

Review

Deployment of machine learning algorithms to predict sepsis: systematic review and application of the SALIENT clinical AI implementation framework

Anton H. van der Vegt ¹, Ian A. Scott², Krishna Dermawan³, Rudolf J. Schnetler⁴, Vikrant R. Kalke⁵, and Paul J. Lane⁶

¹Queensland Digital Health Centre, The University of Queensland, Brisbane, Queensland, Australia, ²Department of Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, Brisbane, Australia, ³Centre for Information Resilience, The University of Queensland, St Lucia, Australia, ⁴School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, Australia, ⁵Patient Safety and Quality, Clinical Excellence Queensland, Queensland Health, Brisbane, Australia, ⁶Safety Quality & Innovation, The Prince Charles Hospital, Queensland Health, Brisbane, Australia

Corresponding Author: Anton H. van der Vegt, PhD, BE, BSc, Queensland Digital Health Centre, The University of Queensland, Princess Alexandra Hospital, 34 Cornwall St, Woolloongabba, Brisbane, QLD 4072, Australia; a.vandervegt@uq.edu.au

Received 19 October 2022; Revised 4 April 2023; Editorial Decision 16 April 2023; Accepted 23 April 2023

ABSTRACT

Objective: To retrieve and appraise studies of deployed artificial intelligence (AI)-based sepsis prediction algorithms using systematic methods, identify implementation barriers, enablers, and key decisions and then map these to a novel end-to-end clinical AI implementation framework.

Materials and Methods: Systematically review studies of clinically applied AI-based sepsis prediction algorithms in regard to methodological quality, deployment and evaluation methods, and outcomes. Identify contextual factors that influence implementation and map these factors to the SALIENT implementation framework.

Results: The review identified 30 articles of algorithms applied in adult hospital settings, with 5 studies reporting significantly decreased mortality post-implementation. Eight groups of algorithms were identified, each sharing a common algorithm. We identified 14 barriers, 26 enablers, and 22 decision points which were able to be mapped to the 5 stages of the SALIENT implementation framework.

Discussion: Empirical studies of deployed sepsis prediction algorithms demonstrate their potential for improving care and reducing mortality but reveal persisting gaps in existing implementation guidance. In the examined publications, key decision points reflecting real-world implementation experience could be mapped to the SALIENT framework and, as these decision points appear to be AI-task agnostic, this framework may also be applicable to non-sepsis algorithms. The mapping clarified where and when barriers, enablers, and key decisions arise within the end-to-end AI implementation process.

Conclusions: A systematic review of real-world implementation studies of sepsis prediction algorithms was used to validate an end-to-end staged implementation framework that has the ability to account for key factors that warrant attention in ensuring successful deployment, and which extends on previous AI implementation frameworks.

Key words: sepsis prediction, systematic review, AI implementation, machine learning, artificial intelligence

INTRODUCTION

Sepsis accounts for nearly 20% of deaths worldwide, killing over 11 million people in 2017.¹ Sepsis has been defined as a “life-threatening organ dysfunction caused by a dysregulated host response to infection”.^{2,3} Early recognition and treatment of sepsis can reduce mortality, and rule-based surveillance systems for detecting sepsis in hospital settings can improve outcomes.^{4,5}

More recently, sepsis prediction algorithms employing artificial intelligence (AI),^{6–8} herein called machine learning algorithms (MLAs), that can detect evolving sepsis in patients earlier than rule-based methods, have proliferated.^{9,10} Most MLA studies assess performance based on static training and testing data collected retrospectively and analyzed *in silico*,¹¹ whereas healthcare providers seek to implement MLAs in dynamic, complex real-world clinical settings using live or near-live data.

Theoretical MLA implementation frameworks^{12–16} have attempted to identify key stages, tasks and contextual factors that warrant consideration, but practical translation into end-to-end MLA implementation in clinical practice is uncertain. While systematic reviews have evaluated pre-implementation studies of sepsis MLAs,^{6–8,11,17} including interviews generating implementation methods,¹⁸ none have focused on MLAs actually implemented. Individual studies of deployed MLAs have revealed barriers and enablers that implementation frameworks must incorporate if they are to fully inform successful end-to-end MLA implementation.^{18–20}

In this article, we identified and appraised studies of clinically applied sepsis MLAs using systematic methods and then map the serial steps in deployment described in these studies to a recently derived AI implementation framework, called SALIENT (reported in a companion paper²¹ and described in brief below). The mapping sought to clarify where and when barriers, enablers, and key decisions arise within the end-to-end AI implementation process and to validate SALIENT’s capability to guide stakeholders involved in end-to-end MLA implementation.

