Censoring-Aware Reinforcement Learning to Optimize Early Risk Alerts from Longitudinal Clinical Data

Qin Weng WENGQIN.WENG@DUKE.EDU
Department of Biostatistics and Bioinformatics
Duke University
Durham, North Carolina, USA

Benjamin A. Goldstein BEN.GOLDSTEIN@DUKE.EDU
Department of Biostatistics and Bioinformatics
Duke University
Durham, North Carolina, USA

Matthew M. Engelhard M.ENGELHARD@DUKE.EDU
Department of Biostatistics and Bioinformatics
Duke University
Durham, North Carolina, USA

Abstract

Early recognition of chronic conditions is critical to ensure patients receive timely interventions and support. Passive surveillance of routine electronic health records (EHRs) provides information about longitudinal health trajectories that can support prompt recognition and inform associated early actions. However, relevant information is acquired at a different rate for each patient, and there is an inherent trade-off between the earliness versus the specificity of diagnosis and related actions. Therefore, determining when to alert providers about a likely chronic condition requires us to weigh the predicted risk at the given time against the anticipated value of future information. To address this challenge, we analyze the optimal timing of early alerts using a Partially Observable Markov Decision Process (POMDP) with asymmetric reward. To learn an optimal alerting policy, we then propose a model-free reinforcement learning (RL) framework tailored to long-term clinical event surveillance from EHRs. Our proposed framework overcomes the pervasive issue of right-censoring in offline EHRs by leveraging a pseudo-label imputation approach. We also analytically demonstrate that entropy-regularized RL enables post-hoc threshold calibration to adapt the learned policy to specific preferences regarding the importance of earliness versus specificity without retraining. Systematic evaluations in synthetic data reveal that the advantage of RL-based look-ahead planning is maximized when diagnostic evidence emerges in predictable information bursts. Finally, real-world validations on two clinical cohorts show that our policy achieves an actionable lead time of 20.9 months prior to Alzheimer's disease diagnosis and 8.7 months prior to autism diagnosis in a pediatric cohort, both at 90% specificity. Our method and results provide a generalizable blueprint for optimal surveillance of chronic disease processes.

1. Introduction

In clinical practice, the definitive diagnosis of a chronic disease often lags substantially behind the actual onset of its underlying pathology. Promptly identifying the emerging disease

state widens the window for protective interventions and often increases their effectiveness. Electronic Health Records (EHRs) allow long-term, passive surveillance of health status by repeatedly capturing snapshots of patient health. However, translating these longitudinal observations into an actionable risk alert introduces an important trade-off between the earliness versus the accuracy of the alert. Because predictive signals are often weakest in the earliest stages of disease progression, premature alerts carry a high risk of false positives, which can lead to alarm fatigue and erode clinician trust (Pierce et al., 2019). This dilemma is particularly pronounced in chronic conditions with extended prodromal phases, such as Alzheimer's disease (Lukkien et al., 2024) or developmental conditions in childhood (Burgess et al., 2025). For instance, poor eye contact at 18 months can be a sign of autism, but can also reflect temporary shyness. However, waiting until school-age for unambiguous behavioral features eliminates this uncertainty, but at the cost of missing a critical window for neurodevelopmental intervention. Consequently, an effective early alerting system must dynamically navigate this trade-off to maximize alert earliness without unduly compromising specificity.

We conceptualize the decision-making task of whether and when to alert a clinician (or trigger subsequent screening) as a Partially Observable Markov Decision Process (POMDP). This framework mirrors longitudinal EHR data in the preclinical phase, where the health system captures noisy observations of a patient's latent health status through routine encounters. Based on observations from each new visit, the early alerting agent updates its belief and must decide between continuing passive monitoring or triggering a proactive alert. Unlike traditional cost-effective early classification tasks that uniformly penalize the step-wise data acquisition cost (Ji and Carin, 2007), this setting features a fundamental asymmetry in temporal cost, because passive surveillance through EHRs is virtually costless for the healthy population. Consequently, the temporal delay cost falls exclusively on individuals on a disease trajectory, representing an irreversible loss of opportunity to intervene during the preclinical phase, when intervention is often most effective.

Drawing on recent advances in deep reinforcement learning (RL) for early sequence classification (Sharma and Singh, 2019; Martinez et al., 2020), we propose a model-free RL framework to directly solve this early alerting POMDP. By utilizing an end-to-end architecture optimized via reward maximization, our approach directly maps longitudinal patient trajectories to optimal discrete actions at each step, namely to (a) continue surveillance, or (b) trigger an alert. Crucially, this model-free paradigm bypasses the computational bottlenecks associated with explicit uncertainty quantification and belief modeling in high-dimensional, long-sequence EHR spaces (Li et al., 2021). Instead, it allows the policy network to implicitly encode complex uncertainty dynamics and anticipate the value of future clinical information directly inferred from raw sequential observations. Ultimately, this yields a context-aware early alerting policy that dynamically adapts its alerting decision based on the patient's unfolding risk profile and uses look-ahead planning to achieve an optimal trade-off across the population.

However, this framework is hindered by the incomplete nature of longitudinal EHR data, where patients frequently exit the monitoring window before a target event is confirmed, resulting in a substantial number of right-censored trajectories. Naively treating unconfirmed trajectories as negatives distorts the reward signal, resulting in an overly conservative policy as the agent overcompensates for the artificially deflated specificity caused

by right-censoring. To address this problem and provide censoring-aware early alerting policies, we propose a novel reward imputation approach. This approach integrates pseudo-labels derived from an auxiliary deep mixture cure model to recover reward signals for right-censored trajectories, while preserving definitive labels for those with observed outcomes. This extension effectively enables the early alerting RL agent to safely leverage the full offline dataset without discarding valuable partial trajectories while mitigating censoring bias. Furthermore, we notice that the clinical priorities for early detection versus controlling FPR vary across different healthcare institutions and resource settings. However, standard RL policies usually encode a single, fixed trade-off preference during training, which requires computationally expensive retraining to accommodate each new preference shift. To overcome this limitation, we introduce a post-calibrated preference adaptation workflow that is grounded by the theoretical properties from entropy-regularized Proximal Policy Optimization (PPO) learning. We mathematically and empirically demonstrate that post-hoc threshold shifting on the predicted logits serves as a principled, first-order approximation for preference interpolation. This approach empowers clinicians to tune the system's sensitivity across diverse clinical settings without preference-specific retraining.

To rigorously evaluate our proposed RL framework, we conduct systematic experiments across both synthetic data and real-world clinical data. Through diverse simulated scenarios, we investigate the specific clinical conditions under which RL-based look-ahead planning outperforms myopic supervised baselines. The simulated trajectories mimic varied information dynamics that might be observed in clinical practice, including gradual evidence accumulation of diagnostic evidence versus predictable information bursts from diagnostic testing. Furthermore, we demonstrate the robustness of our censoring-aware extension across a spectrum of feature-dependent time-to-event dynamics. Finally, we validate the practical utility of our framework on two real-world cohorts for early alerting of Alzheimer's Disease (AD) and Autism Spectrum Disorder (ASD), respectively, establishing its efficacy in delivering actionable and highly specific early alerts across distinct clinical settings.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Balancing earliness against specificity for early risk alerting is a dynamic, patient-specific trade-off. Model-free RL provides a principled framework to navigate this by actively anticipating the value of future diagnostic information. Crucially, this look-ahead planning delivers the greatest clinical benefit when diagnostic evidence arrives at predictable key time points (e.g., scheduled assessments).
- Right-censoring not only biases the learned event risk, but also artificially deflates the perceived disease prevalence, forcing models into an overly conservative, delayed-alerting posture. Adjusting for this bias is mandatory to learn a policy that faithfully reflects actual clinician preferences for earliness versus specificity.

2. Related Work

2.1. Early Sequence Classification

Our problem formulation of early risk alerting falls under the broader paradigm of Early Sequence Classification (ESC), where the fundamental objective is to assign a definitive

label to a streaming sequence as early as possible while navigating the trade-off between accuracy and actionable lead time. Traditional ESC methods generally rely on supervised classifiers equipped with early-trigger mechanisms, such as static posterior thresholds (Xing et al., 2010; Mori et al., 2017). While these paradigms offer tractable early-trigger mechanisms, they share critical structural limitations that inherently conflate irreducible event stochasticity with reducible observation uncertainty. Consequently, they fail to do look-ahead planning for the future value of information (VoI).

