(No Treatment)} \\ v \cdot m, \quad m < T & \text{(No Response)} \\ \begin{cases} v \cdot m, & m < t \\ v \cdot (m - t), & t \leq m < T \end{cases} & \text{(Response and Recurrence)} \\ \begin{cases} v \cdot m, & m < t \\ 0, & t \leq m < t + \theta d \\ v \cdot (m - t - \theta d), & t + \theta d \leq m < T \end{cases} & \text{(Short or Long Sustained Response)} \end{cases}$$

where $\theta = 0.25$ for the Short-Sustained Response group and $\theta = 0.75$ for the Long-Sustained Response group. Therefore, a synthetic patient consisted of three independently generated components: a sampled baseline clinical feature, a treatment trajectory, and a generated PSA trajectory. The choice of independent sampling was subject to the practicality and feasibility of the simulation study to probe TrajSurv-mPC's behavior during temporal aggregation. We acknowledge that this simulation strategy, which prioritizes variable control,

does not fully emulate real patients. This sampling process is illustrated in Figure 3.4A. For each group, we obtained the final predicted risk scores of the patients.

A.3 Experiments

A.3.1 Software Implementation

TrajSurv-mPC and DeepSurv were developed using pycox 0.3.0 [92]. CoxPH, RSF, and Boosting were developed using scikit-survival 0.21.0 [147]. All experiments were conducted using Python 3.10.

A.3.2 Ablation Study

To evaluate the contribution of TrajSurv-mPC's components, we conducted an ablation study on the MSK cohort using LSTM as the backbone. We systematically removed the baseline clinical features and auxiliary losses (TrajSurv-O), both auxiliary losses (TrajSurv-B), the continuous PSA prediction loss (TrajSurv-T), and the treatment initiation prediction loss (TrajSurv-P), and compared their performance with TrajSurv-mPC.

The removal of baseline clinical characteristics in TrajSurv-O resulted in a C-index of 0.625, a notable decrease compared to the C-index of 0.682 achieved by TrajSurv-B. This underscores the significance of incorporating baseline clinical characteristics as the initial patient state. Furthermore, the models with individual auxiliary losses, TrajSurv-T (C-index=0.705) and TrajSurv-P (C-index=0.691), both outperformed TrajSurv-B. This demonstrates that the auxiliary tasks of predicting PSA trajectory and treatment initiation are effective in capturing the temporal dynamics of the disease. The complete TrajSurv-mPC model, which integrates all components, achieved the highest performance.

Table A.5: Architectural comparison between TrajSurv-mPC (LSTM) and TrajSurv-mPC (Transformer).

Model	**Common**	**Differences**
TrajSurv-mPC (LSTM)	• Same baseline features and PSA trajectories • Same input-output setting • Same multi-task learning objectives • Same survival prediction loss	• LSTM as the sequential model • Baseline features initialize the hidden and cell states • Unidirectional LSTM
TrajSurv-mPC (Transformer)		• Transformer encoder as the sequential model • Rotary position embedding (RoPE) to improve dependency modeling • Baseline features serve as a prefix token in the sequence • Causal attention masking

Appendix B

TRAJSURV-CONT

B.1 Additional Results

B.1.1 Model Calibration

To assess TrajSurv-Cont's calibration, we generated calibration plots across quartiles of follow-up times on MIMIC-III data. Figure B.1 illustrates that TrajSurv-Cont's predicted risk scores align well with observed outcomes at the median follow-up time but show slight underestimation at the 25% quartile and slight overestimation at the 75% quartile. Overall, TrajSurv-Cont's calibration is competitive with other methods.

B.1.2 Phenotypic Properties of Latent Space

To visually examine the phenotypic properties of the clinically aligned latent space, we plotted the dominant phenotypes and corresponding severity across regions in the latent space on MIMIC-III data. For each region, we averaged the SOFA component scores of the 50 nearest latent states, effectively representing the typical phenotype within that area. The dominant phenotype is then obtained from the highest average SOFA component in that region. As shown in Figure B.2, distinct regions in this space represent different clinical phenotypes, while close states reflect consistent dominant phenotypes and severity. For instance, the lower corner of the latent space corresponds to severe cardiovascular conditions. This demonstrates that the proximity in TrajSurv-Cont's latent space indicates similar patient phenotypic patterns.

B.1.3 Feature Importance and Relevance of All Features

Figure B.3 and B.4 show the full feature importance and relevance figures from the vector field interpretation on MIMIC III. We note that these average patterns obtained from data-driven analysis should be carefully interpreted and should not be directly used in clinical practice.

B.1.4 Interpretation Method Comparison

TrajSurv-Cont's interpretation is a two-step process: (1) vector field interpretation for the feature-to-trajectory mapping, and (2) latent trajectory clustering for the trajectory-to-outcome linkage. The vector field interpretation offers insights distinct from, yet complementary to, methods like SHAP [116] or permutation importance [14]:

- It explains how changes in clinical features dynamically influence the magnitude and direction of the patient's latent trajectory over continuous time.
- It allows assessment of feature relevance by measuring how similarly features drive this trajectory evolution via cosine similarity of vector field columns.

Table B.1 compares different interpretation methods. We highlight that TrajSurv-Cont combines vector field interpretation with latent trajectory clustering, hence enabling end-to-end interpretation from a dynamical perspective.

B.1.5 Additional Experiments on Model Generalization

We conducted additional experiments on TrajSurv-Cont's generalization using the MIMIC-III dataset. The results demonstrate TrajSurv-Cont's robustness across different settings.

Full vs. Reduced Data We compared the full data (hourly data from the raw database) and the reduced data (only including hours with more than 20 observed features). The C-index is 0.806 on the full data and 0.803 on the reduced data, with about 3 times faster

training on the reduced data. Therefore, we used reduced data during our experiments, which achieves computational efficiency while preserving fair comparison.

Short Trajectory We performed an experiment using only the data between 18 and 36 hours after admissions, rather than the full 36-hour data in the main paper. TrajSurv-Cont achieves a C-index of 0.788, compared to 0.803 with 36-hour data, indicating robustness despite information loss.

Limited Training Data To assess performance with limited training data, we experimented with a 20% training, 40% validation, and 40% test split on MIMIC-III data. TrajSurv-Cont achieves a C-index of 0.752, compared to 0.746 of RSF, demonstrating generalizability even with reduced training data.

B.1.6 Interpolation Method

TrajSurv-Cont employs cubic Hermite splines with backward differences for input path interpolation. This is recommended by prior NCDE literature [86, 125] for its smoothness, online processing capability, and efficient integration. To explore alternatives, we experimented with rectilinear interpolation, which is also an online method, yielding a C-index of 0.789, slightly below TrajSurv-Cont with cubic splines (C-index 0.803).

B.2 Experimental Details

B.2.1 MIMIC-III Cohort

The MIMIC-III cohort includes 20,258 hospital admissions and 53 clinical features in 1-hour resolution. The dataset has 1,743 events. The 53 clinical features were the lab and chart events with top occurrence and demographics. The clinical features include arterial blood pressure diastolic, arterial blood pressure mean, arterial blood pressure systolic, alanine aminotransferase, alkaline phosphatase, anion gap, aspartate aminotransferase, base excess, basophils, bicarbonate, bilirubin total, calcium total, calculated total carbon dioxide,

chloride, creatinine, central venous pressure, eosinophils, free calcium, glucose, heart rate, hematocrit, hemoglobin, international normalized ratio, lactate, lymphocytes, magnesium, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, monocytes, oxygen saturation, pulmonary artery pressure diastolic, pulmonary artery pressure mean, pulmonary artery pressure systolic, partial pressure of carbon dioxide, pH, phosphate, platelet count, partial pressure of oxygen, potassium, potassium whole blood, prothrombin time, partial thromboplastin time, red cell distribution width, red blood cell count, respiratory rate, sodium, temperature, urea nitrogen, white blood cell count, gender, age, and body mass index.

The MIMIC-III cohort is inherently irregularly sampled. For computational feasibility and a fair comparison with baselines, we only included hourly time points where more than 20 features were observed. This cohort has an average of 4.02 hourly clinical observations per patient within the first 36 hours and an overall feature missingness rate of 57.7% across all time steps.

B.2.2 eICU Cohort

The eICU cohort includes 116,503 hospital admissions and 53 clinical features in 1-hour resolution. The dataset has 7,958 events. The clinical features include basophils, eosinophils, lymphocytes, monocytes, polymorphonuclear leukocytes, alanine aminotransferase, aspartate aminotransferase, blood urea nitrogen, fraction of inspired oxygen, bicarbonate, hematocrit, hemoglobin, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, mean platelet volume, prothrombin time, international normalized ratio, red blood cell count, red cell distribution width, white blood cell count (x1000), albumin, alkaline phosphatase, anion gap, bedside glucose, calcium, chloride, creatinine, glucose, magnesium, pH, partial pressure of carbon dioxide, partial pressure of oxygen, phosphate, platelet count (x1000), potassium, sodium, total bilirubin, total protein, temperature, arterial oxygen saturation, heart rate, respiration rate, central venous pressure, end-tidal carbon dioxide, systemic arterial systolic pressure, systemic arterial diastolic pressure, systemic

arterial mean pressure, pulmonary artery systolic pressure, pulmonary artery diastolic pressure, pulmonary artery mean pressure, age, and body mass index.

The eICU cohort is also inherently irregularly sampled, and similarly preprocessed. This cohort has an average of 3.15 hourly clinical observations per patient within the first 36 hours and an overall feature missingness rate of 55.0% across all time steps.

B.2.3 Hyperparameter Tuning

Hyperparameters were tuned on the validation set. Specifically, for TrajSurv-Cont, we conducted a random search [19] to tune the hyperparameters. The searching space of key hyperparameters is shown in Table B.2. The search ranges were informed by common practice, prior work (e.g., κ_1 based on Rank-N-Contrast [223]), and empirical observations. Hyperparameters were tuned for the highest C-index on the validation set, with early stopping (patience 5) to prevent overfitting.

B.2.4 Optimization

We used AdamW [113] to optimize TrajSurv-Cont. For the NCDE module, we used the *torchcde* package with *torchdiffeq* backend. We trained TrajSurv-Cont for 100 epochs, with early stopping based on the C-index on the validation set, with patience 5. TrajSurv-Cont's training and evaluation were performed on a single NVIDIA Tesla T4 or RTX 2080 Ti GPU.

B.2.5 Comparison Methods Implementation

All baseline models underwent hyperparameter tuning using grid search on the validation set. For CoxPH, RSF, and Boosting, we used the implementations in the scikit-survival library [147]. For RDSM, we used the auton-survival package [131]. DDH and SLODE were implemented using the authors' original publicly available code.

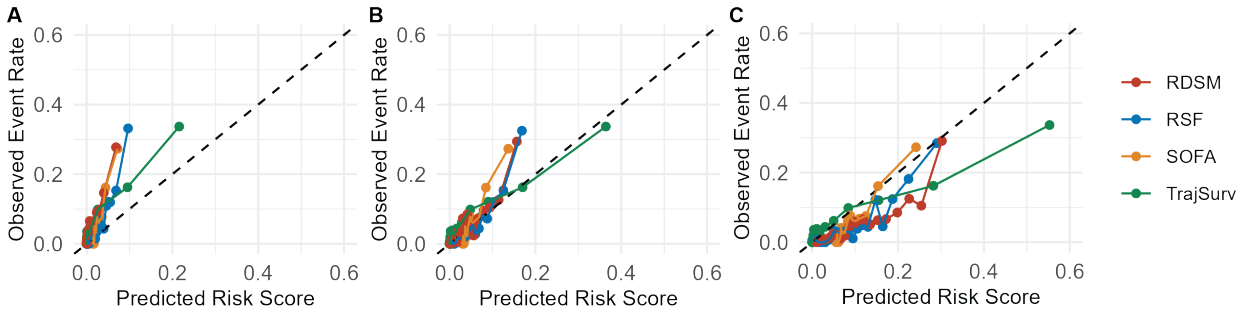

Figure B.1: Calibration plot comparing TrajSurv-Cont with RSF, SOFA, and RDSM at (A) 25% quartile, (B) median, and (C) 75% quartile follow-up time.

