

PERSPECTIVE ARTICLE

Calibrated refresh systems for clinical artificial intelligence: A sociotechnical framework for expert escalation

Arturo Loaiza-Bonilla^{1,2,3*} 

¹Massive Bio, Inc., Boca Raton, Florida, United States of America

²Department of Hematology and Medical Oncology, St. Luke's University Health Network, Bethlehem, Pennsylvania, United States of America

³Lewis Katz School of Medicine at Temple University, Philadelphia, Pennsylvania, United States of America

Abstract

Expert attention is scarce, yet clinical artificial intelligence often escalates by fixed thresholds. Hereby present Calibrated Refresh Systems (CRS), a systems-level design pattern for calibrated expert review rather than a new estimator or validated intervention. CRS accepts a low-resolution action only when calibrated residual bounds cannot cross an action threshold; otherwise, the case receives high-resolution review. It couples selective prediction, learning to defer, conformal calibration, reference-label governance, accepted-case audit, drift arbitration, subgroup safeguards, and never-fade rules. A synthetic capacity-matched mechanism check illustrates how harm-aware budgets can reduce a prespecified weighted disagreement proxy without implying improved patient outcomes. CRS yields testable operational artifacts and a staged validation roadmap, generating artifacts suitable for safety governance and regulatory documentation while preserving attention for rare, high-harm cases. Its claims depend on exchangeability, adequate calibration samples, credible reference processes, workflow capacity, and prospective clinical evaluation.

Keywords: Clinical decision support; Alert fatigue; Learning to defer; Conformal prediction; Selective deferral; Artificial intelligence governance

*Corresponding author:

Arturo Loaiza-Bonilla
(Arturo.loaiza-bonilla@sluhn.org)

Citation: Loaiza-Bonilla A.
Calibrated refresh systems for
clinical artificial intelligence: A
sociotechnical framework for expert
escalation. *J Clin Inform.*
doi: 10.36922/JCI026260008

Received: June 28, 2026

Revised: July 29, 2026

Accepted: October 12, 2026

Published online: October 26,
2026

Copyright: © 2026 Author(s).
This is an Open-Access article
distributed under the terms of the
Creative Commons Attribution
License, permitting distribution,
and reproduction in any medium,
provided the original work is
properly cited.

Publisher's Note: AccScience
Publishing remains neutral with
regard to jurisdictional claims in
published maps and institutional
affiliations.

1. The attention problem in clinical artificial intelligence

A senior emergency physician on a busy overnight shift may receive dozens of sepsis predictions, deterioration warnings, and decision-support alerts. External validations of widely deployed sepsis prediction tools have reported sensitivity below 35%, positive predictive values of approximately 7.6%–12%, and alert burdens that providers learn to tune out.^{1,2} Repeated alerts also interact with workload and complexity, increasing fatigue and override behavior.³ The result is not simply a poor score; it is erosion of the scarce resource these systems are intended to amplify: expert attention.

This is a delegation problem, not only a prediction problem. The same structure may arise in sepsis, intensive care unit (ICU) deterioration, oncology trial matching, radiology,

and chronic-care escalation; however, that generalizability is a design hypothesis, not empirical evidence. Calibrated Refresh Systems (CRS) is applicable only when a domain has an explicit consequential action boundary, an auditable low-resolution output, a credible high-resolution reference process, double-reviewed calibration cases, and capacity to act on refresh requests. Thresholds, labels, residuals, and outcomes must be recalibrated and validated separately for each task, site, and time period.

2. Principle

Calibrated Refresh Systems rest on a simple operating principle: delegate routine background work, refresh high-resolution judgment near consequential boundaries, and protect rare high-harm cases with tighter residual-risk budgets. Senior attention is treated as a constrained clinical intervention rather than an inexhaustible afterthought.

Calibrated Refresh Systems does not replace selective classification, learning to defer, conformal prediction, conformal risk control, or human-in-the-loop governance.⁴⁻¹² Table 1 compares their targets and the specific CRS claim. CRS is not a new estimator. Its proposed contribution is a testable systems-level synthesis: an action-stability gate becomes an institutional work order; reference-label exposure is recorded; accepted cases are audited; monitoring signals follow prespecified precedence rules; never-fade conditions override statistical reassurance; and every change produces traceable governance evidence.

