RECURRENT CLINICAL SAFETY QUEUES REVERSE MULTIMODAL MODEL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

At every decision time, a recurrent clinical safety queue spends capacity on new cases and material updates but not on unresolved carryovers; deferral, dismissal, cooldown, and re-entry then determine the next admissible set. Completed-record sorting and next-admissible selection consequently need not commute. On 9,714 patient-episodes and 1,074 adjudicated events from three held-out trials, completed-record discrimination selects gated fusion (AUROC/AUPRC/PPV@top-5%: 0.90/0.47/65.5), while recurrent review selects ChronoGuard. Under the five-selection per-instant boundary, ChronoGuard recognizes events 1.1 days earlier than gated fusion (95% CI, 0.6–1.6), raises severe-event sensitivity by 0.06 (0.02–0.10) and selected-alert precision by 0.07 (0.03–0.11), and removes 0.31 false alerts per monitor-week (0.17–0.45). AdmitLock makes this reversal testable by locking the evidence filtration, scorer-specific admission, recurrent transition, capacity clock, and endpoint registry, then reporting both Pareto orders, selected-episode overlap, and paired queue effects. Only 12–16 of each model's 25 highest-ranked completed-record episodes occur among its first 25 recurrent selections. Against operational key risk indicators, ChronoGuard improves recognition by 3.5 days, severe sensitivity by 0.14, precision by 0.19, and false-alert burden by 0.95 per monitor-week; day-28 restricted mean recognition time falls from 11.1 to 7.4 days. An independently governed four-trial portfolio reproduces both point-estimate leaders after refitting every learned method, and a 228-monitor-week shadow evaluation preserves ChronoGuard's operating profile under live arrivals. Model selection for scarce clinical review should therefore target the stateful sequence that reaches experts, not a ranking formed after records close.

1 INTRODUCTION

Suppose one review slot remains when a new high-scoring episode and a material update to an open severe case arrive together. Selecting either one defers the other; the chosen case then changes open-case state, later updates, cooldown eligibility, and the candidates competing at the next instant. Sorting both completed records after follow-up cannot reconstruct that trajectory. The relevant invariant is instead queue-local: new cases and material updates consume capacity, unresolved carryovers do not, and deferred cases remain eligible for the next distinct decision time. Thus model quality is a property of a selected sequence generated by evidence chronology and recurrent state, not only of a terminal ranking.

This distinction changes the empirical decision. On three held-out clinical trials, gated fusion ranks first by completed-record AUROC, AUPRC, and top-5% PPV, while ChronoGuard ranks third. Timestamp-censored recurrent queueing reverses those positions: ChronoGuard recognizes events 1.1 days earlier than gated fusion, captures 22 more severe events, raises selected-alert precision by 0.07, and produces 0.31 fewer false alerts per monitor-week. All four paired intervals favor ChronoGuard. The change reaches allocation as well as aggregate performance: depending on the learned scorer, only 12–16 of the 25 highest-priority completed-record episodes appear among its first 25 unique recurrent selections.

AdmitLock turns survival of a completed-record winner into a prespecified test. Its state transition locks evidence availability, scorer-specific admission, same-timestamp batching, per-instant capacity, open-case recursion, updates, cooldown, re-entry, and scorer-independent dispositions. The resulting

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

certificate juxtaposes completed-record and recurrent Pareto orders, measures episode allocation overlap, and attaches paired uncertainty to recognition, severe-event sensitivity, selected-alert precision, and false-alert burden. A selected-sequence preservation result identifies the exact bridge: equal admissible membership, equal ordering at every active capacity boundary, and equal filtration-visible dispositions force equal states and selections by induction. The certificate records the allocation and clinical consequences when that bridge breaks, without choosing an endpoint weighting after results are known.

Clinical-trial safety review makes this comparison consequential because data monitoring committees and sponsor teams must prioritize emerging cases under finite attention (Ellenberg et al., 2002; Slutsky & Lavery, 2004; U.S. Food and Drug Administration, 2006). Operational programs use key risk indicators (KRIs) and manual listings (U.S. Food and Drug Administration, 2013; International Council for Harmonisation, 2016; TransCelerate BioPharma, 2013); modern trial systems additionally expose timestamped trajectories and narratives. We instantiate AdmitLock with four learned scorers—ChronoGuard, structured-only modeling, late fusion, and gated additive fusion—plus the operational KRI program. Every method uses the same trial partitions, time-available inputs for its modalities, calibration data, episode semantics, capacity transition, endpoints, and resampled trial-site clusters.

Contributions. The central technical contribution is a recurrent next-admissible evaluation object whose capacity clock and state make model selection operationally closed. Its preservation criterion and certificate convert chronology, boundary ordering, and repeated-case handling into inspectable model-order and case-allocation claims. The clinical study then subjects the selected model to three primary held-out trials, an independently governed four-trial replication with refitting, all-event censored recognition, timing and onset falsifications, capacity and threshold sensitivity, three training seeds, and a 228-monitor-week shadow window. Comparative replay, prospective operating-profile evidence, and observational reviewer logs retain separate estimands throughout.

2 Related Work

Retrospective multimodal clinical prediction estimates discrimination from completed combinations of notes and structured records (Rajkomar et al., 2018; Zhang et al., 2020; Lyu et al., 2022; Liu et al., 2023). Prequential evaluation instead judges predictions in the order in which information becomes available (Dawid, 1984). Neither object determines a recurrent expert queue: eligibility can change between decisions, a selected case changes later state, and a capacity boundary is encountered repeatedly. AdmitLock treats that stateful selected sequence as the evaluation object and asks whether the completed-record model order survives it.

Resource-constrained net benefit evaluates clinical predictions when only a fixed number of interventions can be delivered (Singh et al., 2023); treatment-allocation ranking and decision-aware top-k forecasting directly optimize a fixed allocation objective (Kamran et al., 2024; Heuton et al., 2025); predict-then-optimize methods likewise train against a downstream decision (Elmachtoub & Grigas, 2022). These formulations require an intervention utility or fixed allocation set. AdmitLock addresses a different model-selection question: already-trained scorers face changing admissible sets, repeated entities, and replenished capacity under one common transition. Its scalarization-free certificate reports whether all prespecified clinical directions agree and which cases create the order change.

Learning-to-defer allocates responsibility between a model and experts (Madras et al., 2018; Mozannar & Sontag, 2020), while alert-stewardship studies quantify burden, repeated notifications, routing, and human interaction (Ancker et al., 2017; Chaparro et al., 2022; Baron et al., 2021; Sutton et al., 2020). Silent clinical evaluations test models under live arrivals without changing care (Kwong et al., 2022). AdmitLock contributes the missing paired comparison on identical held-out cases: completed evidence versus a filtration-censored, stateful, repeatedly replenished queue. Calibration remains a descriptive reliability readout at the locked operating point (Guo et al., 2017; Naeini et al., 2015; Ovadia et al., 2019), rather than a weight used to manufacture a winner.

3 METHODS

3.1 A STATEFUL SELECTION OBJECT AND ITS CERTIFICATE

Let $t = 1, \ldots, T$ index distinct decision timestamps, and let $\mathcal{F}_t$ contain exactly the artifacts and scorer-independent disposition events available to the monitoring service by time t. Scheduled scoring and every artifact sharing an availability timestamp form one batch; capacity is applied once to that batch and replenishes at the next distinct timestamp. We define

$$\mathfrak{A} = (\mathcal{F}, A, B, Q, D, \Psi, \mathcal{M}, \mathcal{E}),$$

where A is the admission rule, $B = (B_t)$ the capacity schedule, Q the ranking and tie-breaking operator, D the recurrent transition, Ψ the candidate and event-linkage map, $\mathcal{M}$ the locked scorer-and-threshold registry, and $\mathcal{E}$ the endpoint registry. Each $m \in \mathcal{M}$ fixes a scorer s_m and phenotype threshold $\tau_{m,\phi}$. A candidate $c = (p, \phi, e)$ identifies a patient, phenotype, and episode. The base admissible set is

$$A_t^m = \{c : s_m(c; \mathcal{F}_t) \geq \tau_{m,\phi(c)} \text{ and } c \text{ has filtration-visible qualifying evidence}\}.$$

The recurrent state $x_t^m = (O_t^m, C_t^m, V_t^m)$ records open entries, cooldown-expiry times, and the last selected evidence version. Let $U_t^m \subseteq A_t^m \cap O_t^m$ contain open entries for which a newly ingested artifact reveals a distinct qualifying episode or raises the rule-derived provisional severity grade; both conditions are measurable from $\mathcal{F}_t$. New and update-eligible candidates are

$$E_t^m = [A_t^m \setminus (O_t^m \cup \{c : t < C_t^m(c)\})] \cup U_t^m.$$

The queue selects

$$N_t^m = Q(E_t^m, B_t) = \mathrm{Top}_{\min(B_t, |E_t^m|)}\{(c, s_m(c; \mathcal{F}_t)) : c \in E_t^m\}, \qquad R_t^m = E_t^m \setminus N_t^m,$$

where R_t^m is deferred, not discarded. The transition

$$x_{t+1}^m = D(x_t^m, N_t^m, r_t, \mathcal{F}_{t+1})$$

opens new selections, carries unresolved entries without another capacity debit, places dismissed entries in seven-day cooldown, records selected updates, and permits re-entry after cooldown only with new qualifying evidence. Here r_t contains dispositions logged by time $t+1$; outcome adjudication is never a transition input. The cap $B_t = 5$ applies to new or update-eligible selected-alert instances at each scoring instant. Weekly burden is an aggregate endpoint and may exceed five because a monitor-week contains multiple decisions.