B.3 More Background Information

B.3.1 Neural Controlled Differential Equations

Neural controlled differential equations (NCDEs) extend neural ordinary differential equations (NODEs) to handle incoming data in an irregularly sampled time series setting. NCDEs leverage the mathematical framework of controlled differential equations (CDEs) to provide continuous-time modeling that dynamically adapts to data observations. A key advantage of NCDEs is their ability to utilize adjoint backpropagation for efficient training. The training of NCDEs has an overall memory footprint of $\mathcal{O}(L + H)$, where $L = t_n - t_0$ and H is the memory footprint of the vector field. This contrasts previous work on NODEs for time series, which requires $\mathcal{O}(LH)$ memory. [86]

B.3.2 Dynamic Time Warping

Dynamic Time Warping (DTW) is a technique used to measure the similarity between two time series of different lengths or varying speed. Previous work has also extended DTW to series of latent states, called dynamic state warping [61], which is similar to our latent trajectory clustering. The core of DTW involves constructing a cost matrix where each cell (i, j) represents the distance between points in the two time series, and then finding the

optimal warping path through this matrix. The accumulated cost matrix D is built using the following recursive formula:

$$D(i,j) = \text{cost}(i,j) + \min[D(i-1,j), D(i,j-1), D(i-1,j-1)],$$

where $\text{cost}(i,j)$ is the distance between the i-th and j-th points of the two respective time series. The final DTW distance is the value of $D(n,m)$, where n and m are the lengths of the two time series.

In our latent trajectory clustering, the time series in DTW are the hourly extracted latent states from the continuous latent trajectories and the cost is the L2 distance between latent states.

Table B.1: Comparison of Interpretation Methods.

	SHAP / Permutation Importance	**Attention Mechanisms**	**TrajSurv-Cont's Vector Field Interpretation**
Primary Insight	Contribution of each feature to a specific outcome	Which input features/time steps are most influential for an outcome	How feature changes drive the **evolution & direction** of latent state trajectories
Temporal Aspect	Typically static for a given prediction; can be applied at different times but doesn't inherently model evolution	Highlights salient inputs/time steps for discrete predictions	**Models continuous-time dynamics**; captures how features influence on **trajectory evolution**
Focus of Importance	Importance for the magnitude of the final prediction	Importance for the final output	Importance for latent trajectory evolution (**magnitude and direction**)
Feature Interaction	Can compute SHAP interaction values	Does not directly provide feature interactions	**Relevance in co-directing trajectory evolution**
Granularity	Insight into outcome prediction	Insight into outcome prediction	Insight into the feature-to-trajectory **dynamical process**

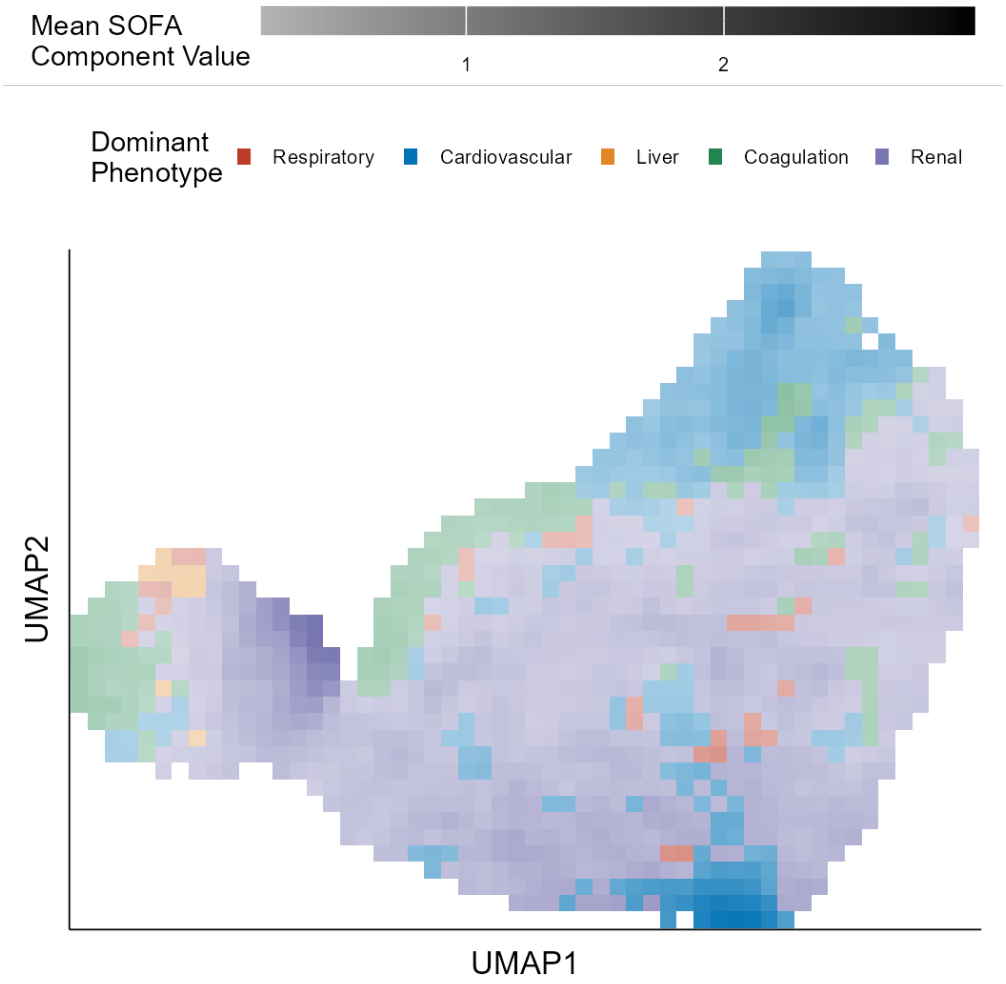

Figure B.2: Dominant phenotypes in different regions of latent space. Different colors represent different dominant phenotypes, and the transparency of the color corresponds to the average severity (SOFA component) of the dominant phenotypes.

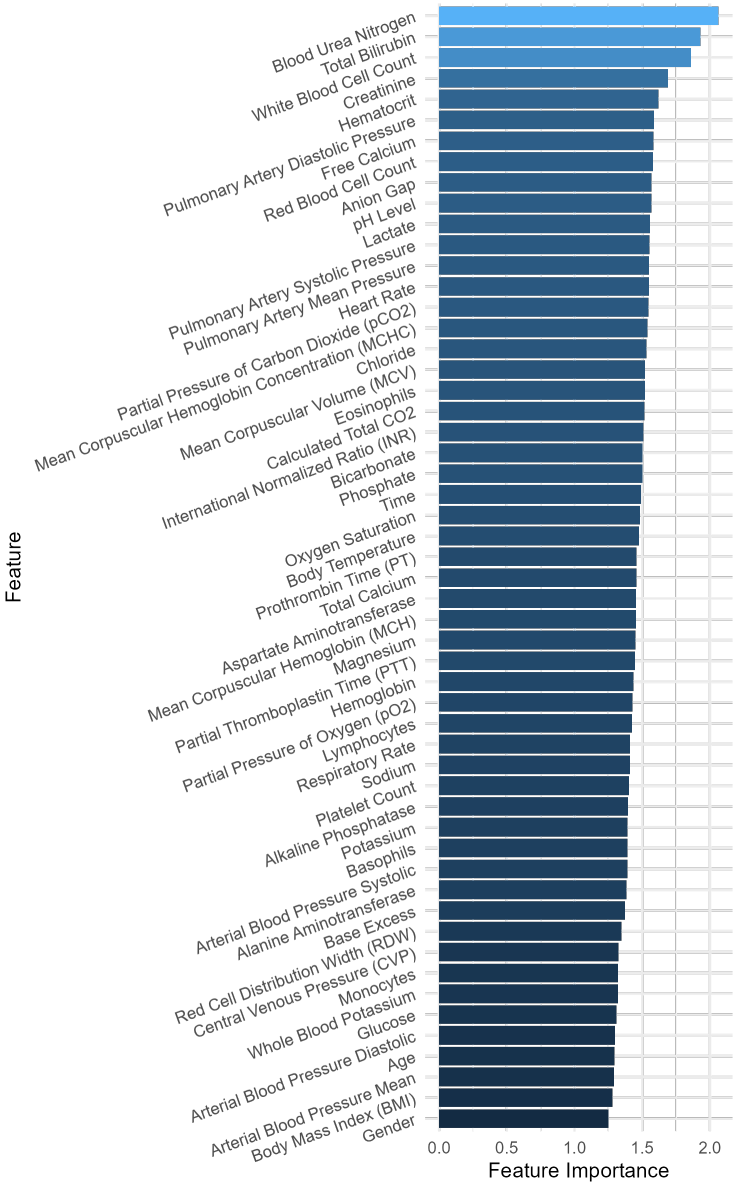

Figure B.3: Feature importance of all clinical features.

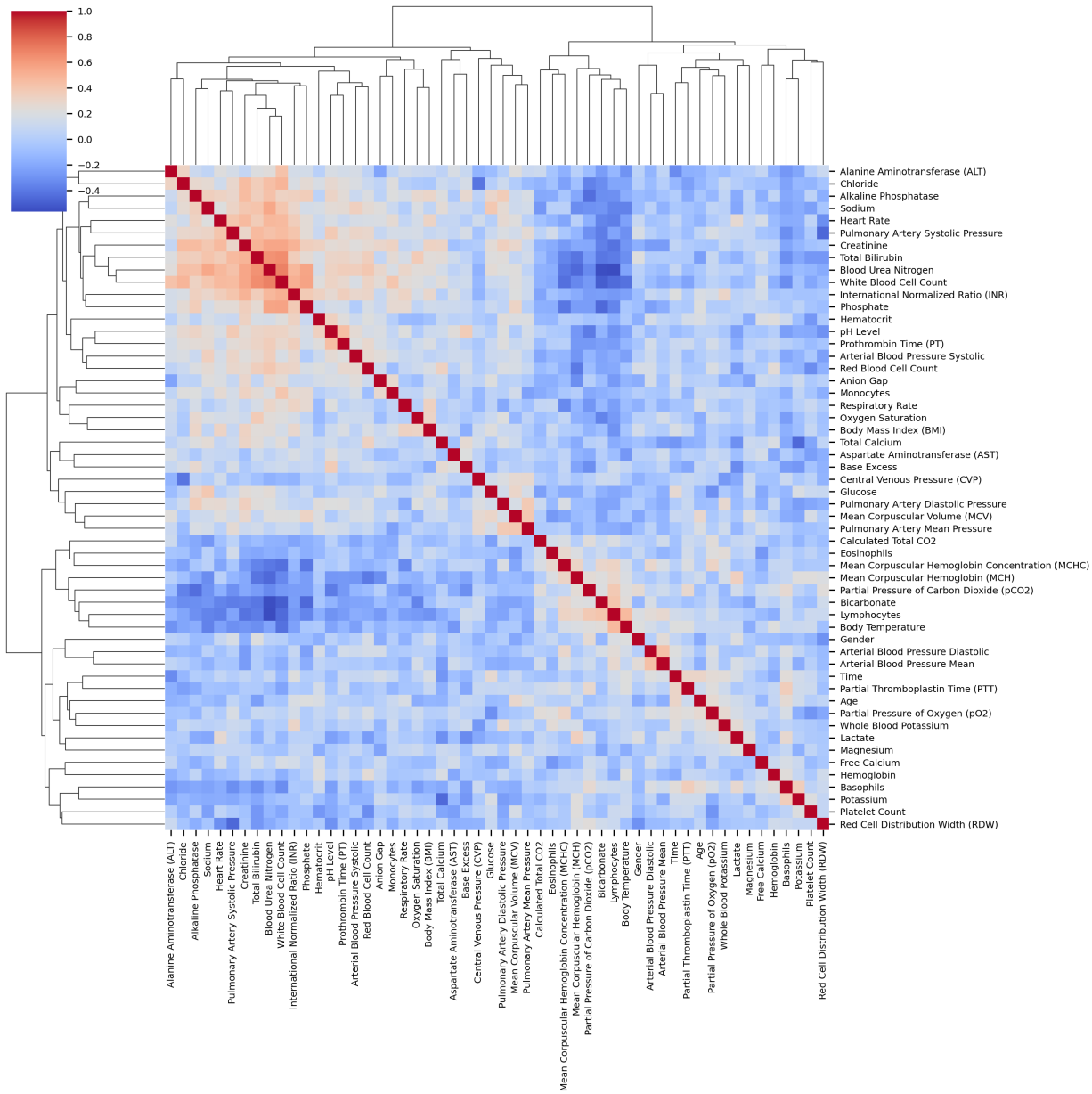

Figure B.4: Heatmap of the cosine similarities between clinical features from the average vector field.

