A Human-centred Generative AI Dashboard for Clinician-in-the-Loop Summarisation in Paediatric Intensive Care

Anonymous

Anonymous Institution

Abstract. Paediatric intensive care units (PICUs) generate dense, fragmented data across wide developmental ranges, while populations, documentation, and reference ranges shift over time and sites. Medical AI trained once and frozen degrades under these shifts. We present a PICU clinical decision support (CDS) system designed for clinician-approved adaptation: a multi-stay timeline, age-adaptive severity encoding, and a source-linked generative-AI summary, extended with a feedback and governance layer. Accept/reject/edit signals, citation clicks, and dwell logs feed versioned prompt and retrieval policies under a human-approval gate, with audit trail, rollback, and a replay buffer, while a frozen source-attribution invariant keeps every statement tied to its EHR row. The current system does not update weights or policies online; it instruments the controls required for adaptation-ready governance. Expert inspection against a static EHR baseline finds 16 lower-severity dashboard violations (mean 1.6, no high-risk) versus 8 for the baseline (mean 2.3, one high-risk), and full situation-awareness support where the baseline has none. We specify an adaptation protocol measuring citation precision, edit rate, and backward transfer across sequential site/time windows.

Keywords: Continual learning · Clinician-in-the-loop · Clinical decision support · Paediatric intensive care · Source-linked summarisation

1 Introduction

A single PICU patient generates on the order of 1,460 data points in 24 hours; across a 22-patient night-shift cohort a physician may face up to 37,960 individual values [2]. This volume strains working memory [19] and accumulates across a shift rather than abating [8]. Unlike adult intensive care, where clinicians draw on stable, well-memorised reference ranges, PICU staff must continuously index vital signs, laboratory values, and medication thresholds against developmental bands spanning neonates to adolescents. Clinical practice is otherwise dynamic: patient populations shift, documentation practices evolve, and reference ranges differ across sites and over time. Most medical AI models, by contrast, are trained under static assumptions and remain fixed after deployment, which leads to performance degradation under distribution shift [16,13]. In PICU specifically, fewer

than 2% of developed AI models progress beyond the prototype stage, partly because deployment-phase validation is rarely reported [1].

Two interface-level failure modes compound this. Temporal data fragmentation and information-gathering failures are main barriers to CDS adoption: in one mixed-methods study, no participant across nine interprofessional simulations used the available in-room CDS tool [21]. Existing ICU dashboards improve situational awareness and pre-rounding efficiency [18,5] but were not designed for the multi-stay longitudinal trajectories that mark complex paediatric admissions, and none pair AI narrative with verifiable source attribution.

We address these gaps with an adaptation-ready CDS artefact. The system joins a temporal view of multi-stay trajectories, age-adaptive severity encoding, and a source-linked generative-AI summary that links each generated statement to its source EHR row. Around this core we specify a feedback and governance layer: clinician accept/reject/edit signals and interaction logs feed versioned prompt and retrieval policies under a human-approval gate, with an audit trail, rollback, and a replay buffer to mitigate forgetting. A frozen source-attribution invariant is designed so that summaries require resolvable source tokens before render, keeping narrative tied to evidence during adaptation. The present study evaluates the pre-adaptation artefact; the feedback loop specifies how continual updates will be governed, but prospective evidence is still required before any claim of learned improvement.

The paper contributes: (1) a PICU trajectory dashboard over multimodal clinical temporal data (vitals, labs, medications, procedures, diagnoses, microbiology); (2) a clinician-feedback and governance architecture for safe continual adaptation of source-linked summaries; and (3) a baseline expert evaluation against a static EHR viewer, plus a specified adaptation protocol.

2 Related Work

Continual learning seeks models that acquire new knowledge without catastrophic forgetting, a setting that maps onto clinical care where populations and protocols shift over time [16]. In medical image segmentation, standard continual methods can fail for medical-specific reasons, including task-order sensitivity and evaluation protocols that mask forgetting [10]. Reviews catalogue scenario taxonomies, stability–plasticity trade-offs, and data drift [13]. Two lessons guide our design: adaptation targets and forgetting metrics must be explicit, and update governance matters as much as the learning algorithm. We treat prompt/retrieval policies, threshold calibration, and alert prioritisation as adaptation targets, and keep source attribution as a non-adaptable safety invariant.

CLiMeM's clinician-in-the-loop theme covers interactive supervision, preference learning, and implicit interaction logs. Workload, explainability, and transparency are the most consistently quantified determinants of AI trust and adoption [7]. Automatically showing full explanations increases trust but also over-reliance on incorrect recommendations, whereas on-demand verification supports more self-directed reasoning [6]. Clinical LLMs raise hallucination risk: CREOLA

Table 1. R1–R6 traceability: literature gap, design decision, system feature.

