PaST-Q: Learning Future-Queryable Patient States for Tool-Augmented Clinical Agents

Anonymous authors
Paper under double-blind review

Abstract

While clinical guidelines provide population-level recommendations, personalized cancer treatment requires considering both longitudinal patient history and future clinical evolution under candidate therapies. Existing approaches remain limited in either the use of longitudinal disease and treatment context or the richness of future predictions. To address these challenges, we propose **PaST-Q**, a patient modeling and multi-agent framework that connects historical context with future simulation for guideline-grounded treatment planning. PaST-Q organizes heterogeneous records into temporally ordered Observation-Action blocks and combines hierarchical set encoding with causal temporal modeling to construct a shared patient state. A unified predictor combines Future Query Tokens, a shared backbone, and role-specific experts to directly forecast future observation and action representations at specified horizons. By incorporating candidate treatments into the patient history, PaST-Q predicts patient-specific clinical evolution under each treatment. Through a shared tool interface, a multi-agent system translates these predictions into clinical evidence, compares candidate treatments, and generates treatment recommendations. We evaluate PaST-Q on the public MMRF CoMMpass dataset and a private cervical cancer cohort. Comparisons of patient history integration and future simulation methods, together with clinician evaluation, demonstrate the value of longitudinal context and patient-specific forecasting for treatment recommendation and clinical decision support.

1 Introduction

Cancer treatment decisions are guided by clinical guidelines and the patient's current disease status, assessed through imaging, pathology, and laboratory examinations (Cibula et al., 2023; Dimopoulos et al., 2025). These guidelines translate population-level evidence into standard treatment recommendations, but the benefits and risks of a treatment vary across individuals (Makarov et al., 2025; Kent et al., 2020). The central challenge is therefore which treatment plan to recommend for an individual patient at the current clinical stage. Answering this question requires looking beyond the current clinical snapshot by **(1) considering the patient's longitudinal history** and **(2) anticipating how their condition may evolve under a candidate treatment in the future** (Makarov et al., 2025; Dimopoulos et al., 2025; Tseng et al., 2017). Recent advances in large language models (LLMs) have enabled AI systems that combine patient records with medical knowledge to assist clinicians in treatment recommendation (Lyu et al., 2026; Berman et al., 2025; Jun et al., 2026). However, translating these capabilities into personalized decision support faces two modeling challenges. First, clinical histories comprise heterogeneous observations and interventions recorded at irregular intervals (Che et al., 2018; Wornow et al., 2025). Static aggregation loses their temporal relationships, while processing entire histories as text faces context-length constraints and difficulties in effectively using distant information, making it challenging to capture long-range dependencies and preserve decision-relevant information (Makarov et al., 2025; Liu et al., 2024). Second, generating treatment recommendations from patient records and medical knowledge does not itself model how an individual patient's condition evolves under treatment. Simulating future responses requires learning patient-specific dynamics that connect longitudinal history, a candidate intervention, and elapsed time to subsequent clinical states (Ding et al., 2026). Without an explicit model of these dynamics, recommendation systems lack patient-specific forecasts for comparing the anticipated outcomes of candidate therapies. We therefore address personalized treatment recommendation by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bringing longitudinal clinical history and anticipated treatment-specific changes into the same decision process, enabling candidate plans to be compared using patient-specific evidence alongside clinical guidelines.

Existing research has explored two complementary directions for personalized clinical decision support: representing patients' clinical histories and simulating their future evolution. One line of work encodes longitudinal electronic health records into patient representations for downstream clinical prediction (Choi et al., 2016; Li et al., 2019; Rasmy et al., 2020), with long-context models extending the amount of historical information that can be incorporated (Wornow et al., 2025). LLM-based medical agents further combine patient information with guideline retrieval (Mohammed et al., 2025; Lyu et al., 2026) and multi-step collaborative reasoning (Chen et al., 2026; Liu et al., 2026) to generate treatment recommendations. These approaches use historical evidence to support task-specific predictions or direct recommendations, but do not expose a patient dynamics model that can be queried to anticipate clinical changes under alternative treatments. Another line of work forecasts future patient trajectories, but existing approaches remain limited in either the richness of predicted clinical information or the use of longitudinal patient history. One class of methods predicts selected clinical variables or predefined outcomes, capturing only a restricted view of the patient's future state (Makarov et al., 2025; Melnychuk et al., 2022; Steinberg et al., 2023; Chandak et al., 2026; Bica et al., 2020). Another class simulates treatment-conditioned disease evolution, but relies primarily on current disease snapshots and summarized clinical profiles rather than explicitly modeling the longitudinal sequence of observations and interventions (Ding et al., 2026; Yang et al., 2025; Mantegna et al., 2026). These approaches therefore either provide a restricted view of the patient's future or only partially represent the disease and treatment history that shapes it. This limits their ability to support comprehensive, history-grounded forecasting of how an individual patient may evolve under candidate treatments. Taken together, these approaches motivate a patient state that goes beyond summarizing historical evidence or forecasting disease outcomes alone: it should encode longitudinal observation-action histories, support queries over both future clinical observations and treatment actions at specified horizons, and provide predictive evidence for guideline-grounded treatment recommendation.

To address these challenges, we propose PaST-Q, a patient modeling and agent framework that integrates longitudinal history and treatment-conditioned future prediction to support patient-specific treatment selection. We organize each patient's history into temporally ordered Observation-Action blocks, preserving the unordered event structure within each block, and use a causal encoder to construct a shared patient state. Building on this state, a Future Query predictor combines explicit role and time conditions with a shared backbone and role-specific experts to forecast future observation and action representations. By incorporating each candidate treatment into the patient history, the model predicts how the patient's clinical state may evolve under that treatment at requested horizons. We expose these capabilities through specialized tools within a guideline-grounded multi-agent system, which coordinates candidate generation, future-state prediction, clinical readout, and treatment comparison to produce treatment recommendations. This framework brings historical disease and treatment dynamics together with anticipated future changes, providing patient-specific evidence for personalized treatment planning.

We evaluate PaST-Q on the public MMRF CoMMpass dataset and a private cervical cancer cohort. MMRF CoMMpass originates from an eight-year longitudinal study, and its treatment landscape includes proteasome inhibitor- and immunomodulatory agent-based regimens and their combinations, supporting the study of patient evolution across different therapeutic strategies (The Multiple Myeloma Research Foundation). Our private cohort comprises cervical cancer patients with a median history span of 7.2 years, and the cohort covers nearly 300 distinct treatment regimens including surgery, radiotherapy, and chemotherapy. Together, these datasets provide complementary settings for evaluating longitudinal patient modeling and future prediction across heterogeneous treatment trajectories.

We summarize our contributions as follows. **(1) Longitudinal patient history modeling.** We encode heterogeneous Observation-Action histories into a shared patient state through hierarchical set encoding and causal temporal modeling, preserving longitudinal disease and treatment context for patient-specific prediction. **(2) Unified future querying.** We develop a predictor with Future Query Tokens and role-specific experts to forecast future Observation and Action representations at requested horizons, enabling history-grounded simulation of clinical changes under candidate treatments. **(3) History-future treatment recommendation multi-agent system.** To the best of our

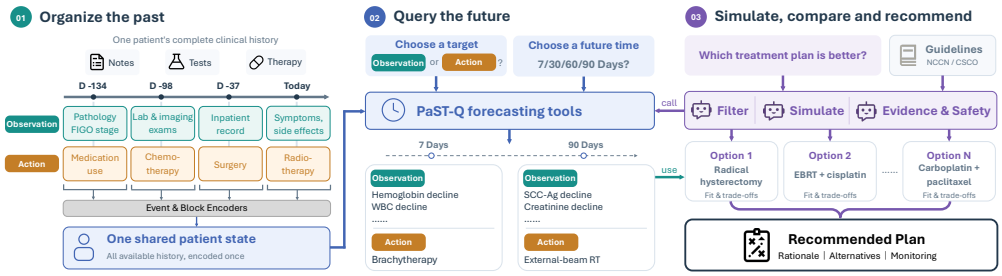


Figure 1: **Illustration of our idea.** PaST-Q connects longitudinal patient history with future simulation for personalized cancer treatment planning. It organizes clinical records into a shared patient state and predicts future clinical changes at requested horizons. A guideline-grounded multi-agent system invokes these forecasting tools to simulate and compare candidate treatments, combining historical context and predicted outcomes to produce recommendations with supporting rationale, alternatives, and monitoring priorities.

knowledge, we are the first to integrate longitudinal Observation-Action states and role and time conditioned future queries through shared tools in a guideline-grounded treatment recommendation multi-agent system, using patient-specific forecasts to compare candidate treatments and support treatment selection. Extensive experiments on MMRF CoMMpass and our private cervical cancer cohort demonstrate the value of history integration and future simulation for treatment recommendation, complemented by clinician assessments of recommendation quality and decision-support value.