Background

The process by which AI interventions are evaluated at any given stage in the implementation cycle is maturing. The recently reported Decide-AI research reporting guidelines depict key stages of algorithm development, evaluation, and implementation,²² (Figure 1) which, in the companion paper to this work,²¹ were mapped to Stead et al’s multi-stage approach to translating medical informatics interventions from the lab to the field.¹² This mapping was used to derive an end-to-end AI implementation framework, called SALIENT (Figure 3 and fully described elsewhere²¹), which accounted for factors found to be missing in many implementation frameworks when subjected to the Stead et al’s taxonomy,^{12,13,16,23,24} that is, components, both technical and clinical, that need to be developed, evaluated, and integrated over several stages.

The resulting SALIENT stages and associated reporting guidelines are: (I) Definition; (II) Retrospective study—TRIPOD(-AI)^{25,26}; (III) Silent trial—TRIPOD(-AI)^{25,26}; (IV) Pilot trial—Decide-AI²²; and (V) Large trial/roll-out—CONSORT(-AI).²⁷ The SALIENT framework integrates all elements of the reporting standards, and, compared to prior frameworks, renders all components of the end-to-end solution, how and when they integrate, and underlying implementation tasks (not shown here) fully visible. However, similar to most prior frameworks, SALIENT has not been validated in its ability to accommodate reported real-world AI implementation stages, barriers, enablers, and decisions.

OBJECTIVE

This study had 2 objectives: (1) conduct a systematic review of real-world implementation studies of sepsis MLAs in clinical practice and extract information into how MLA performance, adoption, and different implementation modes were assessed and impacted clinical care processes and patient outcomes; and (2) map the findings regarding barriers, enablers, and key decision points to the different stages and components of the SALIENT AI implementation framework to assess its potential utility for guiding real-world MLA implementation.

MATERIALS AND METHODS

Systematic review of sepsis MLA implementation studies

Search strategy

The systematic review was performed according to PRISMA guidelines.²⁸ Five databases (Pubmed/Medline, EMBASE, Scopus, Web of Science, and clinicaltrials.gov.) were searched between January 1, 2012 and June 23, 2022 for titles and abstracts published in English using keywords and synonyms for: (1) predict; AND (2) sepsis; AND (3) machine learning; AND (4) trial; and NOT (5) child (see [Supplementary Appendix SA](#) for complete search queries).

A forwards and backwards citation search (snowballing strategy) was then applied to included papers to identify additional articles that reported new MLAs, or, provided further information about a sepsis MLA described in previously included papers. The latter were labeled *linked* papers, describing MLAs at different stages of implementation, but not considered primary articles.

Study selection

Studies of any design were included if: MLAs were applied to adult patients in hospital settings in whom sepsis was formally defined; used live or near-live data; and reported at least one or more algorithm performance metrics (full details in [Supplementary Appendix SB](#)). Covidence software²⁹ supported a 2-stage screening process with screening of articles by 2 independent reviewers (AHvdV and RJS), with conflicts agreed by 3-way consensus (AHvdV, RJS, and KD); and full-text review by 2 independent reviewers (AHvdV and KD), with selection agreed by 3-way consensus (AHvdV, RJS, and KD). Snowballing was then applied to all included papers and any new or linked papers were identified by AHvdV and verified by KD.

Data extraction

Data were extracted independently by 2 authors (AHvdV and KD) using Excel templates, with disagreements resolved by consensus of 2 other authors (RJS and IAS). Extracted data included study meta-data, implementation stage, care setting, MLA details including training and validation datasets, performance metrics, outcome definitions and events, and implementation barriers, enablers, and decision points (see [Supplementary Appendix SC](#) for more details). Decision points were identified when 2 or more studies chose different options at a certain point in implementation. Barriers were defined as pitfalls or problems hindering implementation success and enablers as tips or activities aiding implementation success. Consensus between authors (AHvdV, PL, and IAS) determined which decisions, enablers and barriers to include as found and which to consolidate under a common title to minimize overlap.