Req.	Gap	Design decision	Feature
R1	Fragmented multi-stay trajectories [21]	Unified per-patient journey object on shared temporal axis	Multi-stay timeline
R2	Age-banded reference lookup burden [8]	Inline age-aware severity classification; colour and marker encoding	Age-adaptive encoding
R3	Overplotting and alarm fatigue	Scale-adaptive temporal aggregation; overview, zoom, details	Aggregation engine
R4	No source-linked GenAI narrative [18]	Source-attributed prompting with unique reference IDs	Source-linked summary
R5	Verification overhead drives XAI non-use [3]	Single-click citation-to-source navigation	Citation modal
R6	Prototype-to-deployment gap; safety framing absent [1,4]	Advisory-only framing at prompt and UI level	Advisory label

reported a 1.47% hallucination rate (44% major), with negation hallucinations especially concerning [3]. These findings motivate cheap, user-initiated feedback and a core invariant that every summary statement stays tied to its source EHR row [3,6].

ICU multipatient displays improve situational awareness, yet evidence certainty is low and no reviewed dashboard provided Level 3 (projection) support [18]. PICU situation-awareness dashboards help perception and comprehension [5], but none address multi-stay trajectories with source-linked AI narrative. This leaves a gap between continual-learning governance and bedside PICU CDS that we address with an adaptation-ready, source-linked summarisation architecture.

3 Methods

3.1 Design requirements

We treated the dashboard as a Design Science Research artefact: the contribution is a set of design hypotheses instantiated in a working system and inspected against HCI and clinical-CDS criteria [11]. Six requirements (R1–R6) were derived from a structured literature review and iterative consultations with a paediatric intensivist. Table 1 links each gap to its design decision and feature.

3.2 Dataset and journey object

We used the PIC (Paediatric Intensive Care) database [22], a large-scale deidentified dataset hosted by PhysioNet containing 12,881 patients and 13,499

admissions from five paediatric intensive care units between 2010 and 2018, covering demographics, diagnoses, ICU stays, vital signs, laboratory measurements, medications, procedures, surgeries, and microbiology. Data access followed the PhysioNet Data Use Agreement; no identifiable patient data was processed, and all evaluation was conducted analytically on the artefact or on de-identified data under that agreement, with institutional ethical approval obtained.

At start-up, twelve PIC tables pass through a four-stage pipeline: **record linkage and deduplication** across admissions; **outlier exclusion** (1st–99th percentile per vital type); **age-adaptive severity classification** (Sect. 3.3); and **journey object construction**. The journey object is the single source of truth: patient metadata and data-quality flags; an ordered admission list; a flat event stream indexed by timestamp, type, value, unit, care setting, and severity; and derived analytics (acid–base and oxygenation tags, shock index, organ-dysfunction trajectories, age-banded electrolyte flags) exposed to the summary module as a read-only `clinical_analytics` block, so the model narrates over pre-verified aggregates rather than raw rows.

3.3 Age-adaptive severity encoding

Classification is computed against six developmental bands derived from paediatric advanced life support reference ranges [12]: neonate (0–28 days), infant (1–12 months), toddler (1–2 years), young child (3–5 years), older child (6–11 years), and adolescent ($\geq$12 years). Each classified reading is rendered with redundant colour and marker shape so severity stays legible under colour-vision deficiency. Age-banded danger zones are overlaid on the timeline as shaded regions where sustained instability is recorded, with the applicable band label in the zone tooltip. This converts what would be a memory-retrieval step (recalling the correct reference range for a given age at each inspection point) into a perceptual judgment [8].

3.4 Source-linked generative summarisation

When requested, the selected timeline window is serialised into a structured JSON prompt with per-event reference identifiers (`[E..]` event-level, `[J..]` stay-level, `[H..]` prior-history). Summaries are generated through a chat API (`deepseek-chat`) via an OpenAI-compatible client with temperature 0 decoding, maximising reproducibility. Source linkage relies on the JSON payload and citation rules, not fine-tuning or retrieval augmentation; the same client can target a locally hosted alternative through deployment configuration alone. A renderer validates citation tokens, suppresses unresolved tags, and converts valid citations into clickable markers that open the source row in a modal, implementing R4 and R5 at both prompting and interface layers [3]. An advisory banner enforces R6: output is framed as inspection support, not autonomous direction.