Counterfactual selection never creates its own disposition. Before any scorer is replayed, a total candidate-indexed schedule $\rho_t(c) \in \{\text{open}, \text{dismissed}, \text{resolved}\}$ is constructed from ingestion-time evidence status and routine review-status codes visible in $\mathcal{F}_t$. The construction uses no model score, rank, selected-method identity, or adjudicated outcome. Every replay queries the same schedule through $r_t(c) = \rho_{t+1}(c)$; when no new status is visible, the previous status persists. A selected case therefore has a defined transition even if that scorer chooses a case not chosen by another scorer.

For scorer m, let $H(m)$ collect completed-record discrimination and let $L_{\mathfrak{A}}(m)$ collect recurrent-queue utility. These are two prespecified evaluation objects, not two measurements of one common estimand. For the allocation diagnostic, S_m^H contains the 25 highest-scoring completed-record patient–phenotype episodes, and S_m^L contains the first 25 unique episodes selected by locked replay after ordering by decision time and within-time queue rank. Both are pooled over the same original held-out trials; repeated selections collapse to the first selection of an episode. The certificate is

$$\mathcal{C}_{\mathfrak{A}} = \left(\preceq_H, \preceq_L, \{J(S_m^H, S_m^L)\}_{m \in \mathcal{M}}, \{\Delta L_{\mathfrak{A}}(m, m')\}_{m, m' \in \mathcal{M}}\right),$$

where $\preceq_H$ and $\preceq_L$ are Pareto orders in the prespecified discrimination and queue-utility directions, $J(S^H, S^L) = |S^H \cap S^L| / |S^H \cup S^L|$, and $\Delta L_{\mathfrak{A}}$ collects paired differences in recognition time, severe-event sensitivity, selected-alert precision, and false-alert burden. ECE accompanies the queue endpoints as a descriptive reliability diagnostic and never supplies an ex post utility weight. Conflicting endpoint directions remain a partial order; aligned point estimates induce an order that is explicitly labeled descriptive unless supported by pairwise uncertainty.

Proposition 1 (Recurrent selected-sequence preservation)**.** *Consider two recurrent readings of the same scorers, decision times, initial state, capacity, transition rule, and selected-sequence utility. For every scorer, suppose that at each decision time the readings have (i) identical base-admissible and update-eligible membership, (ii) identical score order for every pair that can straddle the capacity boundary, and (iii) identical filtration-visible disposition inputs. Then their state and selected sequence are identical at every time. Any common separable monotone utility, and therefore every scorer order induced by that utility, is preserved.*

The proof is by induction through D (Appendix S1). A two-candidate construction also shows that completed-record and recurrent-queue readings can prefer different scorers. In the clinical comparison, asynchronous threshold crossings alter membership, arriving evidence alters boundary-relevant score order, and the queue changes the evaluated sequence. AdmitLock measures their joint decision consequence; timing, threshold, capacity, and state analyses then probe the bridge conditions separately.

3.2 Study settings and evidence registry

The original retrospective portfolio contains 17 completed Phase II–III trials from March 2020 through June 2024 in cardiology (9 trials) and oncology (8 trials), totaling 58,420 patient-episodes across 6 reused sites. Trial-level partitions comprise 11 training trials, 3 calibration trials, and 3 held-out trials; the held-out set contains 9,714 patient-episodes and 1,074 adjudicated events, including 369 severe events. The prospective window covers 5 active trials, 4,812 patient-episodes, and 228 monitor-weeks from September 1 through November 23, 2025.

The independent governed portfolio has no sponsor, site, trial, adjudication-team, or calendar overlap with the original portfolio. It uses the same phenotype definitions, architecture, optimization procedure, queue transition, and endpoints with an 8/2/4 train/calibration/held-out trial split. Learned scorers are refit on its eight training trials and recalibrated on its two calibration trials; this is a procedure-level replication rather than fixed-weight transport. Its held-out partition contains 4 trials, 7,862 patient-episodes, 846 adjudicated events, 281 severe events, and 336 monitor-weeks.

Adjudication covers hepatotoxicity, neutropenia, cardiac arrhythmia, thrombocytopenia, and renal toxicity. Event onset is the first evidence satisfying phenotype criteria; recognition is the first selected alert linked to that event. Two physicians label each event with third-review resolution, yielding $\kappa = 0.83$ (95% CI, 0.80–0.86). Evidence layers retain distinct units: held-out trial–site queues estimate comparative efficacy, live selected-alert instances test operating-point and drift stability, and downstream reviewer tickets corroborate human-review process.

Across all layers, model output enters the internal review queue and carries no authority for site contact, protocol amendment, enrollment action, or patient management. Under a no-action policy, the study cleared the documented privacy, data-governance, and operational-use review path covering shadow scoring, reviewer-log analysis, and non-interventional secondary analysis. Sponsor and review-body identities are withheld in the anonymous submission.

3.3 Scorers and matched comparator suite

The learning target at time t is whether a patient–phenotype episode corresponds to an adjudicated event whose onset has occurred by t and has not passed the 28-day evaluation horizon. Six-hour grid points and artifact-triggered scoring points form examples; pre-onset points are negative, and post-onset points are positive until resolution or day 28. Structured inputs use a 30-day timestamped history of laboratory values, value changes, telemetry, exposure timing, missingness indicators, and time since measurement. Narrative inputs contain the most recent 512 WordPiece tokens from investigator progress notes, central-laboratory comments, telemetry annotations, and treatment-exposure notes available within the same history window, ordered by governed ingestion time and masked when absent. Safety-review notes, adjudication forms, endpoint-committee records, medical-monitor dispositions, reviewer comments, and direct event-label text are excluded before tokenization in every partition.

ChronoGuard maps the structured sequence to h_s with a 6-layer temporal transformer (hidden size 256, 8 heads) and the narrative tokens to h_n with 12-layer, 768-dimensional ClinicalBERT (Vaswani et al., 2017; Devlin et al., 2019; Alsentzer et al., 2019). Cross-attention uses structured states as

queries and narrative states as keys and values:

$$z_t = \mathrm{LN}(h_s + \mathrm{MHA}(h_s, h_n, h_n)), \qquad p_{t,\phi} = \sigma(w_\phi^\top z_t + b_\phi).$$

Training minimizes phenotype-weighted binary cross-entropy, with inverse-frequency weights estimated only from training trials. AdamW uses learning rate 2×10^{-4}, batch size 64, weight decay 0.01, and early stopping within 40 epochs on training-only validation folds. All learned methods use three prespecified seeds.

The comparator lock contains four alternatives. The structured-only model trains the temporal encoder and phenotype heads without a narrative branch. Late fusion independently encodes both streams and averages their phenotype logits after fitting one scalar mixing weight on the calibration partition. Gated additive fusion learns a sigmoid gate over the two embeddings before an additive prediction head. These learned comparators share the target, histories, encoders, loss, and checkpoint rule wherever applicable. The operational KRI program uses phenotype-specific structured triggers, including ALT/AST and bilirubin rules for hepatotoxicity, ANC thresholds for neutropenia, QTc and heart-rate rules for arrhythmia, platelet thresholds for thrombocytopenia, and creatinine/eGFR rules for renal toxicity; Appendix S2 gives exact definitions and its queue priority.

Every comparator receives identical timestamp-censored inputs available to its modality, trial splits, calibration partition, scoring cadence, candidate map, phenotype thresholds, deduplication policy, and $B = 5$ queue. Temperature scaling minimizes calibration-partition negative log likelihood; ChronoGuard's fitted temperature is $T = 1.39$. Phenotype-specific deployment thresholds are then chosen on the calibration partition to match each phenotype's KRI candidate rate before the common capacity cap. Neither thresholds nor tie-breaking use held-out outcomes.

Training and replay use Ubuntu 22.04, Python 3.11.8, PyTorch 2.2.1, CUDA 12.1, and four NVIDIA A100 80-GB GPUs. Appendix S3 reports run-level and aggregate compute.

3.4 Frozen queue semantics

Scoring occurs every six hours and immediately after ingestion of a new safety artifact. An artifact enters $\mathcal{F}_t$ at the later of its source-system availability and governed ingestion timestamps; event-time fields never advance availability. A scheduled tick and all artifacts with that same availability timestamp are coalesced before ranking, so processing order within an ingestion batch cannot create extra capacity. Missing narrative input is represented by a modality mask, and no future note, forward fill, or backward imputation is permitted. A patient–phenotype pair becomes a candidate after its calibrated score crosses the frozen phenotype threshold.

At each scoring instant, learned-model candidates are ordered by calibrated probability, then by earliest admissible-evidence time, patient pseudonym, and phenotype code. KRI candidates are ordered lexicographically by trigger severity (grade 4 before grade 3), number of concurrent qualifying triggers, earliest admissible-evidence time, patient pseudonym, and phenotype code. The queue returns all eligible entries when fewer than $B = 5$ are available and defers the remainder when the cap binds. The five slots are a per-instant selection budget, not queue occupancy: they replenish at the next distinct scoring timestamp. One open entry per patient–phenotype episode is allowed. A filtration-visible provisional-grade increase or a distinct qualifying episode is a material update and consumes a new slot; unchanged unresolved entries carry forward without another selected-alert instance. Dismissed entries enter seven-day cooldown and may re-enter afterward only when new qualifying evidence is present.

A monitor-week is one trial-site-week. Selected-alert precision is confirmed selected-alert instances divided by selected-alert instances; false alerts per monitor-week is unconfirmed selected-alert instances divided by monitor-weeks. Severe-event sensitivity is captured grade-≥ 3 events divided by all adjudicated grade-≥ 3 events. The fixed cap is applied after thresholding, so the protocol measures queue quality without changing the scorer to fill capacity.