Quality assessment

Papers reporting all-cause or sepsis-related mortality underwent risk of bias (RoB) assessment, performed independently by 2 authors (AHvdV and VRK), using either the ROBINS-I tool³⁰ for non-randomized studies, or the Cochrane RoB 2 tool³¹ for randomized trials. Mortality was chosen for RoB assessment as it was the most frequently reported and patient-critical measure.

Application of AI implementation framework

The systematic review findings for barriers, enablers, and decision points were mapped by AHvdV to the stages and elements of the SALIENT implementation framework, followed by a review by IAS and adjustments made where discrepancies were found. An item could be mapped to more than one element and where no obvious element was found to map to, it was recorded.

RESULTS

Systematic review of sepsis MLA implementation studies

From 3133 retrieved abstracts, 1126 duplicates were removed, leaving 2007 for screening, from which 12 full-text studies were included for analysis (Figure 1). Most excluded studies were not sepsis prediction studies, or were rule-based rather than AI-based algorithms or were not implemented. An additional 7 articles found by snowballing were selected, yielding a 19 included papers as primary

articles, with further snowballing yielding 11 linked papers, giving a total of 30 articles.^{32–61}

Study characteristics

All 30 studies were published between 2015 and 2022, with 8 algorithm groups (A to H) identified according to the common or named MLA that was the focus of study (Table 1); all were US-based except for Group (C), which was Brazilian. Five groups (A, B, E, F, H) implemented MLAs with a quantitative evaluation (before-after,^{33,50,59–61} randomized controlled trial,⁵⁸ 2-armed cohort study,⁴⁶ prospective observational,^{33,44,48,53} retrospective observational³⁴). Two other groups (C, D) provided case studies^{35,41,43} or qualitative evaluations⁴² and one group (G) reported only post-implementation analyses (retrospective⁵¹ or difference-in-difference⁵²). Groups (B, E) conducted the only multicenter trials with outcomes of more than 10 000 sepsis episodes.^{46,61} Median trial length was 14 months (range 2–79) and median time between publishing a retrospective study on MLA development and an implementation study was 3 years (range 1–7).

Two studies were confined to emergency department settings,^{32,53} 6 involved only intensive care unit (ICU) patients,^{45,49,55–58} and the remainder involved all wards or non-ICU settings. The prevalence of sepsis varied from as low as 2.1%⁴⁷ to as high as 22.7% in non-ICU settings,⁶¹ and from 11.3%⁵⁷ to 32.8%⁵⁸ in ICU settings.