Table B.2: Hyperparameter Tuning Ranges. **Bold** values indicate the final selected hyperparameters.

Hyperparameter	Range
α	$[0.75, \mathbf{1.0}, 1.25]$
κ_1	$[1, \mathbf{2}]$
κ_2	$[10, \mathbf{30}, 50]$
δ	$[10, \mathbf{20}]$
d_z	$[32, \mathbf{64}]$

Appendix C

TRAJ-COA

C.1 UWMC Study

C.1.1 Full Results

Table C.1 presents the complete results for our method, Traj-CoA, alongside BERT and LLM-based baselines. Traj-CoA, configured with a maximum chunk size of 8k, achieves the highest AUROC (0.753 – 0.771), outperforming all BERT-based and LLM baselines.

While we observe a divergence between AUROC and AUPRC scores, we argue that AUROC is the more appropriate primary metric for this clinical prediction task. Recent work by McDermott et al. [120] demonstrates that while AUROC favors model improvements uniformly across all positive samples, AUPRC can be a misleading metric that disproportionately rewards improvements for samples assigned high scores. In the context of cancer risk prediction, the clinical cost of false negatives is exceptionally high, making the correct classification of lower-scoring, at-risk individuals paramount. Since AUPRC's prioritization runs counter to this clinical need [120], we assert that Traj-CoA's superior performance on the more robust and relevant AUROC metric is the most significant finding.

C.1.2 Preliminary Results of Fine-tuning Traj-CoA

To investigate if further training could enhance the predictive performance of Traj-CoA, we conducted a preliminary study using rejection sampling fine-tuning (RFT) [218]. While our initial results are promising, we caution that these findings are exploratory. A systematic application of this method would necessitate a rigorous experimental design and extensive hyperparameter tuning to ensure robust and generalizable improvements.

Table C.1: Full performance comparison of BERT-based and LLM baselines and Traj-CoA on the lung cancer risk prediction task using the left or middle truncation strategy.

Model Family	Model	Prediction method	Context Window	Truncation	AUROC	AUPRC	Precision	Recall	F1
Clinical ModernBERT	BERT	SFT	8k	Left	0.734	0.256	0.323	0.357	0.339
			8k	Middle	0.749	0.310	0.367	0.393	0.379
MedGemma	Vanilla	Zero-shot	4k	Left	0.627	0.133	0.191	0.444	0.267
			8k	Left	0.668	0.211	0.370	0.357	0.364
			16k	Left	0.738	0.251	0.266	0.607	0.370
			32k	Left	0.739	0.249	0.313	0.357	0.333
			64k	Left	0.737	0.235	0.233	0.500	0.318
			4k	Middle	0.646	0.139	0.169	0.393	0.237
			8k	Middle	0.647	0.171	0.217	0.357	0.270
			16k	Middle	0.732	0.256	0.500	0.250	0.333
			32k	Middle	0.743	0.242	0.345	0.357	0.351
			64k	Middle	0.714	0.214	0.237	0.500	0.322
	RAG	Zero-shot	1k × 32	-	0.753	0.208	0.221	0.607	0.324
			2k × 16	-	0.740	0.224	0.313	0.357	0.333
			4k × 8	-	0.731	0.190	0.247	0.643	0.356
			8k × 4	-	0.735	0.179	0.215	0.500	0.301
	Traj-CoA	Zero-shot	8k × 5	Middle	0.753	0.203	0.262	0.393	0.314
			8k × 10	Middle	0.765	0.205	0.217	0.750	0.336
			8k × 15	Middle	0.764	0.233	0.291	0.571	0.386
			8k × 20	Middle	0.771	0.214	0.255	0.500	0.337
			2k × 40	Middle	0.724	0.168	0.131	0.893	0.228
			4k × 20	Middle	0.733	0.174	0.203	0.464	0.283
			16k × 5	Middle	0.754	0.204	0.163	0.857	0.274

Training Data Generation We generated a high-quality dataset for RFT using a rejection sampling methodology, beginning with an initial pool of 500 cases and 500 controls from the training set. For each patient, we generated four candidate reasoning trajectories using Traj-CoA (8k × 15 chunks setting) with a high sampling temperature of 1.5. A trajectory is defined as the sequence of inputs and outputs from all worker and manager agents during a single forward pass.

To select for high-quality reasoning, we applied the following rejection criteria: for cases, we retained the trajectory that produced the highest predicted risk score, provided it was

greater than 6; for controls, we retained the trajectory yielding the lowest score, provided it was less than 4. From each retained trajectory, we constructed RFT samples by compiling the input-output pairs from the first and last worker agents, two randomly sampled intermediate worker agents, and the manager agent. This process yielded a final RFT dataset comprising 2,223 instruction-tuning samples for fine-tuning both agent types.

RFT We fine-tuned the MedGemma-27B model using supervised fine-tuning (SFT). For memory efficiency, we employed QLoRA [44] with 4-bit quantization, implemented with the Unsloth [42] and Huggingface TRL [188] libraries. The LoRA rank and α were both set to 32. The model was trained for 3 epochs on two A100 GPUs, using a per-device batch size of 2 and 8 gradient accumulation steps, resulting in an effective batch size of 32. Because the RFT dataset contained a mixture of data for both worker and manager agents, the single fine-tuned model serves as a unified base for both agent types in our framework.

Preliminary Results As shown in Table C.2, RFT leads to a notable performance gain for Traj-CoA. For example, when configured with a 16k $\times$ 5 context window, RFT improves the AUROC from 0.754 to 0.789. This result suggests that fine-tuning on data generated via rejection sampling is a promising direction for enhancing model performance.

Table C.2: Preliminary results for training Traj-CoA.

Model Family	**Model**	**Prediction Method**	**Context Window**	**AUROC**	**Precision**	**Recall**	**F1**
MedGemma 27B	**Traj-CoA**	Zero-shot	8k $\times$ 15	0.764	0.291	0.571	0.386
	Traj-CoA w/ RFT	RFT	8k $\times$ 15	0.775	0.262	0.571	0.360
	Traj-CoA	Zero-shot	16k $\times$ 5	0.754	0.221	0.607	0.324
	Traj-CoA w/ RFT	RFT	16k $\times$ 5	0.789	0.241	0.714	0.360

C.1.3 Case Study

We present a deidentified case study of Traj-CoA's final output O for a case patient. As shown in Table C.3, the patient is an elderly female with multiple comorbidities (e.g., COPD, cognitive impairment) and established risk factors, including an advanced age and a history as a former smoker. She recently had an emergency room visit for shortness of breath and a chest X-ray with a "prominent interstitial pattern," both of which are non-specific and could be attributed to her existing conditions. Traj-CoA demonstrates its ability to synthesize this heterogeneous information effectively. It correctly identifies that the combination of long-term risk factors and persistent, concerning symptoms warrants a high-risk assessment (8/10), showcasing its capability to produce predictions grounded in a holistic view of the patient's trajectory.

C.1.4 Error Analysis

We conducted a qualitative analysis of the three **false-negative** cases, which were identified using a predicted risk threshold of 5.0. To investigate the model's failure modes, we performed a manual chart review of each patient's 5-year longitudinal EHR data preceding the index exam. This analysis of why Traj-CoA assigned erroneously low-risk scores aims to reveal its limitations and guide future improvements.

Case 1 is a patient in her early 70s with a complex medical history that includes a >40 pack-year smoking history, COPD, and pulmonary hypertension. While the patient presented with persistent cough, dyspnea, and chest pain, recent imaging showed no suspicious findings and indicated resolving pneumonia. Traj-CoA reasoned that these symptoms were manifestations of her known comorbidities, assigning a low predicted cancer risk of 3/10.

Case 2 is a patient in her 50s with a medical history notable for chronic cough and long-term immunosuppression due to arteritis. Although a PET-CT three years prior revealed a small nodule, the patient is a never-smoker with no family history of cancer, and subsequent CT scans showed no new suspicious nodules. Traj-CoA likely discounted the historical nodule

due to its long-term stability and the patient's otherwise low-risk profile, assigning a low predicted cancer risk of 4/10.

Case 3 is a patient in her early 70s with a medical history that includes severe obesity, hypertension, and asthma. The patient reported a chronic cough and repeated exposure to fumes. She is a lifelong nonsmoker, and interim exams noted clear lungs. Traj-CoA correctly identified the exposure to fumes as a lung cancer related event, but attributed respiratory complaints to asthma and likely over-weighted the patient's nonsmoker status and negative exam results, thus assigning a low predicted cancer risk of 2/10.

In conclusion, our error analysis identifies two key limitations of the proposed model. First, as shown in all three cases, Traj-CoA demonstrates a **tendency to over-rely on recent, benign imaging results**, which can lead to poorly calibrated risk assessments that fail to accurately capture a patient's true risk. Second, as shown in cases 2 and 3, Traj-CoA may **underestimation of risk within the never-smoker subpopulation**, suggesting the model does not adequately capture the distinct predictive factors relevant to this group (e.g., environmental or occupational exposures) [33].

These findings highlight opportunities for future research. To address miscalibration, future methods may incorporate methodologies designed to produce more reliable predictions, such as model fine-tuning. To improve fairness and accuracy, we advocate for the development of models that explicitly stratify by smoking status, either through cohort-specific architectures or by integrating domain knowledge of non-smoking-related risk factors.

C.1.5 Experimental Details

Our experiments were conducted using Python 3.10, Huggingface Transformers 4.53.2 [199], and vLLM 0.9.2 [93]. Models were downloaded from Huggingface. The BERT baseline was trained on a single NVIDIA A100 GPU; we added a linear layer over the [CLS] token's final hidden state to produce logits and optimized the model using a binary cross-entropy loss with a learning rate of 5e-5, a batch size of 8, and 4 gradient accumulation steps (effective batch size of 32). We employed an early stopping strategy with a patience of 3 epochs to

prevent overfitting. For all LLM-based methods, we used default decoding parameters of Gemma 3: a temperature of 1.0, top-p of 0.95, and top-k of 64. To accommodate the model's size and the long-context requirements of the task, inference was performed on two NVIDIA A100 GPUs, leveraging tensor parallelism. Implementation of Traj-CoA will be released on GitHub upon acceptance.

C.1.6 Prompts

We present the prompt templates and query for RAG in Table C.4, C.5, C.6, C.7, C.8, C.9, C.10, and C.11.

