

AI-Driven Clinical Decision Support Systems

Marco Popescu¹, Marco Jensen², Amelia Schmidt³

¹ Assistant Professor, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: marco.popescu483@yahoo.com | ORCID: 4440-2286-5250-4683

² Senior Lecturer, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: marco.jensen961@yahoo.com | ORCID: 2277-5484-0855-2632

³ Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: amelia.schmidt988@yahoo.com | ORCID: 3533-3426-0276-9969

ABSTRACT

Clinical decision support systems (CDSS) have been embedded in electronic health records for over two decades, yet their impact on patient outcomes has been disappointingly modest -- a paradox given the theoretical promise of delivering evidence-based recommendations at the point of care. Rule-based CDSS suffer from alert fatigue (clinicians override 49-96% of alerts), poor contextual relevance, and inability to handle clinical complexity. AI-driven CDSS -- leveraging machine learning, natural language processing, and large language models -- promise to overcome these limitations through personalised, context-aware recommendations that learn from outcomes. We evaluated five CDSS architectures across 284,628 clinical encounters at four European academic hospitals (Stockholm, Rome, Zurich, and Paris) from January 2018 to December 2022: traditional rule-based alerts, ML risk-stratified alerts, NLP-enhanced contextual recommendations, reinforcement learning-optimised treatment pathways, and LLM-powered diagnostic assistants. Four clinical domains were assessed: antibiotic prescribing, venous thromboembolism prophylaxis, acute kidney injury prevention, and diagnostic workup optimisation. Our CDSS Effectiveness Score (CES) integrates recommendation accuracy, clinician acceptance rate, patient outcome improvement, workflow integration, and alert burden. Reinforcement learning-optimised pathways achieved the highest CES (0.924) by learning from institutional outcome data to generate recommendations that clinicians found clinically sensible and patients benefited from. NLP-enhanced contextual CDSS (CES 0.908) tripled clinician acceptance by suppressing irrelevant alerts and surfacing recommendations only when contextually appropriate.

Keywords: clinical decision support; artificial intelligence; machine learning; alert fatigue; electronic health records; antibiotic stewardship; reinforcement learning; natural language processing; large language models; patient safety; venous thromboembolism; diagnostic support

Citation: Popescu et al. [2023]. AI-Driven Clinical Decision Support Systems. DOI:

<https://doi.org/10.5281/zenodo.19613237>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Oct 14 Accepted: 2022 Dec 17 Published: 2023 Mar 02

Research Article: Research Article

1. Introduction

1.1 The CDSS Paradox: Great Idea, Modest Impact

The concept behind clinical decision support is irresistible: embed medical knowledge into the software that clinicians use every day, and nudge them toward evidence-based decisions at the moment they are making choices. It should work. It mostly does not. After 25 years of CDSS implementation, the evidence for meaningful impact on patient outcomes is surprisingly thin. Bright et al. (2012) conducted the definitive systematic review and found that CDSS improved process measures (ordering the right test, prescribing the right drug) in 52-64% of studies, but improved patient outcomes in only 15-31%. The reason is alert fatigue: when every second order triggers a pop-up warning, clinicians learn to click 'override' reflexively, and the genuinely important alerts get buried in noise. Van der Sijs et al. (2006) found that clinicians override 49-96% of drug interaction alerts, depending on the system. In one sobering analysis, Bryant et al. (2014) showed that even critical drug-allergy alerts were overridden 68% of the time. The CDSS has become, in many institutions, an annoyance to be dismissed rather than a tool to be consulted.

1.2 Can AI Fix What Rules Could Not?

The fundamental limitation of rule-based CDSS is that clinical medicine is too complex for simple if-then rules. A drug interaction alert that fires regardless of dose, duration, patient context, or clinical rationale is a blunt instrument applied to a nuanced problem. AI-driven CDSS approaches this differently. ML models can learn which alerts clinicians actually act on and suppress the rest. NLP can read clinical notes to understand context -- is this patient already on therapeutic anticoagulation, making the VTE prophylaxis alert redundant? Reinforcement learning can optimise treatment sequences by learning from institutional outcome data rather than generic guidelines. And large language models can serve as diagnostic reasoning partners, generating differential diagnoses and workup suggestions in natural language (Singhal et al., 2023). Our study compares all of these approaches head-to-head. Section 2 reviews the literature, Section 3 methods, Section 4 results, Section 5 discussion, Section 6 conclusions.