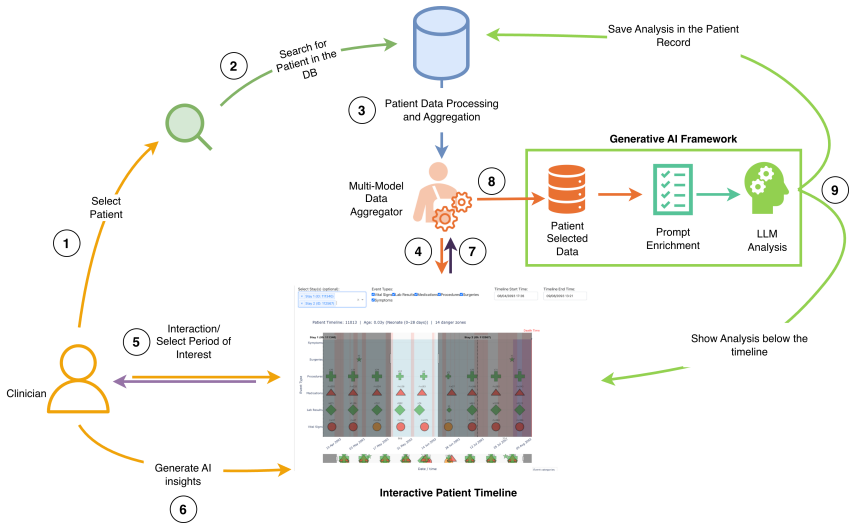

Fig. 1. End-to-end pipeline. Twelve PIC tables are normalised into a per-patient journey object (R1). The selected timeline window is serialised into the summary prompt so narrative and display share the same data (R4, R5). The feedback/governance layer (Sect. 3.5) wraps the summary stage without altering the source-attribution invariant.

3.5 Feedback and adaptation architecture

The summary stage above is static: the prompt policy is fixed and no feedback is captured. To make it adaptation-ready without abandoning source linkage, we wrap it in a feedback and governance layer (Fig. 1). The current implementation instruments the signals and controls below; it does not yet apply online policy updates.

Feedback signals. Each summary statement exposes accept, reject, and edit controls. A reject or edit records the offending statement, its source tokens, and the clinician's correction. Implicit signals are captured passively: citation clicks (which source rows a clinician opens), and window dwell (how long a summary is read before action). All signals are timestamped and stored against the prompt-policy version that produced them.

Adaptation targets. Feedback is mapped to four updateable targets: (i) the summary *prompt policy* (which aggregation patterns and analytic blocks to surface); (ii) a *retrieval memory* of clinician-endorsed summary patterns and corrections, consulted at generation time; (iii) *threshold calibration* for age-adaptive severity bands; and (iv) *alert prioritisation* for the events-per-day density layer. Model weights are not a target in the present design; the system remains prompt-and-retrieval based so adaptation is auditable at the policy level.

Continual-learning safeguards. Five controls govern how updates are applied: (1) a *frozen source-attribution invariant*: every summary statement, pre- and post-adaptation, must carry resolvable source tokens before render, so adap-

Table 2. Feedback and adaptation loop: stage, captured signal, binding safeguard.

Stage	Signal / action	Safeguard
Summarise	Source-linked statements, temp. 0	Frozen source-attribution invariant
Verify	Citation click, dwell	Audit trail
Correct	Reject / edit + correction text	Versioned against producing policy
Accumulate	Feedback windowed by site/time	Replay buffer seeded
Propose	Prompt/retrieval/threshold update	Human-approval gate
Promote	Versioned policy swap	Rollback + replay consolidation

tation cannot promote narrative that lacks evidence linkage; (2) *versioned policies* with a *human-approval gate* before any policy is promoted, since silently changing endorsed output would violate trust [6]; (3) an *audit trail* recording every feedback event and policy change; (4) *rollback* to any prior policy version; and (5) a *replay buffer* of endorsed summary patterns replayed when policies update, to reduce catastrophic forgetting of earlier endorsed behaviour [10,16].

Adaptation loop. The intended loop is: summarise → clinician verifies via citation → accept/reject/edit → signals accumulate → proposed policy update → human-approval gate → versioned promotion with replay-buffer consolidation. Table 2 summarises the stages and the safeguard that binds each.

3.6 Baseline EHR viewer

The control is a static web application providing tabular access to the same PIC patient data across five tabs (Overview, Procedures and Surgeries, Vitals, Labs, Medications), with no interactive analytics, temporal aggregation, severity encoding, or AI features. Both interfaces load the same cohort [22], so observed differences are attributable to interface design.

