Confounding Masquerading as Improvement: A Systematic Evaluation of Offline Reinforcement Learning for Stroke Antithrombotic Treatment in a 129,000-Patient Registry

Kihun Rhee RHEEKH00@SNU.AC.KR
AI Research Center, JLK Inc., Seoul, South Korea

Abstract

Recent offline reinforcement learning (RL) studies report data-driven policies outperform physician decisions by 10–32 % on clinical outcomes. We conduct a systematic, partially crossed evaluation of five offline RL algorithm families and 14 reward designs on 44,894 post-2018 acute ischemic stroke patients from the nationwide stroke registry ($N = 129{,}033$).

Standard Fitted Q-Evaluation yields $V(\text{imp}) = +0.0069$ ($p = 0.048$); adding an Early Neurological Deterioration penalty strengthens the signal to $+0.0101$ ($p = 0.0002$, approximate E-value = 18.37) — estimates that could support a premature positive policy-improvement claim despite relying on a confounded reward function.

We identify *reward-embedded confounding*, where the proxy terminal reward encodes baseline severity/prognosis information in addition to any treatment efficacy signal. E-value analysis does not evaluate this pathway because the confounding is carried through the reward channel. A 2×2 factorial experiment reveals that terminal reward confounding alone more than eliminates the apparent improvement, accounting for 218.6 % of the $r_{\text{END}} \to r_{\text{full}}$ signal change (i.e., removing terminal confounding overshoots null).

After DML-inspired GBM reward residualization, $V(\text{imp})$ attenuates to $+0.0033$ ($p = 0.132$); full deconfounding yields $+0.0025$ ($p = 0.291$). Three independent approaches converge away from a clinically meaningful aggregate policy-improvement claim: FQE-based reward deconfounding, a T-learner showing 75–90 % attenuation with a clinically small residual (< 6 % MCID), and direct ischemic-stroke recurrence analysis (IPW/AIPW). A within-registry 1-year mRS factorial (N=35,744) replicates the pattern: $V(\text{imp}) = +0.0133$ attenuates to -0.0004 after reward residualization (103.3 % attenuation), indicating that the finding is not specific to the 3-month mRS horizon in this registry. We present an empirically motivated 6-step evaluation checklist; applied retrospectively, three highly cited clinical RL papers do not report the FQE confirmation specified by Step 1.

Despite the overall null, National Institutes of Health Stroke Scale (NIHSS)-stratified heterogeneity (T-learner conditional average treatment effect (CATE) ratio 4.6×, corroborated by Q-evaluation 3.4× gradient) identifies a hypothesis-generating subgroup for prospective trial design; hospital-level disagreement did not persist after full reward deconfounding.

Keywords: offline reinforcement learning, reward-embedded confounding, off-policy evaluation, Fitted Q-Evaluation, confounding, antithrombotic therapy, stroke, double machine learning

1. Introduction

Standard FQE initially yielded a positive policy-improvement estimate. Using Fitted Q-Evaluation (FQE) on 44,894 post-2018 acute ischemic stroke patients, our offline reinforcement learning (RL) policy achieved $V(\text{imp}) = +0.0069$ ($p = 0.048$) under standard evaluation and $+0.0101$ ($p = 0.0002$, E-value = 18.37) with an END-augmented reward — two estimates that could support a premature positive policy-improvement claim under current clinical RL evaluation practice. Here $V(\text{imp})$ denotes the FQE-estimated policy-value difference $V(\pi_{\text{RL}}) - V(\pi_{\text{physician}})$. This paper explains why that signal is not robust to reward deconfounding, and why standard external-confounding checks alone do not evaluate this reward-channel pathway.

To understand how this occurred, consider the setting. Selecting antithrombotic therapy after acute ischemic stroke requires decisions that current guidelines leave genuinely uncertain at the individual patient level: when to initiate anticoagulation in atrial fibrillation (AF) patients with large infarcts where early hemorrhagic transformation is a concern, how long to continue dual antiplatelet therapy in non-cardioembolic stroke, and whether to add a statin outside the large-artery atherosclerosis indication. Across the 20 tertiary stroke centers, discharge prescription patterns for the same stroke subtype vary substantially, suggesting practice variation that sequential optimization could in principle reduce. Antithrombotic management for acute ischemic stroke unfolds across acute and discharge decision phases — a natural candidate for offline RL, where an agent observes patient state, selects from a discrete action space, and receives feedback through 3-month functional outcomes. The widespread adoption of electronic health record (EHR) systems has provided the sequential decision trajectories needed to train offline RL policies from observational registries without prospective randomization, attracting substantial clinical interest following claims of 10–32 % improvements over physician decisions (Komorowski et al., 2018; Choi et al., 2024; Ghasemi et al., 2025).

Several highly cited clinical RL papers share three methodological limitations that can inflate apparent policy improvement: (1) reliance on direct Q-value evaluation rather than Fitted Q-Evaluation (FQE), which we find overestimates policy value by 5–131× on our EHR data (Figure 1, inset), consistent with Tang and Wiens (2021); (2) no reward deconfounding, leaving prognostic severity signals embedded in the proxy reward so that sicker patients who receive more intensive treatment drive an apparent improvement signal; and (3) no E-value analysis (VanderWeele and Ding, 2017), providing no bound on the strength of confounding required to explain the observed effect.

We characterize an RL-specific failure mode that falls outside the estimand targeted by E-value sensitivity analysis. Intuitively, the standard reward function based on 3-month modified Rankin Scale (mRS) encodes how sick the patient was at baseline rather than how well the treatment worked: sicker patients systematically receive more intensive treatment and have worse outcomes, and the resulting value estimate can mistake this severity–outcome correlation for a treatment effect. Formally, we term this *reward-embedded confounding*: the proxy reward carries baseline severity/prognosis information in addition to any treatment-efficacy signal, so the RL value estimate may mistake the severity–outcome pathway for treatment benefit. Crucially, r_{END}'s estimate passes an *external* unmeasured-confounding sensitivity check (E-value = 18.37), yet a 2×2 factorial experiment reveals that terminal

reward confounding alone exceeds the full signal magnitude — i.e., removing terminal confounding more than eliminates the observed improvement, because the raw 3-month mRS encodes the severity→outcome pathway that gradient boosting machine (GBM) residualization removes (Figure 2). The E-value is therefore necessary but not sufficient here: it targets *external* confounding, whereas this bias is carried through the reward channel rather than only through treatment assignment.

This paper uses the offline RL evaluation pipeline as a diagnostic instrument — not to deploy a clinical policy, but to systematically expose evaluation pitfalls that can produce premature positive policy-improvement claims. We make five contributions:

1. **Reward-embedded confounding**: We formalize and decompose this failure mode via 2×2 factorial experiments, showing terminal confounding accounts for 218.6 % of the $r_{\text{END}} \to r_{\text{full}}$ net change (Figure 2). A T-learner provides non-RL triangulation through substantial attenuation and a clinically small residual signal.
2. **Limits of E-value analysis for policy improvement**: We apply E-value sensitivity analysis to policy improvement estimates, demonstrating that external-confounding sensitivity checks are insufficient to address reward-embedded confounding (the FQE signal falls from +0.0101 to +0.0025, $p = 0.291$, under full deconfounding).
3. **6-step evaluation checklist**: An empirically motivated checklist with pre-specifiable diagnostic checks (Gottesman et al., 2019) (Table 2). Applied retrospectively, three highly cited clinical RL papers do not report the FQE confirmation specified by Step 1.
4. **Largest-scale FQE vs. direct Q-evaluation**: Largest systematic quantification of direct Q overestimation (5–131× across ∼450 training runs), confirming Tang and Wiens (2021) on real EHR data.
5. **From null to actionable**: Stratification by National Institutes of Health Stroke Scale (NIHSS) severity reveals a 4.6× conditional average treatment effect (CATE) gradient for hypothesis generation and prospective randomized controlled trial (RCT) design.

Related Work

Deep RL for clinical decision-making gained prominence with Komorowski et al. (2018), whose AI Clinician reported weighted importance sampling (WIS)-estimated survival gains for intensive care unit (ICU) sepsis; Choi et al. (2024) and Ghasemi et al. (2025) followed with claims of 10-percentage-point and +32 % improvements, respectively. Methodological audits have since questioned whether these gains reflect true treatment-benefit signals (Luo et al., 2024). Tang and Wiens (2021) demonstrated that FQE achieves Spearman $\rho = 0.89$ for policy ranking while direct Q-value estimates achieve $\rho < 0.3$; our experiments corroborate this on 44,894 real EHR trajectories (5–131× overestimation). Action imbalance further compounds the problem: Nambiar et al. (2023) quantified ratios up to 6,400:1, rendering importance sampling (IS)-based off-policy evaluation (OPE) unreliable before confounding is considered.

Prior work on confounding in sequential OPE (Namkoong et al., 2020; Kallus and Zhou, 2020; Lu et al., 2018) addresses only *external* unmeasured confounders; our focus is the reward-definition problem that arises when an observational clinical outcome is used as the terminal reward. Confounding by indication is well-established in pharmacoepidemiology (Salas et al., 1999), and reward shaping theory establishes conditions under which

reward transformations preserve optimal policies (Ng et al., 1999) — but when the reward itself encodes a confounding pathway, these invariance guarantees do not apply. This mechanism is also related to proxy-reward misspecification: a clinically convenient proxy endpoint may optimize a prognostic correlate rather than the treatment-relevant outcome of interest. We term this RL-specific failure mode *reward-embedded confounding* and provide an empirical factorial decomposition. Double/Debiased Machine Learning (DML) (Chernozhukov et al., 2018) enables causal estimation from observational data; we adapt its outcome-partialling logic to prognosis-residualized reward evaluation as a diagnostic stress test, not as a standalone causal treatment-benefit estimand (see §2.4).

Gottesman et al. (2019) identified seven evaluation pitfalls for RL in healthcare, including inappropriate OPE selection and insufficient confounding analysis. The three highly cited clinical RL publications we review (Komorowski et al., 2018; Choi et al., 2024; Ghasemi et al., 2025) — including one predating Gottesman et al.'s recommendations — do not report the FQE confirmation specified by Step 1 of our checklist. Our 6-step checklist operationalizes Gottesman et al.'s recommendations with concrete numerical diagnostics; external validation of this checklist is future work.

Generalizable Insights about Machine Learning in the Context of Healthcare

1. **Audit the reward channel separately.** When an observational outcome defines the reward, baseline prognosis can remain in the value after standard OPE. Raw-versus-residualized factorials test this pathway, which is distinct from treatment-assignment confounding; residualization is a diagnostic stress test, not a causal remedy.
2. **Align endpoints with the treatment mechanism.** Functional outcome can be a weak proxy for secondary prevention; our direct AIPW analysis found no significant reduction in ischemic-stroke recurrence ($p = 0.12$–0.23), without implying that every clinical endpoint was null.
3. **Match model complexity to the data.** With 98.3 % state invariance, simpler causal methods may be more appropriate than a sequential policy model.
4. **Pre-specify layered evaluation checks.** Our six-step checklist combines FQE confirmation, support, external- and reward-channel confounding, non-RL triangulation, and subgroup stability. It is motivated by one Korean registry; external validation remains future work.

2. Methods

2.1. Dataset

We used the Clinical Research Collaboration for Stroke in Korea (CRCS-K) registry, a prospective multicenter registry of consecutive ischemic stroke patients admitted to 20 tertiary hospitals across South Korea from 2008 to 2023 (Bae and CRCS-K Investigators, 2022). Registry data use was approved under the institutional research agreement governing CRCS-K multicenter data sharing. Registry enrollment and research use of clinical data were approved by the institutional review boards of the participating centers, with written informed consent obtained from patients or their legally authorized representatives. The present secondary analysis of deidentified CRCS-K registry data was approved by the Institutional

Review Board of Seoul National University Bundang Hospital (IRB No. B-2308-845-302). For main analyses we restricted to post-2018 admissions ($N = 44{,}894$), following the 2018 Korean Stroke Society guideline update that standardized antithrombotic practice (Appendix B). Functional outcome was measured by 3-month modified Rankin Scale (mRS, 0–6; 90.2 % coverage; the analyzed cohort has 100 % by construction). Three cohort sizes appear throughout: $N = 44{,}894$ (primary RL cohort); $N = 44{,}974$ (T-learner cohort, 80 additional patients with valid mRS but incomplete MDP episodes; Appendix R); $N = 35{,}744$ (1-year mRS subcohort; Appendix O). Cohort demographics are in Appendix Table 3.

We defined six antithrombotic treatment actions from discharge medication records (distributions in Appendix Table 3): dual antiplatelet (AP_dual, 58.9 %), anticoagulant+statin (AC_statin, 16.2 %), antiplatelet+statin (AP_statin, 15.1 %), conservative (5.6 %), anticoagulation alone (AC, 2.5 %), and antiplatelet monotherapy (AP_mono, 1.7 %). Statin co-prescription is included because it co-varies systematically with antithrombotic choice in post-stroke prescribing (statin rates differ by >20 percentage points across antithrombotic categories). This 35:1 action imbalance has direct implications for IS validity (§2.5; Appendix C), consistent with Nambiar et al. (2023).

2.2. MDP Formulation

We formulated a 2-step episodic Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$ corresponding to the acute ($t = 0$) and discharge ($t = 1$) phases. The *state space* $\mathcal{S} \subset \mathbb{R}^{421}$ consists of 421 clinical features (admission demographics, 14 lab values, NIHSS subitems, pre-stroke medication history, AI imaging features; Appendix L). The *action space* $\mathcal{A} = \{0, \ldots, 5\}$ (hereafter Config A) is the six antithrombotic strategies defined above. The discount factor is $\gamma = 0.99$. The *reward function* $\mathcal{R}$ defines 14 variants (§2.4), six of which are central to our analysis. An anchor experiment confirms that the correction from an earlier 3-step to this 2-step formulation does not materially affect results ($V(\text{imp}) = +0.0024$ vs. $+0.0020$, $p = 0.94$; Appendix I).

