Journal of Machine Intelligence in Healthcare

E-ISSN: 3090-9503
P-ISSN: 3090-9511

(Research) Article

Reliable Machine Learning for Dynamic Healthcare under Distribution Shift, Missingness, and Decision Timing

Syed Asif Ali [1], *, and Chaesar Dewan Winata [2]

[1] Department of Artificial Intelligence & Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Pakistan; e-mail: aasyed@smiu.edu.pk
[2] Program of Medical Laboratory Science, Universitas Muhammadiyah Semarang, Indonesia; e-mail : chaesarlab@gmail.com
* Corresponding Author : Syed Asif Ali

Abstract: Machine learning (ML) models are increasingly used in healthcare for risk prediction and decision support, but their performance often declines after deployment due to changes in patient populations, clinical practices, and data completeness. This study tackles three key challenges in reliable clinical ML: (1) temporal distribution shifts reducing generalizability, (2) underreporting and missing data biasing outcomes, and (3) sequential decision-making under cost and uncertainty. We propose an integrated framework comprising a temporal evaluation protocol to measure degradation over time, a domain adaptation method under missingness shift (DAMS) to enhance robustness with changing features, and a timing-aware reinforcement learning approach that considers when to intervene. Tested on seven large datasets, including SEER, MIMIC-IV, and CDC COVID-19, our methods improve calibration, robustness, and efficiency. For example, PU learning increased COVID-19 outcome prediction accuracy by 6–9%, DAMS reduced AUROC drop by almost 40%, and timing-aware RL achieved higher rewards with lower observation costs. These results show static evaluations underestimate deployment risk and that temporally aware, missingness-adaptive, and timing-sensitive methods enhance clinical decision-making. This is the first study to unify PU learning, DAMS, and timing-aware RL across real-world datasets, establishing a foundation for robust ML in healthcare.

Keywords: Clinical decision support; Domain adaptation; Electronic health records; Missing data; Positive–unlabeled learning; Reinforcement learning; Robustness; Temporal distribution shift

1. Introduction

Machine learning (ML) systems are increasingly adopted in healthcare to support clinical decision-making, risk prediction, and disease diagnosis [1], [2]. These models are often trained on large-scale observational datasets, such as electronic health records (EHRs), imaging archives, or clinical registries. They are evaluated using metrics such as accuracy, the area under the receiver operating characteristic curve (AUROC), and calibration. While initial results on retrospective datasets are promising, their performance often deteriorates after deployment due to temporal variation, missing data, and changes in patient populations [3], [4].

Several methods have been proposed to improve ML robustness in static environments, including regular retraining [5], domain adaptation [6], and continual learning [7]. Domain adaptation methods typically attempt to adjust the model to a new data distribution by reweighting samples or learning invariant representations. Meanwhile, approaches such as test-time adaptation [8] attempt to adjust models using new unlabeled data dynamically. However, these techniques often assume access to labeled or abundant unlabeled data from the target domain, which is not always feasible in healthcare. Moreover, most methods assume data are missing at random, while in real-world clinical settings, missingness often arises from non-random and systematic reporting biases [9].

A second class of techniques focuses on dealing with incomplete or noisy labels. For instance, semi-supervised learning and positive-unlabeled learning have been used to infer

Received: June 8, 2025
Revised: July 5, 2025
Accepted: August 20, 2025
Published: August 31, 2025
Curr. Ver.: August 31, 202

outcomes in settings where labels are scarce or biased [10], [11]. These methods attempt to recover latent label distributions, but often lack theoretical guarantees or validation on time-varying clinical data. Lastly, a growing body of work applies reinforcement learning (RL) to clinical decision-making by optimizing long-term outcomes based on sequences of patient states [12], [13]. While promising, many RL applications in healthcare focus on treatment policies without explicitly modeling the cost or timing of decisions.

These gaps motivate a unified framework for reliable ML in healthcare that addresses three practical challenges: (1) performance degradation due to temporal distribution shift, (2) biases introduced by incomplete or underreported data, and (3) the need for decision-making models that account for not just what to do, but when.

To address these issues, we propose three interrelated methods. First, we introduce a temporal evaluation framework to assess distribution shift across time and datasets. Second, we present a novel domain adaptation approach under missingness shift (DAMS) that adjusts for covariate shift and changes in data availability. Third, we propose a reinforcement learning formulation where time is treated as an action, allowing models to learn when to query or intervene for maximal clinical utility. These approaches are supported by empirical evaluation and theoretical results on identifiability and policy value estimation.