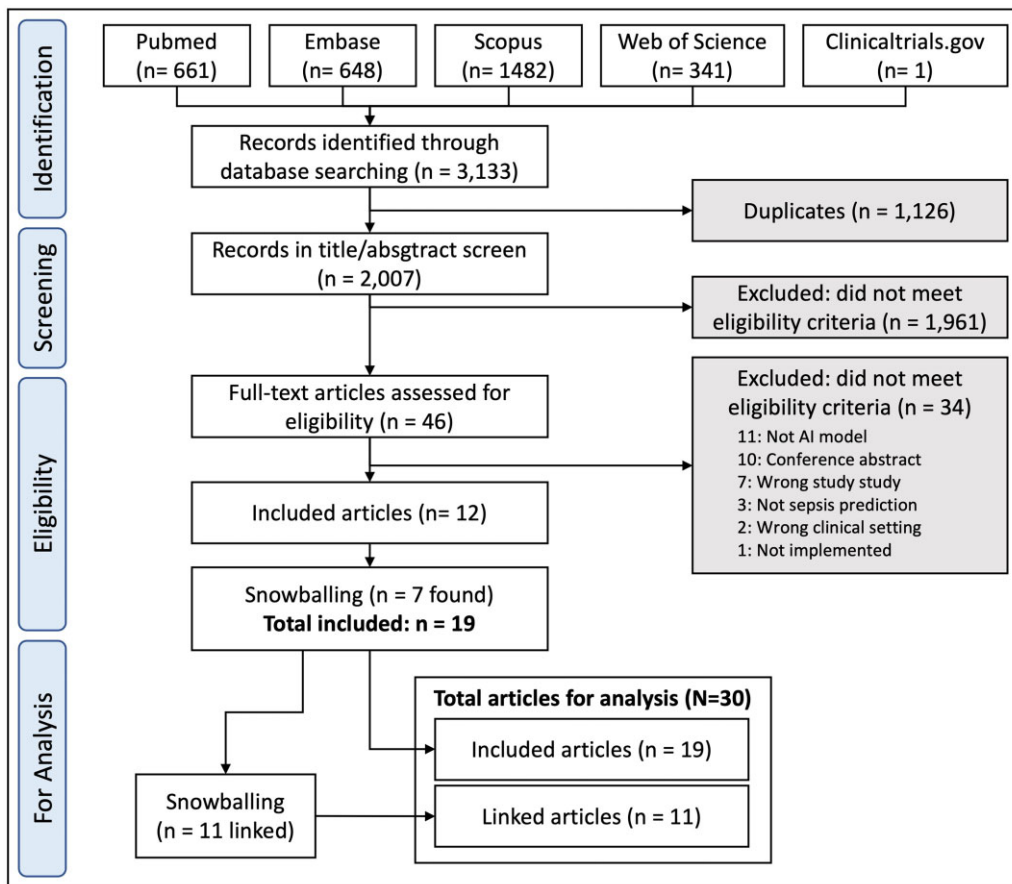


Figure 1. Flowchart for study selection.

Table 1. Listing of studies identified in systematic review, grouped by MLA

Group	Reference	SALIENT stage ^a	Study design (# sites)	Study period	Care location	Outcome ^b count (prevalence)
A (EWS 2.0) University of Pennsylvania. MLA = random forest; 587 features	Taylor et al., 2016 ³²	II	R (3)		ED	1056 (4.7)
	Giannini et al., 2019 ³³	II	R (3)		Non-ICU	347 (3.3)
		III	PO (2)	Jan–Jun’16	Non-ICU	1540
		V	BA (2)	Jul’16–Feb’17	Non-ICU	2137
B (Insight) Dascena Inc and University of California. MLA = various, Inc. Gradient boosted trees ²⁹ and logistic regression ⁴⁵ , 6+ variables (laboratory results optional)	Ginestra et al., 2019 ⁴⁴	Post	PO (1)	Nov–Dec’16	Non-ICU	NA
	Calvert et al., 2016 ⁵⁵	II	R (1)		ICU	159 (11.4)
	Calvert et al., 2016 ⁵⁶	II	R (1)		ICU	270 (0.9)
	Desautels et al., 2016 ⁵⁷	II	R (1)		ICU	2577 (11.3)
	Shimabukuro et al., 2017 ⁵⁸	IV	RCT (1)	Dec’17–Feb’18	ICU	22 (32.8)
	McCoy et al., 2017 ⁵⁹	IV	BA (1)	Nov’16–May’17	ALL	921 (NR)
	Calvert et al., 2017 ⁴⁸	Post	NA		NA	NA
	Mao et al., 2018 ³⁶	II	R (1)		Non-ICU	2142 (2.4)
	Burdick et al., 2018 ⁶⁰	IV	BA (1)	Jul–Aug’17	ED+ICU	1136 (NR)
	Topiwala et al., 2019 ³⁴	V	RO (1)	Feb–Jun’18	ALL	269 (NR)
C (Robot Laura) Brazil; MLA = NR	Burdick et al., 2020 ⁶¹	V	BA (9)	2017–mid’18	Non-ICU	14 166 (22.7)
	Gonçalves et al., 2020 ³⁵	Post	CS (1)	Jan–Jun ‘18	NR	NR
	Scherer et al., 2022 ³⁷	Post	RO (1)	Mar–Sep ‘20	NR	NR
D (Sepsis Watch) Duke University. MLA = recurrent neural network; 86 variables	Futoma et al., 2017 ³⁸	II	R (1)		ALL	10 552 (21.4)
	Futoma et al., 2017 ³⁹	II	R (11)		ALL	10 552 (21.4)
	Bedoya et al., 2020 ⁴⁰	II	R (1)		ALL	813 (18.9)
	Sendak et al., 2020 ⁴¹	III/V	CS (3)	Apr’16–Nov’18	ALL	NA
	Sandhu et al., 2020 ⁴²	Post	Q (1)	Jan–Apr’19	NA	NA
E (TREWScore) Johns Hopkins University. MLA = Cox proportional hazard; 27 features	Sendak et al., 2020 ⁴³	Post	CS (1)		NA	NA
	Henry et al., 2015 ⁴⁵	II	R (1)		ICU	2291 (14.1)
	Adams et al., 2022 ⁴⁶	V	2xAC (5)	Apr’18–Sep’20	Non-ICU	13 680 (2.3)
	Henry et al., 2022 ⁴⁷	II	R (5)	Jan’16–Mar’18	Non-ICU	3858 (2.2)
		Post	PO (5)	Apr’18–Mar’20	Non-ICU	9805 (2.1)
F (Sepsis sniffer) Mayo Clinic. MLA = decision Tree	Henry et al., 2022 ⁴⁸	Post	Q (1)	Oct’18–Apr’19	ALL	NA
	Harrison et al., 2015 ⁴⁹	II	R (1)		ICU	86 (29)
G (ESM) USA, independent. MLA = Log. Reg.	Lipatov et al., 2022 ⁵⁰	IV	BA (1)	Sep’11–May’18	ED + ICU	1096 (9.8)
	Wong et al., 2021 ⁵¹	Post	R (1)	Dec’18–Oct’19	Unclear	2552 (6.6)
H USA, independent. MLA = naïve Bayes; 5 features	Schootman et al., 2022 ⁵²	Post	DiD (15)	Jan’16–Jun’19	ALL	6926 (NR)
	Brown et al., 2016 ⁵³	II	R (2)		ED	549 (0.4)
		Post	PO (1)	Apr’09–Jun’10	ED	352 (0.4)