3.5 Recognition estimands and statistical analysis

The primary endpoint is time from adjudicated onset to first linked selected alert. A selected instance links to an event when patient and phenotype match and selection occurs from onset through day 28 before the next distinct episode; simultaneous phenotypes are evaluated separately. Pre-onset

selections count toward alert precision and burden but do not recognize a later event, repeated post-onset alerts recognize an event only once, and recurrent episodes receive distinct identifiers. The event-population analysis includes all adjudicated events: cumulative recognition is reported at days 3, 7, and 14; events unrecognized by day 28 or the end of observed follow-up are right-censored at the earlier boundary; Kaplan–Meier quantiles summarize median recognition; and restricted mean recognition time (RMRT) integrates time without recognition through day 28.

Secondary endpoints are severe-event sensitivity, selected-alert precision, false alerts per monitor-week, and ECE (National Cancer Institute, 2017; Brier, 1950). Selected-alert precision is an instance-level workload-yield measure; unique events contribute at most once to recognition and severe-event sensitivity, preventing repeated updates from increasing event capture. ECE is descriptive and is computed over each method's slot-consuming selected-alert instances using 10 equal-width probability bins pooled across phenotypes and weighted by bin count; it is not a common-population comparison. Repeated instances count when they consume another slot, whereas carryovers do not. Paired differences use 2,000 trial-site clustered bootstrap replicates (Efron & Tibshirani, 1994); the cluster key is the trial–site pair, every episode and monitor-week under that key remains together, and the same draw is reused across methods. Site-weeks are never resampled as independent units. The prespecified confirmatory family contains the two ChronoGuard-versus-KRI endpoints, recognition time and severe-event sensitivity, with Holm–Bonferroni adjustment (Holm, 1979). Other pairwise intervals and all method orders are descriptive. Intervals are 95% percentile intervals, and seed ranges are minima and maxima over the three learned-model seeds.

Timing falsifications delay every narrative by 1, 2, or 3 days while retaining structured timestamps, and a second analysis anchors onset only to structured evidence. Nine calibration-threshold perturbations apply offsets of -0.04 to $+0.04$ on the calibrated probability scale in 0.01 increments while preserving the same $B = 5$ cap. The live shadow layer uses the identical policy; reviewer-ticket analyses remain ticket-denominated after operational queueing.

4 Results

4.1 RQ1: Does recurrent queueing change model selection?

The three completed-record point estimates share the order gated fusion $>$ late fusion $>$ ChronoGuard $>$ structured only; the gated-fusion bundle is 0.90 / 0.47 / 65.5. The four queue point estimates share a different order, ChronoGuard $>$ late fusion $>$ gated fusion $>$ structured only; ECE is reported separately and is lowest for ChronoGuard. The change does not depend on an endpoint weighting because all point-estimate directions align within each evaluation object. The two 25-episode sets intersect in 15, 13, 12, and 16 episodes for ChronoGuard, late fusion, gated fusion, and structured only, giving Jaccards of 0.43, 0.35, 0.32, and 0.47. Thus the evaluation target changes both the leading model and which episodes reach the highest-priority review set.

ChronoGuard remains the recurrent-queue leader for 3/3 seeds, and the two evaluation objects select different leaders in 8/9 calibration-threshold offsets. This sensitivity analysis preserves the fixed queue transition and capacity while perturbing admission thresholds. Figure 1 and Appendix S8 report the certificate and the prespecified perturbation grid.

4.2 RQ2: How large is the queue-selected advantage?

At $B = 5$ per scoring instant, ChronoGuard leads every queue point estimate (Table 1). Against the completed-record winner, gated fusion, paired differences in recognition days / severe sensitivity / precision / false alerts per week are -1.1 (-1.6 to -0.6) / +0.06 (+0.02 to +0.10) / +0.07 (+0.03 to +0.11) / -0.31 (-0.45 to -0.17). Against KRI, the corresponding differences are -3.5 (-4.4 to -2.7) / +0.14 (+0.08 to +0.20) / +0.19 (+0.13 to +0.25) / -0.95 (-1.18 to -0.72); Holm-adjusted p-values for the two-endpoint confirmatory family are 0.004 / 0.012. Appendix S7 gives the late-fusion and structured-only intervals, the descriptive ECE contrasts, and seed dispersion.

The raw counts recover precision, false-alert burden, and severe sensitivity exactly. Together with paired intervals against every learned comparator, they establish that the recurrent winner improves both case yield and event capture rather than trading one endpoint for another. Appendix S7 isolates the contributions of narrative evidence, temporal interaction, and probability calibration.

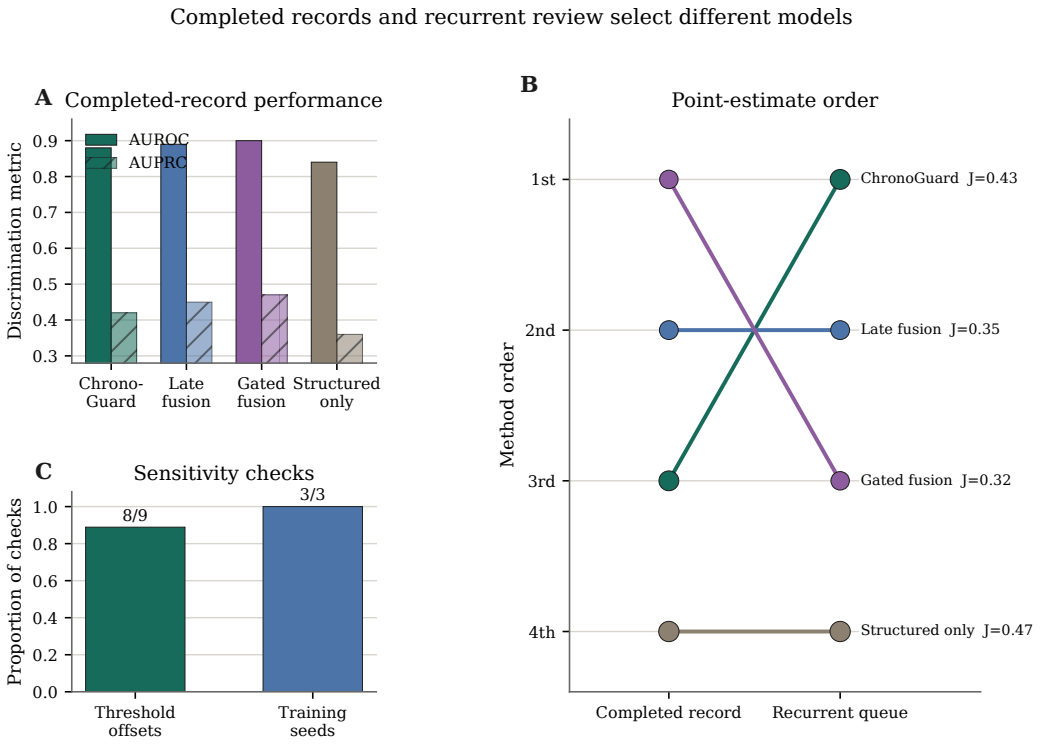


Figure 1: AdmitLock certificate on the original held-out trials. Completed-record discrimination and recurrent-queue utility produce different point-estimate orders; marker size encodes correctly normalized top-25 Jaccard overlap, and the right panel reports threshold and seed sensitivity.

Table 1: Original held-out performance with at most $B = 5$ new or updated selections per scoring instant. Metrics: recognition days / severe sensitivity / selected-alert precision / false alerts per monitor-week / ECE. Counts: selected / confirmed / severe captured / monitor-weeks.

Method	Recurrent-queue metric bundle	Raw-count bundle
ChronoGuard	4.7 / 0.86 / 0.64 / 1.80 / 0.023	2064 / 1321 / 317 / 413
Late fusion	5.4 / 0.82 / 0.59 / 2.01 / 0.029	2035 / 1205 / 303 / 413
Gated fusion	5.8 / 0.80 / 0.57 / 2.11 / 0.033	2040 / 1168 / 295 / 413
Structured only	6.5 / 0.79 / 0.55 / 2.20 / 0.031	2018 / 1110 / 292 / 413
KRI	8.2 / 0.72 / 0.45 / 2.75 / —	2065 / 929 / 265 / 413

4.3 RQ3: Is earlier recognition present across the event population?

The censored analysis includes all 1,074 held-out events (Table 2). By day 7, ChronoGuard recognizes 0.71 (0.66–0.76), compared with 0.58 (0.52–0.64) for structured only and 0.46 (0.40–0.52) for KRI. Its day-28 RMRT is 7.4 (6.8–8.1) days versus 11.1 (10.2–12.0) for KRI, a paired difference of -3.7 (-4.6 to -2.8) days. The cumulative curves separate by day 3 and remain ordered through day 28 (Figure 2).

4.4 RQ4: Do timing falsifications preserve the effect?

Delaying narrative availability by 1, 2, and 3 days moves ChronoGuard's median recognition from 4.7 days to 5.1, 5.6, and 6.1 days, while the KRI median remains 8.2 days. Relative gains are 37.8%, 31.7%, and 25.6%. The structured-onset relabeling in Table S5 yields 5.2 days for ChronoGuard and 8.4 days for KRI, with ChronoGuard precision at 0.62. A separate structured-first slice (Table S11) gives ChronoGuard a queue bundle of 6.0 / 0.78 / 0.60 / 1.49 / 0.028, ahead of late fusion, structured

Table 2: Censored recognition over all 1,074 original held-out events. RMRT is restricted through day 28.