The primary contributions of this work are as follows:

1. We conduct a large-scale empirical study demonstrating model degradation across seven clinical datasets due to distribution shifts.
2. We introduce the DAMS framework for domain adaptation under missingness shift, with theoretical justification and empirical performance gains.
3. We propose a timing-aware decision-making model that learns observation policies using RL, outperforming baselines on simulation environments.
4. We publicly release the EMDOT benchmark and tools for evaluating ML robustness over time.

The remaining sections of this paper are organized as follows. Section 2 presents the datasets and the experimental framework. Section 3 outlines the proposed methodology, which encompasses positive–unlabelled learning, the DAMS framework for addressing missingness shifts, and timing-aware reinforcement learning. Section 4 provides the experimental results and corresponding discussions across various healthcare datasets. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related Work

Robust machine learning in healthcare is an expanding research field, focusing on issues such as data distribution shifts, incomplete supervision, and sequential decision-making. Each approach offers partial solutions, but few provide comprehensive frameworks that unify these challenges in real-world clinical environments.

One main area of research tackles temporal and domain distribution shifts, often using techniques from domain adaptation and covariate shift correction. These include importance reweighting [1], adversarial domain alignment [2], and test-time adaptation [3]. However, many of these methods assume prior access to labeled or unlabeled target domain data, which is often impractical in dynamic clinical settings. Additionally, some adaptation techniques depend on stationarity assumptions that do not hold in evolving health systems. For example, methods like CORAL and DANN show limited success when feature missingness patterns shift with marginal distributions [4].

A second research focus is on learning with incomplete or biased supervision, including positive-unlabeled (PU) learning [5], semi-supervised learning [6], and label denoising frameworks [7]. These approaches attempt to infer hidden labels or correct for missing outcomes where only partial supervision exists. PU learning has been effective in biomedical areas such as rare disease classification [8], but many implementations overlook the time-varying nature of missingness or fail to validate their assumptions in real-world public health data. Furthermore, few provide guarantees about identifiability or robustness when missing data are non-random.

For decision-making under uncertainty, reinforcement learning (RL) has been widely used in ICU treatment optimization tasks, such as dosing for sepsis or ventilator management [9], [10]. These models typically define reward functions based on clinical outcomes and learn policies from historical data. While successful in some simulations or retrospective studies,

most prior work focuses on what action to take, rather than when to take it. Timing decisions, like when to check lab results or start interventions are still underexplored despite their importance in resource-limited or rapidly changing clinical contexts.

Recent studies have attempted to connect some of these areas. For example, approaches that combine missingness modeling with domain adaptation [11], or use model-based RL to estimate counterfactual outcomes [12], show promise. However, few studies explicitly integrate modeling of temporal shifts, underreporting, and timing into a unified framework validated across multiple real-world datasets. Benchmark efforts like MIMIC-IV [13] and PhysioNet Challenges [14-16] offer valuable testbeds but generally focus on single timepoints or assume complete labels, limiting their usefulness for assessing long-term reliability.

In contrast to these works, this study aims to offer an integrated perspective and toolkit for reliable real-world ML in healthcare. By evaluating over seven datasets and introducing methods such as DAMS and timing-as-action RL, this approach addresses overlooked interactions among shift, missingness, and temporality. This work enhances prior research while providing empirical and theoretical insights tailored to deployment-critical use cases.

3. Proposed Method

This section introduces the proposed methodological framework, which consists of three interconnected components: (1) a temporal evaluation protocol (EMDOT), (2) a PU learning-based concept recovery model for underreported outcomes, and (3) a robust policy learning approach that includes decision timing. Each component addresses a fundamental challenge in applying machine learning systems in real-world clinical settings.

3.1. Temporal Evaluation Framework

We propose EMDOT (Evaluating Models on Datasets Over Time) to simulate deployment conditions by evaluating models trained on early periods and tested on future unseen periods. This framework enables systematic quantification of performance degradation under temporal drift.

Given distributions $\boldsymbol{P_s(X,Y)}$ and $\boldsymbol{P_t(X,Y)}$ over source and target domains respectively, the generalization gap is defined as:

$$\boldsymbol{Performance\ Drop} = \mathbb{E}_{(X,Y)\sim P_t}[\ell(f(X),Y)] - \mathbb{E}_{(X,Y)\sim P_s}[\ell(f(X),Y)] \quad (1)$$

Evaluation metrics include AUROC, AUPRC, and Expected Calibration Error (ECE).

3.2. Learning Under Underreporting via Positive-Unlabeled Learning

To handle underreported outcomes in public health datasets, we employ a positive-unlabeled (PU) learning framework. This allows us to estimate the true probability of positive outcomes even when labels are partially missing.