3. Formal skeleton

Let x be a case and c a guarded decision dimension with

action threshold τ_c , such as sepsis activation or ICU upgrade. A low-resolution path produces score $\hat{s}_c(x)$ between 0 and 1; a prespecified high-resolution reference process produces $s^*_c(x)$. For the symmetric teaching case, define the absolute residual:

$$E_c = |\hat{s}_c(x) - s^*_c(x)| \quad (1)$$

The reference process must be specified before deployment. A reference-quality calibration adjudication should be blinded to the artificial intelligence (AI) recommendation whenever feasible and separated from assisted real-time review. CRS records exposure as unexposed, exposed, or unknown; only unexposed cases support the primary residual calibration. A commit-before-reveal interface can preserve independence when predictions are present in the electronic health record. If exposure is unavoidable, an independent second adjudication is required for calibration. Urgent safety overrides blinding but is analyzed separately.

From a double-reviewed calibration set, choose radius $r_{c,g}$ for class c and subgroup g at rank $k = \lceil (ng + 1)(1 - \alpha_{c,g}) \rceil$ among ordered residuals. If $k > ng$, no finite split-conformal radius exists; the conservative radius is infinite unless another prespecified valid method is used. Thus $\alpha = 0.05$ requires at least 19 cases and $\alpha = 0.01$ at least 99 merely for the rank to exist, not for precise subgroup estimation. Coverage depends on exchangeability between calibration and deployment cases. Violations suspend the guarantee and trigger restriction, recalibration, or never-fade review.

To ensure boundary-safe delegation, CRS accepts a low-resolution action only when its residual bound cannot cross

Table 1. Comparison of adjacent frameworks and the specific CRS contribution.

Approach	Primary decision target	Typical statistical or workflow object	Relation to CRS
Selective classification	Predict or abstain	Coverage-risk trade-off for a model	Supplies abstention logic; CRS converts boundary uncertainty into an owned clinical work order.
Learning to defer	Model or expert	Joint error/cost optimization using expert behavior	Supplies allocation methods; CRS adds reference-label integrity, audit, drift precedence, and governance artifacts.
Conformal prediction/risk control	Uncertainty set or bounded loss	Finite-sample coverage or prespecified risk under assumptions	Supplies residual or loss calibration; CRS links it to action thresholds, capacity, subgroups, and never-fade rules.
Human-in-the-loop governance	Clinician oversight, escalation, or override	Roles, workflow controls, accountability, and safety review	Supplies organizational context; CRS adds a calibrated gate, accepted-case audit, exposure status, and monitoring-state precedence.
CRS	Accept routine action or refresh expert review	Closed-loop sociotechnical design pattern	Novelty claim is systems integration and testability, not a new estimator or proof of clinical benefit.

Abbreviation: CRS: Calibrated Refresh Systems.

τ_c and no never-fade rule is present (Figure 1). Absolute residuals are a didactic symmetric baseline, not a universal loss. For asymmetric harm, define $U_c = \max(0, s^*c - \hat{s}c)$ and $D_c = \max(0, \hat{s}c - s^*c)$. Accept a negative action only when $\tau_c - \hat{s}c > r_{FN,c,g} + m_{FN,c}$; accept a positive action only when $\hat{s}c - \tau_c > r_{FP,c,g} + m_{FP,c}$. Separate tail budgets, margins, and clinical weights can protect false negatives and false positives differently; conformal risk control may instead target a bounded clinically weighted loss.¹⁰

For the symmetric case, if A_c denotes acceptance and D_c denotes action disagreement, then $\Pr(A_c \cap D_c) \leq \alpha_{c,g}$ under the calibration assumptions because disagreement outside the boundary band requires a residual exceeding $r_{c,g}$. This does not imply $\Pr(D_c | A_c) \leq \alpha_{c,g}$. Conditional accepted-case risk must be estimated through accepted-case audits or controlled directly. When CRS guards multiple classes and subgroups, institutions must prespecify how local budgets compose through familywise bounds, risk-control procedures, or explicit aggregate governance limits.