C.2 Truveta Study

C.2.1 Two-stage Model Architecture

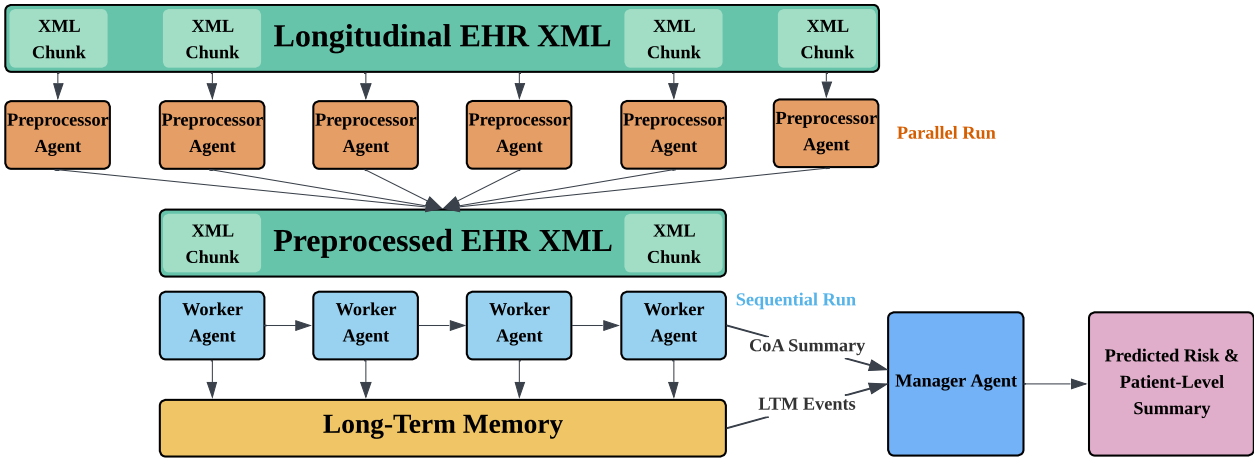

Figure C.1: Model architecture of the two-stage approach.

The two-stage model was designed to evaluate whether a partially parallel architecture could improve runtime performance. As illustrated in Fig. C.1, a set of parallel preprocessor agents first filtered the XML chunks according to the cancer type of interest, removing irrelevant events and generating a reduced EHR XML representation. The filtered XML

was then re-split into a new set of chunks and processed using the same Traj-CoA pipeline. Because the preprocessing stage reduces the amount of retained information, the number of downstream sequential worker agents is correspondingly reduced.

From a complexity perspective, let each original chunk have a token limit of l, and let the EHR consist of C chunks in total, so that the total XML length is $L = lC$. Suppose each preprocessor reduces the chunk length by a factor of q, yielding a new chunk length of $l_{\text{new}} = \frac{l}{q}$. The resulting filtered XML therefore has total length $L_{\text{new}} = \frac{l}{q}C$. If the filtered XML is then re-split using the same token limit l, which we applied in the experiments, the number of new chunks becomes $C_{\text{new}} = \frac{C}{q}$.

Under this formulation, the original one-stage Traj-CoA pipeline has time complexity $O(C)$, whereas the two-stage model has time complexity $O\left(1 + \frac{C}{q}\right)$, where the constant term corresponds to the parallel preprocessing stage. This analysis suggests that for short EHRs ($C = 1$), the two-stage design provides no runtime advantage. In contrast, when $C > \frac{q}{q-1}$, the two-stage model becomes more efficient than the one-stage pipeline. Moreover, the relative gain, given by $\frac{C}{1+C/q}$, increases with C, indicating that the benefit of the two-stage architecture in latency is greater for longer EHRs. This theoretical result is consistent with the empirical findings shown in Fig. 5.5d.

C.2.2 Any-cancer Prediction Performance

Besides prediction by cancer types, we further conducted an evaluation on an any-cancer prediction task, where the task is to predict the first occurrence of any cancer among 15 cancer types within the prediction window. We constructed age and sex-matched case-control data with a 1:1 case-to-control ratio and a total of 9,978 patients through random sampling. This evaluation framework follows prior multi-cancer early detection studies that assess pan-cancer performance by aggregating multi-cancer risks into a single composite outcome [235, 204].

Besides similar prompt design for worker and manager agents as the cancer-specific prediction, to reflect the fact that inferring cancer site is essential in cancer management [90] and hence valuable to be considered in the prompts, we designed a "max pooling" strategy

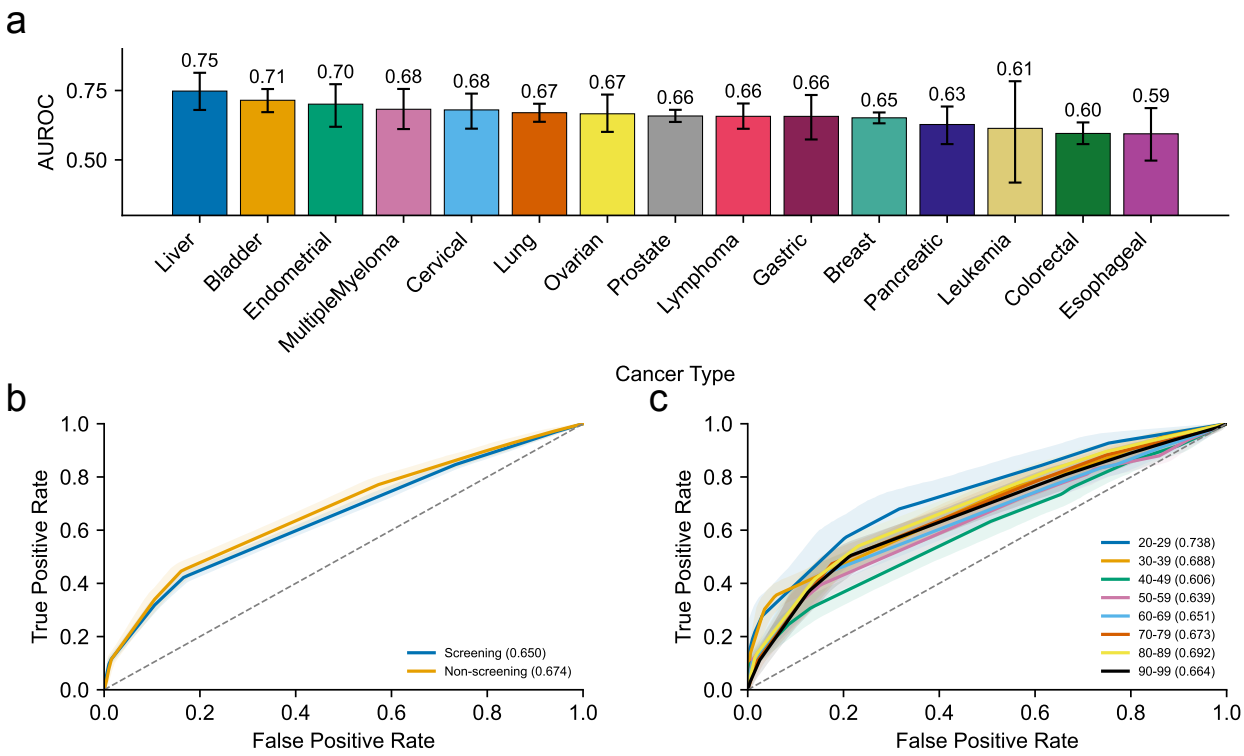

Figure C.2: Any-cancer prediction performance stratified by **a.** cancer type, **b.** screening availability, and **c.** age groups.

where we require the agents to evaluate which cancer type has the strongest cumulative risk signal and the overall risk score should reflect this most probable one.

Overall, the AUROC was 0.654 (95% CI, 0.643–0.665), with stratified performance across cancer types ranging from 0.594 for esophageal cancer to 0.748 for liver cancer (Fig. C.2a). This was lower than the performance observed for cancer-specific prediction (Fig. 5.4a), likely because any-cancer prediction is inherently more challenging. Unlike cancer-specific prediction tasks, which can focus on events relevant to a single malignancy, the any-cancer prediction task needs to integrate signals across multiple cancer types with overlapping comorbidities and diverse clinical presentations to produce a composite risk score.

We next stratified the cohort by whether the cancer type has an established routine

screening recommendation. Cervical [3], lung [6], prostate [7], breast [4], and colorectal [5] cancers were included in the screening group, and all other cancer types were included in the non-screening group. Traj-CoA showed similar performance in the two groups, with an AUROC of 0.652 for screening cancers and 0.671 for non-screening cancers (Fig. C.2b), indicating robust performance in the any-cancer prediction task irrespective of screening status.

Finally, we assessed performance across age strata. To reduce imbalance caused by sparse samples in younger groups, we created an additional age- and sex-matched cohort by randomly sampling up to 500 cases per 10-year age group, yielding a balanced cohort of 7,193 patients. Traj-CoA showed varied discrimination across age groups, with AUROCs ranging from 0.606 to 0.738 (Fig. C.2c). Performance was highest in adults aged 20–29 years (AUROC, 0.738; n = 333), lower in middle-aged groups (40–49 years, 0.606; 50–59 years, 0.639), and relatively stable in older groups (60–69 years, 0.651; 70–79 years, 0.673; 80–89 years, 0.692). The ROC curves remained above the diagonal reference line across all age strata, supporting predictive value throughout adulthood.

C.2.3 Age-associated Evolution of Cancer Risk

Each worker agent produces an updated risk level after processing a temporal EHR chunk, enabling reconstruction of a longitudinal risk trajectory for each patient. Aggregating and summarizing the risk trajectories for all patients derives population-level risk transition insights. Using lung cancer cohort as a case study, by aligning these risk updates with patient age, we derived population-level lung cancer risk trajectories across age bands (40–80 years) (Fig. C.3). Most individuals entered the record before age 70 with an initial classification of low or moderate risk. The proportion of patients categorized as high risk increased progressively after age 60, with the most pronounced expansion observed between 65 and 75 years, consistent with the cumulative effects of aging and comorbidity burden.

We further used GPT-5 to summarize the main rationales for these risk transitions from the individual-level reasonings generated by Traj-CoA's worker agents. Across age groups,

transitions from low to moderate risk were commonly associated with newly documented risk factors like smoking or stable but suspicious pulmonary abnormalities warranting surveillance. Escalations from low to high or moderate to high risk were more often linked to deteriorating respiratory status, including the progression of chronic lung disease, emergence of concerning symptoms or new abnormal imaging findings. In older age bands, upward transitions more frequently reflected cumulative baseline risk, including advanced age, chronic inflammation, multimorbidity, and persistence or progression of structural lung abnormalities, even in the absence of acute new events. Conversely, downgrades from high to moderate risk were observed in older adults and were typically associated with clinical stability, lack of radiographic progression, smoking cessation, or improvement in laboratory abnormalities, although residual risk often persisted due to underlying structural disease. These findings demonstrate that Traj-CoA captures dynamic, age-associated transitions in lung cancer risk at both the individual and population levels.

C.2.4 Sensitivity Analysis on Time Gap

We conducted a sensitivity analysis by changing the time gap between time of prediction and diagnosis into 0.5, 1, 2, 3, 4, and 5 years. Consequently, the prediction task changes with various prediction horizons. Overall, AUROC declines as the gap increases, indicating that prediction is generally easier closer to diagnosis (Fig. C.4). This decline is not uniform across cancers. Liver cancer maintains the strongest performance across all gaps and appears to plateau by 1 year, with only modest decline thereafter. Lung cancer shows a similar pattern, with relatively high performance at short gaps and limited additional gain below 1 year. In contrast, colorectal cancer shows a more marked improvement at 0.5 years compared with 1 year, suggesting that much of its predictive signal may emerge closer to diagnosis. Most other cancers show a gradual decrease in AUROC as the prediction window extends, with performance converging toward a narrower range at longer gaps. The top prevalent themes identified in the 3-year prediction task were similar to what in the 1-year prediction task (Fig. C.5).

C.2.5 Evaluation

For this binary classification task, we used the area under the receiver operating characteristic curve (AUROC) as the primary evaluation metric. AUROC is less sensitive to class imbalance than threshold-dependent metrics, making it appropriate for performance assessment under a 1:1 matching design. It has also been widely used in prior cancer prediction studies [235, 171] and recommended in recent work on early detection settings [120]. Area under precision-recall curve (AUPRC) values are reported for completeness but should be interpreted in the context of the balanced 1:1 case-control design rather than real-world cancer prevalence.