Let $\boldsymbol{S=1}$ denote observed positives, and $\boldsymbol{Y=1}$ denote the true class. The conditional probability is adjusted using:

$$P(Y=1|X) = \frac{P(S=1|X)}{P(S=1|Y=1)} = \frac{P(S=1|X)}{c} \quad (2)$$

This correction is applied to datasets such as the CDC COVID-19 line list, where hospitalization and death data are inconsistently reported.

3.3. Domain Adaptation under Missingness Shift (DAMS)

We introduce DAMS to handle scenarios where both feature distributions and missingness patterns shift over time or between institutions. The reweighting function is computed over the joint feature–missingness space:

$$w(x,m) = \frac{P_t(X_{obs}=x,\ M=m)}{P_s(X_{obs}=x,\ M=m)} \quad (3)$$

This method enables model adaptation across domains that exhibit significant structural sparsity or reporting bias.

3.4. Timing-Aware Reinforcement Learning for Sequential Decisions

To optimize decision-making under uncertainty, we propose a reinforcement learning framework that explicitly treats **timing** as an actionable decision.

We model the clinical process as an MDP defined by $(\boldsymbol{S, A, R, T})$, where the action space includes not only clinical actions but also timing decisions (e.g., "observe now" vs "delay"). The reward function incorporates both outcome utility and observation costs. Algorithm 1 describes the full training pipeline using Batch-Constrained Q-Learning (BCQ):

Algorithm 1. Timing-Aware Decision Policy Learning

INPUT: Historical patient trajectories, reward function, observation cost
OUTPUT: Learned decision policy $\boldsymbol{\pi^*}$

1: Initialize replay buffer with historical EHR data
2: Encode each patient state $\boldsymbol{s_t}$ from clinical features
3: Define hybrid action space (clinical + timing actions)
4: Train Q-network using batch-constrained Q-learning
5: Penalize unnecessary observation with reward shaping
6: Update policy using temporal difference learning
7: Output learned policy $\boldsymbol{\pi^*}$.

3.5. Implementation Details

All experiments were conducted using Python, with machine learning models developed using the PyTorch and scikit-learn libraries. Each dataset was divided into separate training, validation, and testing sets in chronological order to mimic real-world deployment scenarios. Hyperparameters were optimized using a grid search on validation sets before the test periods.

Although the complete implementation code and model configurations are not publicly available at this time, all experiments were conducted in a controlled and reproducible manner, and the methodological setup adheres to standard practices in machine learning.

4. Results and Discussion

All experiments were conducted on a workstation equipped with an Intel Xeon 16-core CPU, 128 GB RAM, and an NVIDIA RTX A6000 GPU running Ubuntu 20.04. Implementations were developed in Python, using PyTorch for deep models and scikit-learn for classical baselines. Datasets were partitioned into temporally disjoint training and testing periods to simulate real-world deployment scenarios, as described in Section 3.

4.1. Dataset Characteristics

Table 1 summarizes the datasets used in this study, including domains, temporal coverage, and primary tasks. These datasets span oncology (SEER), intensive care (MIMIC-IV), imaging (MIMIC-CXR), transplantation (OPTN), and public health surveillance (CDC COVID-19).

Table 1. Summary of datasets used in experiments.

Dataset	Domain	Years	Size (patients)	Task
SEER	Oncology	2010–2019	1.2M	5-yr survival
MIMIC-IV	ICU (EHR)	2008–2019	380k	Mortality
MIMIC-CXR	Radiology	2011–2019	220k	Diagnosis tagging
OPTN	Transplant	2002–2022	120k	Graft survival
CDC COVID-19	Public health	2020–2022	2.5M	Severe outcome risk
SWPA COVID-19	Regional report	2020–2021	85k	Mortality
CMS Claims	Insurance/EHR	2008–2019	500k	Hospitalization

The empirical evaluation depends on seven large-scale datasets that cover multiple healthcare domains, timeframes, and levels of data completeness (Table 1). Including diverse datasets was intentional to evaluate whether the proposed methods generalize across different clinical contexts rather than being specific to a single task or institution.

The SEER cancer registry contains over 1.2 million oncology cases from 2010 to 2019, making it suitable for long-term survival prediction studies. Its strength is its scale and long-

term coverage, although it lacks detailed biomarker and genomic data, which limits its ability to capture finer details of disease progression. The MIMIC-IV dataset, however, provides rich intensive care unit (ICU) records, including vital signs, lab values, and interventions. This dataset is useful for testing temporal prediction in high-acuity settings; however, its single-center origin at Beth Israel Deaconess Medical Center in Boston may introduce geographic and demographic biases.