2 Related Work

2.1 Longitudinal Patient History Modeling

A major line of research encodes historical clinical records into patient representations for clinical prediction and treatment recommendation (Miotto et al., 2016; Rajkomar et al., 2018). RETAIN (Choi et al., 2016), GRAM (Choi et al., 2017), and ConCare (Ma et al., 2019) capture relevant historical information through visit-level attention, ontology-aware concept embeddings, and feature-level temporal modeling. Large-scale EHR encoders, including BEHRT (Li et al., 2019) and Med-BERT (Rasmy et al., 2020), learn reusable representations through contextual pretraining of medical codes. G-BERT (Shang et al., 2019a) and GAMENet (Shang et al., 2019b) further integrate diagnosis and medication information for history-aware medication recommendation. Recent work addresses the practical challenges of longitudinal records (Li et al., 2021; Tipirneni & Reddy, 2022), with mTAN (Shukla & Marlin, 2021) handling irregular observations and Context Clues (Wornow et al., 2025) investigating the benefits of longer historical context. These approaches improve historical information extraction, but using the resulting representations does not itself provide a model of patient evolution. PaST-Q connects historical encoding with future prediction by organizing clinical events into temporally ordered Observation-Action blocks. Hierarchical set encoding and causal temporal modeling preserve within-block event structure and between-block dependencies, producing a shared patient state.

2.2 Future Patient State Prediction

Future patient prediction spans outcome-specific forecasting (Lee et al., 2020), trajectory generation, and treatment-conditioned simulation (Mantegna et al., 2026). MOTOR (Steinberg et al., 2023) and EveryQuery (Chandak et al., 2026) estimate event-time distributions and answer outcome-specific queries, respectively. Treatment-effect models (Melnychuk et al., 2022; Bica et al., 2020; Lim, 2018) and clinical trajectory predictors (Makarov et al., 2025) forecast intervention responses or longitudinal measurements. These methods use patient history but characterize the future through selected variables or outcomes. TransformEHR (Yang et al., 2023), ETHOS (Renc et al., 2024), Delphi (Shmatko et al., 2025), and Event Stream GPT (McDermott et al., 2023) broaden coverage through future-visit and event-sequence generation, while Clin-JEPA (Yang et al., 2026) forecasts latent trajectories autoregressively. Such approaches extend prediction beyond isolated end-

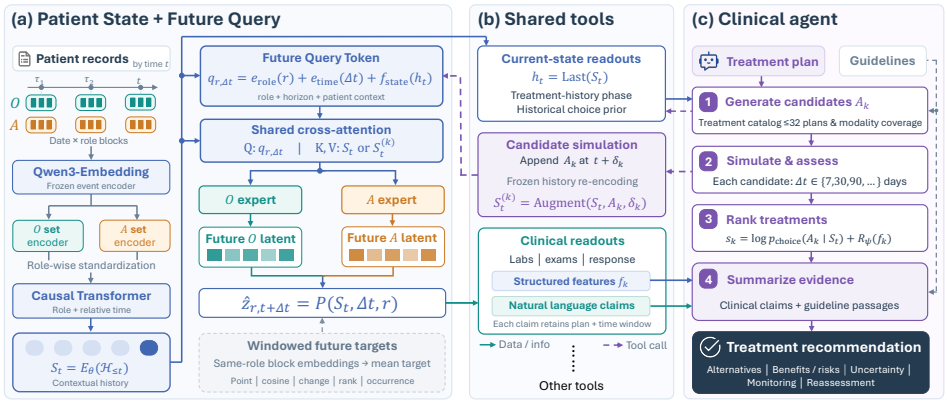


Figure 2: **Architecture of PaST-Q.** Role-specific set encoders and a causal Transformer encode Observation-Action histories into a shared patient state. Role- and horizon-conditioned Future Query Tokens guide shared cross-attention and role-specific experts to predict future latents. Appending candidate Action blocks and re-encoding the history enables treatment-conditioned forecasting. Clinical readouts expose structured predictions through shared tools, enabling agents to integrate guideline evidence, compare treatments, and generate recommendations.

points, but trajectory generation requires sequential rollout rather than direct future-state queries. MeWM (Yang et al., 2025) and CLARITY (Ding et al., 2026) connect treatment-conditioned disease simulation with planning, but start from pre-treatment imaging states, optionally supplemented with clinical profiles, rather than explicitly encoding the preceding longitudinal history. PaST-Q directly predicts future Observation or Action latents at requested horizons from a shared longitudinal patient state, providing history-grounded, reusable forecasts for treatment comparison.

2.3 Clinical Agents for Treatment Planning

Clinical agents support treatment planning through guideline-grounded generation (Zoller et al., 2026) and collaborative reasoning (Wu et al., 2026). NCCN-based Agentic-RAG (Mohammed et al., 2025) combines guideline selection, retrieval, and completeness checking for breast cancer treatment planning, while GatorOnco (Lyu et al., 2026) integrates domain adaptation and agentic retrieval for colorectal cancer treatment planning. MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024) introduce multi-role deliberation and adaptive team organization for medical reasoning. RareAgents (Chen et al., 2026) combines multidisciplinary coordination, visit memory, and medical tools for medication recommendation. EvoMDT (Liu et al., 2026) coordinates treatment, safety, and monitoring agents through retrieval, consensus, and feedback-driven refinement for oncology decision support. These mechanisms improve evidence use and collaboration but do not themselves model patient evolution under candidate treatments. PaST-Q grounds forecasting tools in a shared patient state learned from longitudinal Observation-Action histories, enabling agents to combine patient-specific future predictions with guideline evidence for treatment comparison and recommendation.

3 Method

PaST-Q connects longitudinal patient modeling, future latent prediction, and tool-based treatment recommendation. Dataset-specific adapters convert clinical records into a common Observation-Action representation. A causal encoder constructs the patient state, a Future Query predictor forecasts role-specific clinical content at requested horizons, and specialized tools translate these representations into structured evidence for clinical agents.

3.1 Problem Formulation

Let $\mathcal{H}_{\le t} = (B_1, \ldots, B_{J_t})$ denote a patient's clinical history available by time t, organized into temporally ordered Observation–Action blocks. Each block $B_j = (\mathcal{X}_j, \tau_j, r_j)$ contains an unordered event set $\mathcal{X}_j$, its relative availability time τ_j, and a role $r_j \in \{O, A\}$. Our goal is to encode this

history into a shared patient state S_t and predict future clinical content for a queried role $r \in \{O, A\}$ and time horizon Δt:

$$S_t = E_\theta(\mathcal{H}_{\le t}), \qquad \hat{z}_{r,t+\Delta t} = P_\phi(S_t, \Delta t, r), \tag{1}$$

where $\hat{z}_{r,t+\Delta t}$ represents the predicted future observations or actions. A tool-augmented multi-agent system then uses this predictive interface to assess future clinical changes under candidate treatments and combines these predictions with guideline evidence to generate personalized treatment recommendations.