Method	Day 3 / day 7 / day 14	RMRT (d)	Unresolved at day 28
ChronoGuard	0.48 (516/1074) / 0.71 (0.66–0.76) / 0.82 (881/1074)	7.4 (6.8–8.1)	129
Late fusion	0.43 / 0.65 / 0.77	8.6 (7.9–9.4)	165
Gated fusion	0.40 / 0.62 / 0.75	9.1 (8.3–9.9)	180
Structured only	0.37 / 0.58 / 0.71	9.8 (9.0–10.7)	210
KRI	0.29 / 0.46 / 0.62	11.1 (10.2–12.0)	294

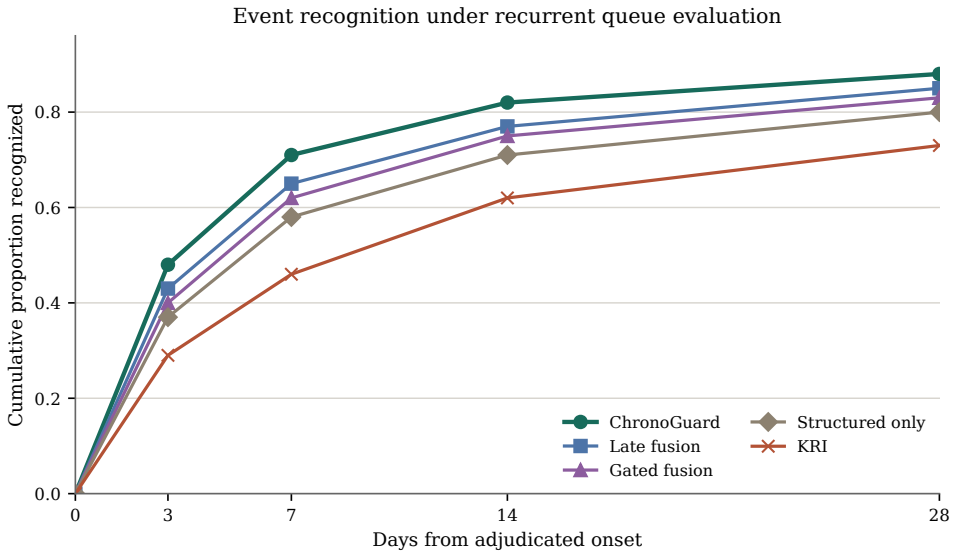


Figure 2: Cumulative recognition on all original held-out events. Curves include unresolved events and use the day-28 censoring rule.

only, and KRI. The initial 3.5-day advantage contracts by 1.4 days under a three-day narrative delay but remains 2.1 days; this graded response is consistent with an early-narrative contribution plus a residual advantage from structured temporal scoring and recurrent selection.

4.5 RQ5: Which findings persist across evidence layers and portfolios?

Across 5 active shadow trials and 228 monitor-weeks, ChronoGuard retains an operating profile close to its original held-out point: 5.1 days to recognition, 0.83 severe-event sensitivity, 0.61 selected-alert precision, 1.92 false alerts per monitor-week, and ECE 0.027. Because only ChronoGuard was shadowed under the study policy, this layer supports operating-point and workload stability, not a live comparator order. Reviewer-ticket observations remain a downstream corroboration layer with separate denominators (Appendix S16).

The independent portfolio refits and recalibrates every learned method under the unchanged procedure, then reproduces the recurrent-queue point-estimate order with attenuated effects (Table 3). ChronoGuard recognizes events 2.5 days earlier than KRI (95% CI, 1.7–3.3), improves severe-event sensitivity by 0.13 (0.07–0.19) and precision by 0.15 (0.09–0.21), and reduces false alerts by 0.66 per monitor-week (0.44–0.88). All 4 held-out trials favor ChronoGuard over KRI. The calendar-disjoint original-portfolio block in Appendix S10 is a separate replay cohort, disjoint from both the primary held-out layer and prospective shadow window.

Table 3: Independent-portfolio held-out performance after procedure-level refitting and recalibration. Queue transition and endpoint bundles follow Table 1.

Method	Locked metric bundle	Selected / confirmed / severe captured / weeks
ChronoGuard	5.6 / 0.82 / 0.59 / 2.02 / 0.031	1651 / 972 / 230 / 336
Late fusion	6.2 / 0.78 / 0.55 / 2.18 / 0.035	1627 / 895 / 219 / 336
Gated fusion	6.4 / 0.77 / 0.53 / 2.25 / 0.038	1609 / 853 / 216 / 336
Structured only	6.9 / 0.74 / 0.50 / 2.39 / 0.042	1606 / 803 / 208 / 336
KRI	8.1 / 0.69 / 0.44 / 2.68 / —	1607 / 707 / 194 / 336

5 Discussion

AdmitLock exposes a model-selection error with direct allocation consequences. In both governed portfolios, gated fusion leads completed-record point estimates while ChronoGuard leads recurrent-queue point estimates. In the original portfolio, ChronoGuard beats gated fusion by 1.1 recognition days, 0.06 severe-event sensitivity, 0.07 selected-alert precision, and 0.31 false alerts per monitor-week, with every paired interval excluding zero. The top-25 sets overlap by only 12–16 episodes, so the evaluation object changes both the model and the cases receiving highest-priority review.

The mechanism evidence maps onto the preservation criterion. Narrative and structured evidence alter scores at different times, threshold crossings alter candidate membership, and open-case transitions alter later eligibility. Narrative delays and the structured-first slice probe chronology; seed and threshold perturbations probe scoring and admission; the $B \in \{1, 2, 4, 5, 8\}$ sweep moves the active selection boundary and yields the expected recognition–precision trade-off. Reporting Pareto orders and paired endpoints separately prevents a post hoc scalar utility from determining the leader.

The evidence layers answer complementary questions. Same-input multimodal comparators isolate interaction policies under the same queue, and censored analysis retains all 1,074 events. The independently refit portfolio carries the procedure-level order across a sponsor, site, adjudication-team, and calendar boundary. The live shadow window then confirms ChronoGuard's calibration and workload profile under new arrivals, while reviewer-ticket observations corroborate the resulting review process with separate denominators.

The certificate estimates the joint consequence of changing from completed records to a filtration-censored, recurrent, capacity-constrained queue; its perturbations localize sensitivity without interpreting any one operator as a causal treatment. Evidence covers two governed portfolios, three primary and four replication held-out trials, five safety phenotypes, and a non-interventional shadow window. Applying the test to another phenotype family or review authority amounts to instantiating its evidence map, transition, and calibration for that setting.

For this regime, the practical consequence is direct: selecting by completed-record AUROC would choose gated fusion, whereas recurrent queue evaluation selects ChronoGuard and yields earlier recognition, greater severe-event capture, higher selected-alert precision, and lower false-alert burden than both gated fusion and the operational KRI program.

6 Conclusion

Completed-record ranking selects gated fusion; recurrent next-admissible evaluation selects Chrono-Guard on the same held-out outcomes. AdmitLock makes that reversal inspectable through an explicit capacity clock and state transition, a selected-sequence preservation result, and a certificate carrying model order, case overlap, paired effects, and uncertainty. The recurrent winner remains stable across same-input comparators, all-event censoring, timing perturbations, three seeds, threshold variation, and an independently refit portfolio, while its live shadow profile matches held-out operation. Clinical safety models should be selected on the chronology, repeated-case state, and scarce expert capacity that determine which episodes are actually reviewed.

Ethics statement

This non-interventional study used governed clinical-trial records and operational review logs under documented privacy, data-governance, and operational-use approval. Model outputs entered only an internal shadow review queue; site, protocol, enrollment, and patient-management actions remained outside the study policy. Access controls, minimum-necessary evidence displays, audit logging, fixed review capacity, and drift monitoring protected participants and reviewers.

Reproducibility statement

Sections 3–4 specify the evaluation tuple, data partitions, comparators, queue semantics, endpoints, uncertainty procedure, and falsifications; the appendix provides the proof, exact KRI rules, compute accounting, raw denominators, seed dispersion, failure analysis, and governance matrix.

AI Use Statement

A GPT-5.6 Terra language model was used in the documented submission-preparation workflow to revise exposition and presentation, normalize formatting, and audit citation metadata. This disclosure is limited to use documented in that workflow; it makes no claim about other stages of the research lifecycle.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, 2019. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909/.

Jessica S. Ancker, Alison Edwards, Sarah Nosal, Diane Hauser, Elizabeth Mauer, and Rainu Kaushal. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making*, 17(1):36, 2017. doi: 10.1186/s12911-017-0430-8. URL https://doi.org/10.1186/s12911-017-0430-8.

Jason M. Baron, Richard Huang, Dustin McEvoy, and Anand S. Dighe. Use of machine learning to predict clinical decision support compliance, reduce alert burden, and evaluate duplicate laboratory test ordering alerts. *JAMIA Open*, 4(1):ooab006, 2021. doi: 10.1093/jamiaopen/ooab006. URL https://doi.org/10.1093/jamiaopen/ooab006.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Juan D. Chaparro, Jonathan M. Beus, Adam C. Dziorny, et al. Clinical decision support stewardship: Best practices and techniques to monitor and improve interruptive alerts. *Applied Clinical Informatics*, 13(3):560–568, 2022. doi: 10.1055/s-0042-1748856. URL https://doi.org/10.1055/s-0042-1748856.

A. Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147 (2):278–292, 1984. doi: 10.2307/2981683. URL https://doi.org/10.2307/2981683.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994. doi: 10.1201/9780429246593. URL https://doi.org/10.1201/9780429246593.