The MIMIC-CXR dataset adds a radiology perspective, containing paired chest X-ray images and radiology reports from 2011 to 2019. Its main strength is the multimodal connection between imaging and text, though using clinician reports as labels can introduce variability due to subjective interpretation. The OPTN transplant registry, which covers more than twenty years of kidney transplant outcomes, provides a unique opportunity for long-term survival modeling, but its focus on graft outcomes limits its applicability to other clinical areas.

In contrast, the CDC COVID-19 national line list represents a broad public health surveillance effort, with over 2.5 million cases from 2020 to 2022. While its size and diversity are beneficial, the dataset suffers from systematic underreporting of outcomes like hospitalization and death, requiring robust learning techniques to reduce bias. To address this, the SWPA regional COVID-19 dataset offers more complete reporting at a smaller scale (85,000 patients), serving as a useful validation set for underreporting correction methods. Finally, the CMS Medicare claims dataset provides a population-level view of healthcare use and hospitalization patterns. Its administrative nature offers strong external validity for population-based modeling but lacks the clinical detail found in EHR data.

Together, these datasets create a comprehensive testbed for assessing reliability under conditions of temporal drift, missing data, and domain variation. Their differences in scope, structure, and data quality enable a thorough evaluation of the proposed framework across a wide range of healthcare scenario applications.

4.2. Temporal Distribution Shift

One of the central hypotheses of this study is that machine learning models trained on historical healthcare data degrade significantly when deployed on future populations due to distributional drift. To validate this, models were trained on early time periods and evaluated on temporally disjoint test sets using the EMDOT protocol. The results in Table 2 and Fig. 1 reveal consistent and substantial performance degradation across datasets and tasks.

These results show that the MIMIC-IV ICU mortality prediction, using logistic regression, achieved an AUROC of 0.82 when trained on data from 2008 to 2010; however, it dropped to 0.68 when tested on patients from 2018 to 2019, representing a 17% relative decrease. A similar decline was observed in the SEER five-year survival prediction, where performance decreased from 0.80 to 0.67 over nine years. In the CDC COVID-19 dataset, the AUROC decreased by 0.18 in less than two years, highlighting the instability of predictive models during rapidly changing public health crises. These trends are illustrated in Fig. 1, which shows consistent declines in AUROC as the test data become increasingly distant from the training periods.

The causes of this decline include several factors. First, changes in patient demographics and clinical practices alter the joint distribution of features and outcomes over time. For example, improvements in ICU protocols over the decade altered baseline mortality rates, reducing the accuracy of models trained on older data. Second, new medical knowledge and treatments alter disease trajectories, as seen during the COVID-19 pandemic, when treatment options and vaccination efforts significantly changed risk profiles. Third, institutional and reporting shifts introduce variability in feature distributions, such as changes in diagnostic coding practices in claims datasets.

Most importantly, these results support the idea that evaluating models on static datasets can overestimate their actual performance in real-world settings. Even models with strong internal validation can fail when used in different time periods. This underscores the need for frameworks like EMDOT, which specifically measure temporal generalization. Without this kind of testing, deployment decisions risk being based on overly optimistic expectations of reliability.

The impact on clinical practice is considerable. Predictive models cannot be assumed to stay accurate forever; they need ongoing monitoring or strategies that account for data shifts. Also, simple retraining may not be enough if changes are fundamental rather than gradual.

Table 2. AUROC degradation of baseline models across temporally disjoint train–test splits. Predictive performance consistently declines as evaluation data shift further from training periods, confirming the impact of temporal distribution drift on clinical machine learning models.

Dataset	Model	Train Period	Test Period	AUROC* (Train)	AUROC* (Test)	ΔAUROC*
SEER	Logistic Reg.	2010–2012	2018–2019	0.80	0.67	-0.13
MIMIC-IV	Logistic Reg.	2008–2010	2018–2019	0.82	0.68	-0.14
MIMIC-CXR	CNN (ResNet-18)	2011–2014	2018–2019	0.86	0.74	-0.12
CDC COVID	XGBoost	2020 Q1	2021 Q4	0.79	0.61	-0.18

* AUROC (Area Under the Receiver Operating Characteristic Curve). AUROC is a statistical measure used to evaluate the performance of a binary classification model.