3.2 Encoding Patient History State

Event organization and encoding. To construct a longitudinal patient timeline for temporal forecasting, we extract atomic clinical events and assign availability timestamps based on laboratory and imaging examinations, pathology reports, and treatment-course completion or summary times. Events are classified as Observations (O) or Actions (A) for role-aware modeling. Observations include clinical notes, diagnoses, examinations, pathology findings, symptoms, and treatment summaries, whereas Actions describe procedures and interventions. Each event is serialized with its clinical content, numerical values, and relevant attributes, then encoded by a frozen Qwen3-Embedding model (Zhang et al., 2025).

Within-block aggregation. We group events sharing the same date and Observation/Action role into blocks, with Action blocks further separated by execution status. Events are unordered within blocks, while blocks retain their temporal order. Within each block, learnable queries adaptively aggregate variable-sized sets of event embeddings through content-dependent attention, and the query outputs are pooled into a fixed-dimensional representation. Separate Observation and Action set encoders are used to capture their distinct clinical semantics. No within-block positional encoding is used, ensuring invariance to event order. We also standardize the resulting block embeddings using fixed training-set statistics for better distinction.

Between-block temporal modeling. These time blocks organize the patient's evolving clinical information into a sequence, providing the basis for temporal forecasting. We arrange block embeddings chronologically and add role and time information to represent event types and intervals between blocks. A causal Transformer integrates the sequence, allowing each position to attend only to the history available at that time. This produces a shared patient state S_t encoding all clinical information available up to t. The state supports both future querying and probe-based readouts for downstream clinical tasks. Future queries attend to the encoded history, while single-vector probes use the current contextual state h_t.

3.3 Future-Query Latent Prediction

Future Query Token design. To support treatment simulation, we introduce a Future Query mechanism for predicting future observations and actions from the current patient state. Each request is represented by a token combining the target role $r \in \{O, A\}$, forecasting interval Δt, and patient-specific context:

$$q_{r,\Delta t} = e_{\text{role}}(r) + e_{\text{time}}(\Delta t) + f_{\text{state}}(\text{Last}(S_t)), \tag{2}$$

where e_{role} selects one of two learnable role embeddings, e_{time} maps continuous time features through a two-layer MLP, and f_{state} projects the final contextual vector $\text{Last}(S_t)$ into the same embedding space. Shared cross-attention layers use this token as the query and S_t as keys and values, allowing different forecasting requests to retrieve relevant information from the same history. The resulting representation passes through the target-role expert. Shared layers capture common longitudinal dependencies, while experts specialize in Observation or Action prediction. All source-target role combinations use the unified interface $\hat{z}_{r,t+\Delta t} = P(S_t, \Delta t, r)$, enabling multiple future predictions from a single historical encoding. Prediction head details are provided in the Appendix.

Training Objective. To accommodate irregular clinical records, we construct targets by averaging standardized target-encoder embeddings of same-role blocks within a window around the requested future time. We jointly optimize

$$\mathcal{L} = \sum_{r \in \{O,A\}} \left(\mathcal{L}^r_{\text{point}} + \lambda_c \mathcal{L}^r_{\text{center}} + \lambda_\Delta \mathcal{L}^r_{\text{change}} + \lambda_r \mathcal{L}^r_{\text{rank}} \right) + \lambda_o \mathcal{L}_{\text{occ}}. \tag{3}$$

The point loss $\mathcal{L}^r_{\text{point}}$ and centered-cosine loss $\mathcal{L}^r_{\text{center}}$ align predicted and target content. The change loss $\mathcal{L}^r_{\text{change}}$ matches target differences across horizons at the same cutoff, while $\mathcal{L}^r_{\text{rank}}$ distinguishes matched targets from other patients' same-role targets. The auxiliary loss $\mathcal{L}_{\text{occ}}$ supervises whether the requested role is recorded in the future window where observation coverage supports labeling. Together, these objectives train the patient state to capture future clinical content and its temporal variation.

3.4 Patient-State Tools and Agent Integration

Candidate treatment generation. To compare treatment options through future simulation, we construct a patient-specific candidate pool from standardized treatment combinations extracted from training records. Candidate generation combines treatment-history phase estimates from a separately trained state readout of S_t with historical treatment-choice probabilities from a proposal classifier. We retain up to 32 plans, reserving representatives from different treatment modalities. Guideline passages are retrieved using each candidate's standardized treatment components. The candidates and associated evidence are then passed to simulation and outcome-assessment tools.

Candidate-conditioned future simulation. We use the future prediction interface in Section 3.3 to forecast the patient's clinical state under each candidate treatment. For candidate A_k, we append its encoded Action to the patient's history at the proposed time and recompute the patient state using the frozen temporal encoder:

$$S_t^{(k)} = \text{Augment}(S_t, A_k, \delta_k), \qquad \hat{z}^{(k)}_{O,t+\delta_k+\Delta t} = P(S_t^{(k)}, \Delta t, O), \tag{4}$$

where δ_k is the treatment-start offset relative to the current cutoff. We query future Observations at $\Delta t \in \{7, 30, 90\}$ days after treatment initiation to capture changes over multiple clinically relevant intervals.

From patient representations to clinical predictions. Candidate simulations produce future latents for different treatments and forecasting horizons. To make this information accessible to the language agent, we train lightweight, task-specific readouts on patient representations for classification, numerical regression, and distribution estimation. Additional tasks require corresponding labels and readout training, without retraining the patient encoder or future predictor. We select deployed readouts by validation performance and label support, covering treatment-history phase, subsequent treatment components, laboratory trends, and clinical events. Each readout produces structured estimates and concise clinical statements, retaining the associated treatment and time window. The shared patient representation thus supports multiple clinical tasks, providing numerical features for treatment comparison and natural language evidence for recommendation synthesis.

Treatment ranking and recommendation. Structured readout predictions support candidate ranking, while their clinical descriptions provide evidence for explaining recommendations. A learned ranker combines simulated outcomes with the current patient state, treatment components, phase compatibility, and a historical treatment-choice prior:

$$s_k = \log p_{\text{choice}}(A_k \mid S_t) + R_\psi(f_k), \tag{5}$$

where f_k contains the structured candidate features and R_ψ is a bounded residual scorer. The ranker is trained with candidate-wise softmax cross-entropy against recorded treatment choices. The execution layer combines rankings, pairwise predicted-outcome differences, readout-derived clinical statements, and guideline evidence to explain differences in predicted response, laboratory changes, and adverse-event risks across horizons. Relevant predictions and guideline requirements also inform monitoring and reassessment priorities. The final output includes preferred and alternative treatments, supporting rationales, predicted benefits and risks, uncertainty, and monitoring and reassessment plans. For supported chemotherapy combinations, a second stage compares predefined starting-dose plans using their predicted outcomes.

4 Experiments

We evaluate our workflow through three experiments. Section 4.2 examines the contribution of patient history integration to agent treatment recommendation by comparing our history encoding method with alternative approaches to incorporating patient history. Section 4.3 assesses the

Table 1: History integration in agent treatment recommendation.

	Private-CC			MMRF		
Method	Prec. ↑	F1 ↑	Acc@1 ↑	Prec. ↑	F1 ↑	Acc@1 ↑
Population state	28.62	26.26	19.50	28.73	20.00	19.23
Recent 4 blocks	60.66	57.90	53.00	65.36	70.48	33.33
Transformer	<u>61.44</u>	<u>60.35</u>	<u>54.50</u>	66.48	<u>72.24</u>	34.62
mTAN	60.65	58.39	51.50	**68.11**	72.41	**35.90**
Mamba	55.11	54.19	46.50	66.49	71.30	34.62
PaST-Q (Ours)	**63.11**	**60.65**	**55.50**	<u>67.72</u>	**72.73**	**35.90**

contribution of future simulation by comparing our simulation method with alternative forecasting methods and a variant without future simulation. A physician study (Section 4.4) further assesses clinical acceptance, complemented by an exploratory study examining support for clinical reasoning and decision confidence.