2.3. Offline RL Algorithms

We trained five offline RL algorithms: Conservative Q-Learning (CQL, $\alpha \in \{0.5, 1, 2, 4, 8\}$) (Kumar et al., 2020) as the primary algorithm, Batch-Constrained Q-Learning (BCQ, $f \in \{0.1, 0.3, 0.5\}$) (Fujimoto et al., 2019), Implicit Q-Learning (IQL, $\tau \in \{0.7, 0.8, 0.9\}$) (Kostrikov et al., 2022), Decision Transformer (DT) (Chen et al., 2021), and Behavior Cloning (BC).

In the common-reward screen, none of the five families showed positive improvement evidence (Table 1; Appendix E): CQL, BCQ, and BC FQE intervals included zero, IQL FQE was below physicians, and DT's direct value estimate collapsed to the physician-value baseline (FQE was unavailable for DT). The design was partially crossed: post-2018 reward deconfounding used only the prespecified primary CQL $\alpha = 4.0$, selected for FQE TD-error convergence across reward variants. Training used `d3rlpy` v2.0 (Seno and Imai, 2022) on one NVIDIA H100 (≈350 GPU-hours; hyperparameters in Appendix K); all 33 experiments are mapped in Appendix D.

2.4. Reward Design

We designed 14 reward variants following a progressive confounding-removal strategy. Eight preliminary variants (raw mRS, NIHSS-adjusted, binary, severity-stratified, etc.) are superseded by the six variants central to our analysis:

- $r_{\mathbf{std}}$ (standard evaluation): intermediate reward $r_t = f(\text{NIHSS}_{24h}, \text{NIHSS}_0)$ (missing in ≈80 % of records; defaults to zero when absent) and terminal reward $r_T = -\text{mRS}_{3\text{mo}}/6 - 0.3 \cdot \mathbb{1}[\text{death}] + 0.1 \cdot \mathbb{1}[\text{mRS} \leq 1]$. This represents typical practice in clinical RL.

- $r_{\mathbf{DML}}$ (DML reward residualization): terminal reward replaced by $r_T = -(\text{mRS}_{3\text{mo}} - \hat{f}(\mathbf{x}_0))/6$, where $\hat{f}$ is a LightGBM model trained on 743 treatment-free baseline features via cross-fitting ($R^2 = 0.70$, calibration slope = 1.03, zero treatment variables in SHAP top-20). We adopt the term r_{DML} for conciseness, noting that our approach applies only the outcome-residualization component of Double/Debiased ML (Chernozhukov et al., 2018); treatment residualization is omitted because near-universal statin prescription renders the treatment-partialling step numerically degenerate. We use the residualized reward as a prognosis-residualized diagnostic stress test, not as a standalone causal treatment-benefit estimand (details in Appendix U). Semi-synthetic calibration suggests 24–26 % absorption, while a more conservative propensity-based bound gives 27 %; main-text claims use the conservative bound (Appendix M).

- $r_{\mathbf{END}}$ ($\lambda = 0.2$): r_{std} terminal reward augmented with intermediate penalty for early neurological deterioration (END): $r_0 = r_0^{\text{std}} - \lambda \cdot \text{END}_{\text{flag}}$, where END is ≥4-point NIHSS worsening within 72 hours.

- $r_{\mathbf{full}}$ (full deconfounding): residualizes both reward components against baseline prognosis. The terminal reward is identical to r_{DML}, $r_T = -(\text{mRS}_{3\text{mo}} - \hat{f}(\mathbf{x}_0))/6$. The intermediate reward is $r_0 = r_{0,\text{std}} - \lambda(\text{END}_{\text{flag}} - \hat{g}(\mathbf{x}_0))$, where $\hat{g}$ predicts END from the same treatment-free baseline features. For the factorial decomposition, r'_{END} denotes the cell with the r_{END} intermediate reward and the r_{DML} terminal reward.

- $r_{\mathbf{1yr}}$ (1-year outcome): terminal reward $r_T = -(\text{mRS}_{1\text{yr}} - \bar{m}_b)/6$, where $\bar{m}_b$ is the cohort-mean $\text{mRS}_{1\text{yr}}$ used to center the reward, with NIHSS 24-hour intermediate reward. Uses 35,744 patients with 1-year follow-up.

- $r_{\mathbf{1yr,DML}}$ (DML-residualized 1-year outcome): terminal reward $r_T = -(\text{mRS}_{1\text{yr}} - \hat{g}(\mathbf{x}_0))/6$, where $\hat{g}$ is a LightGBM model predicting $\text{mRS}_{1\text{yr}}$ (cross-fit $R^2 = 0.75$, calibration slope = 1.03, zero treatment variables in SHAP top-20). $r_{1\text{yr}}$ and $r_{1\text{yr,DML}}$ form a second factorial pair crossing outcome horizon with deconfounding.

The diagnostic mechanism is that baseline severity $\mathbf{X}$ causes both treatment T and outcome Y; the standard reward $r = f(Y)$ conflates the prognostic ($\mathbf{X} \to Y$) and treatment-attributable ($T \to Y$) pathways. GBM residualization subtracts baseline prognosis and tests whether the apparent policy advantage survives removal of this pathway. SHAP validation confirms zero treatment variables in the GBM top-20 (Appendix U).

2.5. Evaluation Hierarchy

We adopt a six-step evaluation checklist (Table 2). As the primary off-policy evaluation method we use **Fitted Q-Evaluation (FQE)** (Tang and Wiens, 2021), which fits a separate Q-function to evaluate a fixed target policy. All results use 15-fold cross-validation (5 folds $\times$ 3 random seeds). The primary metric is $V(\text{imp}) = V(\pi_{\text{RL}}) - V(\pi_{\text{physician}})$ evaluated under FQE.

To quantify confounding robustness we apply the E-value framework (VanderWeele and Ding, 2017) to the observed $V(\text{imp})$.

Structural limitation. The E-value addresses only *external* unmeasured confounders ($U \to T$, $U \to Y$). Reward-embedded confounding operates through observed baseline severity *encoded within the reward itself* — a reward-channel pathway that the E-value does not evaluate. The 2×2 factorial decomposition (§3.2) serves as the primary diagnostic.

Approximate computation. We map $V(\text{imp})$ to RR scale via $\widehat{RR} = \exp(|V(\text{imp})|/\text{SE}_{\text{cons}})$, where $\text{SE}_{\text{cons}} = \text{SD}_{\text{folds}}/\sqrt{3}$ (conservative). This transformation provides a conservative upper bound on confounding strength required to explain away $V(\text{imp})$; derivation and justification appear in Appendix J.

We employ four additional triangulation strategies (doubly robust OPE (DR-OPE), instrumental variable (IV), difference-in-differences (DiD), T-learner / direct recurrence) detailed in §3.2 and Appendices C–H.

2.6. Cross-validation and Statistical Testing

All experiments use patient-level 5-fold cross-validation repeated across 3 random seeds (15 evaluations per configuration; see Appendix V for cross-validation details). Significance of $V(\text{imp})$ is assessed via one-sample t-test against zero (df = 14); factorial effects via paired t-tests ($N = 15$ pairs). We report 95 % CI with $\alpha = 0.05$; Bonferroni correction across 14 reward variants ($\alpha_{\text{adj}} = 0.0036$) would leave r_{END} significant but not r_{std} — our argument does not depend on r_{std} significance. Power analysis and the Nadeau–Bengio variance correction are discussed in §5.2.

3. Results

We present results in three acts that mirror the diagnostic sequence, followed by an analysis of clinical heterogeneity. All $V(\text{imp}) = V(\pi_{\text{RL}}) - V(\pi_{\text{physician}})$ estimates are based on FQE with 15-fold CV unless stated otherwise. Table 1 compactly summarizes the algorithm screen and six key reward variants.

3.1. Act 1: Standard Offline RL Evaluation Yields Positive Signal

MDP structure diagnosis leaves FQE as the least problematic primary evaluator here, not an assumption-free one (including Q-function realizability; §5.3): IS-ESS = 0.21 % (48× below threshold) rules out IS-based OPE (Appendix S; direct Q overestimates FQE by 5–131× across ∼450 runs; Figure 1, inset).

Under standard practice, r_{std} produces $V(\text{imp}) = +0.0069$ (95 % CI $[+0.0001, +0.014]$, $p = 0.048$, E-value = 4.70). Adding an END intermediate reward (r_{END}, $\lambda = 0.2$) strength-

Table 1: Compact evaluation robustness summary: a common-reward screen across five algorithm families, followed by six FQE reward designs for primary CQL α=4.0 (15-fold CV, 95 % CI from t-distribution, 14 df). E-values are approximate upper bounds computed via a non-standard continuous-to-RR mapping (§2.5); Bonferroni-corrected $\alpha_{\text{adj}} = 0.0036$ would leave r_{END} significant but not r_{std}.

Reward	Description	V(imp)	95 % CI	p	E-value
Five families	Common reward (FQE except DT; Appendix E)		No positive evaluated evidence		
3-month outcome (N=44,894)					
r_{std}	NIHSS-bin FQE (standard)	+0.0069	[+0.0001, +0.014]	0.048	4.70
r_{END}	+END intermediate (λ=0.2)	+0.0101	[+0.006, +0.014]	0.0002	18.37
r_{DML}	GBM residualization (DML)	+0.0033	[−0.001, +0.008]	0.132	3.50
r_{full}	Full deconfound. (r_{DML} + END resid.)	+0.0025	[−0.002, +0.007]	0.291	2.65
1-year outcome (N=35,744)					
r_{1yr}	NIHSS-bin, mRS_{1yr} terminal	+0.0133	[+0.006, +0.020]	0.001	11.44
$r_{\text{1yr,DML}}$[†]	GBM residualized mRS_{1yr}	−0.0004	[−0.005, +0.004]	0.843	—[‡]

[†]$r_{\text{1yr,DML}}$ results are from Phase 12 (post-2018 cohort restricted to 2018–2023 admissions with 1-year follow-up, N = 35,744). [‡]E-value not computed (V(imp) < 0; E-value is defined only for positive point estimates).

ens the signal to V(imp) = +0.0101 (p = 0.0002, E-value = 18.37) — an estimate that a reader applying standard practice could interpret as strong evidence for policy improvement.

Direct Q-evaluation yielded estimates 5–131× larger than FQE across ∼450 runs (Figure 1, inset), consistent with Tang and Wiens (2021). As a rough illustration, if our overestimation factor generalized, studies reporting V(imp) = +0.10 to +0.32 via direct Q would report +0.001 to +0.003 under FQE — though this extrapolation is speculative, as the ratio is dataset- and algorithm-dependent. The natural question is whether the positive FQE signal reflects treatment benefit or reward-embedded confounding; Act 2 examines this directly.

3.2. Act 2: Progressive Deconfounding Attenuates the Signal to Null

GBM reward residualization (r_{DML}) replaces the raw terminal reward with the residual after subtracting LightGBM-predicted prognosis (cross-fit R^2 = 0.70, calibration slope = 1.03, 0 treatment variables in SHAP top-20). The signal attenuates: V(imp) = +0.0033 (95 % CI [−0.001, +0.008], p = 0.132), a 52 % reduction from r_{std}. Over-adjustment risk is bounded conservatively at 27 % (Appendix M), and the attenuation-adjusted estimate remains clinically small ($V(\text{imp})_{\text{adj}} \approx +0.0045$, ≤ 2.7 % MCID).

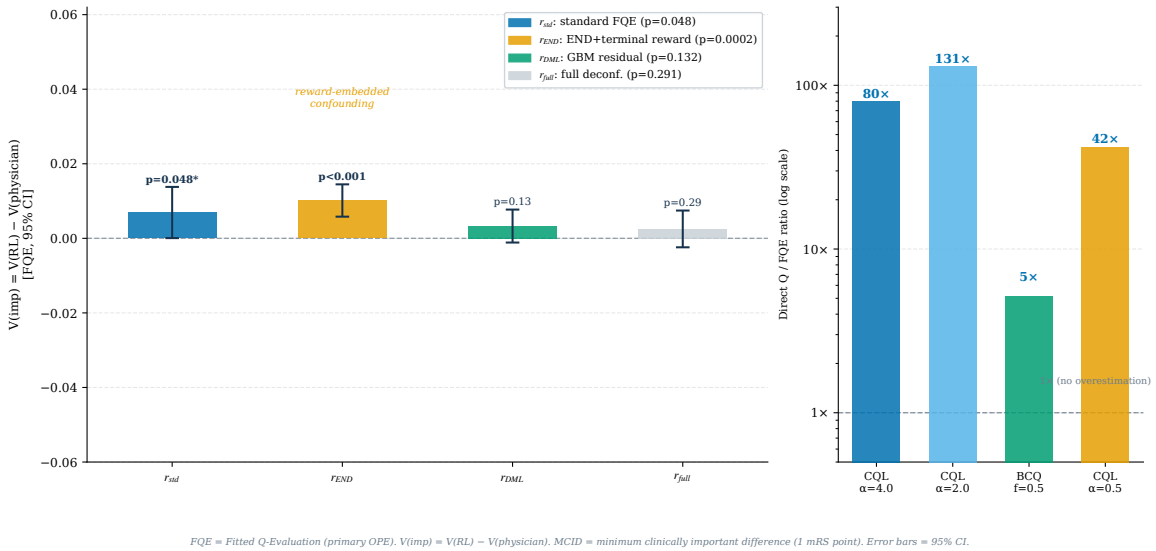
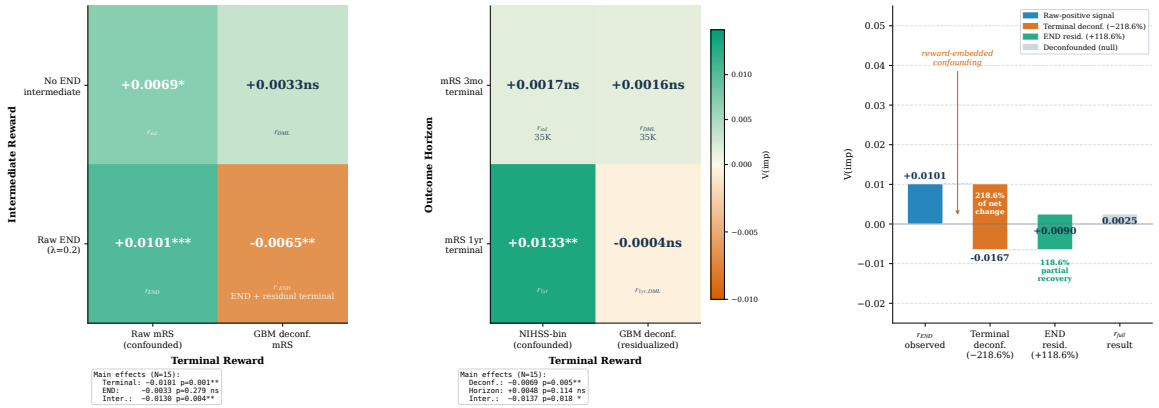


Figure 1: **Confounding cascade across reward designs.** FQE-based $V(\text{imp})$ for four reward variants (CQL α=4.0, 15-fold CV). The apparent improvement attenuates progressively as prognostic confounding is removed ($r_{\text{std}} \to r_{\text{DML}} \to r_{\text{full}}$). r_{END} ($+0.0101$, $p = 0.0002$, E-value = 18.37) falls to $r_{\text{full}} = +0.0025$ ($p = 0.291$) under full reward deconfounding (§3.2). *Inset:* Direct Q overestimates FQE by 5–131× (log scale). Error bars: 95 % CI.