Accordingly, the delegation rule permits acceptance only when the applicable one- or two-sided calibrated bound cannot cross the action threshold and no never-fade rule is present.

4. Residual budgets under harm and capacity

Smaller α widens residual bands and increases refresh burden, so α must be set through a policy-constrained optimization rather than a universal closed-form theorem.

A practical two-stage policy first caps tail-risk budgets using clinical harm and rarity protection, then tunes margins and review capacity while preserving hard floors for high-harm classes. If capacity is insufficient, governance must add review capacity, relax only lower-harm bands, narrow intended use, or assign never-fade status. In a prespecified seeded synthetic mechanism check (Figure 2), high-harm cases comprised 10% of cases and carried a false-negative weight of 20; other cases carried a weight of 2. Equal and harm-aware budgets produced matched refresh fractions (41.85% and 41.86%). Harm-aware allocation increased raw unresolved disagreement from 0.0594% to 0.0664% but reduced the weighted disagreement proxy from 0.002086 to 0.001218 (41.6%). Because refreshed cases were assumed to be resolved perfectly and without delay, this result checks policy mechanics only; it does not demonstrate improved clinical outcomes.

Accept the low-resolution action only when $|\hat{s}_c(x) - \tau_c| > r_{c,g} + m_c$, and no never-fade rule applies

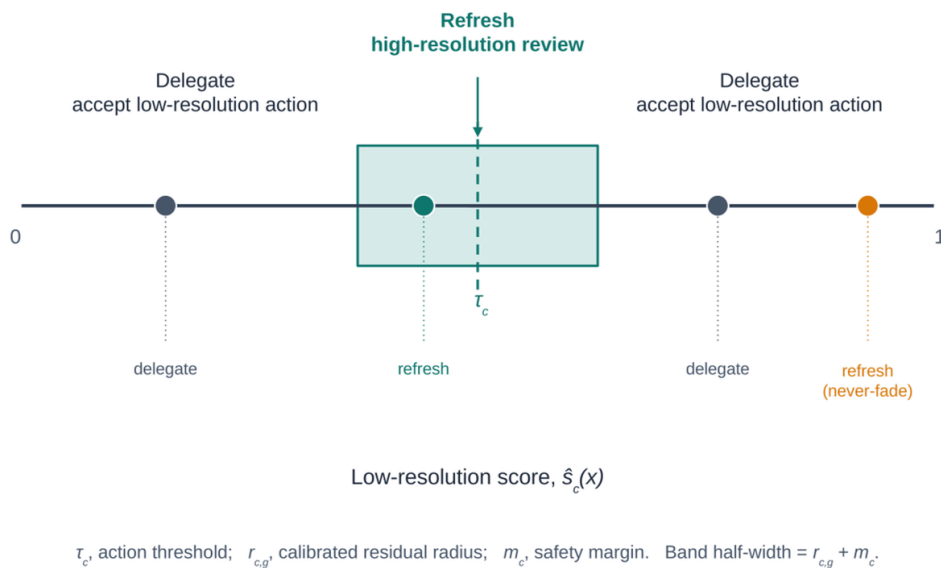


Figure 1. Boundary-band refresh rule. The low-resolution action is accepted only when the applicable calibrated residual bound cannot cross the action threshold; otherwise, high-resolution review is refreshed.