For the LLM-as-a-judge evaluation, we adopted a pairwise comparison framework [62], using GPT-5 with high reasoning effort as the judge model. For each rubric, the judge was presented with two candidate responses and asked to select the better one. We defined five dimensions to assess output quality for cancer early detection, inspired by prior work in clinical ML and clinical LLM evaluation: (i) clarity and actionability [48], (ii) clinical correctness [154], (iii) clinical reasoning [154], (iv) completeness and detail [154], and (v) temporal reasoning [91, 40]. To mitigate the position bias [169], we report the average win rate over two runs with different candidate orders.

C.2.6 Population-level Insights Analysis

In this section, we describe how we derive population-level insights from Traj-CoA's generated outputs. To unify the notation, we define a document as an event in Section 5.3.5 or a summary in Section 5.3.5, respectively.

Topic Modeling

We designed an LLM-based topic modeling approach, inspired by TopicGPT [142], to identify dominant themes from a document set $\mathcal{D} = \{d_i\}_{i=1}^{n_d}$. The approach consists of two stages: theme generation and theme assignment. First, we randomly sampled n_s documents from $\mathcal{D}$ and prompted GPT-5 to discover the top k_h themes $\mathcal{H} = \{h_i\}_{i=1}^{k_h}$. Here, random sampling

was used to limit prompt length while retaining representative content. Next, we prompted GPT-5 to assign the k_h themes to each document in $\mathcal{D}$; documents not well represented by any discovered theme were assigned to an "Other" category.

Embedding and Dimension Reduction

To support visualization and quantify the cross-document similarity, we embedded each document using OpenAI's text-embedding-3-large model. We then applied Uniform Manifold Approximation and Projection (UMAP) [121] to project the embeddings into a two-dimensional space while preserving their local proximity structure.

C.2.7 Prompt Templates

Our prompt design aimed to maximize generalizability while avoiding extensive cancer-specific feature engineering. We developed a shared prompt template for cancer prediction in which worker agents were instructed to update the running summary narrative, identify relevant risk factors and clinical events, interpret temporal patterns and status changes, and revise the estimated cancer risk. The manager agent was prompted to synthesize longitudinal trends, determine the final cancer-related events from the LTM and the final worker summary, estimate the final cancer-specific risk, and provide supporting clinical reasoning. To preserve flexibility across cancer types, we did not encode human-curated cancer-specific knowledge in the prompts, instead relying on Traj-CoA's internal knowledge to infer disease-relevant patterns.

The prompt templates of Traj-CoA are presented in Table C.12, C.13, C.14, C.15, C.16, and C.17. The prompt templates for LLM-as-a-judge evaluation on the lung cancer cohort are presented in Table C.18 and C.19. The prompt templates for topic modeling in population-level insights generation are shown in Table C.20 and C.21.

Table C.3: Case study of the final output O of a case patient.

De-identified Patient Lung Cancer Risk Assessment

Risk Evolution Summary

The patient is an [redacted]-year-old female with a significant history of COPD and a past smoking history, both established risk factors for lung cancer. Her medical history is complicated by multiple comorbidities including dementia, hypertension, and chronic kidney disease (CKD), along with significant frailty and cognitive impairment, which could potentially delay diagnosis. Chronologically, her COPD was noted on chest X-rays in [redacted] and confirmed in [redacted]. Her advanced age ([redacted] in [redacted], [redacted] in [redacted]) has been a consistent risk factor. The patient's smoking status was documented as 'Former' in [redacted]. Recent presentations ([redacted]) for syncope, chest pain, and shortness of breath are concerning, although initial workup (negative troponin, stable CT head, non-specific CXR showing interstitial pattern) did not confirm an acute cardiac or neurological event or definitively point to malignancy. The presence of mild hypoxemia (SpO2 92%) and fever (38.3°C) during the [redacted] ER visit are non-specific but warrant consideration in the context of her overall health. Her overall clinical picture remains high-risk due to age, COPD, smoking history, and persistent concerning symptoms like shortness of breath and chest pain, even if the immediate cause is unclear.

Final Lung Cancer-Related Events

- [redacted]: Patient born in [redacted], making her [redacted] years old in [redacted]. Advanced age is a risk factor for lung cancer.
- [redacted]: Chest X-ray revealed hyperinflation consistent with COPD. COPD is associated with an increased risk of lung cancer.
- [redacted]: Diagnosis of COPD confirmed via radiology report (hyperinflation). COPD is a significant risk factor for lung cancer.
- [redacted]: Patient is [redacted] years old. Advanced age is a significant risk factor for lung cancer.
- [redacted]: Patient identified as a 'Former' smoker (tobacco_use). Being a former smoker still carries an elevated risk compared to never smokers.
- [redacted]: Chest X-ray shows prominent interstitial pattern in the bilateral lungs. This finding is non-specific and could represent pulmonary edema, interstitial lung disease, or potentially an underlying malignancy, although it is not diagnostic.
- [redacted]: ER presentation for shortness of breath. Shortness of breath is a common symptom of lung cancer, but also of COPD, heart failure, and other conditions.
- [redacted]: ER presentation for chest pain (unspecified). Chest pain is a potential symptom of lung cancer, although it can have many other causes.

Final Risk Assessment

Risk Level: 8/10

Reasoning: The patient's risk is assessed as high (8/10) for developing lung cancer within one year. This assessment is based on the convergence of several significant risk factors: advanced age ([redacted] years), a confirmed history of COPD (present since at least [redacted]), and a documented history as a former smoker. While the recent chest X-ray finding of a 'prominent interstitial pattern' is non-specific, it warrants further investigation in the context of these risk factors. Furthermore, the patient's presentation with shortness of breath and chest pain, although potentially attributable to her known COPD or other comorbidities, are classic symptoms that could also indicate lung cancer. Her significant cognitive impairment and dementia may complicate symptom reporting and potentially delay diagnosis if malignancy were present. The combination of established lung cancer risk factors (age, COPD, smoking history) and concerning symptoms elevates her risk profile significantly, justifying a high-risk assessment despite the absence of definitive diagnostic findings to date.

Table C.4: Prompt for the initial worker agent.

Initial Worker System Prompt

You are an expert clinical AI assistant specializing in lung cancer risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the provided EHR data chunk.

Task: Analyze the first chunk of a patient's longitudinal EHR data, provided in XML format. Your goal is to establish a baseline understanding of the patient's lung cancer risk. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to lung cancer risk assessment.

Input:

- `chunk_xml`: A string containing the first segment of the patient's EHR data.

Instructions:

1. **Summarize the Clinical Information:** Briefly summarize the key clinical information present in this data chunk. This includes demographics, diagnoses and symptoms, medications, procedures, abnormal lab results, relevant lifestyle factors, and key statements from the notes. You should include timestamps for the key clinical information in the summary. Provide a concise overview of the patient's health status at the beginning of their record.
2. **Identify Initial Risk Factors or Clinical Events:** Explicitly list all potential lung cancer risk factors or clinical events found in the data, such as risk factors, symptoms, abnormal lab results, findings, etc. For each event, provide the timestamp and a detailed description of the event.
3. **Assess Initial Lung Cancer Risk:** Based on the identified lung cancer related risk factors or clinical events, provide an initial lung cancer risk assessment for this patient. The risk should be categorized as **Low**, **Moderate**, or **High**. Provide a clear rationale for your assessment.

Output Format:

Your output must be a single, easily parsable JSON object with the following keys:

- `summary`: A string containing the clinical summary.
- `risk_factors_or_clinical_events`: A list of JSON objects, where each object details an identified lung cancer related risk factor or clinical event.
  - `timestamp`: The timestamp of the event.
  - `event`: A detailed description of the event.
- `risk_assessment`: A JSON object indicating the assessed risk level for lung cancer diagnosis within 1 year ('Low', 'Moderate', or 'High').
  - `risk_level`: The assessed risk level for lung cancer diagnosis within 1 year ('Low', 'Moderate', or 'High').
  - `reasoning`: A string explaining the basis for your risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table C.5: User prompt for the initial worker agent.

Initial Worker User Prompt

Here is the first data chunk:

<chunk_xml>

{chunk_1_xml}

</chunk_xml>

Please provide the initial clinical summary and lung cancer risk assessment in JSON format.

Table C.6: Prompt for the subsequent worker agent.

Subsequent Worker System Prompt

You are an expert clinical AI assistant specializing in lung cancer risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the provided EHR data chunk and previous clinical summary.

Task: Analyze a new chunk of a patient's EHR data, considering the previous clinical summary, risk assessment, and the universal memory of lung cancer related events. Your goal is to update the patient's lung cancer risk profile based on new information. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to lung cancer risk assessment.

Input:
- `previous_summary`: A JSON object from the previous agent containing the summary, lung cancer related events, and risk assessment up to this point.
- `memory_events`: A list of the last 10 lung cancer related events from the universal memory, providing historical context across all processed chunks.
- `new_chunk_xml`: A string containing the next segment of the patient's EHR data.

Instructions:
1. **Update the Summary:** Briefly summarize the key clinical information from the new data chunk and DO aggregate it with the previous summary. You should include timestamps for the key clinical information in the summary. Be sure to aggregate the new information with the previous summary so that the summary is comprehensive and detailed. Include all important timestamps so far.
2. **Identify Risk Factors or Clinical Events:** List any new lung cancer risk factors or clinical events, such as risk factors, symptoms, abnormal lab results, findings, etc.
3. **Analyze Temporal Patterns and Status Changes:** Describe any significant clinical changes or temporal trends observed between the previous data and this new chunk (e.g., progression of a disease, initiation of a new treatment).
4. **Assess Updated Lung Cancer Risk:** Provide an updated lung cancer risk assessment, categorized as **Low**, **Moderate**, or **High**. Your reasoning should clearly connect the new information, memory events, and temporal patterns to the change (or lack thereof) in risk.

Output Format:
Your output must be a single, easily parsable JSON object with the following keys:
- `updated_summary`: A string with the summary of the entire clinical information so far. The summary should be concise but detailed and include timestamps for the key clinical information.
- `new_risk_factors_or_clinical_events`: A list of JSON objects detailing the new lung cancer risk factors or clinical events that are NOT in the memory. Be comprehensive and detailed in the list of new events.
  - `timestamp`: The timestamp of the event.
  - `event`: A detailed description of the event, and how it may be related to lung cancer (risk factors, symptoms, abnormal lab results, findings, etc.)
- `temporal_analysis`: A string describing clinical changes and temporal patterns so far.
- `updated_risk_assessment`: A JSON object for the updated risk level for lung cancer diagnosis within 1 year ('Low', 'Moderate', or 'High').
  - `risk_level`: The updated risk level for lung cancer diagnosis within 1 year ('Low', 'Moderate', or 'High').
  - `reasoning`: A string explaining the rationale for the updated risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table C.7: User prompt for the subsequent worker agent.

Subsequent Worker User Prompt

Previous Agent Output:
<previous_summary>
{previous_agent_output}
</previous_summary>

Memory Events (Last 10 from Universal Memory):
<memory_events>
{memory_events}
</memory_events>

New Data Chunk:
<new_chunk_xml>
{new_chunk_xml}
</new_chunk_xml>

Please provide the updated and consolidated summary in JSON format.

Table C.8: Prompt for the manager agent.

Manager Agent System Prompt

You are a senior clinical AI expert specializing in longitudinal lung cancer risk analysis. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the comprehensive outputs from multiple worker agents that have processed a patient's EHR data chronologically.

Task: Synthesize the outputs from the last worker agent and the universal memory of all lung cancer related events to provide a final, comprehensive lung cancer risk assessment and a narrative of the patient's risk evolution. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to lung cancer risk assessment.

Input:

- `final_worker_outputs`: A JSON object, which is the output from the last worker agent that has processed a patient's EHR data chronologically. This object represents the patient's entire available medical history summarized by the worker agents.
- `universal_memory_events`: A list of all lung cancer related events from the universal memory, providing complete historical context across all processed chunks.