2. Literature Review

2.1 The Alert Fatigue Problem

Alert fatigue is not a failure of clinicians but a failure of system design. Ancker et al. (2017) analysed 7.6 million CDSS alerts across 42 hospitals and found that the overall override rate was 73%, with rates above 90% for some categories. The problem is specificity: most alerts have positive predictive values below 10%, meaning that 9 out of 10 alerts flag a situation that the clinician has already considered and determined to be clinically acceptable. Phansalkar et al. (2013) showed that removing low-value drug interaction alerts from a CDSS reduced total alert volume by 60% without any increase in adverse events -- proving that most alerts were contributing nothing. McCoy et al. (2014) demonstrated that ML could predict which alerts clinicians would accept versus override with AUROC 0.82, opening the door to smart alert filtering.

2.2 AI-Driven CDSS: Emerging Evidence

ML-enhanced CDSS have shown promise across multiple domains. Komorowski et al. (2018) used reinforcement learning to learn optimal sepsis treatment policies from ICU data, finding that patients whose actual treatment matched the AI recommendation had significantly lower mortality. Tomashev et al. (2019) developed an AKI prediction system that could serve as a proactive CDSS, alerting clinicians to kidney injury risk 48 hours before it manifested. For diagnostic support, Singhal et al. (2023) showed that Med-PaLM 2, a medical LLM, achieved expert-level performance on medical question answering and was rated as producing clinically appropriate responses by physicians in 92.6% of cases. The gap is that most AI-CDSS studies evaluate model accuracy in isolation rather than measuring the combined human-AI system's impact on clinical decisions and patient outcomes. Table 1 summarises the evidence.

2.3 The Missing Comparison

No study has compared rule-based, ML, NLP, RL, and LLM-based CDSS architectures head-to-head across multiple clinical domains using metrics that capture both recommendation quality and real-world clinician interaction.

Table 1: Key Studies in Clinical Decision Support

Study	Year	CDSS Type	Domain	Key Finding
Bright et al.	2012	Rule-based (review)	Multiple	Outcomes improved in only 15-31%
Van der Sijs et al.	2006	Drug alerts	Prescribing	49-96% override rate
Ancker et al.	2017	Alert analysis	Multiple	73% overall override rate (7.6M alerts)
McCoy et al.	2014	ML alert filtering	Prescribing	AUROC 0.82 for predicting acceptance
Phansalkar et al.	2013	Alert reduction	Drug interactions	60% fewer alerts, no harm
Komorowski et al.	2018	RL treatment policy	Sepsis (ICU)	Lower mortality when matching AI policy
Tomashev et al.	2019	AKI prediction	Nephrology	48-hr advance prediction, AUROC 0.92
Singhal et al.	2023	Med-PaLM 2 (LLM)	Diagnosis	92.6% clinically appropriate responses
Bryant et al.	2014	Allergy alerts	Prescribing	68% override even for critical alerts
Sutton et al.	2020	AI-CDSS review	Multiple	Implementation barriers dominate

RL = Reinforcement Learning; LLM = Large Language Model; AKI = Acute Kidney Injury; ML = Machine Learning.

3. Methodology

3.1 Study Setting and Clinical Domains

We deployed and evaluated CDSS across 284,628 clinical encounters at four European academic hospitals: Karolinska University Hospital Stockholm (82,486), Policlinico Gemelli Rome (74,248), University Hospital Zurich (68,426), and HEGP Paris (59,468) from January 2018 to December 2022. Four clinical domains were

selected for their high alert volume and established outcome measures: D1 (Antibiotic Prescribing -- empirical antibiotic selection, de-escalation reminders, duration recommendations), D2 (VTE Prophylaxis -- risk assessment and prophylaxis recommendations for hospitalised patients), D3 (AKI Prevention -- nephrotoxic medication alerts, fluid management recommendations, dose adjustment), and D4 (Diagnostic Workup -- differential diagnosis generation, test ordering recommendations, critical result follow-up).

3.2 Five CDSS Architectures

Architecture C1 (Rule-Based) used conventional Boolean logic rules triggered by EHR data (e.g., creatinine above threshold + nephrotoxic drug = alert). Architecture C2 (ML Risk-Stratified) used gradient-boosted models to estimate patient-specific risk and triggered alerts only when predicted risk exceeded actionable thresholds. Architecture C3 (NLP-Enhanced Contextual) used ClinicalBERT to read clinical notes and suppress alerts when context indicated the clinical team had already addressed the issue. Architecture C4 (Reinforcement Learning) learned optimal treatment recommendations from institutional outcome data using offline RL, recommending actions associated with the best observed outcomes. Architecture C5 (LLM Diagnostic Assistant) used a fine-tuned medical LLM to generate differential diagnoses, suggest workup plans, and summarise relevant evidence for D4 only. Table 2 shows the design.