Full deconfounding (r_{full}) yields $V(\text{imp}) = +0.0025$ (95 % CI $[-0.002, +0.007]$, $p = 0.291$), consistent with a null effect. Figure 1 shows the complete confounding cascade: r_{std} ($+0.007$) $\to$ r_{DML} ($+0.003$) $\to$ r_{full} ($+0.002$).

To identify the mechanism driving r_{END}'s apparent $+0.010$ signal, we designed a 2×2 factorial experiment crossing terminal reward (raw mRS vs. GBM-residualized) and intermediate reward (raw END vs. GBM-residualized). Here r'_{END} denotes the factorial cell that combines r_{END}'s END-augmented intermediate reward with a GBM-residualized terminal reward. The decomposition reveals the magnitude of reward-embedded confounding: the simple effect of terminal residualization within the END-augmented reward ($r_{\text{END}} \to r'_{\text{END}}$) is $\Delta V(\text{imp}) = -0.0167$, **accounting for 218.6 % of the net $r_{\text{END}} \to r_{\text{full}}$ change** (computed from unrounded fold-level estimates: $\frac{V(\text{imp})(r'_{\text{END}}) - V(\text{imp})(r_{\text{END}})}{V(\text{imp})(r_{\text{full}}) - V(\text{imp})(r_{\text{END}})} = \frac{-0.01668}{-0.00763} = 218.6\,\%$). The percentage exceeds 100 % because terminal residualization alone overshoots null; the residualized intermediate reward partially offsets this ($\Delta V(\text{imp}) = +0.0090$, $p = 0.279$). The main effect of terminal deconfounding is -0.0101 ($p = 0.001$); the significant interaction (-0.0130, $p = 0.004$) indicates that END amplifies terminal confounding through a shared prognostic pathway (severity $\to$ END $\to$ mRS; Figure 2).

A second factorial (Phase 12, $N = 35{,}744$) crosses outcome horizon ($\text{mRS}_{3\text{mo}}$ vs. $\text{mRS}_{1\text{yr}}$) with deconfounding. The 1-year reward ($r_{1\text{yr}}$) produces $V(\text{imp}) = +0.0133$ ($p = 0.001$, E-value = 11.44); GBM residualization ($r_{1\text{yr,DML}}$) attenuates the signal by **103.3 %**, reducing to -0.0004 ($p = 0.843$). The deconfounding main effect (-0.0069, $p = 0.005$; from fold-level



(A) Phase 10 factorial (END×Terminal, N=44,894). (B) Phase 12 factorial (Horizon×Deconf., N=35,744). (C) r_END-to-r_full waterfall attribution. Both factorials support a raw-positive, residualized-null pattern.

Figure 2: **Dual 2×2 factorial decomposition.** *(A) Phase 10:* Crossing terminal reward (raw vs. GBM-residualized $\text{mRS}_{3\text{mo}}$) and END intermediate handling on N=44,894. Terminal deconfounding main effect: -0.0101 ($p = 0.001$). *(B) Phase 12:* Crossing outcome horizon with deconfounding on N=35,744. Deconfounding main effect: -0.0069 ($p = 0.005$); $r_{1\text{yr}}$'s $+0.0133$ reduces to -0.0004 after GBM residualization (103.3 % attenuation). *(C) Waterfall:* Terminal confounding accounts for **218.6 %** of the net $r_{\text{END}} \rightarrow r_{\text{full}}$ signal change ($r_{\text{END}} = +0.0101$, $r'_{\text{END}} = -0.0065$, $r_{\text{full}} = +0.0025$).

unrounded estimates) supports the interpretation that the raw-positive, residualized-null pattern is not specific to the 3-month mRS horizon in this registry (Figure 2B).

r_{END}'s E-value of 18.37 is mathematically correct for *external* unmeasured confounders within that sensitivity framework. The confounding is instead *internal* to the reward function: the raw 3-month mRS encodes the severity→outcome prognostic pathway that GBM residualization removes. We term this *reward-embedded confounding*: the proxy reward carries a non-treatment-attributable prognostic pathway, so the learned policy can appear favorable even when the FQE signal is not robust to reward deconfounding. E-values remain informative for external unmeasured confounding but do not address this reward-channel pathway; factorial decomposition or prognosis-residualized reward evaluation is needed as a diagnostic stress test.

Causal triangulation. Two independent approaches corroborate the deconfounding result. A T-learner (Künzel et al., 2019) (N = 44,974; Appendix R) shows 75 % signal attenuation under r_{DML} residualization in the full cohort and 90 % in the overlap zone ($0.1 < e(X) < 0.9$, n = 9,900), with a clinically negligible residual (< 6 % MCID). Direct recurrence analysis via inverse probability weighting (IPW) / augmented IPW (AIPW) (N = 32,583) finds no significant ischemic stroke recurrence reduction for any treatment comparison (AIPW p = 0.12–0.23; Appendix H). DR-OPE serves as a consistency check; IV is uninformative (exclusion restriction violation; Appendix F); DiD excluded for parallel trends violation (Appendix G).

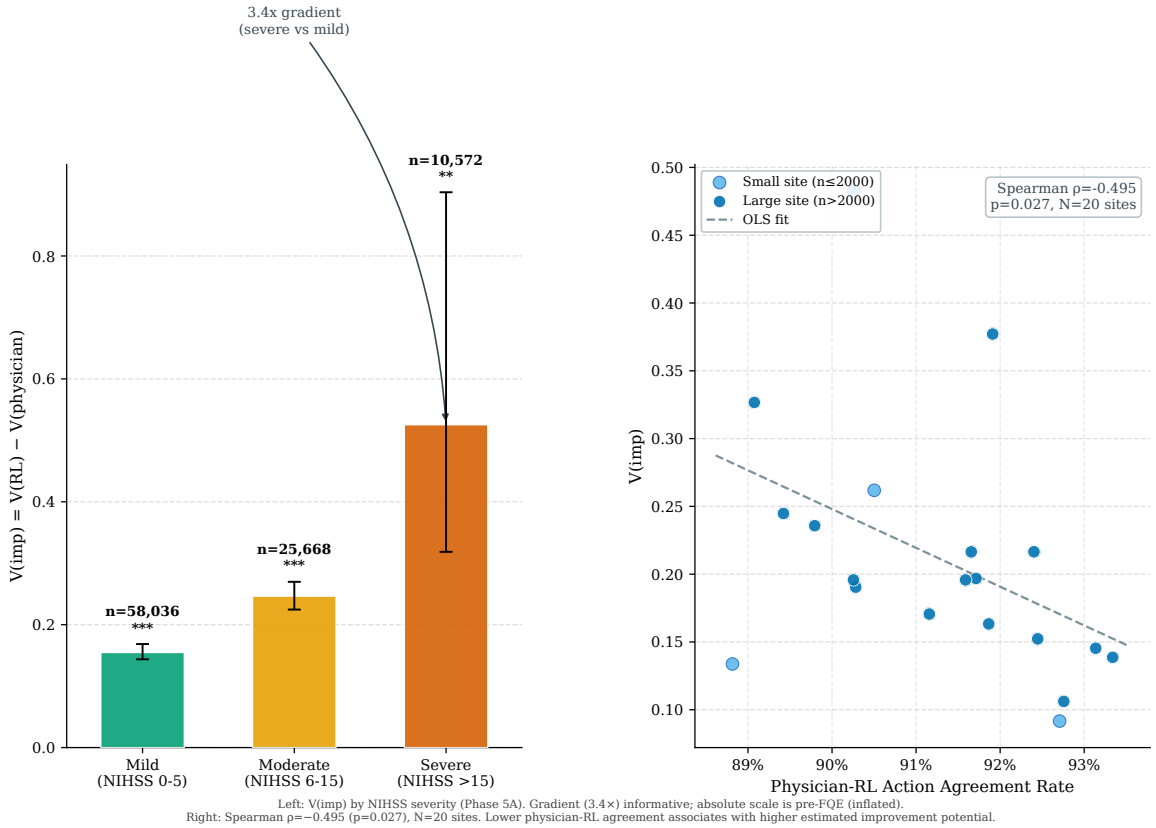


Figure 3: **Hypothesis-generating clinical heterogeneity after deconfounding diagnostics.** *Left:* $V(\text{imp})$ by NIHSS severity (Phase 5A direct Q-evaluation, $N = 116{,}215$; absolute values unreliable but 3.4× relative gradient is directionally consistent with the T-learner's 4.6× CATE ratio). *Right:* Hospital-level $V(\text{imp})$ vs. physician–RL agreement across 20 tertiary centers. The raw association is hypothesis-generating only and does not persist under r_{full} (see §3.4).

3.3. MDP Structure Diagnosis

MDP structure analysis confirms that 98.3 % of state features are invariant across timesteps (Pearson $r > 0.99$), with per-patient Q-value rank $\rho = 0.82$ (Figure 5, left). IS-based OPE is infeasible (ESS = 0.21 %, 48× below threshold; Figure 5, right), leaving FQE as the least problematic OPE method in this setting (see Appendix S).

3.4. Clinical Heterogeneity: Hypothesis Generation for Prospective Evaluation

Having established that the overall policy improvement is null after reward deconfounding, we turn to whether clinically relevant heterogeneity remains as a trial-design signal — not to claim treatment benefit, but to identify subgroups for prospective evaluation.

Despite the overall null result after deconfounding, the NIHSS severity gradient persists as an exploratory, hypothesis-generating pattern.

The T-learner identifies a **4.6× severe/mild CATE ratio** across NIHSS severity subgroups, directionally consistent with a 3.4× $|V(\text{imp})|$ gradient under direct Q-evaluation (Figure 3). Directional consistency across two independent methods measuring different estimands (CATE vs. policy value) supports hypothesis-generating clinical heterogeneity rather than relying on a single policy-value estimate. Hospital-level $V(\text{imp})$ under direct Q-evaluation correlates negatively with physician–RL agreement ($\rho = -0.495$, $p = 0.027$, $N = 20$; jackknife direction-stable; Figure 3); however, this correlation vanishes under the deconfounded reward ($\rho \approx 0$, ns; recomputed estimate $\rho = -0.026$, $p = 0.915$), suggesting an unstable, confounding-driven site signal rather than surviving site-level practice variation. We therefore focus heterogeneity interpretation on the NIHSS-stratified T-learner estimates as the more reliable estimand. Policy disagreement profiling reveals that 57 % of RL–physician disagreements involve statin co-prescription, with AF history as the strongest driver (Appendix T). These patterns suggest that antithrombotic–statin combination decisions in severe-to-moderate NIHSS patients constitute the highest-priority subgroup for prospective evaluation.

4. A 6-Step Evaluation Checklist for Clinical Offline RL

Drawing on the diagnostic sequence revealed in our study, we propose a six-step empirical evaluation checklist that operationalizes Gottesman et al. (2019)'s seven evaluation pitfalls with pre-specifiable diagnostic checks (Table 2). The checklist is intended to identify when a positive $V(\text{imp})$ estimate needs additional evidence before it is interpreted as clinical benefit; external validation of the checklist remains future work.

Retrospective application to three widely cited clinical RL papers (Komorowski et al., 2018; Choi et al., 2024; Ghasemi et al., 2025) shows that Step 1 could not be assessed as satisfied under this checklist: none reports FQE or equivalent model-based OPE. Steps 4–6 were not reported in these papers, leaving positive findings unexamined for reward-channel confounding. We acknowledge this checklist was derived from diagnostic experience rather than pre-registered or externally validated; external validation is future work. The data and code availability statement specifies the permitted release mechanism for analysis code, configuration files, figure-generation scripts, and a synthetic example dataset designed to exercise the analysis pipeline without exposing restricted patient-level registry records.