More recent approaches introduce dynamic halting networks to overcome rigid thresholding. Hartvigsen et al. (2019) design recurrent gating modules to learn adaptive stopping times, and Martinez et al. (2020) explicitly train an auxiliary RL agent alongside the classifier to evaluate whether to request the next observation. However, their halting mechanisms are optimized via time-penalty objectives that are strictly symmetric for positive or negative cases, failing to capture the asymmetric reality of the temporal cost in EHRs surveillance for the healthy population versus patients with diseases. In addition, these RL-based agents require fully observed ground-truth labels for optimization without accounting for right-censored outcomes. In contrast, our framework safely leverages incomplete data to learn a context-aware, globally optimal alerting policy for this asymmetric clinical utility.

2.2. Reinforcement Learning for Clinical Decision Making

Recent works have adapted RL to right-censored outcomes by integrating survival analysis. For instance, some design offline RL frameworks to maximize expected patient survival time (Cho et al., 2023), while others utilize survival models to impute rewards for right-censored trajectories during continuous treatment planning (Lyu et al., 2023). Yet, these censoring-aware RL frameworks are fundamentally tailored to optimize continuous time-to-event outcomes within the Dynamic Treatment Regimes (DTR) paradigm. They typically operate under standard survival assumptions rather than adopting the mixture cure framework, which is important to account for the fraction of the population that will never experience the event. Our work therefore addresses a distinct gap, learning a censoring-robust, preference-conditioned early alert policy for a definitive classification outcome. By integrating an auxiliary deep mixture cure model, we provide a principled reward imputation mechanism that corrects the inherent bias of right-censored trajectories, effectively bridging the gap between censoring-aware RL and early risk alerting.

2.3. Multi-Objective Reinforcement Learning

Multi-Objective Reinforcement Learning (MORL) optimizes a vector-valued reward whose scalarization weights encode user trade-off preferences (Hayes et al., 2022). Existing methods typically fall into two families: multi-policy approaches that train a population of policies to approximate the Pareto front (Xu et al., 2020), and single-policy approaches that condition one universal network on the continuous preference space (Yang et al., 2019; Basaklar et al., 2022). Although the latter is more parameter-efficient, it suffers from optimization challenges caused by destructive gradient interference between conflicting objectives (Liu et al., 2025). We expect this difficulty to be compounded in the clinical setting, where irregularly sampled and partially observed EHR trajectories already yield high-variance policy-gradient estimates. Our workflow therefore adopts a multi-policy strategy, training

a sparse grid of independent anchor policies and then interpolating between anchors at inference time via a theoretically grounded logit shift.

3. POMDP Formulation

We formalize the early alerting task in sequence monitoring as a POMDP, defined by the tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R})$. Each patient i is monitored through a sequence of visits $t = 1, \ldots, T$, yielding a stream of longitudinal clinical snapshots. The objective is to learn an optimal policy π^* that dynamically navigates the intertemporal trade-off between earliness and specificity. The components of our POMDP formulation are detailed below.

3.1. Latent States and Observations

We define the patient's underlying condition as a tuple $s_t = (\mathbf{f}_t, y)$. Here, $\mathbf{f}_t \in \mathcal{F}$ denotes a set of static or evolving latent physiological features at visit t, which collectively govern the patient's intrinsic risk profile $p \in [0, 1]$. The variable $y \in \{0, 1\}$ denotes the onset of a target clinical event along a patient's trajectory, viewed as a stochastic realization of the risk as $y \sim \text{Bernoulli}(p)$. The observations $\mathbf{o}_t \in \mathcal{O}$ are emissions of latent features $\mathbf{f}_t$ at each visit. While we acknowledge that observations might eventually become perfectly predictive of y at the symptomatic stage, the objective of early risk alerting is to identify high-risk patients prior to that. Therefore, we assume that within our tactical observation window, $\mathbf{o}_t$ is only driven by the underlying risk-defining features $\mathbf{f}_t$, making it conditionally independent of the yet-to-be-confirmed outcome y. We emphasize that this formulation explicitly models a hierarchical structure of uncertainty: the uncertainty about the patient state given our observations is reducible by accumulating additional longitudinal evidence, whereas the stochastic nature of the clinical event y represents irreducible aleatoric uncertainty.

3.2. Actions and Transitions

We formulate the action and transition dynamics as a one-sided stopping problem. At each step, the action space is defined as a binary decision $a_t \in \{0, 1\}$, where $a_t = 0$ represents a decision to defer alerting and continue monitoring, whereas $a_t = 1$ represents an irreversible decision to alert and terminate the sequence. The episode thus unfolds until an early alert is triggered or the maximum observation window T is exhausted. We define the termination step as $\tau = \min\{t \mid a_t = 1 \text{ or } t = T\}$. Consequently, the terminal action a_τ serves as the final clinical classification for the entire episode.

3.3. Preference-Aware Objective

Our framework is characterized by a fundamental asymmetry in its optimization objectives. For positive trajectories ($y = 1$), the clinical utility is modeled as a time-decayed positive reward to incentivize early alerts; for negative trajectories ($y = 0$), the utility incurs a time-invariant false-positive penalty. We decompose the reward into a vector $\mathbf{r}_t(y, a_t) = [r_t^+, r_t^-]^\top$, and scalarize these competing objectives through a preference weight vector $\mathbf{w} = [w_+, w_-]^\top$, where both components are strictly positive.

Because the optimal policy is invariant to the absolute scale of the rewards, we normalize $\mathbf{w}$ into a single preference ratio $\kappa = \frac{w_+}{w_-}$. This parameter κ explicitly quantifies the relative

clinical priority of early capturing a true positive against avoiding a false alarm. Accordingly, the agent's objective is to maximize the expected κ-conditioned utility. The optimal policy π^*_κ is thus defined by:

$$\pi^*_\kappa = \arg\max_\pi \mathbb{E}_{a_t \sim \pi(\cdot|\mathbf{o}_{:t})} \left[\sum_{t=1}^{\tau} \left(\kappa r_t^+(y, a_t) \right) + r_\tau^-(y, a_t) \right] \tag{1}$$

We omit the standard discount factor in RL to ensure the critical sparse classification rewards at step τ are correctly propagated back to earlier steps without exponential decay. Furthermore, we explicitly define the reward function to depend solely on the underlying clinical outcome y and the action a_t. The details of each reward component, including techniques to address label ambiguity from right-censoring, are presented in the next section.

4. Censoring-Aware Reward Formulation

Translating the conceptual optimization objective into a computable learning signal requires addressing two practical challenges: (1) formalizing the clinical utility into precise reward functions that correctly encode the trade-off between earliness and specificity, and (2) mitigating label ambiguity caused by right-censoring. This section details the mathematical formulation of these components and introduces a robust imputation strategy designed to calibrate policy learning when training on offline datasets with right-censored patient trajectories.

4.1. Reward Function for Fully Observed Trajectories

Given the preference-aware objective defined in Sec. 3.3, we now specify the mathematical formulation of the reward components r^+ and r^-. While real-world EHR encounters are inherently irregular, we standardize patient trajectories onto a unified discrete temporal grid, and use binary missingness indicators and last observation carried forward (LOCF) imputation to address the temporal sparsity. Assuming the ground true clinical outcome y is fully observed in the training data, the reward vector $\mathbf{r}_t = [r_t^+, r_t^-]$ is defined as:

$$\begin{cases} r_t^+(y, a_t) = y \cdot [a_t R_{tp} - (1 - a_t)\delta_t], \\ r_t^-(y, a_t) = -(1 - y) a_t R_{fp}, \end{cases} \tag{2}$$

where R_{tp}, R_{fp} and δ_t are strictly positive scalars. Here, the positive utility r_t^+ incentivizes the agent to identify positive trajectories early through a terminal anchor positive reward R_{tp} paired with a dense deferral penalty δ_t at each step t, which creates a persistent gradient to maximize proactive lead time. Here we constrain the anchor positive reward to always be larger than the total cumulative delay penalty as $R_{tp} \gg \sum_{t=1}^{T} \delta_t$, ensuring that even a relatively late alert is still preferable to a total miss. Conversely, the negative cost r_t^- applies a sparse terminal penalty R_{fp} solely for false alerts on negative trajectories. Under this asymmetric design, no penalty is incurred for waiting on negative trajectories, as continuous passive monitoring remains the optimal clinical action for healthy individuals.