4.1 Experimental Setup

Dataset. We use two longitudinal clinical cohorts: MMRF-CoMMpass, comprising 995 patients with multiple myeloma, and Private-CC, a private EHR cohort comprising 6,449 patients with cervical malignancy. Records are organized into temporally ordered observation and action blocks. Both cohorts use fixed patient-level training, validation, and test partitions of approximately 80%/10%/10%.

Evaluation metrics. Sections 4.2 and 4.3 share the same evaluation metrics. Precision and F1 measure agreement between the treatment components in the preferred recommendation and the recorded reference treatment, while Acc@1 measures exact agreement of the complete treatment set. Agent evaluation on Private-CC uses a fixed subset of 200 randomly selected test patients, while model-level evaluations use the full test partitions of both cohorts.

4.2 Effectiveness of Patient History Integration

To evaluate the effectiveness of our patient history integration approach, we compare different ways of incorporating patient history into the agent system while keeping the future prediction method fixed. We include two input controls: **Population state**, which replaces the individual patient state with the average training-population state, and **Recent 4 blocks**, which encodes only the four most recent blocks before the decision cutoff. We also compare three different history encoders: **Transformer** (Vaswani et al., 2017), **mTAN** (Shukla & Marlin, 2021), and **Mamba** (Gu & Dao, 2024), which represent the observed history using causal self-attention, time-aware attention with recurrent aggregation, and selective state-space modeling, respectively. Their representations are mapped to the common state space for future prediction.

Table 1 shows that our method achieves the strongest overall treatment agreement across the two cohorts. On Private-CC, PaST-Q leads all three recommendation metrics, achieving 63.11% Precision, 60.65% F1, and 55.50% Acc@1. On MMRF, our method achieves the highest F1 of 72.73% and the joint-best Acc@1 of 35.90%, together with the second-highest Precision of 67.72%. These results support the effectiveness of our patient history integration approach in providing informative patient states for downstream agent treatment recommendation.

4.3 Effectiveness of Future Simulation

To evaluate the effectiveness of our future simulation approach, we keep the patient history encoder fixed and compare different future prediction methods within the agent recommendation workflow. The comparison includes variants that remove future-state simulation or treatment conditioning. **No future simulation** evaluates candidate treatments through direct outcome readouts from the current patient state. **No action conditioning** uses a history-conditioned future prediction shared across candidates, while retaining candidate identity in the downstream outcome readouts. We compare three learned forecasting alternatives: **AR** advances the latent patient state through recurrent transitions; **ClinJEPA-adapted** (Yang et al., 2026) applies an action-conditioned Transformer predictor

Table 2: Future simulation in agent treatment recommendation.

	Private-CC			MMRF		
Method	Prec. ↑	F1 ↑	Acc@1 ↑	Prec. ↑	F1 ↑	Acc@1 ↑
No future simulation	59.80	58.10	52.00	66.12	72.02	34.62
No action conditioning	60.74	58.20	52.50	65.57	71.43	33.33
ClinJEPA-adapted	59.67	56.96	51.50	**67.76**	72.51	34.62
AR	61.67	59.29	54.00	67.55	72.36	<u>35.90</u>
CLARITY-adapted	<u>62.30</u>	<u>59.56</u>	<u>55.00</u>	67.22	**72.89**	**37.18**
AgentEHR-adapted	54.62	46.66	44.50	62.23	66.67	34.62
PaST-Q (Ours)	**63.11**	**60.65**	**55.50**	<u>67.72</u>	<u>72.73</u>	<u>35.90</u>

Table 3: Physician evaluation scores (%). Note that selecting *both* contributes half a point to each method, and selecting *neither* contributes none.

Method	Q1 Clinical plan ↑	Q2 Predictions ↑	Q3 Monitoring ↑	Q4 Decision value ↑	Q5 Errors ↓
AgentEHR-adapted	38.9	38.9	**50.0**	38.9	44.4
PaST-Q (Ours)	**44.4**	**55.6**	**50.0**	**55.6**	**38.9**

to the common state and action representations; and **CLARITY-adapted** (Ding et al., 2026) uses a residual Transformer to predict future states conditioned on patient history, candidate treatment, and elapsed time. **AgentEHR-adapted** (Liao et al., 2026) is an AgentEHR adaptation that generates clinical outcome forecasts from state-derived patient summaries and treatment descriptions, with up to 32 candidates.

Table 2 shows that our method achieves strong overall performance, leading all three treatment-agreement metrics on Private-CC. PaST-Q achieves 63.11% Precision, 60.65% F1, and 55.50% Acc@1 on Private-CC. On MMRF, our method ranks second in Precision and F1 and joint-second in Acc@1, demonstrating competitive performance across the recommendation metrics. Across both cohorts, ours consistently improves all three treatment-agreement metrics over both No future simulation and No action conditioning. These comparisons support the value of simulating future patient states under candidate treatments for agent treatment recommendation.

4.4 Physician Evaluation

We conduct two physician studies to assess the clinical usefulness of our workflow: a comparative evaluation of recommendation quality and an evaluation of its support for clinical reasoning and decision confidence.

Comparative assessment of recommendations. To validate the clinical usefulness of the recommendations generated by our method, we invited physicians to compare our method with the adapted version of AgentEHR, the baseline most closely aligned with our goal of supporting treatment decisions from longitudinal patient records. For each case, physicians reviewed the longitudinal patient record and assessed two outputs in the same layout. The questionnaire covered clinical appropriateness and completeness (Q1), prediction interpretation and comparison (Q2), monitoring and reassessment (Q3), overall decision-support value and specificity (Q4), and important errors, omissions, or overinterpretation (Q5). Three physicians each evaluated six retrospective cervical cancer cases, yielding 18 paired physician-case assessments.

As shown in Table 3, our method outperforms AgentEHR on four of the five criteria and matches it on monitoring and reassessment. The largest gains are in prediction interpretation and comparison and overall decision-support value, both improving from 38.9% to 55.6%. It also achieves better clinical appropriateness and completeness and a lower error score.

Support for clinical reasoning and confidence. We invited physicians to review longitudinal records of retrospective cervical cancer cases, recording treatment plans and confidence (0-100) before and after reviewing our generated recommendations. Two physicians each reviewed four cases, resulting in eight paired assessments. Physicians reported previously overlooked factors related to

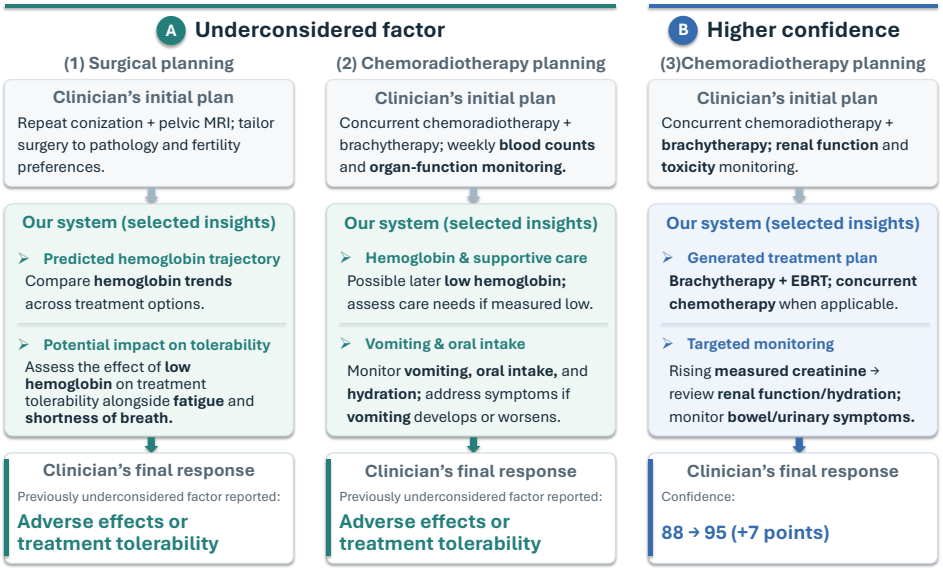


Figure 3: **Case studies of clinical decision support.** The physician reported previously underconsidered adverse-effect or tolerability factors in two cases, and confidence increased from 88 to 95 after reviewing system recommendations in one case.

adverse effects or treatment tolerability in two assessments, and a confidence increase in one assessment. Figure 3 presents these three examples.