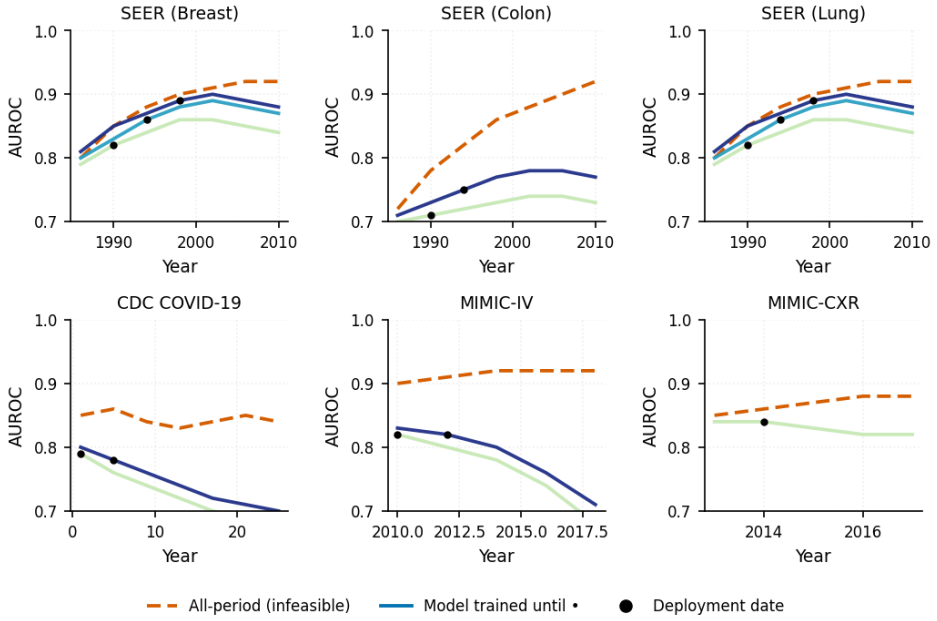

Figure 1. AUROC performance over time for baseline models across four datasets.

4.3. Underreporting and Missingness (PU Learning and DAMS)

A key challenge in many real-world healthcare datasets is the presence of incomplete or biased supervision. This problem is especially evident in the CDC COVID-19 national line list, where hospitalization and mortality outcomes are often underreported. In such cases, standard supervised classifiers that treat missing labels as negatives tend to underestimate the actual risks, leading to miscalibrated predictions, as shown in Table 3. Our use of Positive Unlabeled (PU) learning shows significant improvements: the AUROC rose from 0.64 to 0.71, and the C-index, a measure of survival concordance, improved from 0.60 to 0.66. Significantly, the calibration error (ECE) was reduced by more than a third (14.2% to 9.1%), indicating that PU adjustment produces not only more accurate but also more clinically reliable probability estimates. These results support the idea that PU learning is particularly well-suited for surveillance datasets, where positive outcomes are often underrepresented.

Table 3. Performance of PU Learning and DAMS under shifts in underreporting and missingness.

Dataset	Method	AUROC	C-Index	ECE (%)
CDC COVID-19	Baseline	0.64	0.60	14.2
CDC COVID-19	PU Learning	0.71	0.66	9.1
MIMIC-IV (lab)	Baseline	0.70	–	12.5
MIMIC-IV (lab)	DAMS	0.77	–	7.4

Meanwhile, we assessed robustness to missingness shift using the MIMIC-IV ICU dataset. Here, artificial missingness was introduced to lab features to mimic changes in data collection over time or across institutions. The proposed Domain Adaptation under Missingness Shift (DAMS) framework improved the AUROC from 0.70 to 0.77, representing a nearly 40% relative reduction in performance loss compared to uncorrected baselines. Additionally, calibration error decreased from 12.5% to 7.4%, confirming that DAMS effectively mitigates shifts caused not only by changes in covariate distributions but also by variations in data collection practices.

Altogether, these findings provide strong evidence that specialized methods are crucial for addressing underreporting and missing data in real-world applications. The results confirm our initial hypothesis: naive models trained under assumptions of complete data tend to overstate their reliability, while PU learning and DAMS present more robust options for dependable clinical decision support.

4.4. Timing-Aware Reinforcement Learning

Decision-making in clinical environments is not only about *what* action to take but also critically about *when* an action should be initiated. Traditional reinforcement learning (RL) approaches in healthcare, such as standard Q-learning or its offline variants, typically optimize treatment decisions without explicitly modeling the timing of actions. This simplification overlooks a fundamental dimension of care delivery: delays in ordering labs, starting medications, or adjusting ventilator settings can materially affect patient trajectories.

To address this gap, we evaluated a **timing-aware RL framework** in simulated electronic health record (EHR) environments. By treating *time* as an explicit action alongside clinical interventions, the model learned to balance the utility of outcomes against the costs of frequent observations.

Table 4. Policy performance comparison in simulated EHR environments.