4 Evaluation

4.1 Expert baseline evaluation

We applied a multi-method expert evaluation using two analytical techniques that require no human participants, following design-stage practice in HCI and clinical informatics [14]: (1) a multi-framework heuristic evaluation, and (2) a comparative design artefact analysis. The walkthrough task was: locate critical vital-sign abnormalities across multiple admissions and verify an AI-generated statement against its source row, covering R1–R6. Each violation was logged with a 0–4 severity score [14] and a clinical-consequence rating [4], then mapped to R1–R6.

Table 3. Heuristic evaluation: violations and mean severity by framework.

Framework	**Dashboard**		**Baseline**	
	n	mean sev.	n	mean sev.
Nielsen [14]	4	1.3	5	2.2
TURF [17]	2	1.5	1	2.0
ISO 9241-110	4	1.3	2	2.5
HIMSS CDS [15]	5	2.0	—	—
FDA UX [20]	1	2.0	—	—
All frameworks	**16**	**1.6**	**8**	**2.3**

We applied five frameworks. Nielsen's ten heuristics, TURF [17], and ISO 9241-110 covered both interfaces; HIMSS CDS [15] and FDA UX [20] covered the dashboard only (baseline is not CDS). Across all five, the dashboard accrued 16 violations (mean severity 1.6, no high-risk) versus 8 for the baseline (mean 2.3, one high-risk). The sole Severity-4 violation is the baseline's missing inline age-banded reference ranges: every vital-sign inspection there requires recalling the correct developmental band before a severity judgment, the "death by a thousand cuts" burden [8], which R2 removes by resolving severity at the interface. Under the three shared frameworks counts are comparable (4–5 each) but the dashboard has lower mean severity in each; remaining open violations cluster at the AI-transparency boundary and need no architectural change. Table 3 reports the breakdown.

The comparative design artefact analysis examined both interfaces through three lenses. The baseline is low-mediation (temporal integration, severity classification, and narrative construction remain internal), while the dashboard is high-mediation, moving those operations into the interface. The journey object performs cross-admission stitching at load time, removing a main source of the fragmentation that drives CDS non-use [21]. Under situation awareness [9] the gap is largest: the dashboard supports all three levels (severity-encoded perception, cross-stay comprehension, and GenAI-supported projection), while the baseline supports none. No reviewed ICU dashboard provides Level 3 projection support [18]; when resuscitation cognitive load reaches the 82.3 ± 15.6 NASA-TLX levels documented in PICU [8], the absence of external projection is a plausible risk factor for missed projection under load, not a harm outcome measured in this study.

4.2 Adaptation protocol

The expert evaluation above assesses the pre-adaptation artefact. To test whether the feedback loop in Sect. 3.5 improves summaries without forgetting, we specify the following protocol. Using held-out admissions partitioned into sequential site/time windows, clinician edits are collected on generated summaries at each window. Three metrics are tracked across windows: *citation precision* (fraction

of cited statements a clinician endorses without edit), *edit rate* (fraction of statements edited or rejected), and *backward transfer* (change in citation precision on earlier windows after a policy update on a later window), the standard forgetting diagnostic [10]. Updates pass the human-approval gate only when backward transfer stays within a pre-registered tolerance. This protocol is the prospective evaluation; no adaptation results are reported here.

5 Discussion and Limitations

The expert baseline shows that the dashboard addresses built-in limitations of conventional EHR viewers without introducing high-risk violations, and that source-linked generative summaries target the Level 3 projection support that reviewed ICU dashboards lack [18]. The feedback and governance layer is what makes the contribution continual rather than static: source attribution remains a frozen invariant, and every policy update is bound to a human-approval gate, audit trail, rollback, and replay buffer, so summary behaviour can evolve under clinician supervision rather than degrade silently under distribution shift. Two design principles follow. Age-banded severity should be resolved at the interface, not in clinician memory. AI verification should stay cheap and user-initiated [6]. The main open interface issue is the lack of role-specific defaults; that gap does not require rebuilding the data model.

Several limitations bound these claims. The inspection used a single evaluator, so there is no inter-rater reliability estimate [14]. The PIC cohort is single-country and needs local validation before any site-specific claim. The present study evaluates only the pre-adaptation artefact; Sect. 4.2 remains prospective. The system uses clinical multimodal temporal data without imaging, and adaptation targets prompt/retrieval policies rather than model weights.

6 Conclusion

We presented an adaptation-ready, clinician-in-the-loop architecture for source-linked summarisation of PICU trajectories: multi-stay timeline, age-adaptive severity encoding, source-linked generative summary, and a governed feedback layer bound by a frozen source-attribution invariant. Expert evaluation shows fewer, lower-severity violations than a static EHR baseline and full situation-awareness support. The specified adaptation protocol is the next step toward prospective evidence of governed continual adaptation at the bedside.