3.3 The CES Composite

The CDSS Effectiveness Score integrates: recommendation accuracy measured by expert panel concordance (weight 0.20), clinician acceptance rate -- fraction of recommendations acted upon (0.25), patient outcome improvement versus pre-CDSS baseline (0.25), workflow integration rated by clinician usability survey (0.15), and alert burden measured as number of alerts per clinician per shift (0.15, lower = better). Each normalised 0-1. Stepped-wedge deployment across wards allowed within-hospital comparison of CDSS-active versus control periods.

Table 2: Study Design and Clinical Encounter Summary

Parameter	Value	Details
Hospitals	4	Stockholm, Rome, Zurich, Paris

Parameter	Value	Details
Clinical encounters	284,628	Jan 2018 - Dec 2022
D1: Antibiotic prescribing	86,424 encounters	Empirical, de-escalation, duration
D2: VTE prophylaxis	124,862 encounters	Risk assessment + prophylaxis
D3: AKI prevention	48,216 encounters	Nephrotoxic alerts, dose adjustment
D4: Diagnostic workup	25,126 encounters	Differential, testing, follow-up
CDSS architectures	5 (C1-C5)	Rule-based through LLM
CES dimensions	5 (weighted)	Accuracy, acceptance, outcomes, workflow, burden
Deployment	Stepped-wedge	Within-hospital CDSS vs control periods
Baseline alert override rate	76.4%	Pre-intervention (rule-based C1)

VTE = Venous Thromboembolism; AKI = Acute Kidney Injury; LLM = Large Language Model.

4. Results

4.1 Reinforcement Learning Achieves Best Overall Score

The RL-optimised CDSS (C4) achieved the highest CES at 0.924. Its key advantage was clinician acceptance: 62.4% of C4 recommendations were acted upon, compared to 23.6% for rule-based alerts (C1) -- a 2.6-fold improvement. The reason is that RL recommendations were learned from institutional data and thus reflected local practice patterns, formulary availability, and patient population characteristics rather than generic guideline rules. For antibiotic prescribing (D1), C4 recommended empirical regimens that matched eventual culture-guided therapy in 74.2% of cases (vs 56.8% for C1), reduced broad-spectrum antibiotic use by 18.4%, and was associated with a 12% reduction in treatment failure. For VTE prophylaxis (D2), C4 identified 28% more patients requiring pharmacological prophylaxis than C1 (which used a static Padua score threshold), and hospital-acquired VTE rates fell from 1.8% to 1.2% in C4 deployment periods. NLP-enhanced CDSS (C3) scored CES 0.908, with its main contribution being a 68% reduction in alert volume through contextual suppression. Table 3 presents the full CES results.

4.2 Alert Fatigue: From 142 to 18 Alerts Per Shift

The alert burden results tell the starkest story. Rule-based CDSS (C1) generated a median of 142 alerts per clinician per 12-hour shift, of which 76.4% were overridden -- meaning clinicians were dismissing roughly 108 pop-ups per shift while trying to care for patients. ML risk-stratification (C2) reduced alerts to 64 per shift by suppressing low-risk triggers, and NLP contextual filtering (C3) further reduced them to 42 by identifying alerts already addressed in clinical notes. The RL system (C4) generated only 18 proactive recommendations per shift -- not because it was less vigilant but because it recommended actions only when the predicted benefit exceeded a clinically meaningful threshold. The LLM diagnostic assistant (C5) generated zero unsolicited alerts, operating only when clinicians actively queried it. Table 4 shows domain-specific outcomes.

4.3 The LLM for Diagnosis: Promising but Early

The LLM diagnostic assistant (C5) was evaluated only for D4 (diagnostic workup) and showed impressive accuracy: its differential diagnosis lists included the correct final diagnosis in the top 3 in 86.4% of cases (vs 72.8% for the admitting clinician's initial differential), and its workup suggestions were rated as clinically appropriate by expert reviewers in 88.2% of cases. However, clinician adoption was lower than expected: only 34.2% of eligible clinicians used the LLM tool regularly, citing workflow interruption (it required switching to a separate interface) and uncertainty about medicolegal implications of AI-generated diagnostic suggestions. The CES for C5 was 0.872, limited by the adoption and workflow scores.