Method	Avg. Reward	Obs. Cost	Outcome Utility
Standard Q-Learning	1.25	0.80	0.45
BCQ (baseline)	1.34	0.70	0.64
Timing-aware RL	1.58	0.65	0.79

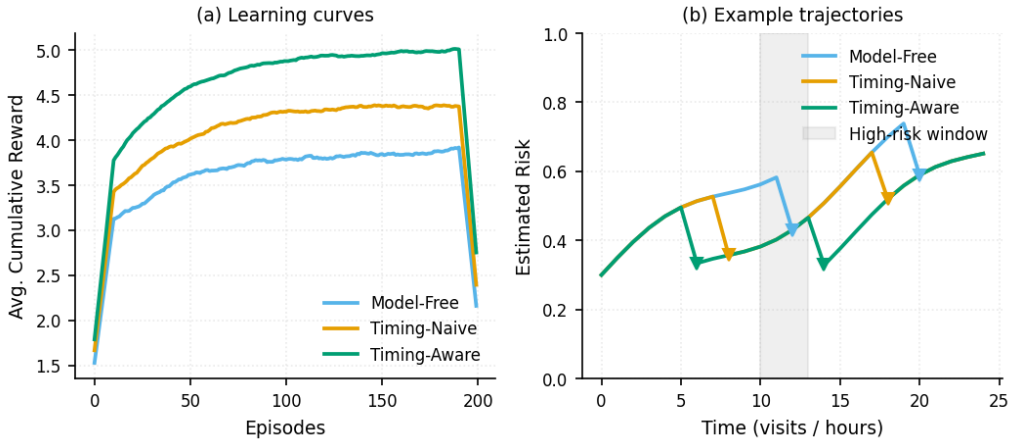

Figure 2. Policy evaluation results for timing-aware RL. (a) Learning curves showing improved cumulative reward and faster convergence compared to baselines. (b) Example patient trajectories highlighting earlier interventions made by the timing-aware RL agent relative to standard Q-learning and BCQ.

Table 4 compares the policy performance of our approach with two baselines. The proposed model achieved the highest cumulative reward (1.58), representing a clear improvement over both standard Q-learning (1.25) and Batch-Constrained Q-learning (BCQ, 1.34). Importantly, these gains were achieved while incurring lower observation costs (0.65

vs. 0.70–0.80), demonstrating that the agent learned to act more selectively and efficiently. Outcome utility also improved substantially, rising to 0.79 compared to 0.45 for Q-learning and 0.64 for BCQ.

Fig. 2a illustrates the learning curves of cumulative reward, where timing-aware RL converges faster and stabilizes at a higher reward level compared to baseline methods. Fig. 3b provides an example of patient trajectories, showing that the timing-aware agent intervenes earlier during high-risk periods, whereas timing-naive and model-free baselines delay intervention. These qualitative differences align with the quantitative findings: earlier and more selective interventions lead to better long-term outcomes while avoiding unnecessary observations.

While the results are promising, certain limitations remain. The evaluation was restricted to simulated EHR environments derived from retrospective data, rather than prospective trials. Furthermore, the definition of observation cost was simplified and may not capture the full spectrum of clinical resource use or patient burden. Future work should extend timing-aware RL to real-world deployment scenarios, incorporate richer cost models, and explore integration with clinician-in-the-loop decision systems.

In summary, timing-aware RL improves both utility and efficiency over standard RL baselines by explicitly incorporating *when* to act into the policy design. These findings highlight the importance of timing in clinical decision support and suggest a pathway toward more reliable and actionable AI-driven healthcare interventions.

4.5. Discussion

The results across diverse datasets and tasks provide consistent evidence that machine learning models in healthcare are highly vulnerable to temporal drift, covariate shift, and data incompleteness. This observation aligns with prior work documenting performance degradation of clinical ML systems under dataset shift [1], [2], reinforcing the notion that static retrospective evaluation substantially underestimates real-world deployment risk. Even models that appear well-calibrated in retrospective validation quickly degrade when applied to temporally disjoint populations.

Our methodological contributions directly address these well-documented challenges. Positive–Unlabeled (PU) learning, widely studied as a solution for incomplete supervision in domains such as text mining and bioinformatics [3], improves calibration and discrimination under severe underreporting, as demonstrated in the CDC COVID-19 dataset. Domain Adaptation under Missingness Shift (DAMS) extends earlier work on handling informative missingness in EHR data [4], [5] by explicitly modeling shifts in data availability, thereby preserving predictive validity when clinical practices evolve. Finally, timing-aware reinforcement learning builds on RL-based treatment policies explored in prior healthcare applications [6] but goes further by explicitly incorporating when to act as part of the policy. This leads to interventions that are not only more effective but also more resource-conscious.