Susan S. Ellenberg, Thomas R. Fleming, and David L. DeMets. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. John Wiley & Sons, Chichester, UK, 2002. doi: 10.1002/0470854162. URL https://doi.org/10.1002/0470854162.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922. URL https://doi.org/10.1287/mnsc.2020.3922.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kyle Heuton, Frederick Muench, Shikhar Shrestha, Thomas J. Stopka, and Michael C. Hughes. Decision-aware training of spatiotemporal forecasting models to select a top-k subset of sites for intervention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23136–23154. PMLR, 2025. URL https://proceedings.mlr.press/v267/heuton25a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979. doi: 10.2307/4615733. URL https://doi.org/10.2307/4615733.

International Council for Harmonisation. Integrated addendum to ICH E6(R1): Guideline for good clinical practice E6(R2), 2016. URL https://database.ich.org/sites/default/files/E6_R2_Guideline.pdf.

Fahad Kamran, Maggie Makar, and Jenna Wiens. Learning to rank for optimal treatment allocation under resource constraints. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3727–3735. PMLR, 2024. URL https://proceedings.mlr.press/v238/kamran24a.html.

Jethro C. C. Kwong, Lauren Erdman, Adree Khondker, Marta Skreta, Anna Goldenberg, Melissa D. McCradden, Armando J. Lorenzo, and Mandy Rickard. The silent trial—the bridge between bench-to-bedside clinical AI applications. *Frontiers in Digital Health*, 4:929508, 2022. doi: 10.3389/fdgth.2022.929508. URL https://doi.org/10.3389/fdgth.2022.929508.

Sicen Liu, Xiaolong Wang, Yongshuai Hou, Ge Li, Hui Wang, Hui Xu, Yang Xiang, and Buzhou Tang. Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(1):504–514, 2023. doi: 10.1109/JBHI.2022.3217810. URL https://doi.org/10.1109/JBHI.2022.3217810.

Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, pp. 719–728, 2022. URL https://pubmed.ncbi.nlm.nih.gov/37128451/.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pp. 6150–6160, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

National Cancer Institute. Common terminology criteria for adverse events (CTCAE), version 5.0, 2017. URL https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events/ctcae-v5-5x7.pdf.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1:18, 2018. doi: 10.1038/s41746-018-0029-1. URL https://doi.org/10.1038/s41746-018-0029-1.

Karandeep Singh, Nigam H. Shah, and Andrew J. Vickers. Assessing the net benefit of machine learning models in the presence of resource constraints. *Journal of the American Medical Informatics Association*, 30(4):668–673, 2023. doi: 10.1093/jamia/ocad006. URL https://doi.org/10.1093/jamia/ocad006.

Arthur S. Slutsky and James V. Lavery. Data safety and monitoring boards. *New England Journal of Medicine*, 350(11):1143–1147, 2004. doi: 10.1056/NEJMsb033476. URL https://doi.org/10.1056/NEJMsb033476.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine*, 3:17, 2020. doi: 10.1038/s41746-020-0221-y. URL https://doi.org/10.1038/s41746-020-0221-y.

TransCelerate BioPharma. Risk-based monitoring methodology position paper, 2013. URL https://www.transceleratebiopharmainc.com/rbminteractiveguide/research-milestones/member-research/.

U.S. Food and Drug Administration. Establishment and operation of clinical trial data monitoring committees, 2006. URL https://www.fda.gov/media/75398/download. Guidance for Clinical Trial Sponsors.

U.S. Food and Drug Administration. Oversight of clinical investigations—a risk-based approach to monitoring, 2013. URL https://www.fda.gov/media/116754/download. Guidance for Industry.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, and Ping Zhang. Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Medical Informatics and Decision Making*, 20(1):280, 2020. doi: 10.1186/s12911-020-01297-6. URL https://doi.org/10.1186/s12911-020-01297-6.

S1 PARTITIONS AND RECURRENT SELECTED-SEQUENCE PRESERVATION

Table S1 fixes the trial-level partitions used for every original-portfolio validity test. Patient episodes from one trial never cross partitions, and the same held-out outcomes support both the completed-cohort and locked-queue readouts.

Table S1: Original retrospective partitions and adjudicated event counts.

Partition	Trials	Patient-episodes	Adjudicated events
Training	11	38,924	4,102
Calibration	3	9,782	1,142
Held-out evaluation	3	9,714	1,074
Total retrospective	17	58,420	6,318

S1.1 PROOF OF THE RECURRENT PRESERVATION CRITERION

First, different completed-record and recurrent-queue orders are possible even at one decision time. Let $B = 1$ with one true candidate a and one false candidate b. For scorer f, set $f_t(b) > f_t(a)$ and $f_T(a) > f_T(b)$; for scorer g, set $g_t(a) > g_t(b)$ and $g_T(b) > g_T(a)$. Completed-record top-one utility prefers f, whereas the time-available top-one selection prefers g. This construction establishes non-equivalence of the two evaluation objects but does not attribute the difference to a single component.

For the recurrent preservation proposition in the main text, fix a scorer and use induction over decision times. The initial states are equal by assumption. Suppose the two readings have equal state at time t. Equal base-admissible and update-eligible membership gives the same E_t; equal boundary-relevant score order and capacity give the same selected set N_t and deferred set R_t. Equal disposition inputs and a common deterministic transition D therefore give equal state at $t + 1$. By induction, the states and selected sequences agree through T. A common separable monotone utility assigns the same value to each scorer under both readings, so every induced pairwise scorer order is preserved. In the clinical instantiation, the certificate records violations of these bridge conditions and the resulting difference between prespecified evaluation objects.

S1.2 SCORER-INDEPENDENT REPLAY DISPOSITIONS

The disposition schedule is constructed once, before any scorer is replayed. For every patient–phenotype episode c and decision time t, the map $\rho_t(c)$ reads only ingestion-time evidence status and routine open, dismissed, or resolved codes already visible in $\mathcal{F}_t$. It is prohibited from reading model scores, queue ranks, selected-method identity, or outcome adjudication. A newly selected case enters the open state; a later visible dismissal starts the seven-day cooldown, a visible resolution closes the episode, and the absence of a new code carries the previous status forward. Material evidence updates are computed separately from filtration-visible episode and provisional-grade changes. Consequently, ρ is total for counterfactual selections and is shared across all learned and KRI replays, even when their selected sets differ.

S2 OPERATIONAL KRI COMPARATOR

The KRI replay uses the operational phenotype rules in Table S2. Each trigger is evaluated only from structured artifacts available at the corresponding AdmitLock scoring instant. When more than five patient–phenotype episodes trigger simultaneously, candidates are ordered by grade-4 before grade-3 severity, then by the number of concurrent qualifying triggers, earliest evidence availability, patient pseudonym, and phenotype code. This deterministic order supplies the KRI score and tie-breaking rule used by the common queue.

Table S2: Phenotype-specific structured KRI trigger definitions.

Phenotype	Trigger definition
Hepatotoxicity	ALT or AST $\geq 5\times$ ULN, or ALT/AST $\geq 3\times$ ULN with total bilirubin $\geq 2\times$ ULN.
Neutropenia	ANC $< 1.0 \times 10^9$/L (grade 3) or $< 0.5 \times 10^9$/L (grade 4).
Cardiac arrhythmia	QTc ≥ 500 ms or ΔQTc ≥ 60 ms; fallback HR ≤ 40 or ≥ 130 beats/min.
Thrombocytopenia	Platelets $< 50 \times 10^9$/L (grade 3) or $< 25 \times 10^9$/L (grade 4).
Renal toxicity	Creatinine $\geq 2\times$ baseline or eGFR decrease $\geq 50\%$ from baseline.

S3 Implementation and compute accounting

Training and replay use Ubuntu 22.04, Python 3.11.8, PyTorch 2.2.1, CUDA toolkit 12.1, NVIDIA driver 550.90.07, and four NVIDIA A100 80-GB GPUs. ClinicalBERT initializes the narrative encoder; all other learned components train from random initialization. The run ledger in Table S3 includes every score-bearing fit, replay, perturbation, and bootstrap analysis.

Table S3: Audited compute for the reported analysis. Bundles give runs / A100 GPU-hours per run / aggregate A100 GPU-hours.

Run family	Run / per-run / aggregate bundle	Configuration
ChronoGuard fits	3 / 11.8 / 35.4	Four A100 80-GB GPUs; early stopping within 40 epochs
Late-fusion fits	3 / 8.6 / 25.8	Same partitions and three seeds
Gated-fusion fits	3 / 9.4 / 28.2	Same partitions and three seeds
Structured-only fits	3 / 6.1 / 18.3	Same temporal encoder and seeds
Ablation and external-portfolio fits	21 / 8.7 / 182.7	Three seeds per learned configuration
Locked, timing, and threshold replays	48 / 0.74 / 35.5	Four A100 80-GB GPUs
Bootstrap and tabulation	612 CPU-h / 18.6 GPU-h	2,000 paired cluster replicates per comparison
Total reported analysis	344.5 A100 GPU-h / 612 CPU-h	Includes all manuscript tables and figures

S4 Capacity sensitivity

Table S4 varies the maximum number of new or materially updated ChronoGuard selections at each distinct scoring timestamp around the $B = 5$ operating point while preserving thresholds and state transitions. Scheduled scoring and artifacts with identical availability timestamps are coalesced into one batch, the budget is spent once on that batch, and all slots replenish at the next distinct timestamp. Deferred candidates are rescored and reconsidered rather than discarded. Lower instantaneous capacity prioritizes precision; higher capacity accelerates recognition while increasing lower-precision workload. The p95 column sums slot-consuming selections over all decisions in a trial-site week, so its value can exceed the per-instant cap. The 7.6 trial-week statistic at $B = 5$ also differs in unit from the reviewer-week statistic in Table S6.