Panel A illustrates how our system broadens consideration of treatment tolerability. In surgical planning, it connected predicted trajectories with low hemoglobin, fatigue, and shortness of breath. In chemoradiotherapy planning, it highlighted possible later low hemoglobin, supportive-care assessment if confirmed by measurement, and vomiting-related concerns. In both assessments, the physician reported previously underconsidered adverse-effect or tolerability factors. Panel B illustrates support for decision confidence through treatment alignment and concrete monitoring guidance. The system situated brachytherapy within combined chemoradiotherapy and specified renal-function and hydration reassessment when measured creatinine increased, alongside bowel and urinary symptom monitoring. These recommendations provided structured support for reviewing the physician's existing strategy. The physician increased confidence from 88 to 95.

Together, these findings demonstrate how our system broadens risk consideration and reinforces decision confidence. By connecting longitudinal predictions with treatment choices and monitoring actions, it shows practical potential to support physicians in clinical planning and patient management.

5 Conclusion

In this work, we presented PaST-Q, a patient modeling and multi-agent framework that connects longitudinal history with future simulation for personalized cancer treatment planning. PaST-Q organizes heterogeneous Observation-Action histories into a shared patient state and uses role and horizon conditioned queries to predict future clinical representations, including patient evolution under candidate treatments. Through a shared tool interface, these predictions are translated into structured clinical information and combined with guideline evidence for treatment comparison and recommendation. Experiments on MMRF CoMMpass and a private cervical cancer cohort support the value of longitudinal patient modeling and future prediction, while clinician assessments illustrate how the system can complement treatment planning with additional considerations about adverse effects and treatment tolerability. Together, these findings highlight the value of a future-queryable patient state that makes accumulated clinical history and anticipated treatment responses jointly accessible to decision-support agents.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refining the conceptual framework, providing feedback on methodology and experimental design, implementing and debugging code. We did not use these tools to synthesize the patient cohorts, and formal theorem proving is not applicable to this work. Additionally, we used generative AI tools to identify and summarize relevant literature, organize and draft manuscript sections, improve readability, and assist with figure preparation. We have reviewed all AI-assisted work, reviewed and tested AI-assisted code, and verified reported results against experimental outputs and clinician assessments. The language models used as components of PaST-Q and the evaluated baselines are described in the corresponding methodology and experimental sections. We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This study uses de-identified clinical data from the Multiple Myeloma Research Foundation (MMRF) CoMMpass study, obtained through the National Cancer Institute's Genomic Data Commons (GDC). We additionally used retrospective cervical cancer records from a collaborating cancer center, whose identity is withheld for double-blind review. The use of the cancer center's records for this study was reviewed and approved by its ethics committee, which waived the requirement for informed consent for the retrospective use of these records. Access to and use of both datasets complied with the applicable data-use policies, agreements, and institutional requirements. The transmission of patient-level data to external services was strictly prohibited; all data processing, including model inference, was therefore performed locally.

REFERENCES

Eliza Berman, Holly Sundberg Malek, Michael Bitzer, Nisar Malek, and Carsten Eickhoff. Retrieval augmented therapy suggestion for molecular tumor boards: Algorithmic development and validation study. *J Med Internet Res*, 27:e64364, Mar 2025. ISSN 1438-8871. doi: 10.2196/64364. URL https://www.jmir.org/2025/1/e64364.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations, 2020. URL https://arxiv.org/abs/2002.04083.

Payal Chandak, Gregory Kondas, Liat Antwarg Friedman, Isaac Kohane, and Matthew McDermott. Everyquery: Zero-shot clinical prediction via task-conditioned pretraining over electronic health records, 2026. URL https://arxiv.org/abs/2603.07900.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.*, 8(1):6085, April 2018.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. RareAgents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *Proc. Conf. AAAI Artif. Intell.*, 40(1):101–109, March 2026.

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. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning, 2017. URL https://arxiv.org/abs/1611.07012.

David Cibula, Maria Rosaria Raspollini, François Planchamp, Carlos Centeno, Cyrus Chargari, Ana Felix, Daniela Fischerová, Daniela Jahnn-Kuch, Florence Joly, Christhardt Kohler, Sigurd Lax, Domenica Lorusso, Umesh Mahantshetty, Patrice Mathevet, Raj Naik, Remi A Nout, Ana Oaknin, Fedro Peccatori, Jan Persson, Denis Querleu, Sandra Rubio Bernabé, Maximilian P Schmid, Artem Stepanyan, Valentyn Svintsitskyi, Karl Tamussino, Ignacio Zapardiel, and Jacob Lindegaard. ESGO/ESTRO/ESP guidelines for the management of patients with cervical cancer - update 2023. *Int. J. Gynecol. Cancer*, 33(5):649–666, May 2023.

Meletios A Dimopoulos, Evangelos Terpos, Mario Boccadoro, Philippe Moreau, María-Victoria Mateos, Sonja Zweegman, Gordon Cook, Monika Engelhardt, Michel Delforge, Roman Hajek, Fredrik Schjesvold, Francesca Gay, Salomon Manier, Katja C Weisel, Martin Kaiser, Niels W C J van de Donk, Elena Zamagni, Paula Rodriguez-Otero, Aurore Perrot, Christoph Driessen, Jelena Bila, Edward Laane, Dominik Dytfeld, Cyrille Touzeau, Meral Beksac, Marc S Raab, Michele Cavo, Mohamad Mohty, Andrew Spencer, Heinz Ludwig, Hermann Einsele, Jesus San-Miguel, and Pieter Sonneveld. EHA-EMN Evidence-Based guidelines for diagnosis, treatment and follow-up of patients with multiple myeloma. *Nat. Rev. Clin. Oncol.*, 22(9):680–700, September 2025.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. Clarity: Medical world model for guiding treatment decisions by simulating context-aware disease trajectories, 2026. URL https://arxiv.org/abs/2512.08029.

Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, Katharina Egger-Heidrich, Marius Bill, Martin M K Schneider, Jan Clusmann, Lejla Kadric, Marcel Oehme, Maximilian Mayrhofer-Schmid, Alexander Oeser, Georg Wölflein, Isabella C Wiest, Jan Moritz Middeke, A John Iafrate, Daniel Truhn, Dirk Jäger, and Jakob Nikolas Kather. Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125):1282–1291, July 2026.

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. Txagent: An ai agent for therapeutic reasoning across a universe of tools, 2025. URL https://arxiv.org/abs/2503.10970.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Hyeji Jun, Yutaro Tanaka, Shreya Johri, Sabrina Y Camp, Erik L Bao, Filipe L F Carvalho, Dan Y Gui, Alexander C Jordan, Chris Labaki, Samantha D Martin, Matthew Nagy, Tess A O'Meara, Theodora Pappa, Erica Maria Pimenta, Eddy Saad, David D Yang, Riaz Gillani, Alok K Tewari, Brendan Reardon, and Eliezer Van Allen. A context-augmented large language model for accurate precision oncology medicine recommendations. *Cancer Cell*, 44(3):676–685.e4, March 2026.