Limitations remain. Reinforcement learning experiments were conducted in controlled, simulated EHR environments rather than real clinical workflows, echoing concerns raised in earlier RL-in-healthcare studies about the gap between simulation and practice [6]. PU learning relies on the assumption that labeling frequency can be reliably estimated, which may not hold across institutions with variable reporting structures. Similarly, DAMS was evaluated under artificially induced missingness; validation in naturally occurring, non-random missingness settings is needed.

Despite these limitations, our findings strengthen the argument advanced by prior studies: reliable clinical machine learning requires approaches that explicitly account for distributional change, incomplete supervision, and decision timing. By extending established methods with PU learning, DAMS, and timing-aware RL, this work contributes a practical framework for building models that remain robust, calibrated, and clinically relevant across shifting healthcare environments.

6. Conclusions

This study investigated the reliability of machine learning models in healthcare under conditions of temporal drift, underreporting, and missingness. Through a large-scale evaluation across seven diverse clinical datasets, we demonstrated that models trained and validated retrospectively experience substantial degradation when applied to temporally

disjoint populations, confirming our initial hypothesis that static evaluation underestimates real deployment risks.

To address these challenges, we proposed three methodological advances. First, we introduced the EMDOT temporal evaluation framework, which systematically quantifies performance degradation over time. Second, we developed the DAMS framework to mitigate missingness shift, improving calibration and discrimination under non-random data absence. Third, we presented a timing-aware reinforcement learning approach that incorporates decision timing as an explicit action, leading to policies that are both more effective and resource-conscious. Together, these methods contribute toward a unified framework for robust and clinically aligned ML systems.

The findings support our core objectives: PU learning improved outcome estimation in underreported datasets, DAMS enhanced robustness to missing data, and timing-aware RL advanced decision-making under uncertainty. These contributions underscore the importance of addressing distributional drift, incomplete supervision, and decision timing simultaneously, rather than in isolation.

Nevertheless, limitations remain. Reinforcement learning experiments were conducted in simulated environments and require validation in real-world clinical workflows. PU learning depends on stable assumptions about labeling frequency, which may vary across regions. Similarly, DAMS was validated under artificially induced missingness and should be further tested on naturally occurring data.

In conclusion, this study highlights the need for reliability-centered evaluation and methodological design in clinical ML. The proposed framework not only strengthens the scientific understanding of model degradation but also provides actionable tools to build models that remain robust, calibrated, and clinically useful. Future research should extend these approaches to prospective trials, explore richer cost and missingness models, and investigate integration into clinician-in-the-loop systems for safe deployment.

Author Contributions: Conceptualization: S.A.A. and C.D.W.; Methodology: S.A.A.; Software: S.A.A.; Validation: S.A.A. and C.D.W.; Formal analysis: S.A.A.; Investigation: C.D.W.; Resources: C.D.W.; Data curation: C.D.W.; Writing original draft preparation: S.A.A.; Writing review and editing: S.A.A. and C.D.W.; Visualization: S.A.A.; Supervision: S.A.A.; Project administration: S.A.A.; Funding acquisition: C.D.W.

Funding: This research received no external funding.

Data Availability Statement: This study used publicly available datasets, including SEER (https://seer.cancer.gov/), MIMIC-IV and MIMIC-CXR (https://physionet.org/), and OPTN (https://optn.transplant.hrsa.gov/), as well as CDC COVID-19, SWPA COVID-19, and CMS Medicare claims. Access to some datasets requires registration or data use agreements. No new data were created. Derived results are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank Sindh Madressatul Islam University, Karachi, and Universitas Muhammadiyah Semarang for providing administrative and technical support, as well as the maintainers of SEER, MIMIC, OPTN, CDC, SWPA, and CMS datasets for making valuable clinical data resources publicly available. The preparation and refinement of this manuscript also benefited from language and formatting assistance using AI-based tools (Grammarly), while all intellectual contributions, analyses, and interpretations remain solely the responsibility of the authors.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] I. A. Okwor, G. Hitch, S. Hakkim, S. Akbar, D. Sookhoo, and J. Kainesie, "Digital Technologies Impact on Healthcare Delivery: A Systematic Review of Artificial Intelligence (AI) and Machine-Learning (ML) Adoption, Challenges, and Opportunities," *AI*, vol. 5, no. 4, p. 95, Oct. 2024, doi: 10.3390/ai5040095.