By aggregating all these reward components, the trajectory-wise cumulative return can be expanded as:

$$R_\kappa(y, \mathbf{a}) = a_\tau \left[\kappa y R_{tp} - (1-y) R_{fp}\right] - \kappa y \sum_{t=1}^{\tau} (1 - a_t)\delta_t. \tag{3}$$

This formulation further decomposes the agent's objective into two terms, which we may define intuitively as (a) a static classification utility determined based on a_τ, and (b) an asymmetric time cost that applies only to positive trajectories.

4.2. Reward Imputation for Right-Censored Trajectories

In real-world studies, patient trajectories are frequently right-censored due to limited follow-up times. In such cases, it is not clear what reward should be given, because we do not know whether the patient may have been diagnosed after the censoring time. Naively conflating censored trajectories with true negative trajectories introduces severe calibration issues by underestimating the event prevalence, while simultaneously biasing the learned risk profile toward features associated with earlier event times. To enable policy learning on retrospective cohorts with incomplete trajectories, we substitute the unobserved y with an imputed pseudo-label $\hat{y}$ that quantifies the expectation of being on a positive trajectory given the entire observed clinical history.

Specifically, we train an auxiliary Mixture Cure Model on the full trajectories to model both the total probability of event occurrence and the time-to-event distribution. The event probability output from this auxiliary model, defined as $\hat{p} = \Pr(y = 1 \mid \mathbf{o}_{:T})$, serves as a point estimate of the marginal event likelihood that is informed by all available observations. Because the cumulative return $R_\kappa(y, \mathbf{a})$ is strictly linear with respect to y, substituting y with a conditional expectation allows the imputed reward to serve as a surrogate for the expected return.

Theoretically, obtaining an unbiased posterior risk in a right-censored patient requires conditioning on the estimated survival distribution. However, estimating survival tails in long-term trajectories introduces significant variance and estimation instability. Therefore, we use the predicted event probability $\hat{p}_i$ as a robust prior estimate for imputation. To further maximize training stability and ground the imputed reward to available clinical truths, we construct the pseudo-label $\hat{y}$ as follows:

$$\hat{y}_i = \nu_i + (1 - \nu_i)\hat{p}_i, \tag{4}$$

where $\nu_i \in \{0, 1\}$ is the event indicator denoting whether the target clinical event is explicitly observed. Acknowledging that right-censored patients may be subsequently diagnosed leads the agent toward a more aggressive alerting policy, since censored patients receive positive reward proportional to $\hat{p}_i$ upon alerting, and the associated false positive penalty is attenuated by $1 - \hat{p}_i$.

5. Preference Adaptation via Decision Margin Analysis

Standard model-free RL policies are bound to a fixed reward structure, therefore adapting a policy to the clinical preferences of each new patient or setting requires retraining, which becomes computationally prohibitive. To overcome this, we introduce a theoretical framework for few-shot preference adaptation. We first formalize the optimal decision margin

to isolate the trade-off between the value of waiting and delay penalties in Sec. 5.1, then prove its strict monotonicity with respect to the preference ratio in Sec. 5.2. Leveraging this monotonicity alongside the analytical properties of entropy-regularized RL, we propose a post-hoc threshold interpolation workflow in Sec. 5.3. This workflow allows our system to adapt to arbitrary clinical preferences at inference time without altering the underlying network weights.

5.1. Optimal Alerting Margin

To establish the theoretical foundation for preference adaptation, we first formalize the agent's decision margin under our asymmetric reward structure. Define the belief state $b_t = \Pr(s_t \mid \mathbf{o}_{:t})$ as the agent's posterior beliefs over the latent state $s_t = (\mathbf{f}_t, y)$. From this, we extract the scalar marginal risk estimate $\hat{p}_t = \mathbb{E}_{s\sim b_t}[y]$, representing the posterior probability of the target outcome. By the Law of Iterated Expectations, the updating process of both the belief state and its scalar projection over sequential observations forms a martingale, satisfying $\mathbb{E}_{b_{t+1}|b_t}[b_{t+1}] = b_t$ and $\mathbb{E}_{b_{t+1}|b_t}[\hat{p}_{t+1}] = \hat{p}_t$.

While the future transitions are governed by the belief state b_t, the expected immediate payoffs for alerting depend only on the scalar risk $\hat{p}_t$. Therefore, we define the preference-scaled action-value functions conditioned on the belief state b_t as:

$$Q_t(a_t = 1, b_t) = \kappa \hat{p}_t R_{tp} - (1 - \hat{p}_t) R_{fp}, \tag{5}$$

$$Q_t(a_t = 0, b_t) = -\kappa \hat{p}_t \delta_t + \mathbb{E}_{b_{t+1}|b_t}[V_{t+1}(b_{t+1})], \tag{6}$$

where $V_t(b_t) = \max_a Q_t(a, b_t)$ represents the optimal expected return.

The agent's decision rule is then dictated by the net advantage function $\Delta Q_t(b_t, \kappa) \triangleq Q_t(a_t = 1, b_t) - Q_t(a_t = 0, b_t)$. Under the optimal policy, the agent crosses the margin and triggers an alert if and only if $\Delta Q_t \geq 0$, which translates to the following inequality:

$$\kappa \hat{p}_t \delta_t \geq \mathbb{E}_{b_{t+1}|b_t}[V_{t+1}(b_{t+1})] - \big(\kappa \hat{p}_t R_{tp} - (1 - \hat{p}_t) R_{fp}\big). \tag{7}$$

This inequality provides an intuitive decomposition of the sequential trade-off: the left-hand side represents the expected step-wise delay cost of choosing to wait, while the right-hand side represents the VoI, quantifying the marginal gain of deferring the decision to accumulate further evidence over acting immediately.

Because the optimal value function $V(\cdot)$ is the maximum of linear functions, it is convex over the belief space. According to Jensen's Inequality and the martingale property, the expected future value is bounded as:

$$\mathbb{E}_{b_{t+1}|b_t}[V_{t+1}(b_{t+1})] \geq V_{t+1}(\mathbb{E}_{b_{t+1}|b_t}[b_{t+1}]) = V_{t+1}(b_t). \tag{8}$$

Since $V_{t+1}(b_t)$ is trivially bounded below by the immediate alerting payoff at step t, given that the optimal policy chose not to terminate the sequence early at step t, the VoI on the right-hand side of Eq. 7 is strictly non-negative. This implies that without the explicit delay penalty ($\delta_t = 0$), the value of deferring a decision to gather more evidence would always equal or exceed the value of acting immediately, causing the agent to wait indefinitely until the window exhausts. Thus, the analysis confirms the intuition that without the temporal cost $\kappa \hat{p}_t \delta_t$, early alerts would never occur.

5.2. Monotonicity of the Preference-Conditioned Margin

To understand how the preference ratio κ calibrates the decision margin in Eq. 7, we analyze the monotonicity of the net advantage function $\Delta Q_t(b_t, \kappa)$ with respect to κ. An increase in κ indicates a shift in priority toward capturing true positives as early as possible, accompanied by a higher tolerance for the clinical burden of false alarms. Intuitively, this asymmetric shift should systematically incentivize the agent to adopt a more proactive alerting policy. We formalize this clinical intuition into a rigorous mathematical guarantee by proving that the net advantage of alerting is strictly monotonically increasing with respect to the preference ratio κ.

Proposition 1 (Strict Monotonicity of Advantage) *Given a valid posterior risk estimate $\hat{p}_t > 0$ and a strictly positive delay penalty $\delta_t > 0$, the net advantage of alerting over waiting is strictly monotonically increasing with respect to κ, bounded from below by the expected temporal cost:*

$$\frac{\partial \Delta Q_t(b_t, \kappa)}{\partial \kappa} \geq \hat{p}_t \delta_t > 0. \tag{9}$$

We establish this bound by differentiating the optimal value function and propagating the derivative backward from the terminal step T via mathematical induction, leveraging the martingale property of the belief updating process. The proof is found in Appendix A.