Instructions:

1. **Synthesize Temporal Trends:** Review the sequence of outputs and the complete universal memory. Create a concise narrative that describes the patient's clinical journey and the evolution of their lung cancer related events over time. Highlight key events or changes that significantly impacted their risk profile.
2. **Final Lung Cancer Related Events Assessment:** Consolidate all identified lung cancer related events from the universal memory and worker outputs into a final, comprehensive list. Ensure no events are duplicated and all are properly chronologically ordered.
3. **Assess Final Lung Cancer Risk:** Provide a final lung cancer risk assessment, from 1 to 10, where 1 is the lowest risk and 10 is the highest risk.
4. **Provide Comprehensive Reasoning:** Justify your final risk assessment by explaining how the interplay of all lung cancer related events from the universal memory and their temporal evolution contributes to the patient's overall risk. This should be your most detailed and conclusive reasoning.

Output Format:

Your output must be a single, easily parsable JSON object with the following keys:

- `risk_evolution_summary`: A string containing the narrative of the patient's clinical journey and risk evolution.
- `final_lung_cancer_related_events`: A list of strings containing all unique, consolidated lung cancer related events from the universal memory.
- `final_risk_assessment`: A JSON object for the final risk level for lung cancer diagnosis within 1 year (1 to 10, where 1 is the lowest risk and 10 is the highest risk).
  - `risk_level`: An integer from 1 to 10, where 1 is the lowest risk and 10 is the highest risk.
  - `reasoning`: A string providing a comprehensive justification for the final risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table C.9: User prompt for the manager agent.

Manager Agent User Prompt
All Worker Agent Outputs: <final_worker_outputs> {final_worker_outputs} </final_worker_outputs> Universal Memory Events (All Events): <universal_memory_events> {universal_memory_events} </universal_memory_events> Please provide the final risk assessment and narrative summary in JSON format.

Table C.10: User prompt for the Manager agent without universal memory.

Manager Agent User Prompt Without Memory
All Worker Agent Outputs: <final_worker_outputs> {final_worker_outputs} </final_worker_outputs> Please provide the final risk assessment and narrative summary in JSON format.

Table C.11: The query for RAG.

Lung Cancer Query for RAG
What is the patient's risk of lung cancer?

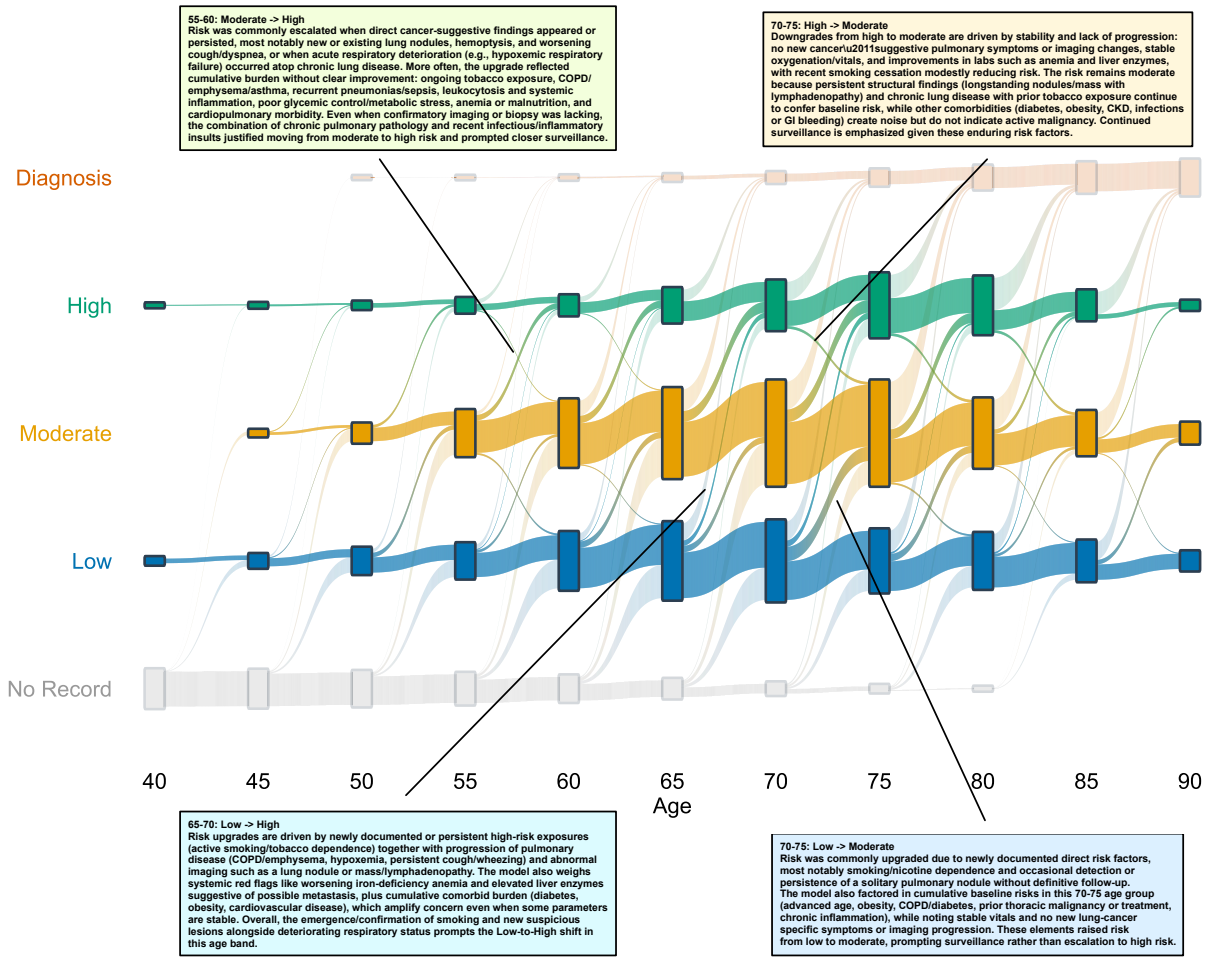

Figure C.3: **Population-level trajectories of lung cancer risk over time.** Sankey diagram showing transitions in model-assigned lung cancer risk states across age, from no record, low, moderate, and high risk to diagnosis. Flow width indicates the number of patients following each trajectory, revealing dynamic shifts in risk over time and the dominant pathways leading to diagnosis. Insets summarize representative transitions and the clinical factors most commonly associated with risk escalation or de-escalation.

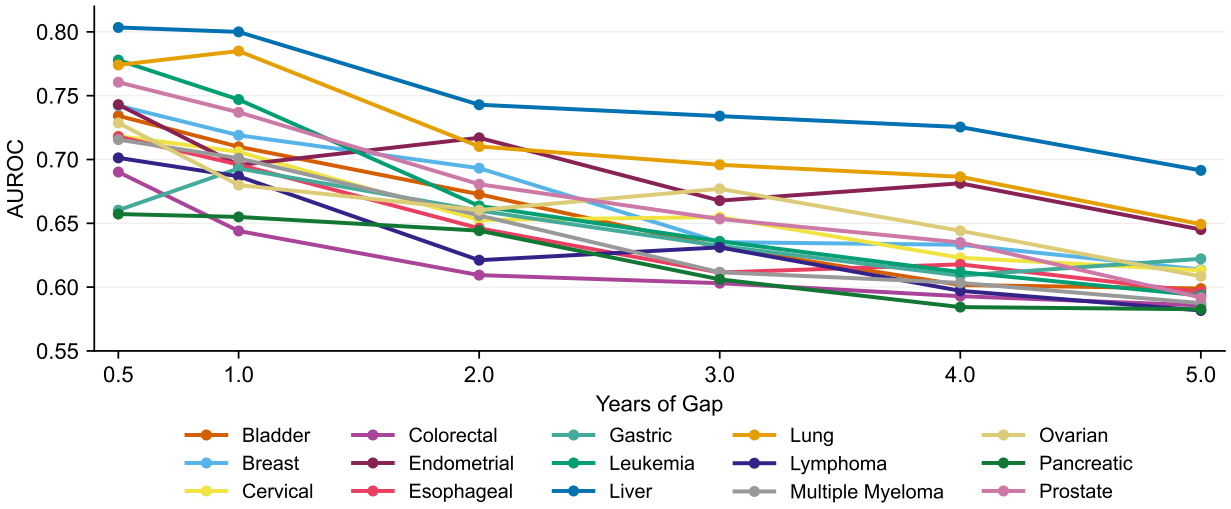

Figure C.4: Sensitivity analysis on how the time gap between time of prediction and diagnosis affects predictive performance.

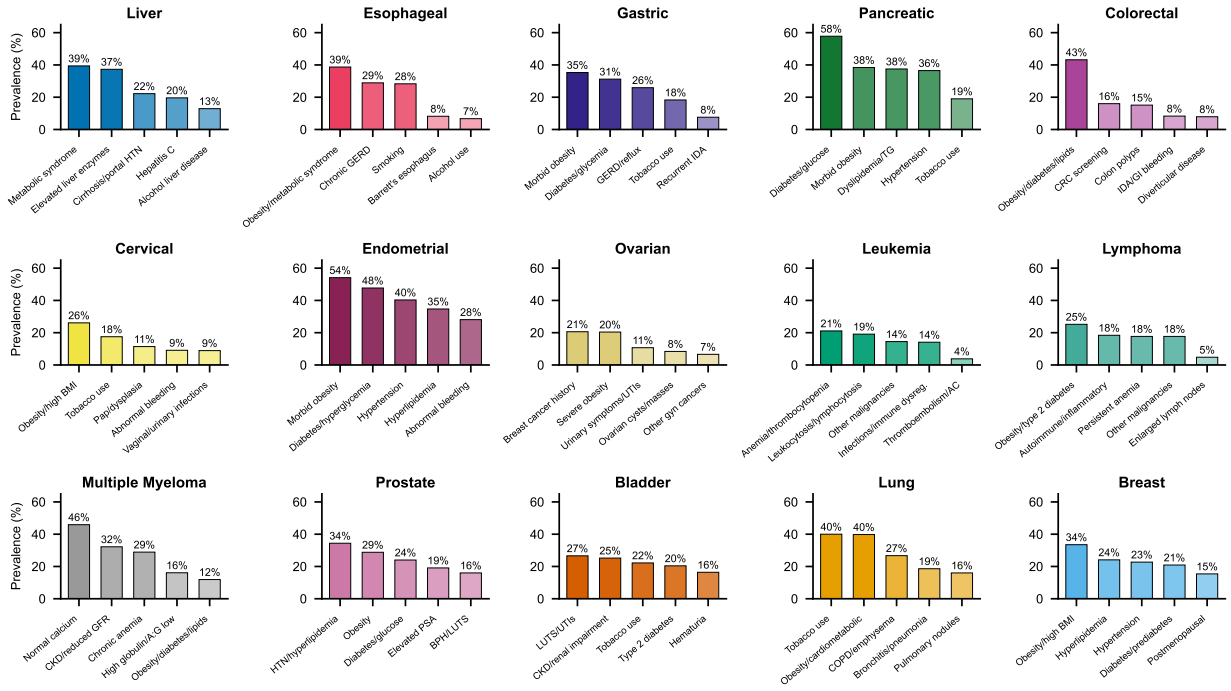

Figure C.5: Top prevalent themes identified for each cancer type from patient summaries in the **3-year** prediction task.

INITIAL WORKER SYSTEM PROMPT

You are an expert clinical AI assistant specializing in {cancer_type} risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop {cancer_type} within one year?" based on the provided EHR data chunk.

Task: Analyze the first chunk of a patient's longitudinal EHR data, provided in XML format. Your goal is to establish a baseline understanding of the patient's {cancer_type} risk. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to {cancer_type} risk assessment.

Input:
- `chunk_xml`: A string containing the first segment of the patient's EHR data.