5. Discussion

5.1. When Can Offline RL Add Value in Clinical Settings?

Our analysis suggests three conditions under which offline RL is most likely to add value beyond simpler causal inference: (a) **reward-channel diagnostics** (proxy rewards should be stress-tested for prognostic confounding; this is the diagnostic condition emphasized here), (b) **meaningful sequential dynamics** (state features must evolve between decision steps; our 98.3 % invariance limits optimization to near-1-step), and (c) **action support balance** (IS-ESS $\geq$ 10 % threshold; our 2.5 % AC prescription rate drives IS-ESS collapse to 0.21 %, making FQE the least problematic OPE choice here). In settings that satisfy all three, RL may provide value beyond simpler methods. In our case, several key findings were discoverable *only* through the RL pipeline: the 5–131× FQE vs. direct Q gap, IS-ESS collapse, the

Table 2: 6-Step Evaluation Checklist for Clinical Offline RL. ✓ = Met; × = Not met or not assessable; NR = Not reported; NI = Not implemented. Steps 1–2: OPE reliability; Steps 3–4: confounding robustness; Steps 5–6: causal triangulation and generalizability.

Step	Check	Threshold	This Study	Existing Papers
S1	FQE validation	Report FQE; if $\lvert V_{\text{Dir}}/V_{\text{FQE}}\rvert \geq 2\times$, do not interpret direct Q alone	✓ Ratio = 5–131× ⇒ FQE primary	NR: WIS only (Komorowski et al., 2018; Ghasemi et al., 2025); matching (Choi et al., 2024); no FQE
S2	IS viability (ESS ≥ 10 %)	ESS ratio ≥ 0.10	× r_{std} ESS=0.21% (48× below threshold); r_{DML} ESS=2.5%[b] ⇒ FQE least problematic	NR
S3	Confounding robustness (E-value)	Report E-value; necessary but not sufficient — see S4	✓ r_{END} E=18.37 (passes external-confounding check but falls under S4; §3.2); × r_{std} E=4.70 (marginal)	NR
S4	Reward deconfounding	DML/GBM residual; $R^2 \geq 0.70$; 0 treatment vars in SHAP top-20	✓ R^2=0.70, calibration = 1.03, 0 treatment vars	NI
S5	Causal triangulation	≥ 2 independent methods agree on direction	✓ FQE + GBM null; T-learner attenuation 75 % full cohort, 90 % overlap-trimmed (n=9,900); IS recurrence AIPW ns (p=0.12–0.23)	NI
S6	Subgroup stability	Pre-specified or hypothesis-generating subgroups; assess deconfounding stability	△[a] NIHSS gradient 4.6× (T-learner), 3.4× (exploratory Q); site signal does not persist under r_{full}	NI

NR = Not reported; NI = Not implemented. [a]S6 CONDITIONAL: subgroups are exploratory; NIHSS gradient is a trial-design signal, not a treatment recommendation; site ρ does not survive full reward deconfounding ($\rho \approx 0$, recomputed $\rho = -0.026$, $p = 0.915$ under r_{full}). [b]r_{DML} ESS= 2.5% is the IS effective sample size ratio under the GBM-residualized reward policy, distinct from the AC action's 2.5% prescription rate (§2.1); both are coincidentally equal.

factorial decomposition, and the 6-step evaluation checklist. Condition (a) reflects what we term *reward-embedded confounding* (§3.2): a mechanism analogous to confounding by indication in pharmacoepidemiology, where sicker patients receive more aggressive treatment and have worse outcomes. In classical OLS settings this makes treatment appear *harmful*; in our offline RL setting the same severity→outcome pathway is encoded within the ter-

minal reward, making the learned policy appear *favorable* — the sign is reversed because the reward is severity-correlated, not the propensity score. In addition to confounding by indication in treatment assignment, the reward itself carries the severity-to-outcome pathway, which is why this pathway falls outside the E-value's external-confounding estimand. For single-decision settings satisfying only condition (a), Causal Forest (Wager and Athey, 2018) or DML (Chernozhukov et al., 2018) directly estimates CATE without sequential assumptions.

5.2. Power Analysis and Effect Size Bounds

Our study has sufficient statistical power to detect clinically meaningful effects. The retrospective minimum detectable effect (MDE) at 80 % power is $V(\text{imp}) \approx 0.010$ for r_{std} and ≈ 0.006 for r_{DML}, corresponding to **5.9 %** and **3.8 %** of the mRS minimal clinically important difference (MCID = 1.0 mRS point = 0.167 in $V(\text{imp})$ scale; Cranston et al. (2017)). The Nadeau–Bengio correction for training-set overlap raises MDE to 6.6 % MCID — still below clinically meaningful thresholds. Even taken at face value, the deconfounded estimates ($r_{\text{DML}} = +0.0033$, $r_{\text{full}} = +0.0025$) correspond to ≈2.0 % and 1.5 % MCID respectively — well below the clinically relevant threshold we are powered to detect (> 3.8 % MCID). The null conclusion therefore reflects a statistically null and clinically small estimated policy advantage, not insufficient power: we had sufficient sensitivity to detect effects at or above this clinically meaningful range, although smaller effects remain possible.

The near-null result reflects a specific property of well-managed tertiary registries: in the post-2018 cohort, >80 % of AF patients receive anticoagulation, reflecting near-universal guideline concordance that leaves limited scope for RL to improve upon current practice. This conclusion is specific to antithrombotic selection in Korean tertiary stroke registries and does not imply that offline RL cannot improve stroke care in settings with greater treatment heterogeneity, different treatment dimensions (thrombolysis timing, endovascular intervention), or richer data modalities.

5.3. Limitations

1. **External validity.** CRCS-K comprises predominantly Korean patients from tertiary referral centers with high guideline adherence; generalizability to other ethnic populations, community hospitals, and healthcare systems with different practice patterns remains unvalidated. The methodological findings (reward-embedded confounding, direct Q overestimation) may apply to EHR-based offline RL studies that use observational outcomes as rewards.
2. **Degenerate MDP.** Our 2-step MDP has 98.3 % state invariance ($r > 0.99$; of 7 dynamic features, 6 are action one-hot encodings and only NIHSS at 24 hours carries clinical information, missing in ≈80 % of records), reducing sequential optimization to near-1-step behavior.
3. **Over-adjustment risk.** GBM residualization may absorb some true treatment-benefit variation. Semi-synthetic calibration suggests 24–26 % absorption, while a conservative propensity-based bound gives 27 %; the adjusted estimate remains clinically small ($V(\text{imp})_{\text{adj}} \approx +0.0045$, ≤ 2.7 % MCID).

4. **FQE realizability.** FQE assumes the Q-function class can represent the true value function. We cannot rule out realizability violations; ensemble-based disagreement tests would strengthen this evidence (Appendix P).
5. **Fairness.** T-learner CATE shows non-monotone sex×age patterns (Appendix N), likely from extrapolation; these should not be interpreted as demographic disparities without prospective validation.

Four additional limitations (power at the margin, outcome endpoint sensitivity, selection bias, post-2018 cohort restriction) are detailed in Appendix P. Action masking enforces clinical guideline constraints (6 % violation rate, all caught at inference; see Appendix Q).

5.4. Recommendations

For researchers applying offline RL to clinical datasets, we recommend the 6-step empirical evaluation checklist (§4) as a set of pre-specifiable diagnostic checks. For clinical trialists, ELAN and OPTIMAS narrowed anticoagulation timing, but largest-infarct estimates remain imprecise and anticoagulation–statin combinations untested (Fischer et al., 2023; Werring et al., 2024; Nash et al., 2026). The T-learner's 4.6× severe/mild CATE ratio therefore motivates stratified prospective evaluation of this recurrence–hemorrhage trade-off, not treatment benefit or site-level targeting. For data infrastructure, clinically useful multi-step RL requires continuous inpatient monitoring data rather than admission-only registry snapshots.

6. Conclusion

Standard offline RL evaluation for stroke antithrombotic optimization yields $V(\text{imp}) = +0.0101$ ($p = 0.0002$, E-value = 18.37), but full deconfounding reduces it to $+0.0025$ ($p = 0.291$), with terminal residualization accounting for **218.6 %** of the net r_{END}-to-r_{full} change. We term this reward-channel pathway *reward-embedded confounding*; external- confounding E-values do not evaluate it. No family produced positive evaluated evidence in the common-reward screen; in primary CQL, residualization attenuated the estimate to null, with T-learner and recurrence AIPW providing triangulation. Whenever observational outcomes define terminal rewards, positive offline RL results should remain hypothesis-generating until reward-channel diagnostics are applied.

Data and Code Availability

Individual-level CRCS-K data cannot be publicly deposited because of informed consent, ethics, data-sharing, privacy, and regulatory restrictions. Qualified researchers may request a de-identified minimum dataset for a legitimate academic proposal, subject to CRCS-K Steering Committee and institutional approvals; the author cannot independently redistribute it. An artifact package is being prepared for a planned arXiv/PMLR release, including code, configurations, figure scripts, and synthetic examples for implementation inspection. The package will not include individual-level CRCS-K data and cannot reconstruct the cohort or reproduce patient-level results without approved data access.

References

Pierre Amarenco, Julien Bogousslavsky, Alfred Callahan, Larry B. Goldstein, Michael Hennerici, Amy E. Rudolph, Henrik Sillesen, Liljana Simunovic, Michael Szarek, K. Michael A. Welch, and Justin A. Zivin. High-dose atorvastatin after stroke or transient ischemic attack. *New England Journal of Medicine*, 355(6):549–559, 2006. doi: 10.1056/NEJMoa061894.

Hee-Joon Bae and CRCS-K Investigators. David G. Sherman lecture award: 15-year experience of the nationwide multicenter stroke registry in Korea. *Stroke*, 53(9):2976–2987, 2022. doi: 10.1161/STROKEAHA.122.039212.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. URL https://fairmlbook.org.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision Transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*, volume 34, pages 15084–15097, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Young Choi, Seungwon Oh, Jin-Won Huh, Han-Tae Joo, Hyeran Lee, Wanki You, Chang Mo Bae, Ji Ho Choi, and Kyung-Jae Kim. Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. *Communications Medicine*, 4(1):245, 2024. doi: 10.1038/s43856-024-00665-x.

Jessica S. Cranston, Brett D. Kaplan, and Jeffrey L. Saver. Minimal clinically important difference for safe and simple novel acute ischemic stroke therapies. *Stroke*, 48(11):2946–2951, 2017. doi: 10.1161/STROKEAHA.117.017496.

Urs Fischer, Masatoshi Koga, Daniel Strbian, et al. Early versus later anticoagulation for stroke with atrial fibrillation. *New England Journal of Medicine*, 388(26):2411–2421, 2023. doi: 10.1056/NEJMoa2303048.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy Deep Reinforcement Learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2052–2062. PMLR, 2019.

Peyman Ghasemi, Matthew Greenberg, Danielle A. Southern, Bing Li, James A. White, and Joon Lee. Personalized decision making for coronary artery disease treatment using offline reinforcement learning. *npj Digital Medicine*, 8:99, 2025. doi: 10.1038/s41746-025-01498-1.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Evaluating reinforcement learning algorithms in observational health settings. *Nature Medicine*, 25(1):16–18, 2019. doi: 10.1038/s41591-018-0310-5. arXiv:1811.11722 (2018).

S. Claiborne Johnston, J. Donald Easton, Mary Farrant, William Barsan, Robin A. Conwit, Jordan J. Elm, Anthony S. Kim, Anne S. Lindblad, and Yuko Y. Palesch. Clopidogrel and aspirin in acute ischemic stroke and high-risk TIA. *New England Journal of Medicine*, 379(3):215–225, 2018. doi: 10.1056/NEJMoa1800410.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 22293–22304, 2020.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007. doi: 10.1214/07-STS227.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR 2022)*, 2022. arXiv:2110.06169.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1179–1191, 2020.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings, 2018. URL https://arxiv.org/abs/1812.10576. arXiv:1812.10576.

Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Reinforcement learning in dynamic treatment regimes needs critical reexamination. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://arxiv.org/abs/2405.18556. arXiv:2405.18556.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. doi: 10.1023/A:1024068626366.

Mila Nambiar, Supriyo Ghosh, Priscilla Ong, Yu En Chan, Yong Mong Bee, and Pavitra Krishnaswamy. Deep Offline Reinforcement Learning for real-world treatment optimization applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023)*, pages 1666–1677, 2023. doi: 10.1145/3580305.3599800. URL https://arxiv.org/abs/2302.07549. arXiv:2302.07549.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 18819–18831, 2020.

Philip S. Nash, Jonathan G. Best, James Lyon, et al. Early versus delayed anticoagulation in acute ischemic stroke with atrial fibrillation according to infarct volume and location: A prespecified subgroup analysis of the OPTIMAS randomized controlled trial. *International Journal of Stroke*, 2026. doi: 10.1177/17474930261441297.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 278–287, 1999.

Maurizio Paciaroni, Giancarlo Agnelli, Nicola Falocci, Valeria Caso, Cecilia Becattini, Simone Marcheselli, Christoph Rueckert, Alessandro Pezzini, Loris Poli, Alessandro Padovani, Laszlo Csiba, Janos Kármán, Jörg Ludicke, and Hansjörg C. Hopf. Early recurrence and cerebral bleeding in patients with acute ischemic stroke and atrial fibrillation: Effect of anticoagulation and its timing. *Stroke*, 46(8):2175–2182, 2015. doi: 10.1161/STROKEAHA.115.008891.

Margarita Salas, Albert Hofman, and Bruno H. Ch. Stricker. Confounding by indication: An example of variation in the use of epidemiologic terminology. *American Journal of Epidemiology*, 149(11):981–983, 1999. doi: 10.1093/oxfordjournals.aje.a009758.

Takuma Seno and Michita Imai. d3rlpy: An offline deep reinforcement learning library. *Journal of Machine Learning Research*, 23(315):1–20, 2022.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC 2021)*, volume 149 of *Proceedings of Machine Learning Research*, pages 2–35. PMLR, 2021.

Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the E-Value. *Annals of Internal Medicine*, 167(4):268–274, 2017. doi: 10.7326/M16-2607.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Yongjun Wang, Yilong Wang, Xingquan Zhao, Liping Liu, David Wang, Chunxue Wang, Chun Wang, Hao Li, Xia Meng, Liying Cui, Jianping Jia, Qiang Dong, Anxin Xu, Jinsheng Zeng, Yansheng Li, Zhenguo Wang, Haiqing Xia, and S. Claiborne Johnston. Clopidogrel with aspirin in acute minor stroke or transient ischemic attack. *New England Journal of Medicine*, 369(1):11–19, 2013. doi: 10.1056/NEJMoa1215340.

David J. Werring, Hakim-Moulay Dehbi, Norin Ahmed, et al. Optimal timing of anticoagulation after acute ischaemic stroke with atrial fibrillation (OPTIMAS): A multicentre, blinded-endpoint, phase 4, randomised controlled trial. *The Lancet*, 404(10464):1731–1741, 2024. doi: 10.1016/S0140-6736(24)02197-4.

Acknowledgments

No external funding was received. This work was conducted as part of the author's employment at JLK Inc. The author is an employee of JLK Inc., a medical AI company. Data access was provided under the CRCS-K multicenter research agreement.

Appendix A. Cohort Demographics

Table 3: Baseline characteristics of the post-2018 CRCS-K cohort (N = 44,894). Continuous variables are reported as mean ± SD; categorical variables as n (%). [a] 3-month mRS coverage is 100 % in this cohort because patients without outcome are excluded from the Phase 8 MDP dataset (see §2.1). [b] TOAST (Trial of Org 10172 in Acute Stroke Treatment) classification is missing for 10.5 % (n = 4,702); percentages are computed among patients with available TOAST (n = 40,192). [c] Discharge antithrombotic regimen, classified using Config A (§2.2); AC = anticoagulant, AP = antiplatelet. One patient contributed two admissions; the duplicate is assigned to AP_dual.

Characteristic	Value
Demographics	
Age, years	68.2 ± 13.4
Female, n (%)	17,730 (39.5 %)
Admission severity	
Initial NIHSS score	4.0 ± 5.2
Vascular risk factors	
Hypertension, n (%)	29,602 (65.9 %)
Diabetes mellitus, n (%)	14,995 (33.4 %)
Atrial fibrillation, n (%)	7,610 (17.0 %)
Previous stroke, n (%)	9,332 (20.8 %)
TOAST stroke aetiology[b]	
Large-artery atherosclerosis (LAA)	14,951 (37.2 %)
Cardioembolism (CE)	7,383 (18.4 %)
Small-vessel occlusion (SVO)	7,114 (17.7 %)
Incomplete evaluation	4,137 (10.3 %)
Mixed aetiology	3,018 (7.5 %)
Other determined	1,914 (4.8 %)
Undetermined	1,675 (4.2 %)
Discharge antithrombotic regimen[c]	
Dual antiplatelet (AP_dual)	26,430 (58.9 %)
Antiplatelet + statin (AP_statin)	6,758 (15.1 %)
Anticoagulant + statin (AC_statin)	7,292 (16.2 %)
Conservative (none)	2,531 (5.6 %)
Anticoagulant alone (AC)	1,122 (2.5 %)
Antiplatelet monotherapy (AP_mono)	761 (1.7 %)
Outcomes	
3-month mRS, mean ± SD[a]	1.7 ± 1.7
Favourable outcome (mRS 0–2), n (%)	31,739 (70.7 %)
In-hospital death, n (%)	744 (1.7 %)

Appendix B. Temporal Cohort Restriction

The primary RL cohort (N = 44,894) is derived from the full CRCS-K registry (N = 129,033) through the following steps:

1. **Ischemic stroke episodes**: 129,033 consecutive admissions across 20 hospitals (2008–2025) → 116,215 complete 3-step MDP episodes after removing records with missing acute or discharge treatment data.
2. **Post-2018 temporal restriction**: 116,215 → ≈49,800 episodes with admission date ≥ 2018-04-05 (following the 2018 Korean Stroke Society guideline update that standardized dual antiplatelet therapy (DAPT) for non-cardioembolic stroke and expanded non-vitamin K oral anticoagulant (NOAC) indications for AF).
3. **Valid 3-month mRS**: ≈49,800 → 44,895 episodes with non-missing 3-month modified Rankin Scale (≈4,900 excluded for missing mRS, who have higher mean NIHSS = 6.8 vs. 4.0, suggesting mild missing-not-at-random (MNAR) bias).
4. **Complete 2-step MDP episodes**: 44,895 → 44,894 after removing 1 patient lacking complete episode data for the 2-step (acute → discharge) formulation.

Including pre-2018 data introduces guideline-era confounding: treatment patterns differ across eras for reasons unrelated to treatment efficacy, biasing the reward signal. Temporal stability analysis shows that a CQL model trained on pre-2018 data achieves only 65.4 % action agreement on post-2018 holdout patients, confirming distribution shift.

Appendix C. Positivity Analysis Details

The effective sample size (ESS) for importance-sampling OPE is defined as

$$\mathrm{ESS} = \left(\sum_i w_i\right)^2 \Big/ \sum_i w_i^2$$

(Kish's effective sample size (Kish, 1965)), where w_i are the cumulative importance weights. In Phase 8B (r_{std}, CQL α=4.0), ESS = 93.6 patients (out of 44,894, ESS ratio = 0.21 %). Cumulative IS weight maximum reached 9.8×10^{14}. A commonly recommended practical threshold for reliable IS-based OPE is ESS ≥ 10 % (Kang and Schafer, 2007); our ESS is 48× below this threshold. Under the undeconfounded r_{std} reward, clip-to-10 weighted doubly robust (WDR) estimation reduces weight variance but produces $|V(\mathrm{imp})_{\mathrm{DR}} - V(\mathrm{imp})_{\mathrm{FQE}}| = 0.055$, far exceeding the 0.02 agreement threshold, so this diagnostic is not met for r_{std}. Under the deconfounded r_{DML} reward, the same statistic is 0.007, satisfying the Step 2 agreement diagnostic (§3.2).

Appendix D. Experiment Overview: Evidence Map

Figure 4 provides a radial overview of experiments across Phases 1–13. Each arc encodes $V(\text{imp})$ (positive outward, negative inward), and color indicates whether the result is null, an apparent positive signal under a non-residualized reward, or a direct causal-analysis result. Positive signals cluster in non-deconfounded reward designs and attenuate after GBM residualization, visually summarizing the reward-embedded confounding diagnostic sequence.

Table 4: Phase legend for Figure 4. The table maps the phase labels used in the evidence map to the cohort, reward or analysis family, and diagnostic purpose.

Phase(s)	Cohort	Reward / analysis	Purpose
1–2	Initial RL cohorts	Algorithm and OPE screening	Select candidate algorithms and quantify direct-Q versus FQE overestimation.
3	Initial RL cohorts	Action-space variants	Test whether finer antithrombotic action encodings change policy-value conclusions.
4	Initial RL cohorts	Terminal reward variants	Assess reward-formulation sensitivity before reward deconfounding.
5–7	Diagnostic subsets	Subgroup, temporal, E-value, and support diagnostics	Identify heterogeneity, guideline-era instability, external-confounding sensitivity, and IS support limits.
8	Post-2018, N=44,894	r_{std} standard FQE and NIHSS strata	Establish the revised 2-step MDP baseline and the raw-positive policy-improvement estimate.
9	Post-2018, N=44,894	r_{DML} prognosis-residualized terminal reward	Stress-test whether the standard positive signal persists after terminal reward residualization.
10	Post-2018, N=44,894	r_{END}, r'_{END}, r_{full}	Decompose terminal and END-channel confounding with the 2×2 factorial and waterfall.
11–12	1-year follow-up, N=35,744	$r_{1\text{yr}}$ and $r_{1\text{yr,DML}}$	Test whether the raw-positive, residualized-null pattern appears beyond 3-month mRS.
13	Direct endpoint cohorts	IPW/AIPW recurrence and safety endpoints	Triangulate the RL reward findings against treatment-aligned clinical endpoints.

Appendix E. Algorithm × Reward Interaction

The experiment grid was partially rather than fully crossed. We first screened five algorithm families under a common initial-cohort reward, then performed the post-2018 reward-deconfounding factorial with the prespecified primary CQL configuration. Tables 5 and 6 report the evaluated cells explicitly; no values are imputed for untested algorithm–reward combinations.

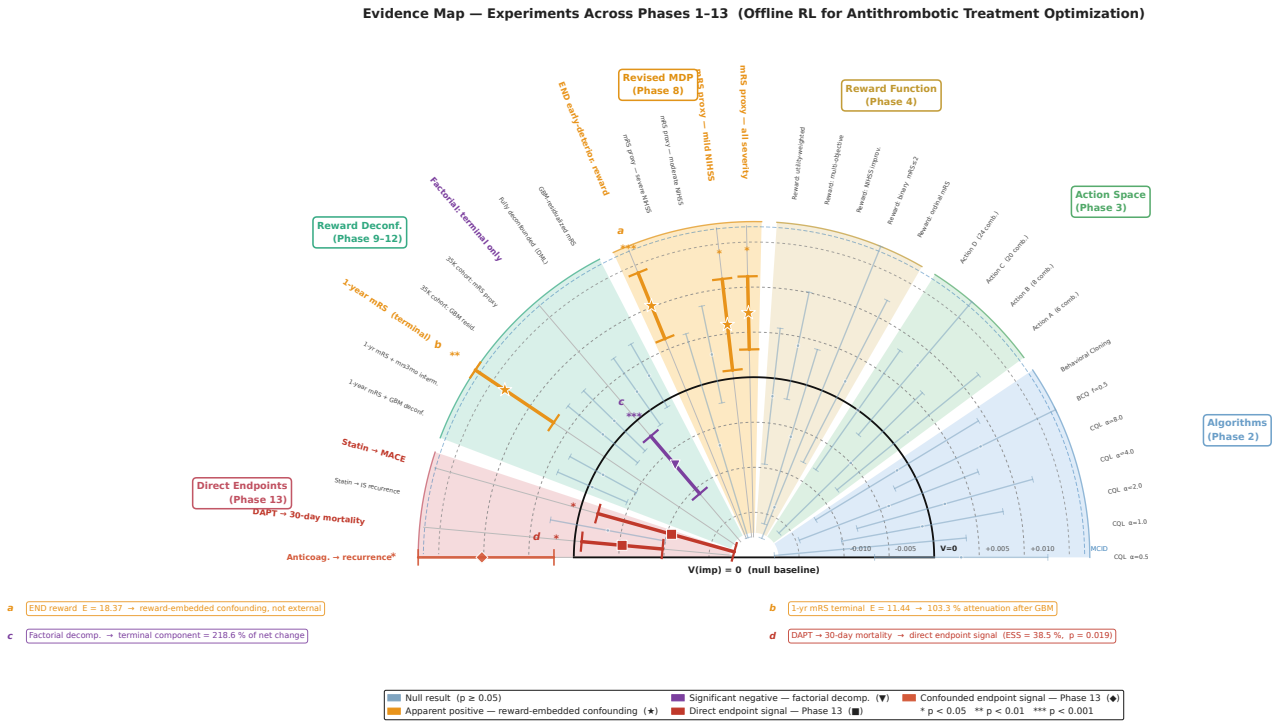


Figure 4: **Evidence map across Phases 1–13.** Each arc represents one experimental comparison; radial position encodes $V(\text{imp})$ (positive outward, negative inward), and color indicates result category. Dashed arc: $V(\text{imp}) = 0$ (physician baseline). Table 4 maps each phase label to its cohort, reward or analysis family, and diagnostic purpose. The map shows that positive signals are concentrated in non-deconfounded reward designs and attenuate after GBM residualization, visually summarizing the reward-embedded confounding diagnostic sequence.

Appendix F. Instrumental Variable Details

We used hospital-level anticoagulation (AC) prescription rate as instrument for AC treatment in the AF subgroup ($N = 6{,}458$). The instrument is strong (first-stage $F = 515.6$) but **fails the exclusion restriction**: 10 of 28 covariate balance checks show significant imbalance ($p < 0.05$), indicating that hospital AC rate correlates with other institutional practices that directly affect mRS outcomes. Because the exclusion restriction is violated, the IV estimate is not causally interpretable; we report it for completeness: LATE $= +0.279$ ($p = 0.712$, 95 % CI $[-1.06, +1.21]$), directionally consistent with the null but unreliable. Endogeneity of treatment is confirmed (2SRI residual test $p < 0.001$), validating the need for causal methods beyond naive OPE.

Table 5: Full common-reward algorithm screen. FQE is reported for CQL, BCQ, IQL, and BC; DT has a direct value estimate because FQE was unavailable. Each configuration has 15 seed–fold evaluations, and none shows positive improvement evidence.

Family	Configuration	V(imp)	95 % CI	Interpretation
CQL	$\alpha = 0.5$	$+0.0025$	$[-0.0071, +0.0122]$	Interval includes zero
CQL	$\alpha = 1.0$	-0.0088	$[-0.0215, +0.0040]$	Interval includes zero
CQL	$\alpha = 2.0$	-0.0012	$[-0.0104, +0.0080]$	Interval includes zero
CQL	$\alpha = 4.0$	$+0.0021$	$[-0.0081, +0.0123]$	Interval includes zero
CQL	$\alpha = 8.0$	-0.0038	$[-0.0146, +0.0070]$	Interval includes zero
BCQ	$f = 0.5$	$+0.0081$	$[-0.0027, +0.0188]$	Interval includes zero
IQL	$\tau = 0.7$	-0.0117	$[-0.0222, -0.0012]$	Below physician value
DT	Default	0.0000	$[0.0000, 0.0000]$	Direct-value collapse; no FQE
BC	Default	-0.0014	$[-0.0120, +0.0091]$	Interval includes zero

Table 6: Full verified reward-stress-test cells for CQL $\alpha = 4.0$. Preliminary rewards use the initial cohort/default MDP; named rewards use the post-2018 two-step cohort. The raw-positive estimates do not persist after prognosis residualization.