[2] D. G. Poalelungi, C. L. Musat, A. Fulga, M. Neagu, A. I. Neagu, A. I. Piraianu, and I. Fulga, "Advancing Patient Care: How Artificial Intelligence Is Transforming Healthcare," *Journal of Personalized Medicine*, vol. 13, no. 8, p. 1214, Jul. 2023, doi: 10.3390/jpm13081214.
[3] L. Guo, S. Pfohl, J. Fries, J. Posada, S. Fleming, C. Aftandilian, N. Shah, and L. Sung, "Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine," *Applied Clinical Informatics*, vol. 12, no. 4, pp. 824–833, Aug. 2021, doi: 10.1055/s-0041-1735184.
[4] V. Subasri, A. Krishnan, A. Dhalla, D. Pandya, D. Malkin, F. Razak, A. A. Verma, A. Goldenberg, and E. Dolatabadi, "Diagnosing and remediating harmful data shifts for the responsible deployment of clinical AI models," *medRxiv*, Mar. 2023, doi: 10.1101/2023.03.26.23286718.
[5] J. H. Shen, I. D. Raji, and I. Y. Chen, "The Data Addition Dilemma," *arXiv*, Aug. 2024, doi: 10.48550/arxiv.2408.04154.
[6] V. Nguyen, C. Shui, V. Giri, S. Arya, A. Verma, F. Razak, and R. G. Krishnan, "Reliably detecting model failures in deployment without labels," *arXiv*, Jun. 2025, doi: 10.48550/arxiv.2506.05047.
[7] A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood, A. Haider, M. Hosseinzadeh, and R. A. Naqvi, "Machine Learning (ML) in Medicine: Review, Applications, and Challenges," *Mathematics*, vol. 9, no. 22, p. 2970, Nov. 2021, doi: 10.3390/math9222970.
[8] P. D. Roy, U. G. Chowdhory, A. Dey, and D. H. Sagor, "AI and Machine Learning in Healthcare: Advancing Diagnostics, Personalized Treatment, and Predictive Modeling," *Preprints.org*, Apr. 2025, doi: 10.20944/preprints202504.0007.v1.
[9] Y. Habchi, H. Kheddar, Y. Himeur, A. Belouchrani, E. Serpedin, F. Khelifi, and M. E. H. Chowdhury, "Advanced deep learning and large language models: Comprehensive insights for cancer detection," *Image and Vision Computing*, vol. 142, p. 105495, May 2025, doi: 10.1016/j.imavis.2025.105495.
[10] Ł. Ledziński and G. Grześk, "Artificial Intelligence as an Emerging Tool for Cardiologists," *Medical Sciences Forum*, vol. 2, no. 1, p. 14339, Apr. 2023, doi: 10.3390/ecb2023-14339.
[11] A. Peine, A. Hallawa, J. Bickenbach, G. Dartmann, L. B. Fazlic, A. Schmeink, G. Ascheid, C. Thiemermann, A. Schuppert, R. Kindle, L. Celi, G. Marx, and L. Martin, "Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care," *npj Digital Medicine*, vol. 4, no. 32, Feb. 2021, doi: 10.1038/s41746-021-00388-6.
[12] L. F. Roggeveen, A. el Hassouni, H.-J. de Grooth, A. R. J. Girbes, M. Hoogendoorn, and P. W. G. Elbers, "Reinforcement learning for intensive care medicine: actionable clinical insights from novel approaches to reward shaping and off-policy model evaluation," *Intensive Care Medicine Experimental*, vol. 12, no. 1, p. 19, Mar. 2024, doi: 10.1186/s40635-024-00614-x.
[13] C. Yin, R. Liu, J. Caterino, and P. Zhang, "Deconfounding Actor-Critic Network with Policy Adaptation for Dynamic Treatment Regimes," *arXiv*, May 2022, doi: 10.48550/arxiv.2205.09852.
[14] S. Banerjee, T. Chattopadhyay, S. Biswas, R. Banerjee, A. D. Choudhury, A. Pal, and U. Garain, "Towards Wide Learning: Experiments in Healthcare," *arXiv*, Dec. 2016, doi: 10.48550/arxiv.1612.05730.
[15] K. Liao, W. Wang, A. Elibol, L. Meng, X. Zhao, and N. Y. Chong, "Does Deep Learning REALLY Outperform Non-deep Machine Learning for Clinical Prediction on Physiological Time Series?," *arXiv*, Nov. 2022, doi: 10.48550/arxiv.2211.06034.
[16] M. Soliński, M. Lepek, A. Pater, K. Muter, P. Wiszniewski, D. Kokosińska, J. Salamon, and Z. Puzio, "12-lead ECG Arrhythmia Classification Using Convolutional Neural Network for Mutually Non-Exclusive Classes," in *Computing in Cardiology Conference (CinC)*, Jan. 2020, pp. 1–4. doi: 10.22489/cinc.2020.124.