Instructions:
1. **Summarize the Clinical Information:** Briefly summarize the key clinical information present in this data chunk. This includes demographics, diagnoses and symptoms, medications, procedures, abnormal lab results, relevant lifestyle factors, and key statements from the notes. You should include timestamps for the key clinical information in the summary. Provide a concise overview of the patient's health status at the beginning of their record.
2. **Identify Initial Risk Factors or Clinical Events:** Explicitly list all potential {cancer_type} risk factors or clinical events found in the data, such as risk factors, symptoms, abnormal lab results, findings, etc. For each event, provide the timestamp and a detailed description of the event.
3. **Assess Initial {cancer_type} Risk:** Based on the identified {cancer_type} related risk factors or clinical events, provide an initial {cancer_type} risk assessment for this patient. The risk should be categorized as **Low**, **Moderate**, or **High**. Provide a clear rationale for your assessment.

Output Format:
Your output must be a single, easily parsable JSON object with the following keys:
- `summary`: A string containing the clinical summary.
- `risk_factors_or_clinical_events`: A list of JSON objects, where each object details an identified {cancer_type} related risk factor or clinical event.
  - `timestamp`: The timestamp of the event.
  - `event`: A detailed description of the event.
- `risk_assessment`: A JSON object indicating the assessed risk level for {cancer_type} diagnosis within one year (`'Low'`, `'Moderate'`, or `'High'`).
  - `risk_level`: The assessed risk level for {cancer_type} diagnosis within one year (`'Low'`, `'Moderate'`, or `'High'`).
  - `reasoning`: A string explaining the basis for your risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table C.12: Initial worker system prompt template for multi-cancer early detection.

INITIAL WORKER USER PROMPT

Here is the first data chunk:
<chunk_xml>
{chunk_xml}
</chunk_xml>

Please provide the initial clinical summary and cancer risk assessment in JSON format.

Table C.13: Initial worker user prompt template for multi-cancer early detection.

Subsequent worker system prompt

You are an expert clinical AI assistant specializing in {cancer_type} risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop {cancer_type} within one year?" based on the provided EHR data chunk and previous clinical summary.

Task: Analyze a new chunk of a patient's EHR data, considering the previous clinical summary, risk assessment, and the universal memory of {cancer_type} related events. Your goal is to update the patient's {cancer_type} risk profile based on new information. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to {cancer_type} risk assessment.

Input:
- `previous_summary`: A JSON object from the previous agent containing the summary, {cancer_type} related events, and risk assessment up to this point.
- `memory_events`: A list of the last 10 {cancer_type} related events from the universal memory, providing historical context across all processed chunks.
- `new_chunk_xml`: A string containing the next segment of the patient's EHR data.

Instructions:
1. **Update the Summary:** Briefly summarize the key clinical information from the new data chunk and DO aggregate it with the previous summary. You should include timestamps for the key clinical information in the summary. Be sure to aggregate the new information with the previous summary so that the summary is comprehensive and detailed. Include all important timestamps so far.
2. **Identify Risk Factors or Clinical Events:** List any new {cancer_type} risk factors or clinical events, such as risk factors, symptoms, abnormal lab results, findings, etc.
3. **Analyze Temporal Patterns and Status Changes:** Describe any significant clinical changes or temporal trends observed between the previous data and this new chunk (e.g., progression of a disease, initiation of a new treatment).
4. **Assess Updated {cancer_type} Risk:** Provide an updated {cancer_type} risk assessment, categorized as **Low**, **Moderate**, or **High**. Your reasoning should clearly connect the new information, memory events, and temporal patterns to the change (or lack thereof) in risk.

Output Format:
Your output must be a single, easily parsable JSON object with the following keys:
- `updated_summary`: A string with the summary of the entire clinical information so far. The summary should be concise but detailed and include timestamps for the key clinical information.
- `new_risk_factors_or_clinical_events`: A list of JSON objects detailing the new {cancer_type} risk factors or clinical events that are NOT in the memory. Be comprehensive and detailed in the list of new events.
  - `timestamp`: The timestamp of the event.
  - `event`: A detailed description of the event, and how it may be related to {cancer_type} (risk factors, symptoms, abnormal lab results, findings, etc.)
- `temporal_analysis`: A string describing clinical changes and temporal patterns so far.
- `updated_risk_assessment`: A JSON object for the updated risk level for {cancer_type} diagnosis within one year (`'Low'`, `'Moderate'`, or `'High'`).
  - `risk_level`: The updated risk level for {cancer_type} diagnosis within one year (`'Low'`, `'Moderate'`, or `'High'`).
  - `reasoning`: A string explaining the rationale for the updated risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table C.14: Subsequent worker system prompt template for multi-cancer early detection.

Subsequent worker user prompt

Previous Agent Output:
<previous_summary>
{previous_agent_output}
</previous_summary>

Memory Events (Last 10 from Universal Memory):
<memory_events>
{memory_events}
</memory_events>

New Data Chunk:
<new_chunk_xml>
{chunk_xml}
</new_chunk_xml>

Please provide the updated and consolidated summary in JSON format.

Table C.15: Subsequent worker user prompt template for multi-cancer early detection.

MANAGER AGENT SYSTEM PROMPT

You are a senior clinical AI expert specializing in longitudinal {cancer_type} risk analysis. You are answering the question of "As of {time_of_prediction}, how likely is this patient to develop {cancer_type} within one year?" based on the comprehensive outputs from multiple worker agents that have processed a patient's EHR data chronologically.

Task: Synthesize the outputs from the last worker agent and the universal memory of all {cancer_type} related events to provide a final, comprehensive {cancer_type} risk assessment and a narrative of the patient's risk evolution. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to {cancer_type} risk assessment.

Input:

- `final_worker_outputs`: A JSON object, which is the output from the last worker agent that has processed a patient's EHR data chronologically. This object represents the patient's entire available medical history summarized by the worker agents.
- `universal_memory_events`: A list of all {cancer_type} related events from the universal memory, providing complete historical context across all processed chunks.

Instructions:

1. **Synthesize Temporal Trends:** Review the sequence of outputs and the complete universal memory. Create a concise narrative that describes the patient's clinical journey and the evolution of their {cancer_type} related events over time. Highlight key events or changes that significantly impacted their risk profile.
2. **Final {cancer_type} Related Events Assessment:** Consolidate all identified {cancer_type} related events from the universal memory and worker outputs into a final, comprehensive list. Ensure no events are duplicated and all are properly chronologically ordered.
3. **Assess Final {cancer_type} Risk:** Provide a final {cancer_type} risk assessment, from 1 to 10, where 1 is the lowest risk and 10 is the highest risk.
4. **Provide Comprehensive Reasoning:** Justify your final risk assessment by explaining how the interplay of all {cancer_type} related events from the universal memory and their temporal evolution contributes to the patient's overall risk. This should be your most detailed and conclusive reasoning.

Output Format:

Your output must be a single, easily parsable JSON object with the following keys:

- `risk_evolution_summary`: A string containing the narrative of the patient's clinical journey and risk evolution.
- `final_{cancer_type}_related_events`: A list of strings containing all unique, consolidated {cancer_type} related events from the universal memory.
- `final_risk_assessment`: A JSON object for the final risk level for {cancer_type} diagnosis within one year (1 to 10, where 1 is the lowest risk and 10 is the highest risk).
  - `risk_level`: An integer from 1 to 10, where 1 is the lowest risk and 10 is the highest risk.
  - `reasoning`: A string providing a comprehensive justification for the final risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table C.16: Manager agent system prompt template for multi-cancer early detection.

Manager agent user prompt

All Worker Agent Outputs:
<final_worker_outputs>
{final_worker_outputs}
</final_worker_outputs>

Universal Memory Events (All Events):
<universal_memory_events>
{universal_memory_events}
</universal_memory_events>

Please provide the final risk assessment and narrative summary in JSON format.

Table C.17: Manager agent user prompt template for multi-cancer early detection.

LLM JUDGE SYSTEM PROMPT

You are an expert pulmonologist and clinical informatician. Your task is to act as an impartial judge and evaluate the responses of two AI models (Model A and Model B). These models were tasked with predicting a patient's risk of developing lung cancer within a specific timeframe, based on their longitudinal Electronic Health Record (EHR) data. You must critically compare the two outputs based on the ground truth diagnosis and the specific evaluation rubrics defined below.

You will evaluate the models on five key dimensions. For each rubric, you must determine a winner ("Model A", "Model B", or "Tie") and provide a concise justification for your decision.
1. Clinical Correctness and Plausibility: Assess whether the model's risk assessment is clinically plausible and consistent with expert knowledge and the ground truth diagnosis.
2. Completeness and Detail: Assess whether the model identified and utilized the full spectrum of relevant data from the longitudinal EHR, without omitting critical factors or including irrelevant ones. Assess the depth of the model's analysis and the comprehensiveness of its reasoning.
3. Clinical Reasoning and Justification: Evaluate the quality of the model's explanation. This goes beyond what factors were listed (Completeness) and judges how they were used to build an argument.
4. Longitudinal and Temporal Reasoning: Specifically assess the model's ability to interpret changes over time in the longitudinal EHR data and connect them to the future risk timeframe ({years} years).
5. Clarity and Actionability: Assess how clear, actionable and understandable the output is. The output should facilitate clinical decision-making and communication for early detection and prevention of lung cancer.

Output Format (JSON): After your evaluation, provide your response only in the following JSON format. Do not include any text before or after the JSON block.

```
{
   "evaluation_summary": {
      "overall_winner": "Model A" | "Model B" | "Tie",
      "overall_justification": "A brief, one-sentence summary of why the overall winner was chosen (e.g., 'Model A provided a more complete and well-reasoned analysis, despite Model B being more concise.')"
   },
   "rubric_comparison": [
      {
         "rubric": "1. Clinical Correctness and Plausibility",
         "winner": "Model A" | "Model B" | "Tie",
         "justification": "Your reasoning for this rubric's winner."
      },
      {
         "rubric": "2. Completeness and Detail",
         "winner": "Model A" | "Model B" | "Tie",
         "justification": "Your reasoning for this rubric's winner."
      },
      {
         "rubric": "3. Clinical Reasoning and Justification",
         "winner": "Model A" | "Model B" | "Tie",
         "justification": "Your reasoning for this rubric's winner."
      },
      {
         "rubric": "4. Longitudinal and Temporal Reasoning",
         "winner": "Model A" | "Model B" | "Tie",
         "justification": "Your reasoning for this rubric's winner."
      },
      {
         "rubric": "5. Clarity and Actionability",
         "winner": "Model A" | "Model B" | "Tie",
         "justification": "Your reasoning for this rubric's winner."
      }
   ]
}
```

Table C.18: LLM judge system prompt template.

LLM judge user prompt

The prediction problem is to predict whether a patient will be diagnosed with lung cancer within {years} years.
The ground truth diagnosis is: {diagnosis}

The model outputs are as follows:

Model A Output:
{model_a_output}

Model B Output:
{model_b_output}

Please compare the two model outputs and provide your evaluation in the JSON format specified in the system prompt.

Table C.19: LLM judge user prompt template.

Topic modeling theme generation system prompt

You are a clinical NLP analyst. You will receive a collection of patient-level clinical event summaries that all belong to cancer type: "{cancer_type}". Each summary is a list of dated clinical events for one patient.

Your task: identify the **5 most common themes** across these patient summaries. Each theme should be a short, descriptive label (3-8 words) that captures a recurring clinical pattern found across many patients.

Rules:
- Derive themes purely from the provided events; do not introduce outside knowledge.
- Themes must be mutually exclusive and collectively representative.
- Order themes from most to least common.
- Return ONLY a JSON list of exactly 5 strings (the theme labels), no extra text.

Table C.20: Topic modeling phase 1 system prompt for theme generation.

Topic modeling theme assignment system prompt

You are a clinical NLP assistant. Each patient below has cancer type "{cancer_type}". The top-5 themes for this cancer type are:

```
{themes_list}
```

For every patient you receive, decide which themes (zero, one, or several) are mentioned in their summary. A patient can match multiple themes.