Conceptual allocation policy and capacity-matched synthetic mechanism check

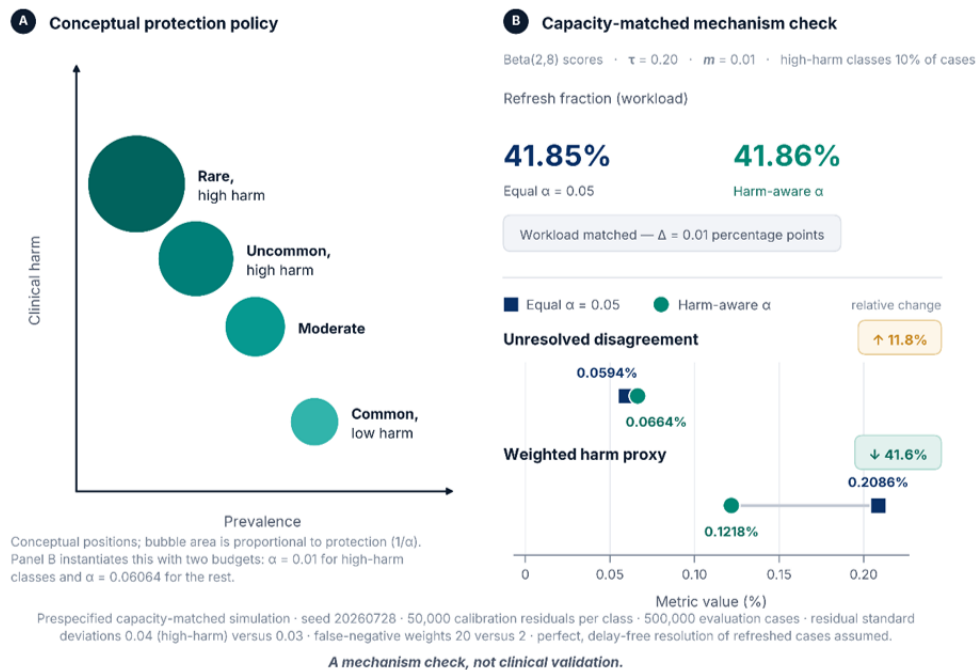
At matched workload the weighted harm proxy falls **41.6%** while unresolved disagreement rises **11.8%** (relative).

Figure 2. Harm-aware residual-risk allocation and synthetic mechanism check. Panel A shows the policy logic: bubble area is proportional to protection ($1/\alpha$), so rare, high-harm classes receive tighter budgets. Panel B reports a prespecified capacity-matched simulation with seed 20260728, 50,000 calibration residuals per class, and 500,000 evaluation cases. Scores followed Beta(2,8), $\tau = 0.20$, and $m = 0.01$. High-harm cases comprised 10%, residual standard deviations were 0.04 versus 0.03, and false-negative weights were 20 versus 2. Equal budgets used $\alpha = 0.05$; harm-aware budgets used $\alpha_{\text{high}} = 0.01$ and $\alpha_{\text{other}} = 0.06064$. Perfect, delay-free resolution of refreshed cases was assumed. The result is a mechanism check, not clinical validation.

5. Six-layer architecture

Calibrated Refresh Systems is implemented as six layers (Figure 3): low-resolution intake; boundary-band gate; high-resolution refresh; accepted-case audit; recalibration; and accountable governance. The individual components—including cumulative-sum (CUSUM) monitoring, population stability index (PSI), audit sampling, and committees—are established. The proposed novelty lies in their closed-loop coupling to an action-stability gate, reference-label integrity, operational ownership, and an auditable state transition. Design should follow sociotechnical and transparency principles,^{13,14} incorporate explainability and process evidence where useful,^{15,16} and anticipate automation bias in human–AI collaboration.^{17,18}

Monitoring mechanisms have prespecified authority and precedence. PSI is a label-free screening signal that can trigger investigation or enriched audit but cannot authorize relaxation. CUSUM, accepted-case residual bounds, and serious-event review may trigger restriction when labels are available. Audits generate evidence; they do not approve

changes. Governance authorizes changes and records the rationale. Operational states are normal, watch, restricted, and suspended. Hard safety, data-pipeline, never-fade, or subgroup breaches override reassuring aggregate metrics; when signals conflict, the more conservative state governs.

Table 2 lists the minimum governance artifacts needed to keep CRS operationally accountable.

Never-fade rules are institutional commitments that specified decisions will not be made without high-resolution review. Examples include unstable physiology, altered mental status, suicidality, child safety, pregnancy with red flags, repeat visits with concerning features, subgroup calibration failure, and explicit clinician concern. Practical deployment also requires staffed service levels across nights and weekends, queue limits, exposure logging in the electronic health record, a manual fallback when reviewers or computation are unavailable, training and compensation, and named medico-legal ownership for accepted and refreshed decisions.