Table S4: ChronoGuard capacity sensitivity on the original held-out trials; B is the per-scoring-instant cap.

B	Time (d)	Sensitivity	Precision	False/wk	p95 alerts/trial-week
1	7.0 (4.9–10.7)	0.63 (0.57–0.69)	0.79 (0.73–0.84)	0.27 (0.22–0.31)	1.0
2	5.8 (3.3–9.5)	0.75 (0.69–0.80)	0.72 (0.66–0.77)	0.56 (0.48–0.64)	2.5
4	5.0 (2.5–8.8)	0.84 (0.79–0.88)	0.66 (0.60–0.71)	1.31 (1.15–1.48)	5.9
5	4.7 (2.2–8.1)	0.86 (0.81–0.90)	0.64 (0.59–0.68)	1.80 (1.62–1.97)	7.6
8	4.1 (1.8–7.2)	0.90 (0.86–0.93)	0.55 (0.49–0.60)	3.15 (2.83–3.47)	11.2

S5 Documentation-timing falsifications

Narrative-delay perturbations produce a graded change in recognition while preserving the ordering against KRI. Structured-onset relabeling retains a 3.2-day median advantage and 0.62 precision.

Table S5: Original held-out timing perturbations and structured-onset analysis.

Scenario	ChronoGuard time	KRI time	Relative gain	ChronoGuard precision
No injected delay	4.7 (2.2–8.1)	8.2 (4.8–15.3)	42.7%	0.64 (0.59–0.68)
Narratives +1 day	5.1 (2.5–8.8)	8.2 (4.8–15.3)	37.8%	0.63 (0.58–0.67)
Narratives +2 days	5.6 (2.9–9.4)	8.2 (4.8–15.3)	31.7%	0.61 (0.56–0.66)
Narratives +3 days	6.1 (3.2–10.0)	8.2 (4.8–15.3)	25.6%	0.59 (0.54–0.64)
Structured-only onset	5.2 (2.6–8.9)	8.4 (5.0–15.7)	38.1%	0.62 (0.57–0.67)

S6 Live drift and evidence-pack controls

Prospective monitors remained inside calibration and severe-sensitivity guardrails. Documentation and reviewer-load breaches were sparse and triggered the predefined operational review path. Reviewer load uses reviewer-weeks; its corrected p95 is 7.4, distinct in value and unit from the 7.6 alerts per trial-week in Table S4.

Table S6: Prospective monitors across 228 observed monitor-weeks.

Monitor	Observed summary	Trigger	Breaches
Weekly ECE	median 0.027; max 0.056	ECE > 0.06	0/228
Documentation latency	median 1.8 d; p95 4.7 d	site median > 3.0 d	13/228
Missing narratives	median 8.7%; worst-site median 14.1%	site median $> 10\%$	11/228
Site severe sensitivity	range 0.74–0.88	any site < 0.70	0/228
Reviewer load	p95 7.4 alerts/reviewer-week	> 8 alerts/reviewer-week	4/228

Each selected case is displayed through a common evidence-pack schema: trial and site pseudonyms, patient pseudonym, phenotype, score and queue rank, structured trajectory, timestamped narrative citations, and reviewer checklist. Role-based access, 180-character minimum-necessary snippets, trial-specific date shifting, direct-identifier masking, and open-event audit logs enforce the review boundary.

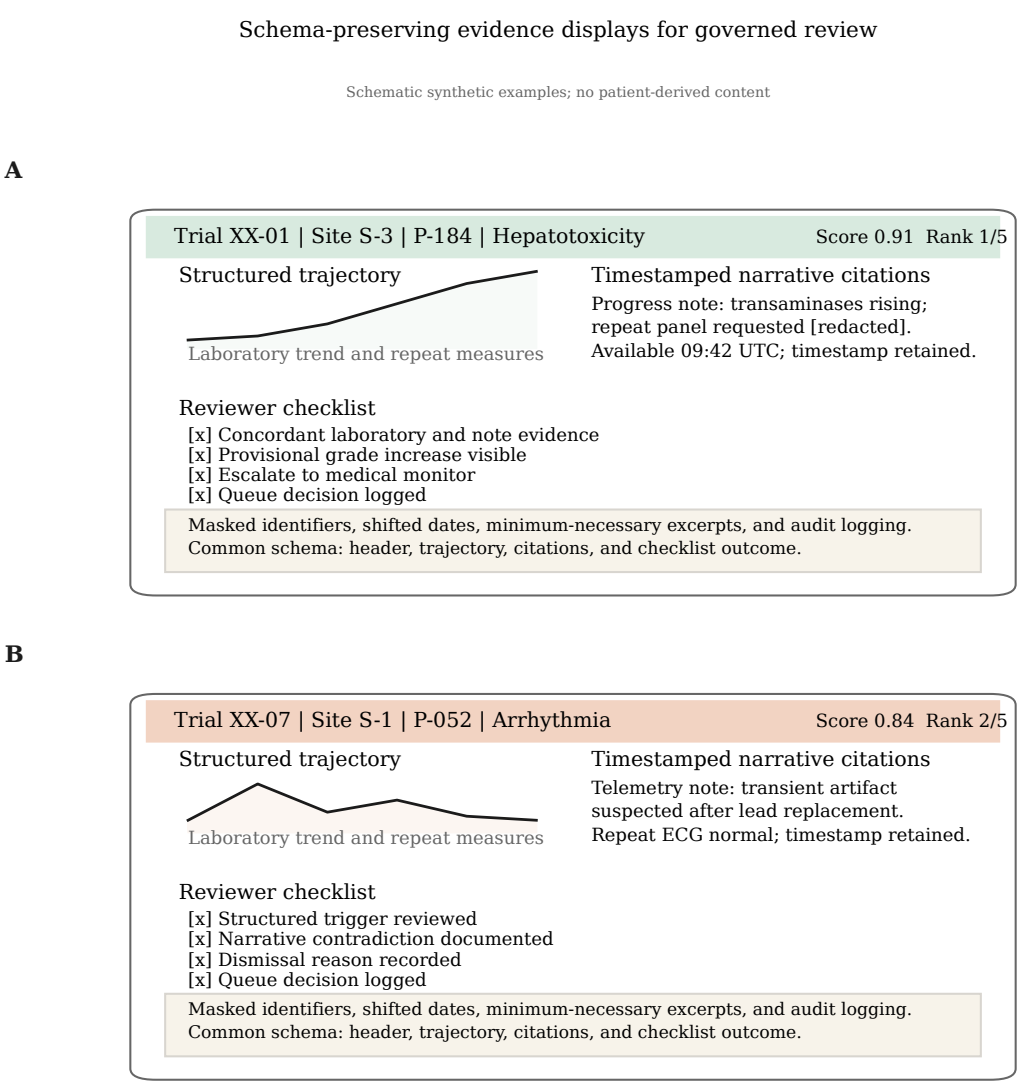


Figure S1: Schema-preserving redacted evidence-pack exemplars used for governed human review.

S7 SAME-INPUT COMPARATORS AND PAIRED UNCERTAINTY

Late fusion and gated additive fusion implement materially different multimodal interaction policies while preserving the AdmitLock comparator lock: structured trajectories, ClinicalBERT inputs, trial partitions, timestamps, calibration partition, threshold rule, and $B = 5$ queue are identical. The learned-model ranges summarize three prespecified seeds. Paired rows report ChronoGuard minus comparator for sensitivity and precision, and ChronoGuard minus comparator for time, false-alert burden, and ECE; negative time, burden, and ECE favor ChronoGuard.

The ablations use the same recurrent readout. The no-note ablation masks every narrative token at inference in a trained ChronoGuard checkpoint; unlike the separately trained structured-only comparator, it preserves the multimodal checkpoint and missing-modality pathway. It yields 6.2 days, 0.78 sensitivity, 0.54 precision, 2.24 false alerts per week, and ECE 0.029. The no-temporal-attention ablation replaces temporal self-attention by masked mean pooling before cross-attention and yields 5.6, 0.81, 0.58, 2.04, and 0.027. The no-temperature-scaling check applies unscaled probabilities and re-estimates phenotype thresholds by the same calibration-partition rate-matching rule. Because temperature scaling is monotone within a phenotype, this check preserves score order, candidate rate, and the selected sequence: recognition, sensitivity, precision, and burden remain 4.7, 0.86, 0.64,

Table S7: Original held-out same-input comparison. Metric bundles give time / severe sensitivity / precision / false alerts per week / ECE; count bundles give selected / confirmed / severe captured / monitor-weeks.