David M Kent, Jessica K Paulus, David van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John P A Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, Gowri Raman, Joseph S Ross, Harry P Selker, Ravi Varadhan, Andrew Vickers, John B Wong, and Ewout W Steyerberg. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann. Intern. Med.*, 172(1):35–45, January 2020.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making, 2024. URL https://arxiv.org/abs/2404.15155.

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, 2020. doi: 10.1109/TBME.2019.2909027.

Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records, 2019. URL https://arxiv.org/abs/1907.09538.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records, 2021. URL https://arxiv.org/abs/2106.11360.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. ReflecTool: Towards reflection-aware tool-augmented clinical agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13507–13531, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.663. URL https://aclanthology.org/2025.acl-long.663/.

Yusheng Liao, Chuan Xuan, Yutong Cai, Lina Yang, Zhe Chen, Yanfeng Wang, and Yu Wang. Agentehr: Advancing autonomous clinical decision-making via retrospective summarization, 2026. URL https://arxiv.org/abs/2601.13918.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

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/.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Goh Kim Huat, Hyeokkoo Eric Kwon, Feng Gao, Xianfu Sun, Zhitao Ying, and Guangliang Qiang. EvoMDT: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *NPJ Digit. Med.*, 9(1):124, January 2026.

Mengxian Lyu, Cheng Peng, Tim Jang, Ang Li, Mengyuan Zhang, Ziyi Chen, Leighton Elliott, Tianshi Liu, Lidice Galindo, Chiranjeevi Sainatham, Oscar F. Borja-Montes, Kaleb E. Smith, Ying Zhang, Lichao Sun, Jiang Bian, Gloria Lipori, Duane A. Mitchell, Elizabeth A. Shenkman, Yi Guo, Thomas J. George, and Yonghui Wu. An agentic generative large language model for treatment planning of colorectal cancer, 2026. URL https://arxiv.org/abs/2608.09142.

Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiantao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context, 2019. URL https://arxiv.org/abs/1911.12216.

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 Digit. Med.*, 8(1):588, October 2025.

Massimiliano Mantegna, Elena Mulero Ayllón, Alice Natalina Caragliano, Francesco Di Feola, Claudia Tacconi, Michele Fiore, Edy Ippolito, Carlo Greco, Sara Ramella, Philippe C. Cattin, Paolo Soda, Matteo Tortora, and Valerio Guarrasi. Longitudinal nsclc treatment progression via multimodal generative models, 2026. URL https://arxiv.org/abs/2603.06147.

Matthew B. A. McDermott, Bret Nestor, Peniel Argaw, and Isaac Kohane. Event stream gpt: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events, 2023. URL https://arxiv.org/abs/2306.11547.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15293–15329. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.*, 6(1):26094, May 2016.

Abdul M. Mohammed, Iqtidar Mansoor, Sarah Blythe, and Dennis Trujillo. Developing an artificial intelligence tool for personalized breast cancer treatment plans based on the nccn guidelines, 2025. URL https://arxiv.org/abs/2502.15698.

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 Digit. Med.*, 1(1):18, May 2018.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction, 2020. URL https://arxiv.org/abs/2005.12833.

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 Digit. Med.*, 7(1): 256, September 2024.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5953–5959. International Joint Conferences on Artificial Intelligence Organization, 7 2019a. doi: 10.24963/ijcai.2019/825. URL https://doi.org/10.24963/ijcai.2019/825.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination, 2019b. URL https://arxiv.org/abs/1809.01852.

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*, 647(8088):248–256, November 2025.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series, 2021. URL https://arxiv.org/abs/2101.10318.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records, 2023. URL https://arxiv.org/abs/2301.03150.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

The Multiple Myeloma Research Foundation. About the MMRF CoMMpass study. MMRF Virtual Lab Documentation. https://themmrf.github.io/docs-vlab/mmrf-resources/mmrf-commpass/about-commpass/. Accessed: 2026-09-25.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series, 2022. URL https://arxiv.org/abs/2107.14293.

Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K. Ten Haken, and Issam El Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics*, 44(12):6690–6705, 2017. doi: https://doi.org/10.1002/mp.12625. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12625.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs, 2025. URL https://arxiv.org/abs/2412.16178.

Yichen Wu, Kailong Fan, Sangjoon Park, Yuhan Liu, Zhiyi Shi, Sekeun Kim, Dania Daye, Hana Farzaneh, Xiang Li, Raul Uppot, Yujin Oh, and Quanzheng Li. Comma: Contribution-aware medical multi-agents for decentralized oncology decision support, 2026. URL https://arxiv.org/abs/2602.09159.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model: Generative simulation of tumor evolution for treatment planning, 2025. URL https://arxiv.org/abs/2506.02327.

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, 2026. URL https://arxiv.org/abs/2605.10840.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat. Commun.*, 14(1):7857, November 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Julian Zoller, Michael Kalz, Xuewei Wu, Sven Cuntz, Linus Kruk, Niklas Kehl, Julius J Michel, Cornelius Funk, Sarah Richter, Tobias Tix, David Sedloev, Jonathan Naboschni, Jan H Frenking, Silvia Barbosa, Oliver L Saldanha, Alanna Kirschner, Anna D Metzler, Antonia Schach, René Onken, Tim R Wagner, Martin Dugas, Andreas Trumpp, Martin Dreyling, Michael von Bergwelt-Baildon, Kai Rejeski, Tim Sauer, Peter Dreger, Marc S Raab, Carsten Müller-Tidow, Jakob N Kather, and Mirco J Friedrich. Clinical decision support in hematological malignancies using a case-grounded AI agent. *Nat. Med.*, 32(9):3224–3234, September 2026.

Table 4: Private-CC cohort and retained model input. Quartiles are computed over patients.

Statistic	All	Train	Validation	Test
Patients	6,449	5,159	645	645
Source rows	9,996,727	7,990,607	1,047,295	958,825
Retained O events	9,547,202	7,630,045	1,006,330	910,827
Retained A events	169,477	135,722	16,930	16,825
Blocks/patient, median	66	66	67	66
Blocks/patient, IQR	37–94	37–93	38–97	36–92
Recorded span (years), median	7.20	7.23	7.28	6.89
Recorded span (years), IQR	4.30–9.82	4.31–9.87	4.31–9.72	4.20–9.60

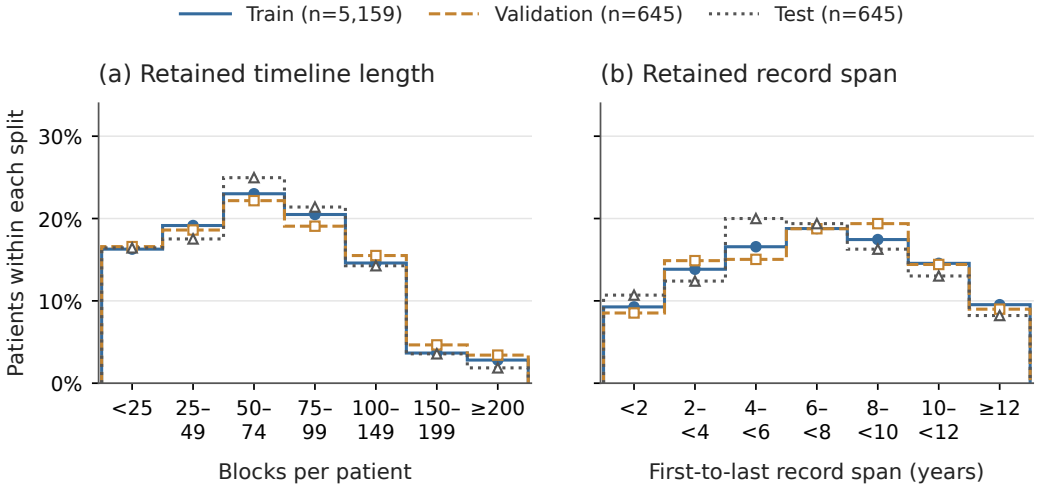


Figure 4: Distribution of retained longitudinal records in Private-CC. Each curve shows the fraction of patients within its own split. (a) Number of retained blocks, grouped by local calendar day. (b) Time between the first and last retained records, including outcome summaries. The horizontal axes show ordered bands, with all upper-tail values included in the last band.