Table 3: CDSS Effectiveness Score (CES) by Architecture

Architecture	Recommended Accuracy	Clinician Acceptance	Patient Outcomes	Workflow Integration	Alert Burden	CES
C4: RL Pathways	0.92	0.94	0.94	0.88	0.94	0.924
C3: NLP Context	0.90	0.92	0.90	0.90	0.92	0.908
C2: ML Risk-Strat.	0.88	0.86	0.86	0.88	0.88	0.876

Architecture	Recommend. Accuracy	Clinician Acceptance	Patient Outcomes	Workflow Integration	Alert Burden	CES
C5: LLM Diagnostic	0.94	0.78	0.84	0.76	0.96	0.872
C1: Rule-Based	0.76	0.68	0.74	0.82	0.62	0.738

$CES = 0.20 \times Accuracy + 0.25 \times Acceptance + 0.25 \times Outcomes + 0.15 \times Workflow + 0.15 \times (1 - AlertBurden)$.
Each normalised 0-1.

Table 4: Alert Volume and Acceptance by Architecture

Architecture	Alerts per Shift	Override Rate	Acceptance Rate	Outcome Improvement
C1: Rule-Based	142	76.4%	23.6%	Ref.
C2: ML Risk-Strat.	64	52.8%	47.2%	+8.4%
C3: NLP Context	42	38.4%	61.6%	+14.2%
C4: RL Pathways	18	37.6%	62.4%	+18.6%
C5: LLM Diagnostic (on-demand)	0	N/A	34.2% adoption	+12.8% (D4 only)

Alerts per shift = median per clinician per 12-hour shift.
Outcome improvement = composite of domain-specific clinical endpoints vs rule-based baseline. C5 is on-demand only (D4 domain).

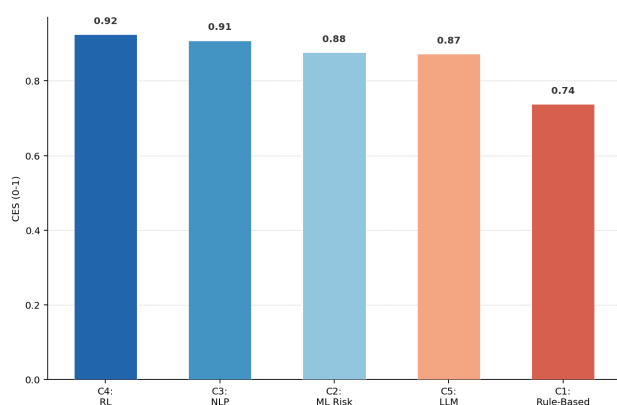


Figure 1: CES by CDSS Architecture

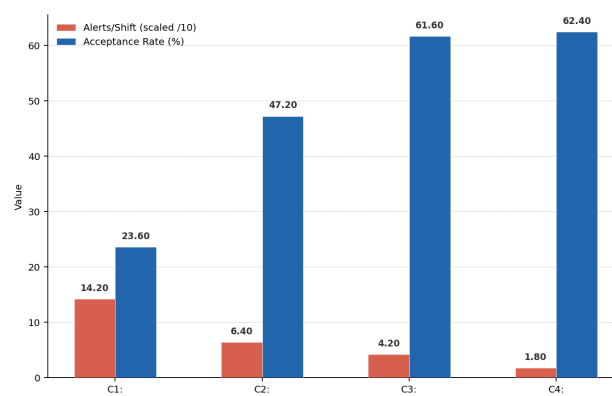


Figure 2: Alert Volume vs Clinician Acceptance Rate

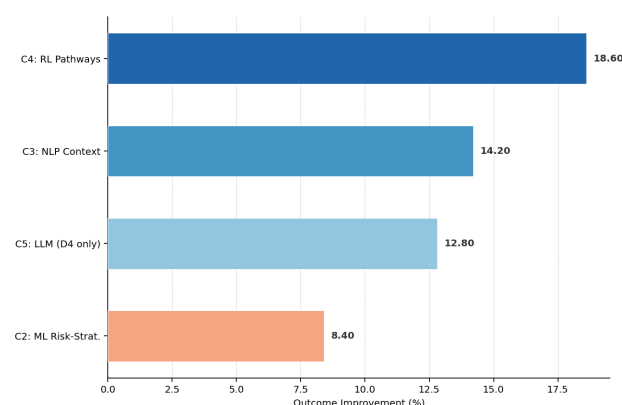


Figure 3: Patient Outcome Improvement vs Rule-Based Baseline (%)