Method or difference	Locked metric bundle	Count bundle	Three-seed range
ChronoGuard	4.7 / 0.86 / 0.64 / 1.80 / 0.023	2064 / 1321 / 317 / 413	4.6–4.9 / 0.85–0.87 / 0.63–0.65
Late fusion	5.4 / 0.82 / 0.59 / 2.01 / 0.029	2035 / 1205 / 303 / 413	5.3–5.6 / 0.81–0.83 / 0.58–0.60
Gated fusion	5.8 / 0.80 / 0.57 / 2.11 / 0.033	2040 / 1168 / 295 / 413	5.7–6.0 / 0.79–0.81 / 0.56–0.58
Structured only	6.5 / 0.79 / 0.55 / 2.20 / 0.031	2018 / 1110 / 292 / 413	6.4–6.7 / 0.78–0.80 / 0.54–0.56
KRI	8.2 / 0.72 / 0.45 / 2.75 / —	2065 / 929 / 265 / 413	deterministic
ChronoGuard − KRI	-3.5 (-4.4 to -2.7) / +0.14 (+0.08 to +0.20) / +0.19 (+0.13 to +0.25) / -0.95 (-1.18 to -0.72)	paired clusters	95% CI
ChronoGuard − structured only	-1.8 (-2.5 to -1.1) / +0.07 (+0.02 to +0.12) / +0.09 (+0.04 to +0.14) / -0.40 (-0.59 to -0.21) / -0.008 (-0.012 to -0.004)	paired clusters	95% CI
ChronoGuard − late fusion	-0.7 (-1.1 to -0.3) / +0.04 (+0.01 to +0.07) / +0.05 (+0.02 to +0.08) / -0.21 (-0.34 to -0.08) / -0.006 (-0.010 to -0.002)	paired clusters	95% CI
ChronoGuard − gated fusion	-1.1 (-1.6 to -0.6) / +0.06 (+0.02 to +0.10) / +0.07 (+0.03 to +0.11) / -0.31 (-0.45 to -0.17) / -0.010 (-0.014 to -0.006)	paired clusters	95% CI

and 1.80, while ECE increases from 0.023 to 0.089. The first two contrasts isolate narrative use and temporal interaction; the monotonicity check isolates probability reliability without attributing a selection change to calibration.

S8 Evaluation-object certificate and stability

Table S8 compares both evaluation objects for all four learned models. AUROC, AUPRC, and top-5% PPV share one completed-record point-estimate order, while recognition time, severe sensitivity, precision, and false-alert burden share a different recurrent-queue point-estimate order. ECE remains a descriptive reliability diagnostic and is lowest for ChronoGuard. No endpoint weighting is needed to identify the leading point estimate within either object.

The cross-attention, late-fusion, and structured-only subset yields the same leading-model change: ChronoGuard moves from 2nd / 1st, late fusion from 1st / 2nd, and structured only from 3rd / 3rd under completed-record / recurrent-queue evaluation. Adding gated fusion strengthens the comparator suite and produces the four-method orders in Table S8.

Table S8: AdmitLock certificate for four learned methods on identical original held-out trials. Completed-record bundles give AUROC / AUPRC / PPV@top-5%; queue bundles follow Table S7. Both overlap sets contain 25 unique episodes and report Jaccard / intersection.

Method	Completed-record bundle	Recurrent-queue bundle	Completed / queue rank; overlap
ChronoGuard	0.88 / 0.42 / 61.7	4.7 / 0.86 / 0.64 / 1.80 / 0.023	3rd / 1st; 0.43 / 15 of 25
Late fusion	0.89 / 0.45 / 64.3	5.4 / 0.82 / 0.59 / 2.01 / 0.029	2nd / 2nd; 0.35 / 13 of 25
Gated fusion	0.90 / 0.47 / 65.5	5.8 / 0.80 / 0.57 / 2.11 / 0.033	1st / 3rd; 0.32 / 12 of 25
Structured only	0.84 / 0.36 / 55.9	6.5 / 0.79 / 0.55 / 2.20 / 0.031	4th / 4th; 0.47 / 16 of 25

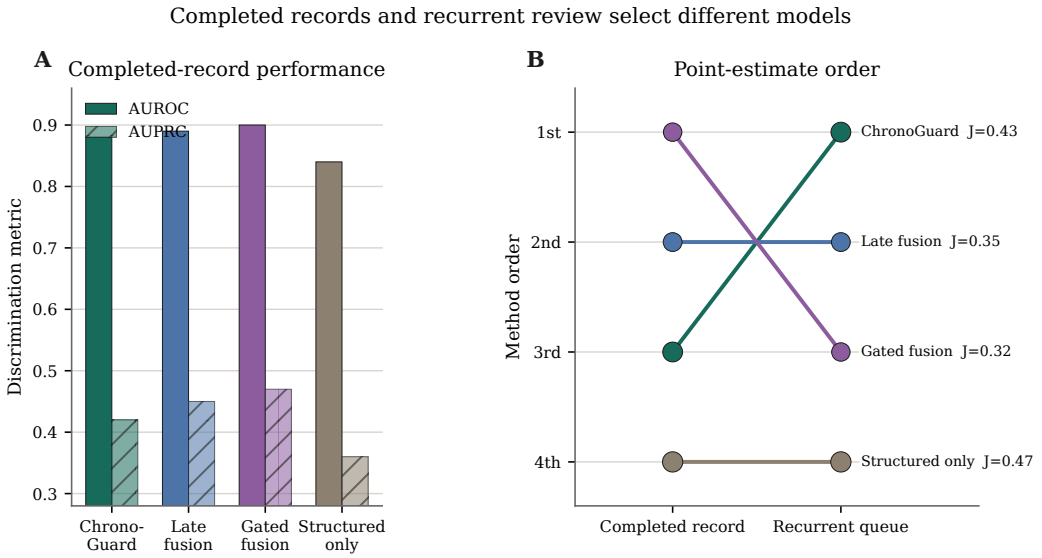


Figure S2: Completed-record and recurrent-queue point-estimate orders with top-25 selected-set overlap. Crossing lines identify order changes; marker size scales with the corrected Jaccard.

S9 Original held-out heterogeneity

All three held-out trials and all five phenotypes retain positive ChronoGuard differences against KRI for recognition time, severe sensitivity, and precision. The varying effect sizes expose within-portfolio heterogeneity while preserving direction.

Table S9: Original held-out heterogeneity. Time is comparator minus ChronoGuard, whereas severe-sensitivity and precision changes are ChronoGuard minus comparator; positive values favor ChronoGuard. Counts give events / severe events / ChronoGuard selections / weeks.

Stratum	ChronoGuard vs KRI	ChronoGuard vs structured only	Counts
Held-out trial A	+4.8 / +0.15 / +0.18	+1.7 / +0.07 / +0.09	351 / 119 / 692 / 128
Held-out trial B	+3.0 / +0.11 / +0.20	+1.3 / +0.05 / +0.08	366 / 131 / 715 / 142
Held-out trial C	+2.1 / +0.07 / +0.12	+0.6 / +0.02 / +0.05	357 / 119 / 657 / 143
Hepatotoxicity	+3.9 / +0.12 / +0.17	+1.2 / +0.05 / +0.09	223 / 77 / 438 / 80
Neutropenia	+1.8 / +0.09 / +0.11	+0.5 / +0.03 / +0.04	214 / 68 / 406 / 82
Cardiac arrhythmia	+2.5 / +0.10 / +0.14	+0.7 / +0.04 / +0.06	205 / 74 / 394 / 79
Thrombocytopenia	+3.4 / +0.16 / +0.13	+1.0 / +0.06 / +0.05	218 / 69 / 404 / 86
Renal toxicity	+2.9 / +0.08 / +0.10	+0.9 / +0.03 / +0.04	214 / 81 / 422 / 86

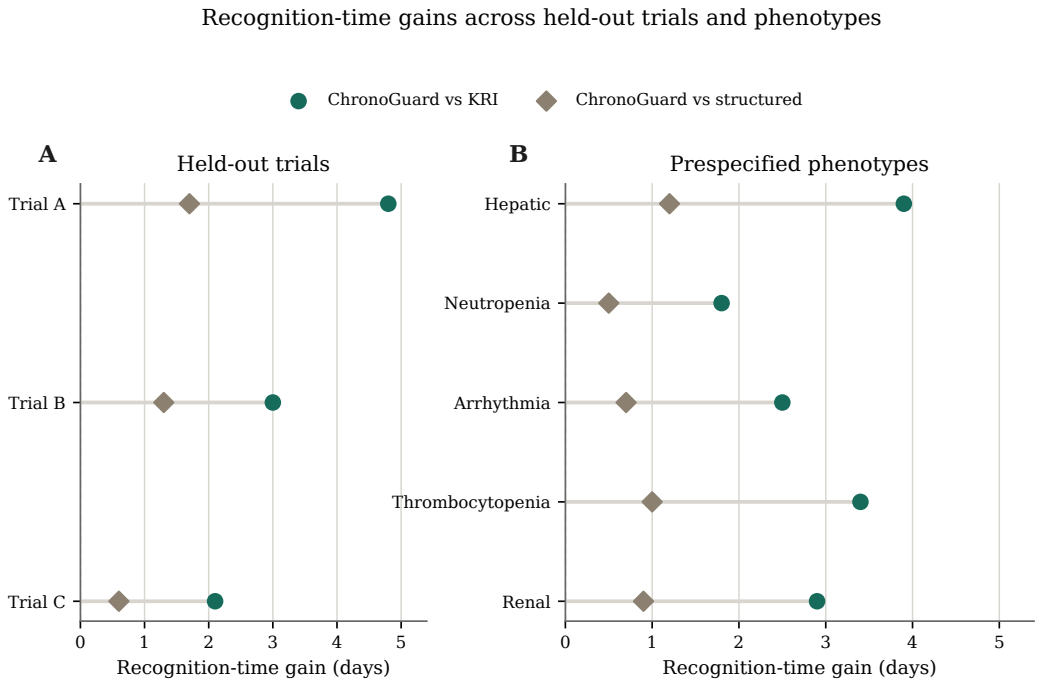


Figure S3: Recognition-time gains by held-out trial and phenotype. Positive values favor ChronoGuard.

S10 Calendar-disjoint validation

The unchanged original-portfolio protocol is replayed on a calendar block with no overlap in trials, sites, or adjudication window relative to the main held-out layer. The effect attenuates while retaining the complete method ordering.