5.3. Post-Hoc Threshold Interpolation

The decision margin in standard model-free RL is not an explicit parameter but is implicitly embedded in the state-action mapping, making it difficult to directly target this margin for recalibration. Instead, we leverage an established analytical structural properties of entropy-regularized RL (Haarnoja et al., 2018), namely that the optimal policy naturally converges to an energy-based Boltzmann distribution: $\pi^*(a \mid s) \propto \exp(Q(s,a)/\alpha)$, where the entropy coefficient α acts as a temperature parameter. Consequently, the logit difference z_t produced by the policy network is proportional to the net advantage:

$$z_t = \log \pi(a_t = 1) - \log \pi(a_t = 0) = \frac{1}{\alpha} \Delta Q_t(b_t, \kappa). \tag{10}$$

The natural decision boundary sits where it is indifferent between the Q-value of waiting or alerting. According to Eq. 10, the boundary condition corresponds to a zero logit difference ($z_t = 0$) and a zero net advantage ($\Delta Q_t(b_t, \kappa) = 0$). Since ΔQ_t is strictly monotonic in κ (Proposition 1), we can apply a first-order Taylor expansion directly to the indifference point. Within a local neighborhood of a trained anchor preference κ_0, a slight preference shift $\Delta\kappa$ induces an approximately linear shift in the net advantage, which analytically translates to a linear scalar shift in the learned action logit z_t. This shows that applying a post-hoc threshold shift to the predicted logits acts as a consistent first-order approximation for preference adaptation. Building on this analytical property, we propose a few-shot, preference-conditioned deployment workflow as follows:

1. **Offline Anchor Training:** Train a set of anchor policies via entropy-regularized RL across a discrete sparse grid of κ values to capture global, non-linear behavioral shifts.

2. **Calibration Mapping:** On a validation set, we empirically map continuous target specificities to optimal (Anchor Policy, Logit Threshold Shift) pairs.

3. **Zero-Shot Deployment:** To align with a specific clinical preference at inference time, we select the nearest anchor policy and apply the pre-calculated threshold shift to the logit output, enabling continuous adaptation without altering network weights.

6. Simulations

We design simulation experiments to address two pivotal questions: (1) under what clinical information dynamics does the look-ahead planning in RL provide the most significant advantage, and (2) how effectively does our imputation mitigate right-censoring biases?

To investigate these questions, we construct a data-generating process (DGP) that simulates longitudinal patient trajectories governed by underlying latent physiological states. For each synthetic patient, this latent state dictates both their ground-truth event risk and their specific time-to-event, while emitting noisy clinical observations across a discrete monitoring horizon. We systematically manipulate two core components of the DGP. We control the *information dynamics*, where we modulate the temporal distribution of observation noise to emulate clinical scenarios ranging from (a) gradual accumulation of evidence over time, to (b) predictable "information bursts", wherein substantial clinical evidence is obtained at specific times that can be inferred from earlier observations. We then focus on time-to-event dependencies, where we adjust the signal-to-noise ratio (SNR) of the event times in positive trajectories to test our censoring-aware agent's robustness under varying levels of survival predictability in the presence of right-censoring. Full mathematical formulations of the DGP are deferred to Appendix B.

We train our early alerting RL agent using Recurrent PPO and compare it against three supervised baselines: a static MLP, an RNN-GRU, and a censoring-aware mixture cure model. This mixture cure model also serves as the auxiliary model for pseudo-label imputation in our censoring-aware RL variants. To ensure a fair comparison between the single-output baselines and our multi-preference RL policy sets, we apply the same post-hoc calibration workflow to all methods on a validation set; implementation details are in Appendix C. We then construct the Pareto Fronts of the average lead time (ALT) versus FPR, with overall performance quantified as the corresponding Hypervolume (HV). To balance computational efficiency and interpolation quality, all RL evaluations in Section 6 and 7 utilize 5 anchor policies (see Appendix D.1 for ablation), as shown in Figure 1a. All simulations maintain event and censoring rates at 0.5, with results averaged over five random seeds.

6.1. Decision-Making under Varied Information Dynamics

To evaluate the agent's capacity and advantage for strategic look-ahead planning, we design three distinct observation dynamics that simulate diverse clinical monitoring regimes as outlined below.

- **S1: Uniform Information.** We first establish a scenario where the noise level of each observation remains identical, and the marginal information gain diminishes

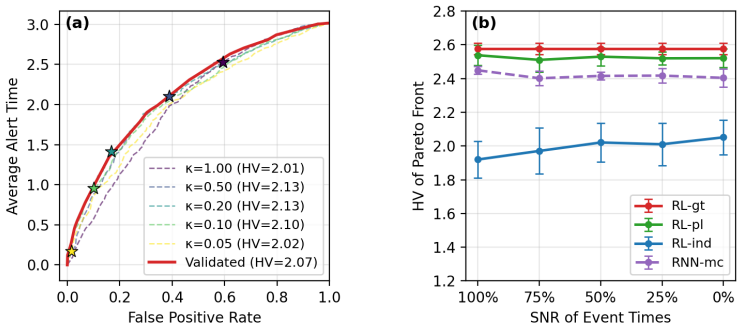

Figure 1: (a) Illustration of how the Pareto front is constructed from a policy set via calibration and selection, where each star represents an individual policy trained under a specific preference κ with naive decision threshold 0.5. (b) Ablation study of pseudo-label imputation under right-censoring. The curves show the average Pareto Front HV under controlled signal-to-noise ratio (SNR) in generating event times, varying from 100% (fully deterministic) to 0% (fully stochastic). The lines compare our proposed pseudo-label imputation (RL-pl) against the ground-truth oracle (RL-gt), the naive indicator approach (RL-ind), and the supervised mixture cure model baseline (RNN-mc).

smoothly as the agent's belief regarding the patient's latent state saturates. Specifically, we maintain a constant noise variance across all timesteps for all trajectories in the DGP. This scenario simulates routine clinical monitoring where diagnostic evidence is collected at a steady rate with consistent measurement precision.

- **S2: Information Bursts.** We then simulate a scenario where highly informative observations come at a cohort-wide key time point. This mirrors clinical regimes where key evidence is observed after a specific physiological stage or through standardized screening protocols. We simulate this by maintaining observation noise at a high-variance plateau during in early stages, then sharply reducing it as the trajectory approaches the key time point, after which the noise is small and constant.

- **S3: Scheduled Assessment.** Finally, we simulate a scenario where key evidence arrives at a known, patient-specific time, mimicking scheduled imaging studies or lab tests. To model this, we introduce an auxiliary binary feature that substantially affects the risk function. This feature remains masked until a patient-specific timestep to reflect heterogeneous clinical timelines across the cohort. We augment the agent's observation space with an additional feature indicating the time to unmasking to reflect the predictability of clinical schedules.

To evaluate the models' architectures without the interference of right-censoring, we first trained all models using uncensored ground-truth labels. Under the Uniform Information scenario (S1) in Figure 2a, both RL and RNN achieve comparable HV and significantly outperform the static MLP. This is because the MLP without memory is vulnerable to

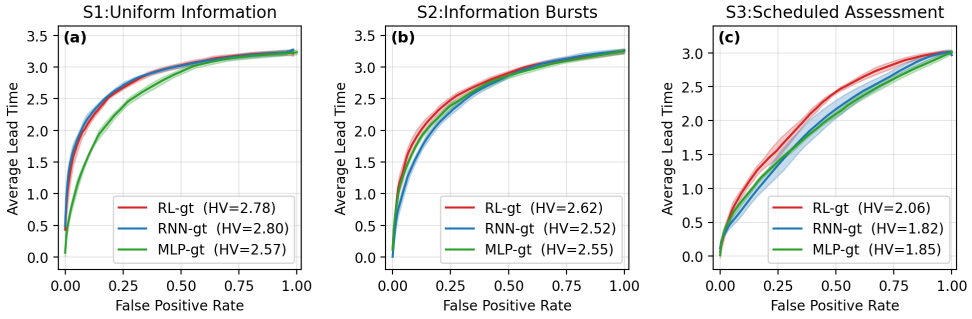

Figure 2: Pareto front analysis across varied information dynamics in simulation experiments. Panels (a)-(c) compare RL and baselines' performance under different scenarios: (a) Uniform Information, where measurement precision remains constant across all timesteps; (b) Information Burst, where observation noise undergoes a sharp reduction at a fixed, cohort-wide key time point; and (c) Scheduled Assessment, where key predictive features are resolved at patient-specific but predictable future time points. Solid lines and shaded areas represent the mean and standard deviation across 5 random seeds.

noise at early time points, whereas the recurrent nature of the RL and RNN algorithms allows them to benefit from the accumulation of information over time. The result shows that when information accumulates steadily and marginal VoI strictly diminishes, look-ahead planning yields no additional benefit. The best strategy is simply to react to the current risk posterior prediction, which both RL and RNN do well.