References

1. Almontasir Abdallah, et al.: The role of artificial intelligence in pediatric intensive care: A systematic review. Cureus **17**(3), e80142 (2025). https://doi.org/10.7759/cureus.80142
2. Asgari, E., et al.: Impact of electronic health record use on cognitive load and burnout among clinicians: Narrative review. JMIR Med Inform **12**, e55499 (2024). https://doi.org/10.2196/55499
3. Asgari, E., et al.: A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine **8**(1), 274 (2025). https://doi.org/10.1038/s41746-025-01670-7
4. Borycki, E., Kushniruk, A., Carvalho, C.: A methodology for validating safety heuristics using clinical simulations: Identifying and preventing possible technology-induced errors related to using health information systems. Computational and Mathematical Methods in Medicine **2013**, 526419 (2013). https://doi.org/10.1155/2013/526419
5. Boudreault, L., et al.: Situation awareness-oriented dashboard in ICUs in support of resource management in time of pandemics. IEEE Journal of Translational Engineering in Health and Medicine **11**, 151–160 (2023). https://doi.org/10.1109/JTEHM.2023.3241215
6. Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: 2015 International Conference on Healthcare Informatics (ICHI 2015). pp. 160–169. IEEE (2015). https://doi.org/10.1109/ICHI.2015.26
7. Choudhury, A., Shamszare, H.: Human factors influencing trust in healthcare artificial intelligence: Systematic literature review. IISE Trans Occup Ergon Hum Factors **14**(2), 116–131 (2026). https://doi.org/10.1080/24725838.2026.2625655
8. Ehrmann, D.E., et al.: Cognitive load in pediatric critical care medicine: Tsunamis and a thousand cuts. Crit Care Explor **7**(10), e1329 (2025). https://doi.org/10.1097/CCE.0000000000001329
9. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors **37**(1), 32–64 (1995). https://doi.org/10.1518/001872095779049543
10. Gonzalez, C., Lemke, N., Sakas, G., Mukhopadhyay, A.: What is wrong with continual learning in medical image segmentation? arXiv preprint arXiv:2010.11008 (2021). https://doi.org/10.48550/arXiv.2010.11008
11. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly **28**(1), 75–105 (2004)
12. Kleinman, M.E., et al.: Part 14: Pediatric advanced life support. Circulation **122**(18_suppl_3), S876–S908 (2010). https://doi.org/10.1161/CIRCULATIONAHA.110.971101
13. Kumari, P., et al.: Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. arXiv preprint arXiv:2312.17004 (2024). https://doi.org/10.48550/arXiv.2312.17004
14. Nielsen, J.: Usability Engineering. Academic Press, Boston, MA (1994)
15. Osheroff, J.A., et al.: Improving Outcomes with Clinical Decision Support: An Implementer's Guide. Healthcare Information and Management Systems Society (HIMSS), Chicago, IL, 2nd edn. (2012)
16. Qazi, M.A., et al.: Continual learning in medical imaging: A survey and practical analysis. arXiv preprint arXiv:2405.13482 (2024). https://doi.org/10.48550/arXiv.2405.13482

17. Staggers, N., Kobus, D., Brown, C.: Nurses' evaluations of a novel design for an electronic medication administration record. Comput Inform Nurs **25**(2), 67–75 (2007). https://doi.org/10.1097/01.NCN.0000263981.38801.be
18. Strechen, I., et al.: Centralized multipatient dashboards' impact on intensive care unit clinician performance and satisfaction: A systematic review. Applied Clinical Informatics **15**(3), 414–427 (2024). https://doi.org/10.1055/a-2299-7643
19. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive Science **12**(2), 257–285 (1988). https://doi.org/10.1207/s15516709cog1202_4
20. U.S. Food and Drug Administration: Applying human factors and usability engineering to medical devices: Guidance for industry and food and drug administration staff. Tech. rep., U.S. Food and Drug Administration, Center for Devices and Radiological Health, Silver Spring, MD (2016), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices
21. Zackoff, M., et al.: Understanding clinical decision support failures in pediatric intensive care units via applied systems safety engineering and human factors problem analysis: Insights from the DISCOVER learning lab. J Patient Saf **21**(7 Suppl), S21–S28 (2025). https://doi.org/10.1097/PTS.0000000000001358, https://doi.org/10.1097/PTS.0000000000001358
22. Zeng, X., Yu, G., Lu, Y., et al.: PIC, a paediatric-specific intensive care database. Sci Data **7**, 14 (2020). https://doi.org/10.1038/s41597-020-0355-4