Table 2. Minimum CRS governance artifacts

Artifact	Owner	Cadence	Evidence generated
Threshold registry	Clinical AI governance committee	Quarterly or after workflow change	Action thresholds, rationale, intended use, affected populations
Residual calibration report	Data science/informatics	Monthly and after drift trigger	rc,g values, accepted-and-disagree rate, conditional audit estimates
Never-fade list	Clinical leadership, ethics, safety	Semiannual and after serious event	Case categories requiring mandatory refresh regardless of score
Audit sample plan	Quality/safety office	Monthly	Random plus subgroup-enriched accepted-case review
Incident review pathway	Risk/safety office	As needed	Root-cause review for delegated decisions associated with harm
Reference-review exposure log	Clinical operations and informatics	Per review; monthly audit	Unexposed/exposed/unknown status; commit-before-reveal record; urgent override reason
Monitoring-state ledger	Safety and AI governance	At each trigger; monthly review	Detector evidence, precedence decision, operating state, approver, and rationale

Abbreviations: AI: Artificial intelligence; CRS: Calibrated Refresh Systems.

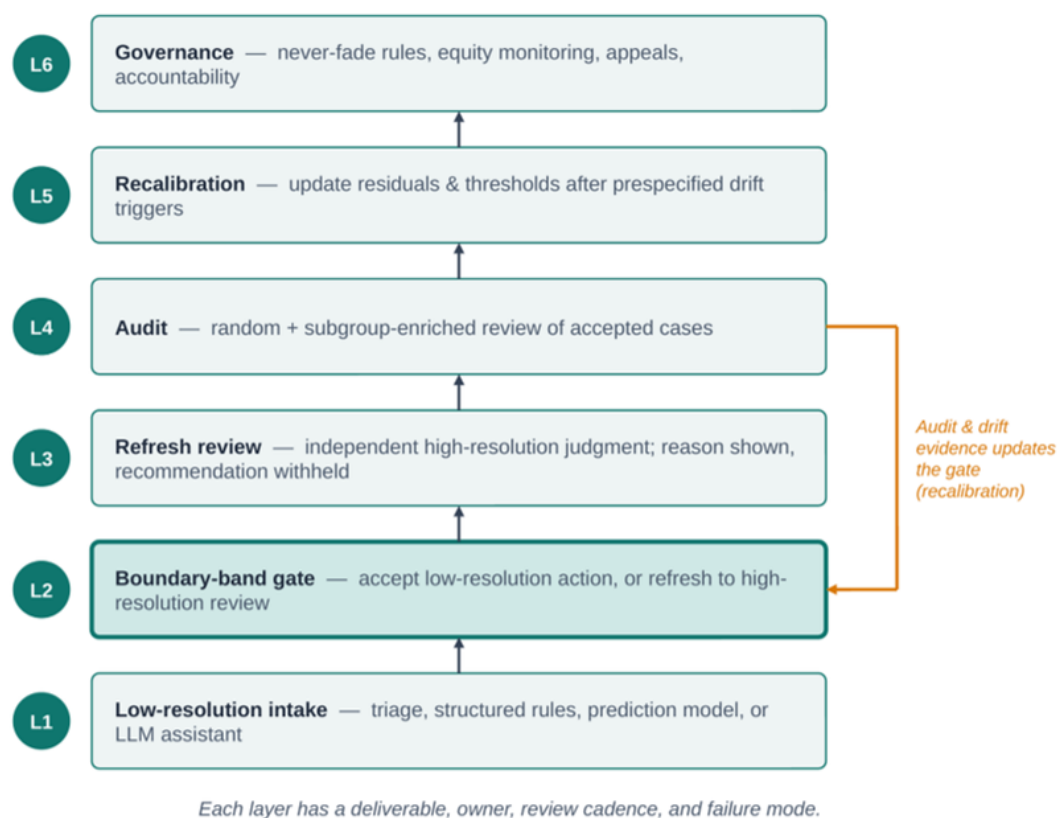


Figure 3. Six-layer Calibrated Refresh Systems architecture. Audit and drift evidence feed back into the boundary-band gate through documented recalibration and governance review.