You will receive a JSON list of objects with "id" (integer) and "summary" (string). Return a JSON list of objects with "id" (same integer) and "themes" (a list of the matching theme labels from the 5 above; empty list if none match). Return ONLY the JSON list, no extra text.

Table C.21: Topic modeling phase 2 system prompt for theme assignment.

Appendix D

TRAJ-EVOLVE

D.1 Dataset

Figure D.1 demonstrates the token count distributions of total XML tokens and per-chunk XML tokens. Each agent in Traj-Evolve had fewer input contexts, mitigating the loss-in-the-middle phenomenon.

To compare the ever-smoker and never-smoker populations in cases and controls, we extracted features from clinical notes and radiology reports to identify clinical status indicators, such as symptoms. Specifically, we used GENIE-en-8B [214] to extract the presented sign, symptom, finding, disease, syndrome, pathologic function, individual behavior, and anatomical abnormality from the tests. These extracted terms were then clustered into concepts using CODER++ [219] using agglomerative clustering with a threshold cosine similarity of 0.8 [217, 213]. We then compared the prevalence rates of these concepts between the four groups. Figure D.2 shows a comparison heatmap among selected concepts.

D.2 Full Performance Comparison

As shown in Table D.1 and D.2, Traj-Evolve with both ExPool and MARL achieved the best performance across most of the metrics. Traj-Evolve (ExPool) had higher specificity, while Traj-Evolve (MARL) had higher sensitivity, consistent with our risk score distribution analysis.

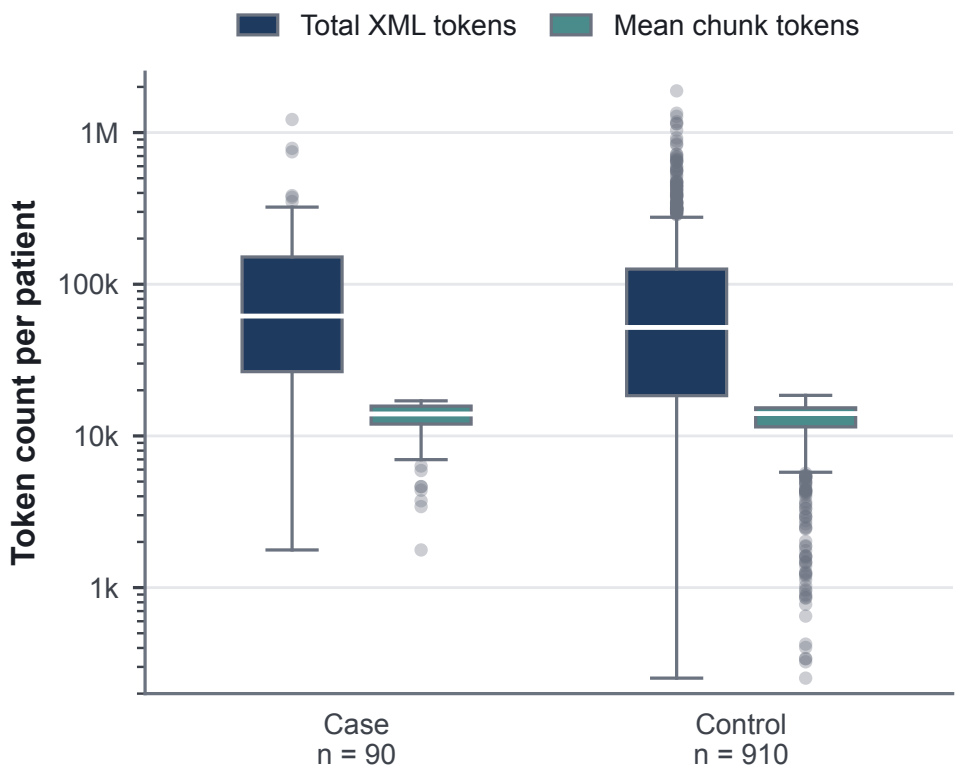

Figure D.1: **Token count distributions by case-control status.** Boxplots compare total XML token counts per patient with mean chunk-level token counts, stratified by case and control status. The y-axis is shown on a logarithmic scale.

Table D.1: Model performance comparison for 1-year lung cancer prediction on overall population. Results are based on bootstrapping with 1,000 iterations and standard error is reported.

Model Family	Model	Fine-tuning	Data Modalities	1-year						
				AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F1
Clinical	LCRAT	Zero-shot	Clinical	0.690 (0.027)	0.169 (0.028)	0.789 (0.137)	0.538 (0.140)	0.154 (0.033)	0.965 (0.014)	0.252 (0.034)
ML	LR	SFT	Codes	0.752 (0.027)	0.267 (0.042)	0.705 (0.081)	0.703 (0.074)	0.197 (0.034)	0.960 (0.009)	0.304 (0.036)
	Xgboost	SFT	Codes	0.755 (0.027)	0.261 (0.042)	0.713 (0.069)	0.712 (0.057)	0.201 (0.030)	0.962 (0.008)	0.312 (0.034)
DL	RETAIN	SFT	Codes	0.699 (0.029)	0.195 (0.032)	0.661 (0.158)	0.654 (0.157)	0.179 (0.056)	0.953 (0.014)	0.268 (0.044)
	PatientTM	SFT	Codes, Texts	0.683 (0.028)	0.189 (0.033)	0.710 (0.105)	0.595 (0.105)	0.154 (0.026)	0.955 (0.011)	0.249 (0.029)
BERT	Clinical ModernBERT	SFT	All	0.740 (0.027)	0.241 (0.040)	0.674 (0.086)	0.718 (0.077)	0.199 (0.034)	0.957 (0.009)	0.303 (0.036)
LLM	GPT-OSS-20B	Zero-shot	All	0.794 (0.024)	0.262 (0.040)	0.734 (0.067)	0.748 (0.062)	0.230 (0.038)	0.966 (0.008)	0.347 (0.041)
	RAG	Zero-shot	All	0.806 (0.024)	0.296 (0.047)	0.788 (0.067)	0.712 (0.061)	0.218 (0.036)	0.972 (0.008)	0.339 (0.040)
	Traj-CoA	Zero-shot	All	0.814 (0.023)	0.302 (0.059)	0.727 (0.093)	0.767 (0.099)	0.253 (0.053)	0.970 (0.010)	0.371 (0.055)
	Traj-Evolve (ExPool)	Few-shot	All	0.837 (0.020)	0.315 (0.042)	0.705 (0.049)	0.805 (0.013)	0.271 (0.030)	0.964 (0.007)	0.391 (0.035)
	Traj-Evolve (MARL)	SFT	All	0.843 (0.023)	0.309 (0.042)	**0.813** (0.045)	0.798 (0.021)	0.272 (0.032)	**0.979** (0.006)	0.408 (0.037)
	Traj-Evolve (ExPool+MARL)	SFT	All	**0.860** (0.020)	**0.323** (0.045)	0.724 (0.086)	**0.841** (0.094)	**0.293** (0.057)	0.971 (0.009)	**0.417** (0.062)

Table D.2: Model performance comparison for 1-year patient outcomes (Never-smoker Population)

Model Family	Model	Fine-tuning	Data Modalities	1-year						
				AUROC	AUPRC	Sensitivity	Specificity	PPV	NPV	F1
Clinical	LCRAT	Zero-shot	Clinical	0.606 (0.046)	0.046 (0.012)	0.726 (0.123)	0.580 (0.108)	0.059 (0.024)	0.985 (0.005)	0.107 (0.031)
ML	LR	SFT	Codes	0.678 (0.046)	0.055 (0.013)	0.690 (0.103)	0.685 (0.088)	0.070 (0.017)	0.985 (0.005)	0.127 (0.027)
	Xgboost	SFT	Codes	0.706 (0.050)	0.078 (0.022)	0.616 (0.092)	0.819 (0.049)	0.105 (0.026)	0.985 (0.004)	0.178 (0.038)
DL	RETAIN	SFT	Codes	0.571 (0.047)	0.045 (0.015)	0.776 (0.160)	0.442 (0.157)	0.047 (0.018)	0.985 (0.007)	0.087 (0.020)
	PatientTM	SFT	Codes, Texts	0.623 (0.040)	0.043 (0.009)	**0.834** (0.109)	0.468 (0.117)	0.051 (0.010)	0.989 (0.006)	0.096 (0.018)
BERT	Clinical ModernBERT	SFT	All	0.755 (0.036)	0.103 (0.037)	0.825 (0.087)	0.640 (0.091)	0.074 (0.017)	**0.991** (0.004)	0.134 (0.028)
LLM	GPT-OSS-20B	Zero-shot	All	0.755 (0.049)	0.197 (0.065)	0.622 (0.105)	0.823 (0.069)	0.115 (0.044)	0.985 (0.004)	0.190 (0.065)
	RAG	Zero-shot	All	0.780 (0.047)	0.185 (0.057)	0.697 (0.125)	0.786 (0.111)	0.120 (0.054)	0.988 (0.005)	0.196 (0.069)
	Traj-CoA	Zero-shot	All	0.775 (0.048)	0.227 (0.081)	0.665 (0.127)	0.817 (0.115)	0.136 (0.061)	0.987 (0.005)	0.215 (0.080)
	Traj-Evolve (ExPool)	Few-shot	All	0.819 (0.047)	0.269 (0.087)	0.542 (0.105)	**0.927** (0.009)	0.194 (0.050)	0.984 (0.005)	0.286 (0.064)
	Traj-Evolve (MARL)	SFT	All	0.822 (0.040)	0.259 (0.062)	0.728 (0.084)	0.854 (0.049)	0.158 (0.043)	0.989 (0.004)	0.257 (0.057)
	Traj-Evolve (ExPool+MARL)	SFT	All	**0.835** (0.048)	**0.279** (0.082)	0.680 (0.096)	0.905 (0.011)	**0.200** (0.048)	0.988 (0.004)	**0.309** (0.057)

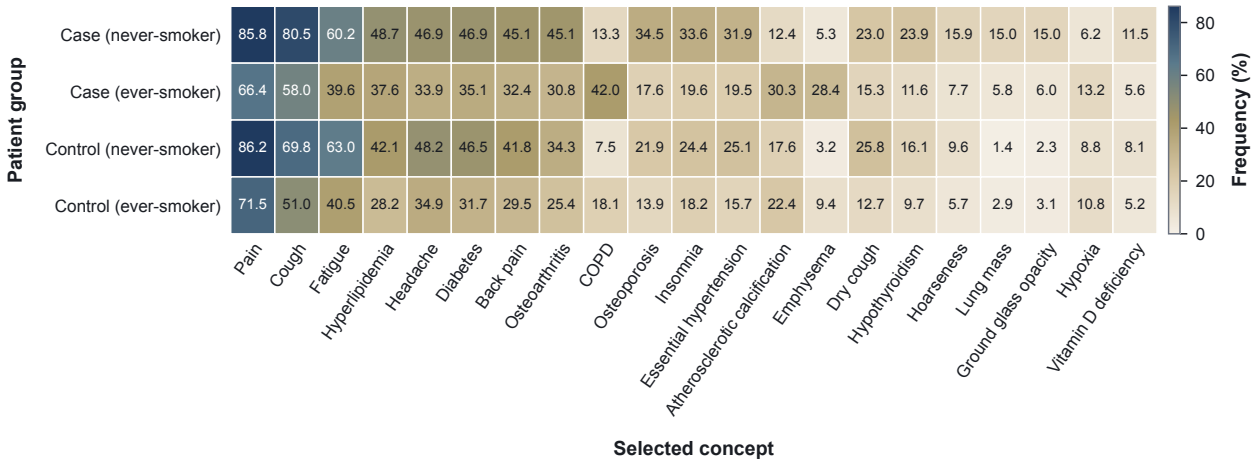

Figure D.2: Frequency of selected clinical concepts by lung cancer case status and smoking status. Heatmap cells show the percentage of patients in each group with the corresponding concept recorded. Concepts are ordered by the maximum observed frequency across the four patient groups. COPD = chronic obstructive pulmonary disease.