Cohort	Reward design	V(imp)	95 % CI	Interpretation
Initial	Ordinal mRS terminal	$+0.0021$	$[-0.0081, +0.0123]$	Interval includes zero
Initial	Binary mRS ≤ 2	-0.0140	$[-0.0453, +0.0173]$	Interval includes zero
Initial	NIHSS-change terminal	-0.0063	$[-0.0166, +0.0039]$	Interval includes zero
Initial	Multi-objective composite	$+0.0042$	$[-0.0020, +0.0104]$	Interval includes zero
Initial	mRS utility weights	-0.0020	$[-0.0095, +0.0055]$	Interval includes zero
Post-2018	r_{std} standard reward	$+0.0069$	$[+0.0001, +0.0140]$	Borderline raw-positive
Post-2018	r_{END} with raw END	$+0.0101$	$[+0.0058, +0.0145]$	Raw-positive, confounded
Post-2018	r_{DML} residual terminal	$+0.0033$	$[-0.0011, +0.0077]$	Interval includes zero
Post-2018	r'_{END} residual terminal/raw END	-0.0065	$[-0.0111, -0.0019]$	Factorial diagnostic cell
Post-2018	r_{full} residual terminal and END	$+0.0025$	$[-0.0024, +0.0075]$	Interval includes zero
1-year	$r_{1\text{yr}}$ raw terminal	$+0.0133$	$[+0.0062, +0.0205]$	Raw-positive
1-year	$r_{1\text{yr,DML}}$ residual terminal	-0.0004	$[-0.0050, +0.0040]$	Interval includes zero

Appendix G. Difference-in-Differences Details

We use the following abbreviations: two-way fixed effects (TWFE), average treatment effect on the treated (ATT), wild cluster bootstrap (WCB), and randomization inference (RI).

DAPT experiment: 39,947 patients, 10 sites, staggered DAPT adoption 2010–2020. TWFE ATT = $+0.134$ ($p = 0.170$); wild cluster bootstrap (WCB) $p = 0.247$; randomization inference $p = 0.438$. Parallel trends violated ($F = 20.58$, $p < 0.001$) — results are exploratory.

NOAC experiment: 13,387 AF patients, 15 sites, NOAC adoption 2015–2016. TWFE ATT = $+0.211$ ($p = 0.071$); WCB $p = 0.104$. Binary mRS$\leq$2 outcome: TWFE $p = 0.021$, WCB $p = 0.110$, RI $p = 0.757$ — estimate unstable across inference methods. Seven sensitivity checks: 3 pass (placebo, binary, statin), 2 mixed (age stratification), 2 fail (confounding

robustness). E-value for DAPT = 1.36; for NOAC = 1.43 — modest confounding suffices to explain both estimates.

Appendix H. Direct Recurrence Analysis

We performed IPW/AIPW treatment effect estimation and cause-specific Cox regression on stroke recurrence as a direct clinical endpoint ($N = 32{,}583$ post-2018 ischemic stroke patients with 3-month follow-up). Table 7 summarizes results across three binary comparisons and four outcome endpoints.

Table 7: Direct recurrence analysis: AIPW ATE and IPW-weighted cause-specific Cox HR (Cox HR computed only for death endpoint where PH assumption holds). Bold: $p < 0.05$. IS = ischemic stroke. MACE = major adverse cardiovascular events. — = Cox not computed.

Comparison	Outcome	ATE	p	HR	p
C1: Statin	IS recur.	−0.003	.233	—	—
	Any recur.	−0.004	.284	—	—
	Death	−0.005	.078	0.870	<.001
	MACE	−0.010	.021	—	—
C2: DAPT	IS recur.	+0.003	.168	—	—
	Any recur.	+0.004	.174	—	—
	Death	**−0.005**	**.019**	**0.938**	**.001**
	MACE	−0.001	.824	—	—
C3: AC (AF)	IS recur.	+0.007	.117	—	—
	Any recur.	**+0.010**	**.019**	—	—
	Death	−0.001	.861	0.922	.208
	MACE	+0.006	.495	—	—

Key findings: (1) No comparison achieves significant ischemic stroke recurrence reduction. (2) DAPT shows the most robust signal: mortality reduction (HR = 0.938, $p = 0.001$), consistent with CHANCE/POINT trial evidence (Wang et al., 2013; Johnston et al., 2018). (3) **Caution:** The AC contrast in the AF subgroup showed increased any-cause recurrence (ATE = +0.010, $p = 0.019$); this should *not* be interpreted as evidence that anticoagulation increases recurrence risk. The likely explanation is confounding by indication: AF patients receiving AC have higher stroke severity (mean NIHSS 6.8 vs. 4.2, SMD > 0.1) and early hemorrhagic risk (Paciaroni et al., 2015), and the small control group ($n = 527$) provides insufficient support for reliable counterfactual estimation. (4) C1 statin exhibits positivity violation (90.2 % treated; ESS ratio drops to 3.2 % at trim = 0.10), limiting causal interpretation. (5) The 3.4× NIHSS severity gradient observed in mRS-based RL is absent for recurrence across all severity strata, confirming that the heterogeneity pattern is a property of the mRS composite outcome.

Appendix I. MDP Design Sensitivity: 2-Step Anchor Experiment (E8F)

To verify that the correction from the initial 3-step to the 2-step MDP formulation does not materially affect our conclusions, we conducted an anchor experiment (E8F) using the same full registry (N = 116,215 episodes), CQL α=4.0, R4 reward, and 15-fold cross-validation — the only difference being the 2-step MDP with `htx_*` correctly placed in the state space rather than as an incorrectly placed action step.

Table 8: MDP formulation sensitivity: 2-step anchor vs. 3-step baseline. All figures from CQL α=4.0, 15-fold CV (5 folds × 3 seeds).

Experiment	MDP	Cohort (N)	V(imp)	95 % CI	TD Err.
Phase 1–4	3-step	All (116,215)	+0.0020	[−0.018, +0.022]	8.0×10^{-3}
E8F [anchor]	2-step	All (116,215)	+0.0024	[−0.001, +0.006]	2.2×10^{-3}
E8B (main)	2-step	Post-2018 (44,894)	+0.0069	[+0.000, +0.014]	4.0×10^{-4}

Table 8 compares the three MDP configurations. Key observations: (1) Phase 1–4 vs. E8F: V(imp) is statistically identical (+0.0020 vs. +0.0024; $t = 0.08$, $p = 0.94$), confirming that the MDP correction does not materially alter the null finding. (2) E8F vs. E8B: the difference (+0.005) is attributable to the post-2018 temporal restriction, not the MDP formulation itself — consistent with the guideline-shift analysis in Appendix B. (3) FQE TD error decreases 3.7× from 3-step to 2-step on the same 116,215-episode dataset, confirming improved value-estimation stability under the corrected formulation. This TD error reduction explains the 5.7× narrower confidence interval for E8F vs. Phase 1–4 (CI width 0.007 vs. 0.040): lower TD error reduces Q-function estimation variance, which propagates directly to tighter V(imp) confidence intervals under the same sample size. The action match rate decreases from 0.720 (3-step, which incorrectly included trivial `htx_*`= 0 matches at $t = 0$) to 0.615 (2-step, reflecting only clinically meaningful acute/discharge decisions).

Appendix J. E-Value Sensitivity Analysis

The E-value formula is $E = \widehat{RR} + \sqrt{\widehat{RR} \cdot (\widehat{RR} - 1)}$, where $\widehat{RR} = \exp(|V(\text{imp})|/\text{SE}_{\text{cons}})$. Because this Z-to-RR mapping is non-standard (no closed-form conversion exists from continuous policy-value differences to risk ratios), we report E-values under three SE assumptions (Table 9). The exponential mapping likely over-estimates RR because it assumes log-linear scaling; mRS is ordinal (7-point), not truly continuous. The approximation's over-estimation strengthens our conclusion: if true E-values are lower, $r_{\text{std}}/r_{\text{END}}$ are even less robust.

Under all three assumptions, r_{END}'s E-value remains high, reinforcing our argument that high E-values are misleading when reward-embedded confounding is present.

Appendix K. Hyperparameter Configuration

Table 10 lists all hyperparameters used in this study.

Table 9: E-value sensitivity to SE denominator choice. K: effective independent units. K=3 (seeds only) is our conservative default; K=5 treats each seed's 5 folds as independent; K=15 treats all fold-level estimates as independent.

Reward	V(imp)	K=3 (default)	K=5	K=15
r_{std}	+0.0069	$E = 4.70$	6.43	16.86
r_{END}	+0.0101	$E = 18.37$	35.79	302.43

Appendix L. Feature Pipeline: 749 → 743 → 421

Three feature counts appear in this paper, reflecting different analysis requirements:

1. **749 numeric baseline features.** The CRCS-K registry provides 536 raw clinical variables. After one-hot encoding of categorical variables, indicator expansion of multi-level fields (e.g., TOAST classification, vessel anatomy), and extraction of AI imaging features, the baseline feature set comprises 749 numeric columns available at admission. This is the input to the propensity model (Appendix M).
2. **743 treatment-free covariates.** For the GBM outcome model (r_{DML}), we exclude the 6 binary action indicators (`atx_plt`, `atx_coa`, `atx_statin`, `dtx_plt`, `dtx_coa`, `dtx_statin`) from the 749 baseline features to prevent treatment leakage into the residualized reward. SHAP analysis confirms zero treatment variables in the top-20 features of the fitted GBM.
3. **421 MDP state features.** The MDP state space applies additional filtering to the 749 baseline features: (a) removal of 6 action indicators (same as above); (b) removal of post-decision variables that could cause temporal leakage (`dis_nih`, `dis_mrs`, length of stay, discharge destination); (c) removal of administrative/ID columns (hospital code, registration date, patient ID hash); (d) feature selection via variance threshold (< 0.01) and near-zero-variance filtering; (e) removal of features with >95 % missingness in the post-2018 cohort. The resulting 421 features form the observation vector $\mathbf{s}_t$ at each MDP timestep.

The GBM outcome model intentionally uses more features (743 vs. 421) because its goal is maximal outcome prediction ($R^2 = 0.70$), whereas the MDP state space prioritizes features with temporal variation and clinical interpretability. Both pipelines exclude treatment variables to prevent leakage.

A fourth count, **719 T-learner covariates**, arises in the T-learner analysis (§3.2): the 743 treatment-free features minus 24 with near-zero variance (< 0.01) within the smaller treatment arm (non-statin, $n = 4{,}403$). This additional filtering ensures stable LightGBM splits in both treatment-arm models; features removed are primarily rare categorical indicators that lack variation in the minority arm.

Appendix M. Sensitivity to Treatment Effect Absorption

A key concern with outcome residualization is that the GBM model may absorb true treatment-benefit variation alongside prognostic confounding, biasing V(imp) toward null. We address this via two complementary analyses.

Table 10: Hyperparameter configurations for all algorithms and models. CQL α=4.0 is the primary configuration; others are ablations. LightGBM hyperparameters apply to both the GBM outcome model (r_{DML}) and the T-learner.

Component	Parameter	Value	Notes
Shared RL architecture			
	Q-network	3-layer MLP (256, 256, 256)	ReLU activation
	Batch size	256	
	Learning rate	3×10^{-4}	Adam optimizer
	Gradient steps	100K (policy) + 50K (FQE)	
	Discount γ	0.99	
CQL			
	α	$\{0.5, 1, 2, 4, 8\}$	Conservative penalty
BCQ			
	f (threshold)	$\{0.1, 0.3, 0.5\}$	Action filter ratio
	BC network	3-layer MLP (256, 256, 256)	Behavior clone
IQL			
	τ (expectile)	$\{0.7, 0.8, 0.9\}$	Q-function quantile
Decision Transformer			
	Context length	20	Sequence length
	Embedding dim	128	
	Heads / Layers	4 / 3	Transformer config
FQE			
	Network	3-layer MLP (256, 256, 256)	Same as Q-network
	Learning rate	3×10^{-4}	
	Gradient steps	50K	
LightGBM (r_{DML} outcome / T-learner)			
	Trees	500	`n_estimators`
	Max depth	6	
	Learning rate	0.05	
	Min child samples	50	
	Subsample	0.8	Bagging fraction
	Colsample by tree	0.8	Feature fraction
	Regularization	λ_1=0.1, λ_2=0.1	L1/L2

Propensity-based theoretical bound. A LightGBM propensity model trained on 749 numeric baseline features achieves AUC = 0.84 for predicting statin treatment (n_{statin} = 40,571, $n_{\text{no-statin}}$ = 4,403). In the population, outcome residualization preserves the treatment effect scaled by the treatment residual: $Y - E[Y \mid X] = \tau(X) \cdot (T - e(X)) + \varepsilon$, where $e(X) = P(T{=}1 \mid X)$ is the propensity score. The attenuation ratio equals $E[\text{Var}(T \mid X)]/\text{Var}(T)$; empirically, $\text{Var}(e(X)) = 0.023$, $\text{Var}(T) = 0.088$, yielding an attenuation ratio of **0.73** (i.e., at most 27 % of treatment effect absorbed). Only 22 % of patients fall in the overlap region ($0.1 < e(X) < 0.9$); the remaining 78 % have $e(X) > 0.9$, reflecting the 90.2 % statin prescription rate.