The inherent limitations of myopic supervised classifiers emerge in scenarios characterized by sudden information gains in Predictable Information Bursts (S2) and its more extreme counterpart, Scheduled Assessment (S3), as shown in Figure 2b and Figure 2c. In these scenarios, the RNN's hidden state accumulates early neutral beliefs, making it slow to react when predictive signal suddenly emerges, which in turn degrades performance in high-specificity regions where decision thresholds are set at a high level. Conversely, the MLP's lack of history allows it react more effectively, but it is vulnerable to false alarms from early noise unless its threshold is set restrictively high.

Our RL framework overcomes these limitations by balancing patience and responsiveness via looking-ahead planning. In S2, the agent learns to implicitly capture the VoI within its value network. By recognizing the temporal patterns for noise dissipating at predictable timesteps, it strategically defers decisions in the noisy phase, then immediately shifts to zero-delay alerting once the signal becomes clear. This look-ahead capability is even more beneficial in the Scheduled Assessment scenario (S3). The presence of an "information countdown" feature allows the agent to accurately estimate VoI, leading it to raise the decision threshold and thereby delay alerts for patients who will soon be assessed. This provides a robust advantage over our supervised classifiers across all specificity regions.

6.2. Robustness under Right-Censoring

Building upon the S2 scenario in the previous section, we further investigate the framework's robustness against censoring bias. We systematically vary the ratio between the deterministic component and the stochastic noise in generated event times, while fixing the total variance. This allows us to control the feature dependency of event times, thereby modulating how strongly the right-censoring mechanism for positive cases correlates with latent features. To demonstrate the contribution of our proposed censoring-aware approach, we perform an ablation analysis comparing the pseudo-label imputation variant (RL-pl) against a ground-truth oracle trained on uncensored data (RL-gt) and a naive agent that treats all right-censored trajectories as negative (RL-ind). Additionally, we include a GRU-based Mixture Cure model (RNN-mc) as a censoring-aware supervised learning baseline.

Figure 1b shows the HV value across a spectrum of event-times feature dependency, ranging from fully feature-determined (100%) to entirely stochastic (0%). The Oracle agent RL-gt establishes a stable, optimal performance ceiling irrespective of the survival dynamics. In contrast, the naive agent RL-ind shows severely degraded performance, because treating right-censored trajectories as negatives leads to a misidentification of features associated with late events as features indicating low risk. As event times become more stochastic, this specific correlation bias weakens, allowing RL-ind to recover somewhat. However, a massive performance gap remains even at zero dependency. This highlights the fundamental limitation of RL-ind, namely that it underestimates the event prevalence. As a result, the agent's value estimation becomes systematically miscalibrated, distorting its trade-off optimization logic regardless of the underlying time-to-event mechanisms.

Our proposed pseudo-label reward imputation approach successfully mitigates this degradation, with RL-pl recovering near-oracle performance across all settings. It performs best under high feature dependency, where the auxiliary model can most precisely map latent features to true event probabilities. Furthermore, sensitivity analysis (Appendix D.2) demonstrates that RL-pl remains highly robust to minor errors in the auxiliary model, although severe misspecification will eventually degrade its performance. Crucially, while both RL-pl and RNN-mc utilize the same Mixture Cure loss for censoring awareness, RL-pl consistently outperforms RNN-mc across the entire survival dependence spectrum. This confirms that our imputation strategy mitigates the bias from censoring, yet preserves the RL framework's capacity for look-ahead planning.

7. Early Alerting in Real-World Clinical Cohorts

We validate the clinical utility of our proposed RL framework by applying it to early alerting for Alzheimer's Disease (AD) and Autism Spectrum Disorder (ASD), respectively. Both tasks encapsulate the challenges of preclinical surveillance, as characterized by regular sequential clinical encounters, delayed diagnostic unmasking, and prevalent right-censoring. The AD early alerting task uses the public National Alzheimer's Coordinating Center (NACC) dataset with 8,803 patients and an incident AD prevalence of 14.2%. Clinical follow-ups are discretized into 12-month intervals over a 60-month window. At each step, the agent observes 20 core clinical features, and Clinical Dementia Rating (CDR) subscores are excluded to prevent target leakage. The ASD early alerting task uses a proprietary pediatric cohort of 29,357 children with a rarer event rate of 3.2%. Encounters are aligned

to the recommended well-child visit schedule over 15 decision steps, which is dense through infancy and sparse thereafter. Each well-child visit is represented by a 768-dimensional encounter-level embedding of its ICD-10 diagnosis codes and demographic data obtained from a pretrained EHR foundation model (Sun et al., 2026). Finally, to more rigorously evaluate our censoring-aware method's robustness while overcoming the inherent right-censoring in both registries, we employ two specialized strategies: (1) administrative early-truncation, which creates a temporal gap between training histories and extended evaluation trajectories; and (2) an age-filtered subcohort (AD: age ≥ 85; ASD: age ≥ 8), which serves as a highly reliable ground-truth surrogate. Detailed procedures regarding cohort construction, time-grid mapping, early truncation, data imputation, and validation cohort design are provided in Appendix E.

Table 1: Average Hypervolume (HV) of the ALT vs. FPR for AD and ASD early alerting. Results are reported as Mean $\pm$ Std over 5 random seeds.

	Complete Outcome			Truncated Outcome				
Cohort (Test Set)	**RL-gt**	**RNN-gt**	**MLP-gt**	**RL-pl**	**RNN-mc**	**RL-ind**	**RNN-ind**	**MLP-ind**
AD (Full Set)	$\mathbf{38.36 \pm 0.51}$	38.25 ± 0.63	37.89 ± 0.62	36.78 ± 0.70	36.79 ± 0.80	$\mathbf{37.25 \pm 0.79}$	36.99 ± 0.74	37.19 ± 0.74
AD (Age-filtered)	$\mathbf{35.26 \pm 1.18}$	35.09 ± 1.65	33.32 ± 1.71	$\mathbf{35.16 \pm 1.19}$	34.50 ± 1.09	34.83 ± 0.89	31.61 ± 0.99	33.00 ± 1.91
ASD (Full Set)	$\mathbf{29.52 \pm 0.24}$	28.56 ± 0.36	28.08 ± 0.48	$\mathbf{29.09 \pm 0.48}$	26.28 ± 0.72	28.68 ± 0.24	27.84 ± 0.48	27.84 ± 0.36
ASD (Age-filtered)	$\mathbf{37.80 \pm 0.24}$	35.40 ± 0.72	35.76 ± 0.72	$\mathbf{37.08 \pm 0.36}$	32.64 ± 1.32	36.60 ± 0.84	35.28 ± 0.72	34.44 ± 1.20

The HV results for both early alerting tasks are summarized in Table 1. Under the empirical oracle setting where models are all trained on complete outcome labels, our RL framework consistently achieves the highest performance across both cohorts. When investigating the robustness under right-censoring bias, RL-pl's margin over naive RL-ind appears modest or even slightly negative on the full test set. This is largely an artifact of evaluation bias because the empirical "ground truth" labels in the full set remain heavily right-censored. Consequently, censoring-aware models that correctly predict impending yet unobserved events are wrongly penalized as false positives. This evaluation gap is effectively resolved in the age-filtered test set, which provides a reliable ground-truth surrogate with substantially completed clinical trajectories. Here, RL-pl outperforms both the naive RL-ind agent and its own auxiliary imputation baseline RNN-mc in both cohorts. We also notice that the advantage over supervised baselines is more pronounced in ASD than in NACC. We conjecture that this is because milestone-anchored well-child visits deliver development information in predictable bursts, where look-ahead planning gains most over myopic classifiers.