A Appendix

A.1 The Private Cervical Cancer Cohort

Private-CC. Private-CC is a private longitudinal electronic health record (EHR) cohort for cervical malignancy. Patients were included when at least one diagnosis-name entry exactly matched the term for cervical malignancy. The resulting cohort contains 6,449 unique patients and 9,996,727 source-table rows across 22 table schemas, covering diagnoses, laboratory results, imaging and pathology reports, inpatient and outpatient documentation, and treatment records. We retain a fixed, patient-disjoint training/validation/test split of 5,159/645/645, keeping each patient's longitudinal records in one split. After the frozen preprocessing pipeline, the model input contains 9,547,202 Observation events and 169,477 Action events, with a median of 66 blocks per patient, grouped by date, role, and execution status. The median first-to-last retained-record span is 7.20 years. Natural language clinical records are used by the upstream patient encoder and for labels and evaluation. Downstream models and agents obtain patient clinical state only through the encoded state, with age and sex permitted separately. Summary statistics and timeline distributions are provided in Table 4 and Figure 4.

Preprocessing and temporal availability. The frozen adapter removes invalid dates, normalizes and deduplicates events, and applies limits of 96 O events and 64 A events per day within each execution-status group. Clinical text fields are limited to 320 characters per field and 900 characters in total per event. Pathology results require report dates, and retrospective summaries become available at their documented completion or summary time.

Table 5: Inventory of source tables in the selected cohort. Coverage means at least one source row for a patient. Patients overlap across tables.

Source table	Rows	Patients	Coverage
Admission records	31,856	6,433	99.8%
Anesthesia events	1,860	1,547	24.0%
Anesthesia medications	36,307	1,520	23.6%
Brachytherapy sessions	8,848	2,250	34.9%
Death events	23	23	0.4%
Diagnosis records	240,554	6,449	100.0%
Discharge records	31,821	6,430	99.7%
Follow-up records	90,822	6,278	97.3%
Imaging reports	154,449	6,373	98.8%
Other invasive procedures	31	20	0.3%
Laboratory results	8,935,720	6,436	99.8%
Outpatient encounters	208,837	6,377	98.9%
Pathology reports	21,719	6,449	100.0%
Patient outcome summaries	6,449	6,449	100.0%
Periodic course summaries	11	9	0.1%
Radiotherapy courses	12,766	2,939	45.6%
Rescue events	128	93	1.4%
Surgery-anesthesia links	1,899	1,547	24.0%
Surgery components	48,718	4,775	74.0%
Surgery events	36,510	4,775	74.0%
Systemic therapy administrations	83,269	5,653	87.7%
Systemic therapy courses	44,130	5,469	84.8%

Coverage and interpretation. Table 5 reports source-specific row counts, unique patient counts, and patient coverage. Coverage indicates that a patient has at least one record of the corresponding type; patients can appear in multiple tables, and source availability does not establish complete outcome ascertainment.

Availability of future-state supervision. Figure 5 reports the fraction of fixed patient–cutoff queries with an eligible future-state target at each prediction horizon, separately for the training, validation, and test splits. Eligibility requires recorded coverage through the target time, a new record between the cutoff and the target time, and a most recent record no more than 30 days old. The denominator counts queries, with O/A requests for the same future state counted once. In the test split, coverage decreases from 82.1% at 30 days to 18.5% at 365 days. This reduced availability of supervision can make prediction at longer horizons more challenging.

A.2 Detailed Method

Prediction head details. The predictor combines the queried historical information with a continuation estimate based on the latest same-role block. Let $c_{t,r}$ denote this recent content. The prediction takes the form

$$\hat{z}_{r,t+\Delta t} = b_{r,\Delta t}(c_{t,r}) + R_r(S_t, u_{r,\Delta t}, c_{t,r}, \Delta t), \tag{6}$$

where $b_{r,\Delta t}$ is an affine persistence estimate fitted on training data, and $u_{r,\Delta t}$ is the role-specific query output. The residual term R_r combines the query output with a learned correction conditioned on the current state, recent content, and horizon. This design captures both the continuation of recently recorded clinical content and changes indicated by the wider patient history.

Implementation details. We implement the models in PyTorch. A frozen Qwen3-Embedding-4B encoder produces 1,024-dimensional event embeddings, which are pooled into 512-dimensional block representations. The predictor uses a four-layer, eight-head causal Transformer and a two-layer future-query decoder. We train predictors separately for each cohort. Our tools and agent obtain patient clinical information through the encoded state. All agent comparisons use Qwen3.8-27B with common inference budgets. Hyperparameters and checkpoints are selected using training and validation data.

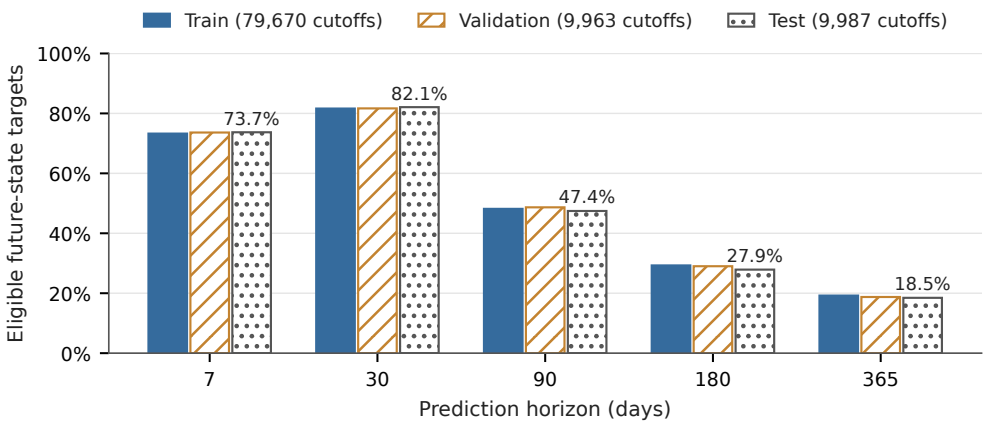


Figure 5: Coverage of future-state supervision under the frozen query protocol. The denominator is the number of fixed patient-cutoff requests in each split, with O/A requests for the same future state counted once. Eligible targets require a new record after the cutoff, recorded coverage through the target time, and a most recent record no more than 30 days old. Test percentages are annotated.

A.3 Supplementary Experiments

The main experiments evaluate how patient history integration (Section 4.2) and future simulation (Section 4.3) affect treatment recommendation within the agent workflow, followed by physician assessment (Section 4.4). The experiments below examine the intermediate capabilities underlying these results. We first evaluate future prediction in the latent space, then assess whether the representations support useful clinical tool outputs, and finally examine complete agent systems. These evaluations connect representation quality, tool performance, and recommendation behavior, while distinguishing the evidence provided by each level.

Evaluation scope. The model-level and tool-level evaluations use MMRF CoMMpass and Private-CC, with the patient-level partitions described in the main experimental setup. The supplementary agent benchmark uses Private-CC. This agent benchmark reports treatment-selection metrics on 179 labeled patients, whereas Completion is evaluated over all 200 tasks.

A.3.1 Future-Query Latent Forecasting

The history and future-simulation experiments in the main text evaluate representations through their downstream effect on treatment recommendation. Here we test the forecasting capability directly by examining whether the model predicts the clinical content recorded at a requested future horizon. This evaluation assesses the predictive information supplied to the clinical readouts without involving the language agent or treatment ranker. It therefore complements the end-to-end comparisons by examining whether the shared patient state supports accurate and discriminative future representations.