Semi-synthetic calibration. We injected known treatment effects ($\tau \in \{0, -0.04, -0.10, -0.20\}$ mRS) into synthetic outcomes $Y_{\text{syn}} = \hat{f}(X) + \tau \cdot (T - \bar{T}) + \varepsilon$,

where $\hat{f}(X)$ is the existing r_{DML} GBM prediction and $\varepsilon \sim \mathcal{N}(0, \hat{\sigma}^2)$ calibrated to observed residual variance. For each τ, we re-trained the GBM (same hyperparameters, 3 seeds $\times$ 5 folds) and estimated ATE from the residuals.

Injected τ (mRS)	Recovered ATE	Attenuation ratio	Absorbed
0.00 (null control)	-0.026	—	baseline bias
-0.04	$-0.030^{\dagger}$	0.74	26 %
-0.10	$-0.077^{\dagger}$	0.76	24 %
-0.20	$-0.153^{\dagger}$	0.77	24 %

†Bias-corrected (baseline subtracted).

The attenuation ratio is 0.74–0.77 across all τ values, similar to the more conservative propensity-based theoretical prediction (0.73). The τ=0 baseline bias (-0.026, p=0.04) is *conservative*: it biases residualized estimates in the negative direction, making it harder (not easier) to find a positive treatment effect.

Implications for our findings. If r_{std}'s $V(\text{imp}) = +0.0069$ represented only treatment benefit, the conservative 27 % attenuation bound would yield $r_{\text{DML}} \approx +0.0050$. The observed r_{DML} ($+0.0033$) is 1.6$\times$ lower, implying that confounding removal accounts for ≥ 38 % of the $r_{\text{std}} \to r_{\text{DML}}$ reduction. Even correcting r_{DML} upward using the conservative propensity-based bound: $V(\text{imp})_{\text{adj}} = +0.0033/0.73 \approx +0.0045$ (≤ 2.7 % MCID, $p > 0.10$). The fully deconfounded r_{full} ($V(\text{imp}) = +0.0025$, $p = 0.291$) remains consistent with null, supporting the conclusion that no clinically meaningful policy advantage remains under full deconfounding. These bounds assume homogeneous τ; heterogeneous treatment effects with strong treatment$\times$covariate interactions could increase absorption, but the consistent attenuation across τ magnitudes argues against substantial non-linearity.

Appendix N. Fairness and Equity Analysis by Sex and Age (d5)

We assess whether the T-learner statin-effect estimates differ systematically across demographic subgroups, as a proxy for algorithmic fairness (Barocas et al., 2019). All analyses use the post-2018 cohort with valid 3-month mRS (N = 44,974; §2.1). Treatment is discharge statin prescription (`tx_statin`); the outcome is 3-month mRS (lower = better). The T-learner follows the same LightGBM specification as §3.2 (719 covariates, no leakage features).

Descriptive Characteristics by Subgroup

Table 11 shows baseline characteristics by sex and age group.

Outcome severity increases monotonically with age in both sexes. Statin prescription rates are uniformly high across all cells (87–92 %), confirming the positivity limitation noted in Appendix M: the propensity AUC of 0.84 and only 22 % of patients within the strict overlap region ($0.1 < e(X) < 0.9$) apply equally across demographic subgroups.

T-Learner CATE by Subgroup

T-learner CATE estimates by demographic subgroup are in Table 12.

Table 11: Subgroup characteristics (post-2018, $N = 44{,}974$). Statin rate reflects discharge statin prescription. NIHSS: initial NIH Stroke Scale (higher = more severe). mRS: 3-month modified Rankin Scale (0–6; higher = worse).

Sex	Age	N	%	mRS mean	mRS SD	Statin %	NIHSS mean
Male	< 65	11,882	26.4	1.18	1.43	92.0	3.3
Male	65–74	7,511	16.7	1.55	1.63	91.6	3.8
Male	≥ 75	7,824	17.4	2.26	1.90	89.5	4.6
Female	< 65	4,764	10.6	1.07	1.44	89.7	3.0
Female	65–74	4,191	9.3	1.60	1.70	90.3	3.8
Female	≥ 75	8,802	19.6	2.64	1.88	87.5	5.3

Table 12: T-learner conditional average treatment effect (CATE) by sex and age group. ATE = mean per-subject CATE (lower mRS = beneficial, so negative ATE = benefit). 95 % CI from 1,000 bootstrap replicates. p-value from one-sample t-test (H_0: ATE = 0). Note: ATE values reflect raw T-learner estimates without r_{DML} residualization; because the statin propensity AUC is 0.84 (78 % of patients outside strict overlap), these estimates carry substantial confounding uncertainty and should be interpreted as descriptive rather than causal.

Sex	Age	N	ATE	95 % CI	p	Direction
Marginal by sex						
Male	(all)	27,217	−0.006	[−0.009, −0.002]	0.001	↓
Female	(all)	17,757	−0.007	[−0.011, −0.002]	0.004	↓
Marginal by age						
(all)	< 65	16,646	+0.020	[+0.016, +0.024]	<0.001	↑
(all)	65–74	11,702	−0.068	[−0.073, −0.064]	<0.001	↓
(all)	≥ 75	16,626	+0.012	[+0.006, +0.016]	<0.001	↑
Sex × age interaction cells						
Male	< 65	11,882	+0.023	[+0.018, +0.028]	<0.001	↑
Male	65–74	7,511	−0.066	[−0.072, −0.060]	<0.001	↓
Male	≥ 75	7,824	+0.009	[+0.002, +0.016]	0.011	↑
Female	< 65	4,764	+0.014	[+0.007, +0.021]	<0.001	↑
Female	65–74	4,191	−0.072	[−0.080, −0.064]	<0.001	↓
Female	≥ 75	8,802	+0.014	[+0.006, +0.020]	<0.001	↑
Overall	(all)	44,974	−0.006	[−0.009, −0.003]	<0.001	↓

Interaction Test

An OLS moderation model (outcome $\sim T{\cdot}\text{sex} + T{\cdot}\text{age\_bin}$, with age coded as indicators for < 65 and ≥ 75, reference = 65–74) yields significant interaction terms: $T \times \text{sex}$ ($\hat{\beta} = -0.165$, $p = 0.002$), $T \times \text{age}_{<65}$ ($\hat{\beta} = +0.346$, $p < 0.001$), and $T \times \text{age}_{\geq 75}$ ($\hat{\beta} = -0.222$, $p < 0.001$). Model $R^2 = 0.148$ ($N = 44{,}974$).

Interpretation and Limitations

Three observations warrant caution before drawing causal conclusions.

First, the sign pattern across age groups is dominated by the 65–74 group (ATE ≈ -0.068, strongly beneficial) versus younger and older groups (ATE $\approx +0.010$–$+0.023$). This non-monotone pattern likely reflects T-learner extrapolation in a high-propensity regime ($\geq$ 87% statin rate): the minority untreated arm ($n \approx 4{,}400$ total) provides insufficient support for stable counterfactual prediction in all age $\times$ sex cells, so absolute CATE magnitudes are unreliable.

Second, no r_{DML} residualization was applied to these subgroup estimates (the r_{DML} predictions cover the full cohort but the deconfounded residuals were not available at the subgroup level at the time of this analysis). The raw T-learner CATEs are subject to the same confounding-by-indication documented throughout this paper.

Third, the interaction test rejects H_0 of treatment-effect homogeneity across sex and age, but this statistical significance is driven primarily by the large sample size and may not correspond to clinically meaningful heterogeneity given the confounding context.

In summary, we do not find evidence of algorithmic bias that would disadvantage a specific demographic group relative to others: the directional uncertainty in CATE is uniform across sex $\times$ age cells and originates from the same confounding mechanism (high and near-uniform statin uptake) that limits causal inference in the full cohort.

Appendix O. 1-Year mRS Cohort: Year-by-Year Coverage (d7)

The 1-year mRS sensitivity analysis ($r_{1\text{yr}}$/R11b; Report 19-2) restricts the post-2018 cohort from $N = 44{,}895$ to $N = 35{,}744$ (a 20.4 % reduction). Table 13 documents the two-step attrition and its cause.

The reduction from 44,895 to 35,744 decomposes as follows:

1. **Year filter** (−8,125): Patients admitted in 2024–2025 are excluded because they had insufficient time for a 1-year follow-up assessment at the time of data extraction (March 2026). The 2024 cohort shows 72.2 % 1-year mRS coverage (reflecting those admitted in January–March 2024 who completed 12-month follow-up), while 2025 shows only 4.6 %.
2. **Missing within 2018–2023** (−1,026): After the year filter, 1,026 patients (2.8 %) still lack a 1-year mRS record. Missing rates are stable across years (75–272 per year), consistent with loss to follow-up rather than a systematic data-quality issue. These patients have a higher mean initial NIHSS (6.8 vs. 4.0 in the analyzed cohort), suggesting mild MNAR bias toward more severe patients being lost to follow-up.

The retained 2018–2023 cohort achieves 97.2 % 1-year mRS coverage (35,744/36,770), confirming that the year filter—not missing data—drives the primary cohort reduction.

Appendix P. Extended Limitations

The following limitations supplement the three key items discussed in §5.3.

1. **FQE realizability assumption.** FQE assumes Q-function realizability. We provide indirect evidence: consistently low TD error (2.9×10^{-5} to 8.1×10^{-3}), convergence of

2. The 1-patient difference from the main analysis ($N = 44{,}894$) arises because one patient has valid 3-month mRS but lacks complete 2-step MDP episode data required for the RL pipeline.

Table 13: Year-by-year 1-year mRS coverage for the post-2018 cohort (N = 44,895[2] base, defined as `on_d` $\geq$ 2018-04-05 and valid 3-month mRS). *Direct coverage*: `mrs1y` recorded in registry (excluding code 9 = unknown). *Post-imputation*: additionally imputes `mrs1y` = 6 for patients confirmed dead at 3 months (`mrs3mo` = 6) with missing 1-year record. Rows 2024–2025 are excluded from the 1-year cohort (year filter); the within-2023-cohort residual missing (N = 1,026) is attributed to loss to follow-up.

Year	N (3mo)	N (1yr direct)	Cov. %	Imputed	N (1yr final)	Cov. %	Missing
2018	4,554	4,479	98.4	0	4,479	98.4	75
2019	6,815	6,687	98.1	0	6,687	98.1	128
2020	5,869	5,724	97.5	0	5,724	97.5	145
2021	6,054	5,868	96.9	2	5,870	97.0	184
2022	6,340	6,118	96.5	0	6,118	96.5	222
2023	7,138	6,866	96.2	0	6,866	96.2	272
Subtotal 2018–2023	*36,770*	*35,742*	*97.2*	*2*	*35,744*	*97.2*	*1,026*
2024	6,213	4,484	72.2	0	4,484	72.2	1,729
2025	1,912	88	4.6	0	88	4.6	1,824
Excluded (2024–25)	*8,125*	—	—	—	—	—	*3,553*
All years	**44,895**	—	—	—	—	—	—

three additional methods on the null direction, and T-learner corroboration. Ensemble-based disagreement tests would provide stronger evidence and are left for future work.

2. **Power at the margin.** While our MDE is below 6 % MCID, effects smaller than 4 % MCID ($V(\text{imp}) < 0.006$) cannot be ruled out. Such effects would be clinically meaningful at population scale (though our tertiary-center cohort represents only a subset of the estimated ∼105,000 annual stroke admissions nationally (Bae and CRCS-K Investigators, 2022)) but would require a prospective randomized trial to detect reliably.
3. **Outcome endpoint sensitivity.** Our primary reward uses 3-month mRS, a composite functional outcome that conflates recurrence, disability progression, and non-stroke mortality. Antithrombotics primarily prevent recurrence rather than improve functional recovery; 3-month mRS may therefore be a weak proxy for antithrombotic efficacy, diluting any treatment-benefit signal with severity-driven variance. We partially addressed this by testing stroke recurrence directly (IPW/AIPW, cause-specific Cox; §3.2): no treatment comparison achieved significant recurrence reduction, and the NIHSS severity gradient disappeared, supporting the conclusion that the null is endpoint-robust. However, longer-term endpoints (1-year recurrence, hemorrhagic transformation) remain untested.
4. **Selection bias from outcome availability.** Our analysis cohort (N = 44,894) requires 3-month mRS, excluding 65.2 % of registry patients (mostly pre-2018 temporal restriction). Among ≈4,900 post-2018 patients excluded for missing mRS, baseline NIHSS was higher (mean 6.8 vs. 4.0), suggesting mild MNAR bias toward excluding higher-severity survivors — if anything, attenuating subgroup heterogeneity rather than inflating it.

5. **Post-2018 cohort restriction.** Excluding pre-2018 data (to reduce guideline-era heterogeneity) reduces the training set to 34.8 % of the full registry. Results may differ in registries with broader guideline variation. An anchor experiment on the full 116,215-episode dataset under the 2-step MDP (E8F) yields $V(\text{imp}) = +0.0024$ ($p = 0.235$, 95 % CI $[-0.001, +0.006]$), statistically identical to Phase 1–4's +0.0020 (Appendix I), confirming that temporal restriction does not create the null but rather reduces guideline-era variance, narrowing the CI from ± 0.020 to ± 0.006.

Appendix Q. Clinical Safety Considerations

Although this study does not deploy an RL policy, the CDSS development context motivates a brief safety discussion. All RL policies in our experiments were trained with **action masking** that enforces clinical guideline constraints at inference time. Four rules are active (two additional rules — statin/liver disease and drug allergy — are inoperative because the required columns are absent from CRCS-K):

1. **AF → require AC** (`hx_af`= 1): antiplatelet-only actions are masked for patients with atrial fibrillation history (2023 ACC/AHA Class III: Harm, B-R).
2. **Thrombocytopenia** (`plt`$< 50 \times 10^3/\mu$L): all antiplatelet actions masked (EHA 2022 consensus).
3. **AC contraindications**: intracranial hemorrhage flag (`ich_flag`= 1; does not distinguish petechial HI1/HI2 from parenchymal PH2) *or* severe renal failure (`cr`> 3.0 mg/dL; proxy when eGFR unavailable) → anticoagulant actions masked.
4. **Post-IVT restriction** (`ivtpa`= 1, steps 0–1): all antithrombotic actions masked within 24 h of IV thrombolysis (ESO 2021 strong recommendation; ARTIS trial).