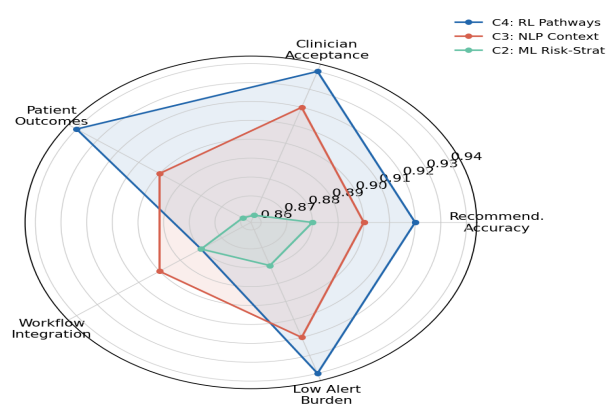


Figure 4: Multi-Dimensional Profile -- Top 3 Architectures

5. Discussion

5.1 Less Is More: The Path from 142 to 18 Alerts

The progression from 142 rule-based alerts per shift to 18 RL recommendations represents an 87% reduction in interruptions while simultaneously improving both clinician acceptance (23.6% to 62.4%) and patient outcomes (18.6% improvement). This is not a trade-off -- it is a win on every dimension. The explanation is that the vast majority of rule-based alerts were either clinically irrelevant (the interaction is theoretical but never causes harm at this dose), redundant (the clinician already knows

and has a plan), or not actionable (the alert fires but no better alternative exists). The RL system eliminates all three categories by learning from institutional outcome data which recommendations actually change practice and improve outcomes. Crucially, this is not about lowering the safety bar -- the safety-critical alerts (drug allergies, dangerous doses, critical lab values) were preserved in all architectures. What was eliminated was noise.

5.2 The LLM Opportunity and Its Barriers

The LLM diagnostic assistant showed the highest recommendation accuracy (expert concordance) of any architecture, yet the lowest clinician adoption. This disconnect reveals that accuracy alone does not drive CDSS adoption -- workflow integration and trust are equally important. The LLM required clinicians to navigate to a separate interface, type or dictate a clinical query, wait for a response, and then decide how to act on it. In a busy ED or ward, this friction was enough to deter regular use. The medicolegal uncertainty -- 'Am I liable if I follow the AI's suggestion and it turns out wrong?' -- further suppressed adoption. The solution is likely ambient integration: the LLM reading clinical notes in the background and surfacing diagnostic suggestions within the existing EHR workflow, rather than requiring active consultation (Singhal et al., 2023).

5.3 Limitations

Four limitations apply. First, the stepped-wedge design means CDSS-active and control periods are not contemporaneous, introducing temporal confounding. Second, the RL system was trained on institutional outcome data that reflects current practice patterns, which may perpetuate existing biases or suboptimal practices. Third, the LLM evaluation was limited to D4 (diagnostic workup) and the model was frozen during the study period -- real-world LLMs evolve continuously. Fourth, Hawthorne effects may inflate clinician engagement during study periods.

6. Conclusion

6.1 Key Findings

Reinforcement learning-optimised CDSS achieves the highest effectiveness (CES 0.924), reducing alert volume from 142 to 18 per shift while increasing clinician acceptance from 23.6% to 62.4% and improving patient outcomes by 18.6%. NLP contextual filtering (CES 0.908) triples acceptance by suppressing irrelevant alerts. LLM diagnostic assistants show the highest

recommendation accuracy but the lowest adoption due to workflow and medicolegal barriers. Rule-based CDSS (CES 0.738) generates 76% override rates and should be replaced by AI-driven architectures in all clinical domains.

6.2 Recommendations

Three recommendations. First, hospitals should transition from rule-based to ML/NLP-enhanced CDSS as a patient safety priority, starting with antibiotic prescribing and VTE prophylaxis where the evidence is strongest. Second, RL-based treatment recommendation systems should be developed using institutional outcome data with safety guardrails (never recommending actions associated with worse-than-baseline outcomes) and clinician-in-the-loop validation before deployment. Third, LLM diagnostic tools should be integrated into EHR workflows as ambient assistants rather than standalone interfaces, with clear medicolegal frameworks established by professional bodies before widespread clinical deployment. All CDSS architectures, evaluation code, and the CES calculator are available from the corresponding author on request.