Abbreviation: LLM: Large language model.

6. Subgroup calibration and equity

A marginal residual radius can be safe on average and unsafe for a subgroup.¹⁹ CRS therefore prespecifies clinically relevant and protected strata, monitors missingness and provenance, enriches accepted-case audits for historically harmed groups, and reports calibration, refresh rate, accepted-case disagreement, latency, appeals, and clinical utility. When reference labels are defensible, false-negative and false-positive rates can assess equalized-odds components.^{20,21} Demographic parity is reported only when appropriate to the clinical allocation question because base rates and legitimate need may differ; metric conflicts must be disclosed rather than collapsed into a single fairness claim.

Missing or inconsistently recorded protected attributes form an explicit unknown/not-recorded stratum; sensitive attributes are not inferred for real-time routing. Imputation may be used only in aggregate sensitivity analyses. If a stratum lacks enough double-reviewed cases, CRS uses the most conservative valid radius, an infinite radius, enriched auditing, or never-fade review. Partial pooling for sparse intersections is exploratory unless its guarantee is established for the method actually deployed.

7. Evidence traceability and regulatory mapping

CRS does not automatically generate, satisfy, or replace any mandatory regulatory document. Its threshold registry, residual reports, exposure logs, audit records, change approvals, and incident files can supply traceable evidence for human-authored intended-use, risk-management, validation, change-control, post-market monitoring, corrective-action, and incident documentation. These artifacts map to the National Institute of Standards and Technology AI Risk Management Framework,²²

may support—but do not constitute—an authorized Predetermined Change Control Plan under United States Food and Drug Administration guidance,²³ and may support European Union technical documentation and human-oversight duties.²⁴ They do not replace 510(k), De Novo, premarket approval, quality-system, or European Union conformity-assessment obligations. Trustworthy-AI guidance likewise emphasizes lifecycle evaluation and organizational accountability.²⁵

8. Emergency department walk-through

For illustration only, consider a community emergency department using a triage score as the low-resolution path and attending review as the refresh path. Suppose $\tau = 0.20$, $r = 0.06$, and $m = 0.01$; these values are not estimates from patient data. The symmetric boundary band spans 0.13 to 0.27. A score of 0.17 triggers refresh because it lies inside the band. A score of 0.08 with no never-fade condition remains on the standard path. A score of 0.31 with altered mental status triggers refresh because the never-fade rule overrides the score.

Evaluation should proceed in stages: retrospective specification and stress testing; silent deployment; shadow-mode workflow testing; limited comparative prospective evaluation, using a stepped-wedge or cluster-randomized design when feasible; and scaled deployment with stopping rules. Core indicators are accepted-case risk, refresh fraction, attending-minutes, queue latency, calibration stability, subgroup-specific false-negative and false-positive harm, override and appeal outcomes, patient outcomes, and cost-effectiveness. Sepsis, deterioration, and emergency triage studies illustrate settings and endpoints but do not validate CRS in those domains.^{26–28} Early-stage and implementation-evaluation guidance can structure prospective study plans.^{29,30} Table 3 lists candidate use cases, each requiring task-, site-, and time-specific validation.

Table 3. Candidate clinical CRS use cases and prospective evaluation endpoints

Domain	Low-resolution path	Refresh path	Never-fade examples	Primary endpoints
Emergency department sepsis	Triage, vitals, chief complaint, sepsis score	Attending or rapid-response review	Unstable vitals; altered mental status; staff concern	Refresh rate; time to antibiotics; missed activation; subgroup coverage
Inpatient deterioration	Early-warning score or TREWS-like system	Rapid response or ICU consult	Lactate breach; hemodynamic instability; family escalation	Unplanned ICU transfer; alarms per clinician-hour; accepted-case disagreement
Oncology trial matching	LLM plus structured eligibility screen	Trial navigator and physician review	Pediatric; pregnancy; rare biomarker; unreported toxicity	Eligibility disagreement; screen-failure rate; time to trial option

Notes: These examples are not validated applications; each requires task-, site-, and time-specific calibration and prospective evaluation.