Table S10: Calendar-disjoint original-portfolio validation. Bundles follow Table S7; denominators give events / severe / selected / confirmed / severe captured / weeks.

Method or difference	Locked metric bundle	Denominator bundle	Seed range or paired 95% CI
ChronoGuard	5.3 / 0.81 / 0.57 / 1.94 / 0.030	301 / 103 / 574 / 327 / 83 / 127	5.2–5.5 / 0.80–0.82
Late fusion	5.8 / 0.78 / 0.54 / 2.05 / 0.032	301 / 103 / 567 / 306 / 80 / 127	5.7–6.0 / 0.77–0.79
Structured only	6.3 / 0.75 / 0.51 / 2.16 / 0.034	301 / 103 / 560 / 286 / 77 / 127	6.2–6.5 / 0.74–0.76
KRI	7.6 / 0.69 / 0.43 / 2.47 / —	301 / 103 / 551 / 237 / 71 / 127	deterministic
ChronoGuard − late fusion	-0.5 (-0.9 to -0.1) / +0.03 (0.00 to +0.06) / +0.03 (0.00 to +0.06) / -0.11 (-0.23 to +0.01) / -0.002 (-0.006 to +0.002)	paired clusters	95% CI

S11 Structured-first timing isolation

This slice requires the earliest admissible onset anchor to arise from laboratory or telemetry evidence before any narrative evidence. ChronoGuard retains its lead with a smaller margin, matching the expected contribution of genuinely early narrative signal.

Table S11: Structured-first slice. Bundles and denominators follow Table S10.

Method or difference	Locked metric bundle	Denominator bundle	Seed range or paired 95% CI
ChronoGuard	6.0 / 0.78 / 0.60 / 1.49 / 0.028	146 / 54 / 295 / 177 / 42 / 79	5.9–6.2 / 0.77–0.79
Late fusion	6.5 / 0.75 / 0.56 / 1.63 / 0.031	146 / 54 / 293 / 164 / 41 / 79	6.4–6.7 / 0.74–0.76
Structured only	6.7 / 0.74 / 0.54 / 1.72 / 0.033	146 / 54 / 296 / 160 / 40 / 79	6.6–6.9 / 0.73–0.75
KRI	8.1 / 0.69 / 0.46 / 2.18 / —	146 / 54 / 319 / 147 / 37 / 79	deterministic
ChronoGuard − late fusion	-0.5 (-0.9 to -0.1) / +0.03 (0.00 to +0.06) / +0.04 (+0.01 to +0.07) / -0.14 (-0.25 to -0.03) / -0.003 (-0.007 to +0.001)	paired clusters	95% CI

S12 Governance and auditability matrix

Table S12 connects the protected object, credible misuse capability, trust assumption, enforced property, boundary, fallback comparator, held-out or live setting, benign overhead, failure mode, and paper evidence. The controls preserve the fixed scientific object across replay and live scoring.

Table S12: Governance and auditability matrix for AdmitLock.

Protected object and capability	Trust, property, and boundary	Comparator, setting, and benign overhead	Failure mode and evidence
Timestamped trial records; retrospective peeking or off-protocol reuse	Source and ingestion logs preserve timestamp and role integrity; filtration and least privilege are enforced at governed storage and scoring	KRI listings; held-out and shadow; 1.8 s p95 scoring / 3.4 s queue materialization	Delayed source documentation changes admissibility; Methods 3.4, Tables 1 and S6
Reviewer queue; threshold drift, duplicate flooding, or unauthorized escalation	Thresholds, $B = 5$, cooldown, and no-action authority are locked at the queue service	Structured KRI queue; held-out and shadow; 0.7 min median evidence-pack review overhead	Changed capacity or escalation authority breaks comparability; Table 2 and Table S6
Reviewer tickets and logs; dismissal-taxonomy bypass or log reconstruction	Immutable audit events and predefined reason codes enforce denominator traceability at the review system	Routine KRI ticket process; live observational layer; 832/845 versus 990/1,058 complete logs	Ticket mix supports process corroboration under observed assignment; Table S16
Evidence displays; overexposure or context-free scores	Redaction, shifted dates, masked identifiers, and open-event logging hold at display time	Manual chart review; shadow; 180-character citation cap	Note latency can remove narrative context; Section S6 and Figure S1

S13 INDEPENDENT GOVERNED PORTFOLIO

The second portfolio applies unchanged phenotype definitions, model architectures, optimization, queue semantics, and endpoints under a disjoint sponsor, site set, adjudication team, and calendar. Learned methods are refit on eight training trials and recalibrated on two calibration trials before evaluation on four held-out trials. The split therefore tests procedure-level replication across a governance boundary, not transport of fixed model weights.

The four trial-level ChronoGuard-versus-KRI recognition gains are 3.1 / 2.7 / 2.2 / 1.9 days. The varied magnitudes quantify attenuation across trials, while the completed-record and recurrent-queue point estimates again select gated fusion and ChronoGuard, respectively.

Table S13: Independent-portfolio held-out results. Metric and count bundles follow Table S7.

Method or analysis	Locked metric bundle	Count bundle	Paired effect or validation detail
ChronoGuard	5.6 / 0.82 / 0.59 / 2.02 / 0.031	1651 / 972 / 230 / 336	4/4 trials favor ChronoGuard
Late fusion	6.2 / 0.78 / 0.55 / 2.18 / 0.035	1627 / 895 / 219 / 336	three seeds
Gated fusion	6.4 / 0.77 / 0.53 / 2.25 / 0.038	1609 / 853 / 216 / 336	three seeds
Structured only	6.9 / 0.74 / 0.50 / 2.39 / 0.042	1606 / 803 / 208 / 336	three seeds
KRI	8.1 / 0.69 / 0.44 / 2.68 / —	1607 / 707 / 194 / 336	deterministic
ChronoGuard − KRI	-2.5 (-3.3 to -1.7) / +0.13 (+0.07 to +0.19) / +0.15 (+0.09 to +0.21) / -0.66 (-0.88 to -0.44)	paired clusters	95% CI
Evaluation-object certificate	gated 1st / ChronoGuard 1st	0.39 / 14 of 25	completed-record / queue leader

S14 Censored recognition counts

Table S14 exposes raw recognized counts at each landmark and unresolved counts at day 28. These are event denominators, distinct from selected-alert counts used for precision and burden.

Table S14: Recognition counts among all 1,074 original held-out events. Bundles give recognized by day 3 / day 7 / day 14 / unresolved at day 28.

Method	Raw-count bundle	Day-28 RMRT (95% CI)
ChronoGuard	516 / 763 / 881 / 129	7.4 (6.8–8.1)
Late fusion	462 / 698 / 827 / 165	8.6 (7.9–9.4)
Gated fusion	430 / 666 / 806 / 180	9.1 (8.3–9.9)
Structured only	397 / 623 / 763 / 210	9.8 (9.0–10.7)
KRI	311 / 494 / 666 / 294	11.1 (10.2–12.0)

S15 Failure taxonomy

We assign each false alert and each event unrecognized by day 14 to one primary failure category using the adjudication record and timestamp trace. Table S15 reports original false alerts (denominator 743), original late or missed events (193), independent-portfolio false alerts (679), and independent-portfolio late or missed events (171), in that order.

No site contributes more than 18.6% of original failures or 17.9% of independent-portfolio failures; the largest phenotype share is 23.4%. The redacted trajectories show complementary mechanisms: one hepatotoxicity event entered late after a delayed narrative linked rising transaminases to treatment, while one arrhythmia false alert followed a transient telemetry abnormality that normalized before the next queue cycle. This distribution rules out concentration in one site, phenotype, or error mechanism.

Table S15: Failure composition. Each bundle gives original false n (%) / original late-or-missed n (%) / independent false n (%) / independent late-or-missed n (%).

Primary category	Count and percentage bundle
Note latency	194 (26.1) / 54 (28.0) / 166 (24.4) / 43 (25.1)
Missing structured evidence	129 (17.4) / 34 (17.6) / 122 (18.0) / 33 (19.3)
Transient laboratory abnormality	157 (21.1) / 27 (14.0) / 145 (21.4) / 28 (16.4)
Duplicate or recurrent alert	78 (10.5) / 22 (11.4) / 83 (12.2) / 24 (14.0)
Calibration-threshold error	63 (8.5) / 15 (7.8) / 57 (8.4) / 14 (8.2)
Documentation gap	71 (9.6) / 24 (12.4) / 64 (9.4) / 18 (10.5)
Comparator disagreement	51 (6.9) / 17 (8.8) / 42 (6.2) / 11 (6.4)

S16 Observational reviewer-ticket layer

Reviewer-ticket metrics are measured after operational queueing and log capture. They use 845 ChronoGuard tickets and 1,058 routine KRI tickets and remain separate from selected-alert denominators.

Table S16: Reviewer-ticket observations from the prospective window. Times are medians with IQRs.

Metric	ChronoGuard tickets	Routine KRI tickets
Complete review log, n/N (%)	832/845 (98.5)	990/1,058 (93.6)
Review completed within 24 h (%)	83.1	44.2
Trigger to first review (h)	6.2 (2.4–17.6)	31.7 (15.4–61.9)
Onset to first review, confirmed events (d)	5.4 (2.9–9.3)	9.1 (5.5–16.8)
Escalation PPV (%)	81.2 (74.6–86.5)	68.7 (60.9–75.7)
False escalation rate (%)	1.6 (0.9–2.8)	3.8 (2.4–5.6)
Predefined dismissal reason captured (%)	97.9	72.4