To operationalize our framework for real-world deployment, we analyze how the abstract objective preference parameter κ aligns with interpretable clinical metrics. For this analysis, we report results from the full test set because its larger sample size provides greater statistical power, and its inherent censoring bias yields a more conservative estimate of the performance gains of RL-pl over RL-ind. Figure 3 and Appendix Figure 5 illustrate the FPR and ALT across varying preference levels after post-hoc calibration to specific preferences. Importantly, under a strict constraint of 90% specificity, our censoring-aware RL-pl agent achieves an ALT of 20.85 months prior to formal AD diagnosis and 8.71 months prior to ASD diagnosis, approaching the empirical oracle RL-gt (AD: 22.42 months; ASD: 8.93

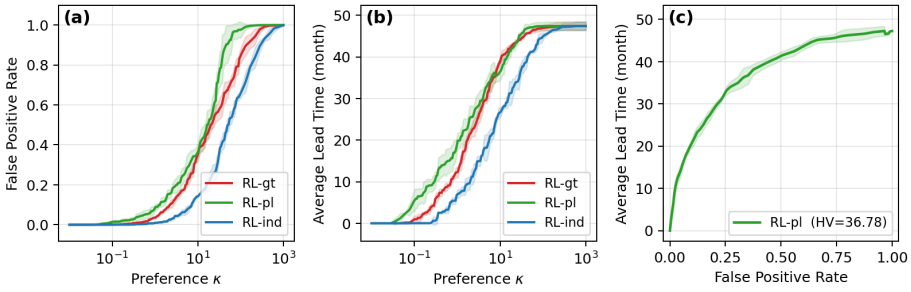

Figure 3: Preference sensitivity and Pareto front tested on the full AD cohort. Panels (a)-(b) illustrate the variation of FPR and ALT calibrated to the preference ratio κ spectrum. Panel (c) shows the resulting Pareto front from the optimal RL-gt agent, achieving an ALT of 20.85 months with 90% specificity.

months) despite being trained on truncated trajectories. We also find that behavior significantly diverges between censoring-aware RL-pl and naive RL-ind agents in both cohorts. Because RL-ind fails to account for right-censoring during calibration, it underestimates event prevalence and learns an overly conservative policy, delaying alerts unless given a high false-positive tolerance. In contrast, RL-pl closely approximates the empirical oracle RL-gt dynamics across the preference spectrum. This confirms that our strategy ensures faithful internal risk calibration by preventing bias from right-censoring from influencing the learned policies.

8. Discussion

In this work, we present an RL framework to address the inherent trade-off between the earliness versus the specificity of early alerts for clinical events. This work is motivated by the need to promptly identify patients at elevated risk of chronic conditions based on passive surveillance of routine EHRs, thereby ensuring they receive timely support and intervention, in turn leading to improved outcomes. Our approach allows us to reason effectively from a patient's risk profile based on longitudinal health trajectories updated after each encounter, and to adapt to specific health system preferences regarding early recognition versus false-positive reduction to decide whether to trigger an alert or continue monitoring. Validated on extensive simulations and real-world clinical data, our framework demonstrates that look-ahead planning that integrates the anticipated value of future information outperforms myopic event classification.

A central contribution of our work is elucidating the intrinsic mechanisms that drive the superiority of RL-based look-ahead planning. Through theoretical analysis, we find that the agent learns an implicit decision margin that only decides to issue immediate alerts when the expected step-wise delay cost outweighs the expected marginal VoI gain associated with waiting for more evidence; which allows the agent to strategically defer alerts when early signals are ambiguous. Crucially, our simulations demonstrate that the advantage of this look-ahead behavior is maximized under two specific conditions: (1) when

diagnostic evidence emerges in distinct "information bursts", and (2) when the timing of these bursts can be predicted from available features. In real-world EHRs, relevant information is acquired at a different rate for each patient, meaning the accumulation of clinical evidence is rarely smooth and does tend to occur in bursts. We also find that look-ahead planning becomes more essential as information accumulation becomes more irregular, and the timing of the "information bursts" becomes more patient-specific. By anticipating the value of future information, the agent avoids premature false alarms while successfully capturing the patient's true pathological trajectory.

Furthermore, our findings underscore the necessity of censoring-aware policy learning. While existing literature frequently emphasizes right-censoring as a source of feature bias (Lyu et al., 2023; Cho et al., 2023), we expose its more profound impact on the systematic underestimation of true event prevalence. Early risk alerting tasks are exceptionally sensitive to the underlying event prevalence. We demonstrate that failing to account for the pervasive issue of right-censoring systematically underestimates this prior prevalence. This underestimation distorts both the policy learning phase and subsequent preference calibration. Consequently, naïve models develop an overly conservative policy, which is reluctant to trigger alerts in order to artificially protect specificity, ignoring that true specificity is inherently higher due to unobserved positives. Ultimately, this unwanted behavior defeats the purpose of early surveillance, leaving a high proportion of high-risk cases without timely and effective intervention. Our pseudo-label-based imputation approach effectively overcomes this challenge, restoring the true prevalence distribution and ensuring the policy remains clinically actionable.

To bridge the gap between algorithm development and real-world clinical application, we introduced a highly efficient post-hoc threshold calibration workflow. Different patients and healthcare systems hold specific preferences regarding the importance of earliness versus specificity based on their available resources and intervention capacities. We analytically and empirically demonstrate that entropy-regularized RL allows for a lightweight, training-free calibration mapping. This workflow enables institutions to seamlessly adapt a base policy to their specific operating points instantly, highlighting the immense practical efficiency of decoupling policy learning from preference alignment. Building on this, a promising avenue for future research is the exploration of single-policy MORL frameworks, where conditioning a universal network directly on preference parameters could eventually provide an even more elegant, end-to-end mechanism for continuous preference adaptation.

Limitations Despite these promising results, our framework also has several limitations. First, model-free RL inherently lacks explicit interpretability. The exact internal logic driving the neural policy remains opaque, which may present barriers to clinical trust and verification. Second, the effectiveness of our censoring-aware extension is bounded by the performance of the auxiliary mixture cure model. If this auxiliary model fails to accurately capture the true survival distributions due to data sparsity or unobserved confounding, the pseudo-label imputation could inject erroneous reward signals, potentially degrading the RL policy rather than improving it. Finally, to facilitate learning, our current design relies on a binary action space and cohorts with relatively regular visit patterns. Extending this framework to accommodate more complex action spaces and highly irregular observation dynamics represents a critical step for broader clinical translation.

Acknowledgments

This work was supported by the U.S. National Institutes of Health, including the National Institute of Mental Health (grant no. K01 MH127309), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant no. P50 HD093074).

References

Toygun Basaklar, Suat Gumussoy, and Umit Ogras. Pd-morl: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations*, 2022.

Andrea Burgess, Carly Luke, Michelle Jackman, Jane Wotherspoon, Koa Whittingham, Katherine Benfer, Sarah Goodman, Rebecca Caesar, Tiffney Nesakumar, Samudragupta Bora, et al. Clinical utility and psychometric properties of tools for early detection of developmental concerns and disability in young children: A scoping review. *Developmental Medicine & Child Neurology*, 67(3):286–306, 2025.

Hunyong Cho, Shannon T Holloway, David J Couper, and Michael R Kosorok. Multi-stage optimal dynamic treatment regimes for survival outcomes with dependent censoring. *Biometrika*, 110(2):395–410, 2023.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. Adaptive-halting policy network for early classification. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 101–110, 2019.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning: Cf hayes et al. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Shihao Ji and Lawrence Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, 2007.

Yikuan Li, Shishir Rao, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Gholamreza Salimi-Khorshidi, Mohammad Mamouei, Thomas Lukasiewicz, and Kazem Rahimi. Deep bayesian gaussian processes for uncertainty estimation in electronic health records. *Scientific reports*, 11(1):20685, 2021.

Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. Efficient discovery of pareto front for multi-objective reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dirk RM Lukkien, Sima Ipakchian Askari, Nathalie E Stolwijk, Bob M Hofstede, Henk Herman Nap, Wouter PC Boon, Alexander Peine, Ellen HM Moors, and Mirella MN

Minkman. Making co-design more responsible: case study on the development of an ai-based decision support system in dementia care. *JMIR Human Factors*, 11(1):e55961, 2024.

Lingyun Lyu, Yu Cheng, and Abdus S Wahed. Imputation-based q-learning for optimizing dynamic treatment regimes with right-censored survival outcome. *Biometrics*, 79(4): 3676–3689, 2023.

Coralie Martinez, Emmanuel Ramasso, Guillaume Perrin, and Michèle Rombaut. Adaptive early classification of temporal sequences using deep reinforcement learning. *Knowledge-Based Systems*, 190:105290, 2020.

Usue Mori, Alexander Mendiburu, Eamonn Keogh, and Jose A Lozano. Reliable early classification of time series based on discriminating the classes over time. *Data mining and knowledge discovery*, 31(1):233–263, 2017.

Karen Pierce, Vahid H Gazestani, Elizabeth Bacon, Cynthia Carter Barnes, Debra Cha, Srinivasa Nalabolu, Linda Lopez, Amanda Moore, Sarah Pence-Stophaeros, and Eric Courchesne. Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. *JAMA Pediatrics*, 173(6): 578–587, 2019. doi: 10.1001/jamapediatrics.2019.0624.

Anshul Sharma and Sanjay Kumar Singh. Early classification of time series based on uncertainty measure. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6. IEEE, 2019.

Minghui Sun, Haoyu Gong, Xingyu You, Jillian Hurst, Benjamin Goldstein, and Matthew Engelhard. Nest: Nested event stream transformer for sequences of multisets. *arXiv preprint arXiv:2602.00520*, 2026.

Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, 12(1):40–48, 2010.

Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*, pages 10607–10616. PMLR, 2020.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

Appendix A. Proof of Proposition 1: Monotonicity of Advantage

We prove the strict monotonicity of the net advantage function by bounding the partial derivative of the optimal value function $V_t(b_t)$ with respect to κ via backward induction. Since V_t is the maximum of linear functions, it is differentiable almost everywhere. At points of non-differentiability, the inequalities hold for all subgradients. For brevity, we proceed using standard derivatives.

Base Case ($t = T$): At the terminal step T, the observation window is exhausted, and the agent must choose between alerting and forced termination. The value function is $V_T(b_T) = \max\{\kappa\hat{p}_T R_{tp} - (1-\hat{p}_T)R_{fp}, -\kappa\hat{p}_T\delta_T\}$.

Taking the partial derivative with respect to κ yields:

$$\frac{\partial V_T(b_T)}{\partial \kappa} \in \hat{p}_T R_{tp}, -\hat{p}_T\delta_T.$$

Since R_{tp} and δ_T are strictly positive constants, we establish a definitive upper bound for the derivative at the terminal step:

$$\frac{\partial V_T(b_T)}{\partial \kappa} \leq \hat{p}_T R_{tp}.$$

Inductive Step:

Assume inductively that the upper bound $\frac{\partial V_{t+1}(b_{t+1})}{\partial \kappa} \leq \hat{p}_{t+1} R_{tp}$ holds for step $t+1$. We differentiate the continuation value $Q_t(a_t = 0, b_t)$ for step t:

$$\frac{\partial Q_t(a_t = 0, b_t)}{\partial \kappa} = -\hat{p}t\delta_t + \mathbb{E}_{b_{t+1}|b_t}\left[\frac{\partial V_{t+1}(b_{t+1})}{\partial \kappa}\right].$$

Applying the inductive hypothesis, we bound the expectation:

$$\frac{\partial Q_t(a_t = 0, b_t)}{\partial \kappa} \leq -\hat{p}t\delta_t + \mathbb{E}_{b_{t+1}|b_t}\big[\hat{p}_{t+1} R_{tp}\big].$$

Recalling the martingale property of the risk estimate update, $\mathbb{E}_{b_{t+1}|b_t}[\hat{p}_{t+1}] = \hat{p}_t$, the bound simplifies to:

$$\frac{\partial Q_t(a_t = 0, b_t)}{\partial \kappa} \leq \hat{p}tR_{tp} - \hat{p}_t\delta_t.$$

Evaluating the Net Advantage:

Next, we evaluate the derivative of the immediate alert value:

$$\frac{\partial Q_t(a_t = 1, b_t)}{\partial \kappa} = \hat{p}tR_{tp}.$$

Finally, we evaluate the derivative of the net advantage function $\Delta Q_t(b_t, \kappa)$:

$$\begin{aligned}\frac{\partial \Delta Q_t(b_t,\kappa)}{\partial \kappa} &= \frac{\partial Q_t(a_t = 1, b_t)}{\partial \kappa} - \frac{\partial Q_t(a_t = 0, b_t)}{\partial \kappa} \\ &\geq \hat{p}tR_{tp} - (\hat{p}tR_{tp} - \hat{p}_t\delta_t) \\ &= \hat{p}_t\delta_t.\end{aligned}$$

Given that $\hat{p}_t > 0$ for trajectories that have not converged to definitive negative certainty, and assuming a strictly positive temporal penalty $\delta_t > 0$, it follows that:

$$\frac{\partial \Delta Q_t(b_t, \kappa)}{\partial \kappa} \geq \hat{p}_t \delta_t > 0.$$

This confirms that the net advantage is strictly monotonically increasing with respect to κ, completing the proof.

Appendix B. Data Generating Process in Simulations

We formulate a data-generating process (DGP) that simulates patient trajectories with explicitly defined latent states, noisy observation dynamics, clinical outcomes, and time-to-event distributions. For each synthetic patient i, the underlying physiological state is represented by a randomly sampled latent feature vector $\mathbf{f}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. These base features deterministically govern the patient's intrinsic risk profile p_i via a logistic function $p_i = \sigma(\mathbf{w}_p^\top \mathbf{f}_i)$, where $\mathbf{w}_p$ are the assigned feature weights. The ground-truth of clinical event onset $z_i \in \{0, 1\}$ is then drawn as a stochastic realization of this risk profile as $z_i \sim \text{Bernoulli}(p_i)$.

We discretize the active monitoring trajectory into a fixed horizon of $T = 8$ clinical visits, and enforce a strict temporal separation between the observation window and the possible event horizon to mimic the situation of the preclinical phase. Each clinical snapshot $\mathbf{o}_t$ is generated as a noisy observation to the latent features, following $\mathbf{o}_t \sim \mathcal{N}(\mathbf{f}_i, \Sigma_t)$.

For patients who experience the event ($z_i = 1$), we model their event times T_i^* using an Accelerated Failure Time (AFT) formulation as $T_i^* = \exp(\mathbf{w}_t^\top \mathbf{f}_i + \epsilon_i)$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the stochasticity in event times. The censoring times C_i are feature-independent and randomly sampled from a fixed log-normal distribution, which is tuned to match the scale and variability of the generated event times. For each trajectory, the observed time is defined as $T_i = \min(T_i^*, C_i)$, where events occurring after C_i are treated as right-censored and masked during the offline training phase.

The simulation is driven by a 6-dimensional latent feature vector $\mathbf{z} \in \mathbb{R}^6$, wherein the first four features determine the event risk and the remaining two govern the event distribution. For information dynamics control, in S1 and S3, the observation noise is maintained at a constant variance of $\sigma^2 = 1.5$ throughout the simulation duration, while in S2 the observation noise is initialized at $\sigma^2 = 3.0$ and follows a decay trajectory, falling below $\sigma^2 = 0.5$ at step 5, after which it remains constant at this reduced level.

Appendix C. Training and Evaluation Details

We use a Recurrent PPO as the backbone model for our RL agents, and compare it against two supervised learning (SL) baselines: (1) a static MLP using only current clinical snapshots $\mathbf{o}_t$, and (2) an RNN-GRU incorporating full patient history $\mathbf{o}_{:t}$ up to current time step. Both baselines are trained as binary classifiers for event prediction, and their decision policies are derived by applying a static threshold to the output probabilities at each timestep. We also train a GRU-based mixture cure model optimized via a discrete time-binned survival loss, which serves as a strong, censor-aware supervised baseline. This mixture cure model is also utilized as the auxiliary model for pseudo-label imputation in censor-aware

RL variants. For each RL variant, we train a discrete set of anchor policies across varying κ to comprehensively cover the entire FPR spectrum.

To ensure a fair comparison between our multi-preference RL policy set and the single-model SL baselines, we evaluate all methods by constructing Pareto fronts over the objective space of ALT versus FPR. We follow a unified evaluation protocol evaluated across a dense grid of preferences κ. For the RL variants, we apply the preference calibration workflow detailed in Sec. 5.3 to map each κ to its optimal (Anchor Policy, Logit Threshold Shift) pair, as shown in Figure 1a. For the SL baselines, we perform an equivalent workflow by sweeping decision thresholds over the output probabilities on the validation set. For all methods, only the specific configuration that maximizes the empirical validation utility for a given κ is selected to form the final frontier. This guarantees that both frameworks are evaluated strictly at their optimal empirical trade-off curves. The overall performance is then quantified by Hypervolume (HV). All simulations maintain event and censoring rates at 0.5, with results averaged over five random seeds.

Appendix D. Additional Robustness and Sensitivity Analyses

D.1. Impact of Anchor Policy Quantity on Post-hoc Interpolation

We conduct an ablation study on the number of anchor policies to evaluate the robustness of our proposed post-hoc threshold interpolation workflow and justify our hyperparameter selection. We perform this evaluation using the empirical oracle method RL-gt under the S2: Information Bursts simulation setting (detailed in Section 6.1). We vary the number of anchor policies (N) across the set $\{1, 2, 3, 5, 8\}$. For each N, we train independent anchor policies across a discrete grid of preference ratio κ. The specific κ subsets allocated for each N are detailed in Table 2.

Table 2: Anchor policy quantities and their corresponding preference ratio (κ) grids.