Task and metrics. Given the patient state S_t, a target role $r \in \{O, A\}$, and a horizon Δt, the predictor estimates the future latent $\hat{z}_{r,t+\Delta t}$. Targets are constructed from encoded same-role blocks in the future window. RMSE and MAE measure error against these encoded targets. For retrieval, we rank the full set of distinct target candidates of the requested role by cosine similarity to the prediction. R@10, R@20, and R@30 measure whether the paired target appears among the top-ranked candidates, while MRR measures its reciprocal rank. Results aggregate Observation and Action predictions at the query level. Latent retrieval measures paired-target discriminability.

Comparators. We compare with copying the last available same-role latent, a causal Transformer with a conditional prediction head, mTAN for irregular time series (Shukla & Marlin, 2021), the Mamba-based Context Clues model (Wornow et al., 2025), and autoregressive (AR) latent rollout.

Table 6: Future-query latent forecasting on both cohorts. Recall values are percentages, whereas MRR is reported on a 0-1 scale. Bold indicates the best value within each dataset and metric.

Method	RMSE ↓	MAE ↓	R@10 ↑	R@20 ↑	R@30 ↑	MRR ↑
MMRF CoMMpass						
Last same-role latent	1.355	1.059	4.51	8.73	14.31	0.025
Causal Transformer	0.726	0.558	35.11	53.78	63.33	0.178
mTAN	0.731	0.561	35.03	53.10	**65.81**	0.190
Context Clues	0.747	0.567	37.39	53.06	61.95	0.188
AR latent rollout	**0.721**	0.553	36.52	54.27	63.49	0.185
PaST-Q	0.722	**0.546**	**40.49**	**55.53**	65.62	**0.201**
Private-CC						
Last same-role latent	1.084	0.811	3.51	4.71	5.58	0.017
Causal Transformer	0.537	0.394	4.79	7.92	10.59	0.023
mTAN	0.528	0.388	6.01	9.69	12.58	0.028
Context Clues	0.527	0.386	5.02	8.28	11.01	0.024
AR latent rollout	0.538	0.395	5.43	9.01	11.58	0.025
PaST-Q (Ours)	**0.508**	**0.370**	**6.51**	**10.04**	**12.74**	**0.031**

Table 7: Treatment candidate retrieval. All values are reported on a percentage scale. Bold indicates the best value within each dataset and metric.

	Private-CC			MMRF		
Method	Recall@5 ↑	nDCG@5 ↑	MRR ↑	Recall@5 ↑	nDCG@5 ↑	MRR ↑
Copy + readout	56.25	44.15	45.32	62.11	48.24	47.27
Direct-state Linear	58.31	47.30	48.31	38.14	32.21	33.15
Direct-state MLP	61.18	49.07	49.99	55.93	43.07	43.15
mTAN + readout	61.15	51.03	52.78	59.39	45.43	44.68
Mamba + readout	60.82	50.55	52.33	54.86	43.41	43.54
PaST-Q (Ours)	**61.40**	**51.22**	**53.22**	**65.05**	**49.73**	**47.56**

mTAN and Context Clues are adapted to future-latent regression. The copying baseline tests persistence, the history models test alternative ways of extracting predictive information from longitudinal records, and AR rollout compares iterative prediction with direct future querying.

Results. Table 6 shows that PaST-Q achieves the best value on all metrics on Private-CC. On MMRF, PaST-Q achieves the lowest MAE and the highest R@10, R@20, and MRR. The results support the predictive utility of future querying across both cohorts.

A.3.2 Clinical Utility of Representation-Based Tools

The future-simulation experiment in Section 4.3 measures whether predicted futures improve the final recommendation. Here we examine the information passed from patient representations to the agent: treatment candidates and anticipated laboratory changes. These evaluations test whether predictive representations can be translated into clinically interpretable outputs, bridging latent forecasting and end-to-end agent evaluation.

Treatment candidate retrieval. This task evaluates whether the tool places recorded treatment choices in a short candidate list before outcome assessment and final ranking. Recall@5 measures candidate coverage, while nDCG@5 and MRR assess ranking quality. Copy plus readout provides a persistence control; direct-state Linear and MLP readouts bypass future-representation prediction; mTAN and Mamba (Gu & Dao, 2024) provide alternative predictors with task-specific readouts.

As shown in Table 7, PaST-Q leads all three retrieval metrics on both cohorts. These results support the tool's role in proposing relevant candidates for subsequent comparison.

Table 8: Future observation readout for laboratory changes and trends. Macro-F1 and selective accuracy are percentages. Bold indicates the best value within each dataset and metric.

	Private-CC			MMRF		
Method	Δ-nMAE ↓	Macro-F1 ↑	Sel. Acc. ↑	Δ-nMAE ↓	Macro-F1 ↑	Sel. Acc. ↑
Copy + readout	0.965	34.82	62.06	4.733	33.05	57.48
Direct-state Linear	1.058	36.96	61.75	5.202	32.17	57.20
Direct-state MLP	0.954	35.33	61.97	4.391	32.18	58.46
mTAN + readout	0.966	34.97	62.07	4.828	35.70	57.52
Mamba + readout	0.967	34.79	62.17	4.951	33.41	53.96
PaST-Q (Ours)	**0.920**	**38.03**	**62.60**	**4.320**	**36.14**	**61.47**

Table 9: Supplementary Private-CC agent benchmark. Rates are percentages, and LLM Tokens is the mean task total in thousands. Acc@1 includes dosage agreement. Bold indicates the best value, including ties.

Method	Precision ↑	Treatment F1 ↑	Acc@1 ↑	Completion ↑	LLM Tokens (k) ↓
Text-only Qwen	0.56	0.56	0.00	1.50	41.991
Text + Guideline RAG	0.93	1.01	0.00	3.50	47.955
ReAct + Same Tools (Yao et al., 2023)	51.12	49.80	41.90	97.00	488.332
TxAgent-style (Gao et al., 2025)	23.65	24.87	10.61	81.50	587.287
AgentEHR (Liao et al., 2026)	31.84	30.90	24.58	38.50	206.140
ReflecTool (Liao et al., 2025)	29.05	28.44	20.11	74.00	411.842
MIRA (Ferber et al., 2026)	8.38	8.03	6.15	16.50	606.021
PaST-Q (Ours)	**62.57**	**61.36**	**54.19**	**100.00**	**1.756**

Future observation readout. This task evaluates the laboratory forecasts used to explain anticipated changes and suggest monitoring priorities. We measure change magnitude using training-IQR-normalized MAE (Δ-nMAE). Three-class Macro-F1 evaluates decreasing, stable, and increasing trends before abstention. Selective accuracy evaluates the directions exposed to the agent after allowing an explicit "uncertain" outcome. The baseline families match those used for treatment candidate retrieval.

Table 8 shows that PaST-Q achieves the lowest Δ-nMAE and the highest Macro-F1 and selective accuracy on both cohorts. These results support the utility of the predicted representations for estimating both the magnitude and direction of laboratory changes. They complement the main future-simulation experiment by directly evaluating the clinical readouts that supply the agent with predictive information for monitoring and explanation.

A.3.3 Extended Agent Benchmark and Inference Cost

The main history and future-simulation experiments vary individual modeling components within a common recommendation workflow. The supplementary benchmark compares complete systems, asking whether the combination of patient-state tools and their orchestration supports treatment selection and task completion at a practical language-model token cost. It complements the controlled comparisons with a broader system-level view.

Results. Table 9 shows that PaST-Q achieves the highest Precision, Treatment F1, and Acc@1 and completes all tasks. Relative to ReAct with the same tools, F1 improves by 11.56 percentage points and Acc@1 by 12.29 points, while completion increases from 97% to 100%. Mean LLM token use decreases from 488.332k to 1.756k, a ratio of approximately $278.1\times$. These findings supplement the main component comparisons with evidence about system execution and language model usage.