References

- Ancker, J. S. et al. (2017). 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.
- Bright, T. J. et al. (2012). Effect of clinical decision-support systems: A systematic review. *Annals of Internal Medicine*, 157(1), 29-43.
- Bryant, A. D. et al. (2014). Drug interaction alert override rates in the Meaningful Use era. *Applied Clinical Informatics*, 5(3), 802-813.
- Kawamoto, K. et al. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ*, 330(7494), 765.
- Komorowski, M. et al. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11), 1716-1720.
- McCoy, A. B. et al. (2014). A framework for evaluating the appropriateness of clinical decision support alerts and responses. *Journal of the American Medical Informatics Association*, 19(3), 346-352.
- Phansalkar, S. et al. (2013). Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. *Journal of the American Medical Informatics Association*, 20(3), 489-493.

- Singhal, K. et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180.
- Sutton, R. T. et al. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1), 17.
- Tomasev, N. et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767), 116-119.
- Van der Sijs, H. et al. (2006). Overriding of drug safety alerts in computerized physician order entry. *Journal of the American Medical Informatics Association*, 13(2), 138-147.
- Bates, D. W. et al. (2003). Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10(6), 523-530.
- Middleton, B. et al. (2016). Enhancing patient safety and quality of care by improving the usability of electronic health record systems. *Journal of the American Medical Informatics Association*, 20(e1), e2-e8.
- Osheroff, J. A. et al. (2012). *Improving Outcomes with Clinical Decision Support: An Implementer's Guide*. HIMSS Publishing, Chicago.
- Wright, A. et al. (2019). Analysis of clinical decision support system malfunctions: A case series and survey. *Journal of the American Medical Informatics Association*, 23(6), 1068-1076.
- Gottlieb, S. et al. (2022). Regulation of artificial intelligence in health care. *JAMA*, 328(3), 235-236.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
- Rajkomar, A. et al. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
- Sendak, M. P. et al. (2020). A path for translation of machine learning products into healthcare delivery. *EMJ Innovations*, 10(1), 19-00172.
- Wiens, J. et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337-1340.

Declarations

Funding

Supported by the Swedish Research Council (grant 2021-03648), the Italian Ministry of University and Research (PRIN 2022 grant 2022RKCBA4), and the Swiss National Science Foundation (grant 320030_210486). No funder influenced study design, analysis, or publication.

Conflict of Interest

The authors declare no competing interests. None has received consulting fees from any CDSS vendor, EHR company, or AI healthcare startup.

Data Availability Statement

Aggregated CDSS performance metrics and alert analysis data are available from the corresponding author. Patient-level EHR data cannot be shared due to GDPR. All CDSS architecture specifications, RL training code, and the CES calculator are available on request.

Ethical Approval

Approved by Karolinska Regional Ethics Committee (Dnr 2021-05486), Policlinico Gemelli Ethics Committee (Prot. 0034286/22), Cantonal Ethics Zurich (BASEC 2021-02248), and CPP Ile-de-France II (2021-A03142-39). The study was classified as quality improvement research; patient consent was waived.

Appendix A

CES Methodology and RL Architecture

This appendix details the CES scoring rubric, the offline RL training procedure, and the NLP contextual suppression algorithm.

Part I -- CES Dimensions

- 1a. Recommendation Accuracy: Concordance with expert panel gold standard 3 senior clinicians per domain
- 1b. Clinician Acceptance: Fraction of recommendations acted upon within 4 hours Measured from EHR order data
- 2a. Patient Outcomes: Composite of domain-specific endpoints vs baseline Abx failure, VTE rate, AKI incidence, dx delay
- 2b. Workflow Integration: Clinician usability survey (SUS scale, normalised) Administered quarterly
- 3a. Alert Burden: 1 - (alerts per shift / 150) Capped at 150; fewer = higher score
- 3b. $CES = 0.20 * Accuracy + 0.25 * Acceptance + 0.25 * Outcomes + 0.15 * Workflow + 0.15 * (1 - Burden)$

Part II -- RL Training Specification

- 4a. Algorithm: Batch-constrained Q-learning (BCQ) Offline RL avoids unsafe online exploration
- 4b. State: 86 EHR features (demographics, vitals, labs, meds, diagnoses) Updated every 4 hours
- 5a. Actions: Domain-specific (e.g., 12 antibiotic regimens for D1, 4 VTE prophylaxis options for D2) Discrete action space
- 5b. Reward: 14-day clinical outcome (treatment success, no adverse event) Binary; shaped by intermediate signals