Number of Anchor Policies (N)	**Preference Ratio (κ) Set**
1	$\{1.0\}$
2	$\{0.8, 1.5\}$
3	$\{0.5, 1.0, 5\}$
5	$\{0.5, 0.8, 1.0, 1.5, 5.0\}$
8	$\{0.5, 0.8, 1.0, 1.5, 2.0, 3.0, 5.0\}$

As shown in Figure 4(a), attempting to interpolate the entire Pareto front with a single policy ($N = 1$) yields suboptimal performance. As the number of anchor policies increases, the interpolation quality improves significantly before eventually converging. Because the computational cost of the offline anchor training phase scales linearly with N, we uniformly adopt $N = 5$ across all experiments, which establishes a practical balance between computational efficiency and post-hoc threshold interpolation quality.

D.2. Sensitivity to Auxiliary Model Errors

To assess the robustness of our framework to potential misspecification of the auxiliary mixture cure model discussed in Section 6.2, we conducted a sensitivity analysis. In the S2 simulation setting, we inject Gaussian noise ($\sigma \in \{0, 0.1, 0.3, 0.5\}$) directly into the imputed pseudo-labels during the RL-pl training phase.

As shown in Figure 4(b), the HV performance is largely preserved under minor noise ($\sigma = 0.1$), indicating that the learned policy is not brittle to small imputation errors. However, the steeper degradation observed at higher noise levels ($\sigma = 0.3$ and $\sigma = 0.5$) indicates that the proposed method genuinely relies on the accurate censoring-aware signal provided by the auxiliary mixture cure model. These results emphasize the practical necessity of carefully calibrating and validating the auxiliary model before freezing its weights in the RL pipeline.

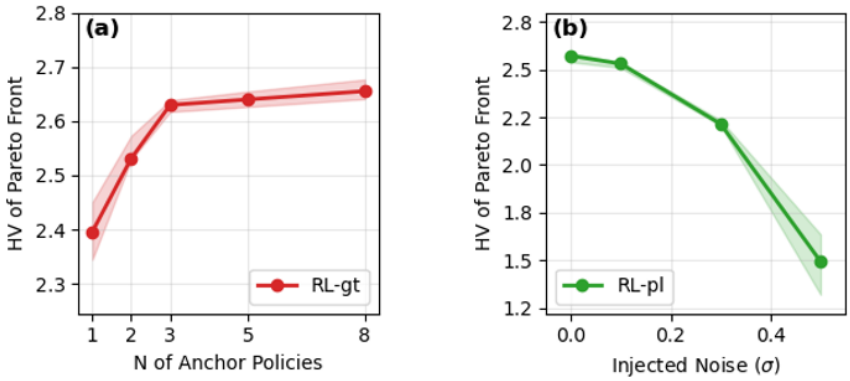

Figure 4: Sensitivity and robustness analyses of system components. (a) Impact of the number of anchor policies on the Pareto front interpolation quality, evaluated via the oracle method RL-gt. (b) Sensitivity of the censoring-aware method (RL-pl) to errors in the auxiliary mixture cure model, assessed by injecting Gaussian noise (σ) into the imputed pseudo-labels.

Appendix E. Additional Details of Real-World Experiments

E.1. Early Alerting for AD

We used NACC (UDS v3) to train an early alerting policy for Alzheimer's disease. The training cohort included patients with ≥ 4 longitudinal visits. It excluded those with prevalent AD at their first visit and those who developed non-AD dementias to prevent confounding trajectories. This yielded a cohort of 8,803 patients with overall incident AD prevalence of 14.2%. To handle the irregularity of real-world clinical follow-ups, patient trajectories were mapped to a discrete time grid. We viewed a patient's first UDS v3 encounter as the baseline ($t = 0$), then established expected follow-up slots at 12-month intervals over a 60-month observation window. Actual visits were mapped to the closest expected slot within a $\pm$6-month tolerance. Slots without a corresponding valid visit were marked as missing

and imputed via Last Observation Carried Forward (LOCF). At each grid step, the agent's snapshot observation was derived from 20 core clinical features, spanning demographics, genetic risk, cognitive battery scores, and baseline comorbidities. Importantly, we excluded all Clinical Dementia Rating (CDR) subscores from the input space, as these instruments are used directly by clinicians to determine the diagnostic label.

Given that the raw NACC dataset is subject to right-censoring, we employed two validation strategies to ensure a fair evaluation and address the ambiguity of ground-truth labels in censored test trajectories. First, we imposed an administrative early-truncation at Jan 2021 to mask all subsequent clinical visits in the training set. This creates a gap between the observation histories available to the model and the observed diagnostic outcomes, allowing us to test our framework's capacity to handle right-censoring bias. Second, we established a specific evaluation sub-cohort comprising elderly patients (age ≥ 85). Because patients in this advanced age bracket have a significantly lower probability of remaining right-censored in the raw NACC registry, their recorded outcomes serve as a highly reliable "ground-truth surrogate" for unbiased performance evaluation. The age-filtered subcohort comprises $n = 1731$ patients with a 23.6% incident AD event rate and a mean age at last follow-up of 90.83 years. The elevated event rate relative to the 14.2% prevalence of the full cohort reflects both the reduced censoring in this age bracket and the higher underlying risk of the oldest-old stratum.

E.2. Early Alerting for ASD

We utilized a proprietary pediatric Electronic Health Record (EHR) cohort to train an early alerting policy for Autism Spectrum Disorder (ASD). The cohort comprised children born between 2016 and 2023 who had at least five well-child visits recorded between 0 and 24 months of age. To ensure that the absence of a diagnosis reflects genuine clinical follow-up rather than loss to follow-up, we excluded children whose records were censored before age 2, requiring their last observed encounter to occur beyond 24 months of age. This selection criteria yielded a final cohort of 29,357 children with an overall ASD prevalence of 3.2%.

To address the inherent irregularity of real-world pediatric encounters, patient trajectories were mapped to the standard recommended well-child visit schedule. Target observation slots were established at 0, 2, 4, 6, 9, 12, 15, 18, 24, and 30 months, followed by annual visits thereafter. This schedule is deliberately dense through infancy and sparse during later childhood, mirroring the milestone-anchored structure of developmental surveillance. Each actual clinical encounter was assigned to the nearest target slot within a predefined acceptance window. Slots lacking a corresponding encounter were flagged as missed and imputed via LOCF. To allow the agent to appropriately discount stale observations and account for off-schedule care, we augmented the state space with a missingness indicator and the signed age deviation between the actual encounter and the target slot. Each visit is represented by a 768-dimensional encounter-level embedding—derived from a pretrained EHR foundation model (Sun et al., 2026), capturing demographic data and ICD-10 diagnosis codes.

Because the raw cohort is subject to natural right-censoring, we employed two distinct validation strategies mirroring those applied to the NACC dataset. First, we induced synthetic right-censoring by retroactively shifting the study endpoint back to December 2023, masking all ASD diagnoses recorded after this date and truncating subsequent trajectory

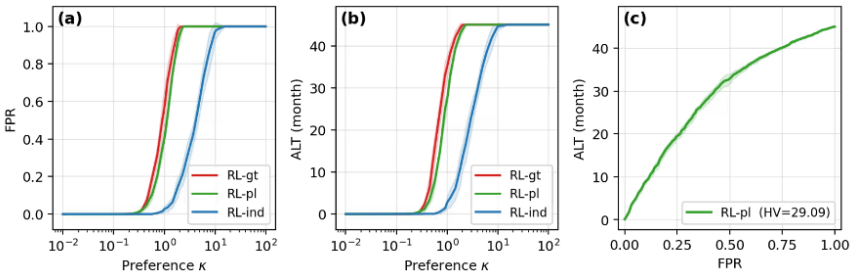

Figure 5: Preference Sensitivity and Pareto Front tested on the full ASD cohort. Panels (a)-(b) illustrate the variation of FPR and ALT calibrated to the preference ratio κ spectrum. Panel (c) shows the resulting Pareto front from the optimal RL-gt agent, achieving an ALT of 8.71 months before formal ASD diagnosis with 90% specificity.

data. This intervention reduces the observed ASD event rate from 3.2% to 1.7%—nearly halving the apparent prevalence—thereby directly instantiating the prior underestimation that our censoring-aware formulation is explicitly designed to correct. Second, to establish a reliable evaluation subcohort, we isolated children with sufficiently long longitudinal follow-up (age ≥ 8). For this subcohort, the primary diagnostic window for ASD is considered substantially closed, allowing their recorded outcomes to serve as a robust ground-truth surrogate for evaluation.