Abbreviations: CRS: Calibrated Refresh Systems; ICU: Intensive care unit; LLM: Large language model; TREWS: Targeted Real-time Early Warning System.

9. Limitations and open questions

Calibrated Refresh Systems remains a conceptual perspective supported only by a synthetic mechanism check. It has not demonstrated patient benefit, workload reduction, or cost-effectiveness. It requires representative double-reviewed data, a credible reference process, reliable protected-attribute recording, timely reviewers, and exchangeability or a justified alternative. Reference bias, unavoidable AI exposure, missingness, distribution shift, review delay, capacity failure, gaming, and legal ambiguity can invalidate or reverse intended benefits. A safe deployment must prespecify restrictive states, manual fallback, outcome-discordance review, and prospective stopping criteria.

10. Conclusion

Clinical AI should not only calibrate predictions; it should calibrate delegation. CRS is a testable sociotechnical design pattern for deciding when routine actions can proceed and when expert judgment must be refreshed. Its statistical components are established; the contribution claimed here is their closed-loop integration with reference-label integrity, audits, drift arbitration, subgroup safeguards, never-fade rules, and accountable governance. Prospective evidence—not conceptual coherence—must determine whether CRS improves clinical care.

Acknowledgments

The author thanks colleagues in clinical informatics, emergency medicine, oncology, and operations research for discussions on escalation, alert fatigue, and the calibration of expert attention.

Funding

None.

Conflict of interest

The author is Co-Founder and Chief Medical AI Officer of Massive Bio, Inc., which develops AI-enabled clinical-trial matching and oncology decision-support products; holds AI-related patents filed in 2025; and is Co-Founder of Aequis AI Inc. None of these entities funded or directed this work, and Calibrated Refresh Systems as described is a general framework not tied to any product. The author declares no other competing interests related to this manuscript.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable. This Perspective did not involve human participants, human material, human data, animals, or cell lines, and no ethics approval or consent to participate was required.

Consent for publication

Not applicable.

Availability of data

Not applicable for clinical data. No patient or clinical datasets were generated or analyzed. [Figure 2](#) reports a fully specified, fixed-seed synthetic mechanism check; all parameters and assumptions are described in the manuscript. No supplementary materials accompany this Perspective.

Further disclosure

During manuscript preparation, the author used a large language model (OpenAI GPT-5.5 Pro) solely to improve the readability and language of the manuscript and to assist with formatting. The tool was not used to produce scientific insights, analyze or interpret data, or draw scientific conclusions. The author reviewed and approved all content, references, and claims and takes full responsibility for the manuscript.

References

1. Wong A, Otles E, Donnelly JP, *et al.* External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* 2021;181(8):1065-1070.
doi: 10.1001/jamainternmed.2021.2626
2. Ostermayer DG, Braunheim B, Mehta AM, *et al.* External validation of the Epic sepsis predictive model in 2 county emergency departments. *JAMIA Open.* 2024;7(4):ooae133.
doi: 10.1093/jamiaopen/ooae133
3. Ancker JS, Edwards A, Nosal S, *et al.* Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med Inform Decis Mak.* 2017;17(1):36.
doi: 10.1186/s12911-017-0430-8
4. Geifman Y, El-Yaniv R. SelectiveNet: A deep neural network with an integrated reject option. *Proceedings of the International Conference on Machine Learning.* PMLR. 2019;97:2151-2159. Accessed June 27, 2026. <https://proceedings.mlr.press/v97/geifman19a.html>
5. Madras D, Pitassi T, Zemel R. Predict responsibly: Improving