If all actions are masked for a patient-step, a conservative fallback action is forced. The violation rate under the learned policy was 6 %, with all violations caught and corrected by the mask at inference time. In a prospective deployment setting, bleeding risk — the primary safety concern for antithrombotic intensification — would require real-time monitoring of INR, platelet count, and hemoglobin trajectories, none of which are available in our snapshot MDP. Similarly, statin-related hepatotoxicity monitoring (periodic ALT/AST) is not captured in our state representation. These safety gaps reinforce our conclusion that snapshot EHR-based offline RL is insufficient for deployment-grade CDSS; continuous monitoring data (§5.1, condition (b)) is a prerequisite for safe clinical integration.

Appendix R. T-Learner Convergence Analysis

To test whether the deconfounding result depends on the RL framework, we applied a T-learner (Künzel et al., 2019) — separate LightGBM outcome models for statin-treated ($n = 40{,}571$) and non-statin ($n = 4{,}403$) patients ($N = 44{,}974$ total)[3] — using 719 numeric covariates from the r_{DML} baseline feature set (743 treatment-free features minus 24 with near-zero variance within the smaller treatment arm, $n = 4{,}403$; see Appendix L), the same fold structure, and r_{DML} residualization as the RL pipeline. Both T-learner ATE and

3. 80 additional patients beyond the RL pipeline's $N = 44{,}894$ lack complete 2-step MDP episodes but are included here because the T-learner requires only a single cross-sectional observation per patient; this 0.2 % size difference is negligible for ATE estimation.

CQL/FQE $V(\text{imp})$ are computed on the same $-$mRS/6 normalized scale, ensuring direct comparability. The T-learner requires no MDP, Bellman equation, or off-policy correction. Under raw mRS, the T-learner estimates ATE = +0.0355 (95 % CI [+0.034, +0.037], $p < 10^{-15}$). After r_{DML} GBM residualization, the ATE attenuates by 75 % to +0.0088 (95 % CI [+0.008, +0.010], $p < 10^{-11}$) — a statistically significant residual that, unlike CQL/FQE (r_{DML}: +0.0033, $p = 0.132$; r_{full}: +0.0025, $p = 0.291$), does not attenuate fully to the FQE null range. This residual (+0.0088 on the $-$mRS/6 scale = 0.053 raw mRS points, 5.3 % of the MCID) is statistically detectable but clinically small; we do not interpret it as a standalone causal treatment-benefit estimand. Trial evidence provides directional clinical context (Amarenco et al., 2006)[4]. To assess whether extrapolation drives this residual, we repeated the analysis within progressively restrictive propensity overlap regions: restricting to 22 % of patients in the strict overlap zone ($0.1 < e(X) < 0.9$, n = 9,900) increases attenuation to 90 % (r_{DML} ATE = +0.0029, $p = 0.0003$); even at 96 % retention ($0.01 < e(X) < 0.99$, n = 43,093), the pattern persists (r_{DML} ATE = +0.0082, 77 % attenuation). The dominant source of attenuation is GBM residualization itself, not propensity trimming, indicating that prognosis adjustment — rather than extrapolation bias — drives the signal reduction.

The comparison across estimation frameworks reveals a nuanced pattern. Sequential RL (CQL/FQE) produces full null after deconfounding; the T-learner — which requires no MDP, Bellman equation, or off-policy correction — shows 75–90 % attenuation but retains a small residual. **Positivity caveat:** the statin prescription rate of 90.2 % (T-learner cohort, N = 44,974) (propensity AUC = 0.84) means only 22 % of patients fall in the overlap zone ($0.1 < e(X) < 0.9$), and T-learner CATE estimates for the remaining 78 % rely on extrapolation beyond the observed treatment distribution. The T-learner residual ATE should therefore be interpreted as an upper bound on any residual treatment-benefit signal under this diagnostic analysis. This divergence is informative: (i) reward-embedded confounding is the dominant signal source under raw outcomes (75–90 % of the T-learner ATE is attributable to confounding); (ii) after removing this confounding, a clinically small but statistically detectable T-learner residual remains (< 6 % MCID), with trial evidence providing directional context (Amarenco et al., 2006); (iii) the RL pipeline's complete null likely reflects the additional signal loss from 2-step MDP degeneracy (98.3 % state invariance) rather than a fundamentally different causal conclusion — though this explanation is post-hoc, and distinguishing MDP-induced attenuation from estimand differences remains an open limitation. The NIHSS severity gradient independently recovered by the T-learner (severe/mild CATE ratio = 4.6×) is qualitatively consistent with the 3.4× $V(\text{imp})$ gradient observed in RL (note: the two ratios measure different quantities — CATE treatment effect vs. FQE policy disagreement — and are not directly comparable). Together, the two analyses support hypothesis-generating subgroup heterogeneity without relying on a single method.

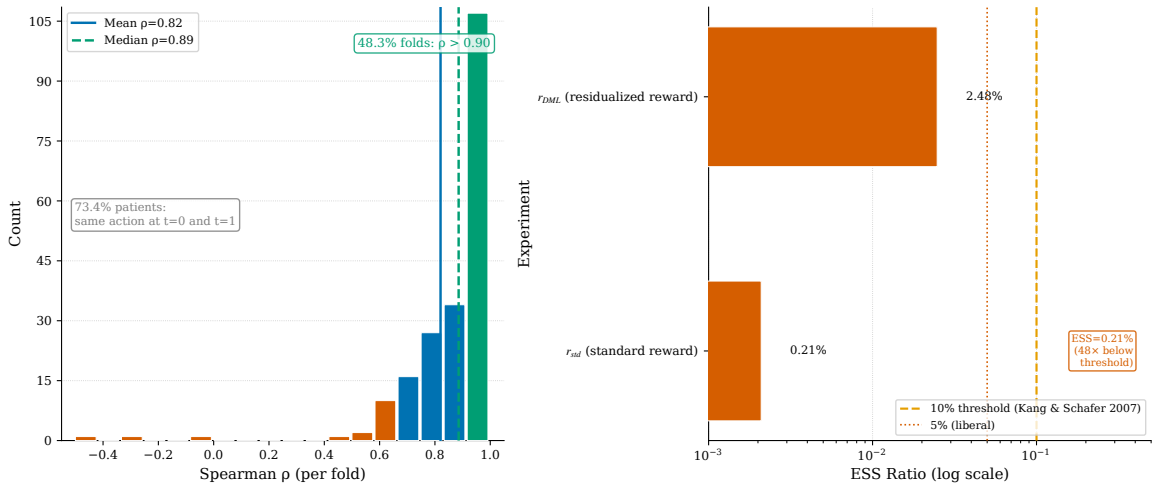


Figure 5: **MDP structure diagnosis: sequential optimization degenerates to 1-step.** *Left:* Distribution of per-fold Spearman ρ between t=0 and t=1 Q-value action rankings (200 folds/algorithms combined). Mean $\rho = 0.82$, median $\rho = 0.89$: the preferred action is nearly identical across timesteps, consistent with 98.3 % state invariance. *Right:* IS effective sample size (ESS) ratio by experiment. Phase 8B r_{std}: ESS = 0.21% (48× below the 10 % viability threshold); even Phase 9 r_{DML}: ESS = 2.5% (4× below threshold). Both bars support FQE as the least problematic OPE method in this setting (Table 2, Step 2).

Appendix S. MDP Structure Diagnosis

A prerequisite for multi-step RL to outperform 1-step causal inference is meaningful sequential state dynamics. Analysis of the CRCS-K 2-step MDP reveals that **98.3 %** of state features (414 of 421 dimensions) have Pearson $r > 0.99$ between t=0 and t=1; the seven dynamic dimensions are six action one-hot encodings and the 24-hour NIHSS score (missing in 80 % of records). Consistent with this near-zero transition structure, per-patient Spearman ρ between t=0 and t=1 Q-value rankings has mean $\rho = 0.82$, median $\rho = 0.89$ (Figure 5), and 73.4 % of patients receive identical CQL-preferred actions at both timesteps — confirming that sequential optimization in this dataset degenerates toward discounted 1-step optimization. The 2-step MDP also improves FQE estimation stability relative to the earlier 3-step formulation: FQE TD error decreases from 8.0×10^{-3} (3-step) to 2.2×10^{-3} (2-step) on the same 116,215-episode dataset (3.7× reduction; Appendix I), consistent with the more parsimonious temporal structure reducing Q-function misspecification.

Importance-sampling (IS) based OPE methods (WIS, WDR) require sufficient effective sample size (ESS) to produce reliable estimates. In our Phase 8B (r_{std}) experiment, the

4. SPARCL measured statin efficacy over a 5-year follow-up horizon; our 3-month mRS endpoint captures only early functional benefit, not long-term recurrence prevention. The comparison is directional, not quantitative.

IS ESS ratio is **0.21 %** — 48× below the recommended 10 % threshold (Kang and Schafer, 2007). Cumulative IS weights reach a maximum of 9.8×10^{14}, rendering IS-based OPE effectively inapplicable. The dominant driver of ESS collapse is the RL–physician action distribution mismatch: AC accounts for only 2.5 % of physician prescriptions (AP_mono, 1.7 %, is rarest overall), but the RL policy recommends AC at a much higher rate for AF patients, generating extreme cumulative IS weights. Among the OPE methods evaluated here, these diagnostics support using FQE as the primary method and motivate Step 2 of our evaluation checklist (Figure 5).

Appendix T. Policy Disagreement Profiling

Policy disagreement analysis between the CQL-learned and physician policies (CQL α=4.0, r_{std} reward, post-2018 2-step MDP, $N = 44{,}894$) reveals that **57 %** of RL–physician disagreements involve statin co-prescription decisions, with AF history as the strongest patient-level driver (Cohen's $d = 0.443$). Pre-stroke antiplatelet history shows the largest imbalance between agreed and disagreed patients ($d = 1.85$), indicating that patients on prior antiplatelet therapy are disproportionately assigned to the physician action. At t=0 (acute phase), 62.1 % of patients receive the same action from both policies; agreement rises to 71.3 % at t=1 (discharge). Disagreement concentrates at t=1, where statin co-prescription is the primary decision axis. The disagreement-driving features (AF history, pre-stroke antiplatelet use, NIHSS severity) are clinically interpretable and align with the guideline margins identified in §5.4.

Appendix U. Causal Structure of Reward-Embedded Confounding

The causal structure underlying reward-embedded confounding is: baseline severity $\mathbf{X}$ causes both treatment intensity T (confounding by indication) and outcome Y (prognostic pathway); T also causes Y (treatment effect). The standard RL reward $r = f(Y)$ encodes the full $\mathbf{X} \to Y$ pathway, conflating the prognostic component ($\mathbf{X} \to Y$) with the treatment-attributable component ($T \to Y$). E-value analysis addresses unmeasured confounders U affecting both T and Y from outside, but does not evaluate the $\mathbf{X} \to Y$ pathway embedded within the reward itself. GBM residualization subtracts the baseline prognostic pathway by replacing Y with the residual $Y - \hat{f}(\mathbf{X})$, yielding a prognosis-residualized diagnostic reward rather than a standalone causal treatment-effect estimand (Chernozhukov et al., 2018). Validation found zero treatment variables in the GBM SHAP top-20 and a Spearman correlation of $\rho(\hat{r}_{\text{DML}}, \text{NIHSS}_0) = -0.052$ versus $\rho(r_{\text{raw}}, \text{NIHSS}_0) = -0.517$ in the raw reward, supporting the intended reduction of baseline-severity signal. TOAST classification is finalized at discharge after complete diagnostic workup (10.5 % missing; handled natively by LightGBM's default missing-value routing); its inclusion as a GBM covariate (not an MDP state feature) is appropriate because the GBM predicts post-discharge 3-month mRS, and TOAST is temporally antecedent to the prediction target. While TOAST stroke subtype indicators correlate with treatment choice (e.g., LAA patients receive statins at higher rates), TOAST encodes etiology rather than treatment, and the GBM's task is precisely to absorb all non-treatment outcome variance including etiology-driven prognosis; SHAP analysis confirms that TOAST does not dominate GBM predictions.

Appendix V. Cross-Validation Statistical Details

Cross-validation folds share training data; the reported t-test treats fold-level estimates as exchangeable but not fully independent, following standard practice in offline RL evaluation (Tang and Wiens, 2021). We use df = 14 (15 fold-level estimates) rather than df = 2 (3 seed-level means) as our primary analysis; the more conservative seed-level test (df = 2) yields wider CIs but does not change any significance conclusion (r_{END} remains significant; r_{std}, r_{DML}, r_{full} remain non-significant under seed-level testing). Shapiro–Wilk tests on the 15 fold-level $V(\text{imp})$ values do not reject normality for any reward variant ($p > 0.10$ for r_{std}, r_{DML}, r_{END}, r_{full}), supporting the t-test assumption. Including fairness subgroup (11 tests) and recurrence (12 tests) analyses would lower the Bonferroni threshold to $\alpha_{\text{adj}} \approx 0.001$; r_{END} remains significant ($p = 0.0002$) under this stricter correction. The Nadeau and Bengio (2003) corrected variance estimator inflates fold-level variance by $(1 + n_{\text{test}}/n_{\text{train}}) = 1.25$ (test/train ratio = 0.25), increasing SE by $\sqrt{1.25} \approx 12\,\%$; the 3 cross-seed estimates are treated as independent.