- fairness and accuracy by learning to defer. *Adv Neural Inf Process Syst.* 2018;31:6147-6157. <https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html>
6. Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. *Proceedings of the International Conference on Machine Learning.* PMLR. 2020;119:7076-7087. Accessed June 27, 2026. <https://proceedings.mlr.press/v119/mozannar20b.html>
7. Verma R, Nalisnick E. Calibrated learning to defer with one-vs-all classifiers. *Proceedings of the International Conference on Machine Learning.* PMLR. 2022;162:22184-22202. Accessed June 27, 2026. <https://proceedings.mlr.press/v162/verma22c.html>
8. Vovk V, Gammerman A, Shafer G. *Algorithmic Learning in a Random World.* Springer; 2005.
doi: 10.1007/b106715
9. Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. *Found Trends Mach Learn.* 2023;16(4):494-591.
doi: 10.1561/22000000101
10. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. Conformal risk control. *International Conference on Learning Representations.* In: *The Twelfth International Conference on Learning Representations (ICLR 2024).* Vienna, Austria; 2024. Accessed June 27, 2026. https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html
11. Straitouri E, Wang L, Okati N, Gomez Rodriguez M. Improving expert predictions with conformal prediction. *Proceedings of the International Conference on Machine Learning.* PMLR. 2023;202:32633-32653. Accessed June 27, 2026. <https://proceedings.mlr.press/v202/straitouri23a.html>
12. Verma R, Barrejón D, Nalisnick E. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. *Proceedings of the International Conference on Artificial Intelligence and Statistics.* PMLR. 2023;206:11415-11434. Accessed June 27, 2026. <https://proceedings.mlr.press/v206/verma23a.html>
13. Sittig DE, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual Saf Health Care.* 2010;19(Suppl 3):i68-i74.
doi: 10.1136/qshc.2010.042085
14. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *npj Digit Med.* 2020;3:41.
doi: 10.1038/s41746-020-0253-3
15. Caroprese L, Vocaturo E, Zumpano E. Argumentation approaches for explainable AI in medical informatics. *Intell Syst Appl.* 2022;16:200109.
doi: 10.1016/j.iswa.2022.200109
16. Guzzo A, Rullo A, Vocaturo E. Process mining applications in the healthcare domain: A comprehensive review. *WIREs Data Min Knowl Discov.* 2022;12(2):e1442.
doi: 10.1002/widm.1442
17. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med.* 2020;26(8):1229-1234.
doi: 10.1038/s41591-020-0942-0
18. Reverberi C, Rigon T, Solari A, et al. Experimental evidence of effective human-AI collaboration in medical decision-making. *Sci Rep.* 2022;12:14952.
doi: 10.1038/s41598-022-18751-2
19. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.
doi: 10.1126/science.aax2342
20. Teeple S, Smith A, Toerper M, et al. Exploring the impact of missingness on racial disparities in predictive performance of a machine learning model for emergency department triage. *JAMIA Open.* 2023;6(4):ooad107.
doi: 10.1093/jamiaopen/ooad107
21. Anderson JW, Visweswaran S. Algorithmic individual fairness and healthcare: A scoping review. *JAMIA Open.* 2024;8(1):ooae149.
doi: 10.1093/jamiaopen/ooae149
22. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology; 2023. NIST AI 100-1.
doi: 10.6028/NIST.AI.100-1
23. US Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff. 2025. Accessed June 27, 2026. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>
24. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union* 2024;L:2024/1689. 2024. Accessed June 27, 2026. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
25. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI Consortium. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554.

- doi: 10.1136/bmj-2024-081554
26. Adams R, Henry KE, Sridharan A, *et al.* Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med.* 2022;28(7):1455-1460.
doi: 10.1038/s41591-022-01894-0
27. Henry KE, Adams R, Parent C, *et al.* Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat Med.* 2022;28(7):1447-1454.
doi: 10.1038/s41591-022-01895-z
28. Levin S, Toerper M, Hamrock E, *et al.* Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the Emergency Severity Index. *Ann Emerg Med.* 2018;71(5):565-574.e2.
doi: 10.1016/j.annemergmed.2017.08.005
29. Vasey B, Nagendran M, Campbell B, *et al.* DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med.* 2022;28(5):924-933.
doi: 10.1038/s41591-022-01772-9
30. van de Sande D, Chung EFF, Oosterhoff J, van Bommel J, Gommers D, van Genderen ME. To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *npj Digit Med.* 2024;7:58.
doi: 10.1038/s41746-024-01064-1