Study Designs: R: Retrospective; RO: Retrospective Observational; PO: Prospective Observational; BA: Before-after study; 2xAC: Two-Arm Cohort; RCT: randomized control trial; Q: qualitative study; CS: case study; DiD: difference-in-difference analysis.

^aStages in the SALIENT framework: I = problem definition; II = retrospective development; III = Silent trial; IV = Pilot trial; V = Large trial/Roll-out; Post= Post deployment study.

^bOutcome definitions are itemized in [Supplementary Appendix SF, Table F2](#).

NA: not applicable; NR: not reported; ED: emergency department; ICU: intensive care unit; ALL: patients from all wards.

Most studies reported MLA evaluations at stage II (retrospective),^{32,33,36,38–40,45,48,49,55–57} 2 at stage III (silent trial),^{33,41} 4 at stage IV (pilot trial),^{50,58–60} 3 at stage V (large trial/roll-out),^{33,46,61} and 11 reported post-deployment evaluations.^{35,37,42–44,47,48,51–53,62}

Quality assessment

Eight papers from 5 groups (A, B, E, F, G) were assessed for risk-of-bias (RoB) (see [Supplementary Appendix SD](#)). Overall, RoB was serious for 2 groups (A, F), moderate for 2 (E, G) and moderate to critical for one (B). Major sources of bias were potential confounding from additional sepsis control co-interventions, such as staff training in sepsis recognition and management, and other changes in non-sepsis conditions and patient characteristics potentially impacting all-cause mortality at the trial sites. Only 2 trials controlled for these differences in before-after cohorts.^{46,52}

Implementation evaluation

Eighty-five distinct metrics were identified across 26 (87%) papers, grouped into 5 evaluation categories: (1) algorithm performance; (2) algorithm adoption; (3) clinical process effects; (4) patient outcome effects; and (5) financial impact. All metrics reported are listed in [Supplementary Appendix SE](#).

Algorithm performance and adoption. Of 27 performance metrics, sensitivity and positive predictive value were most commonly reported (7 groups, 18 and 9 papers, respectively), closely followed by area under the receiver operating curve (AUROC) and specificity (6 groups, 17 and 13 papers). Most (66%) post-implementation studies did not report real-world MLA performance ([Figure 2](#)), and of the 3 that did, one reported improved MLA performance⁵⁸ while the other 2 showed marked declines,^{34,50} and similarly for the external validation study of the EPIC tool (Group G).⁵¹

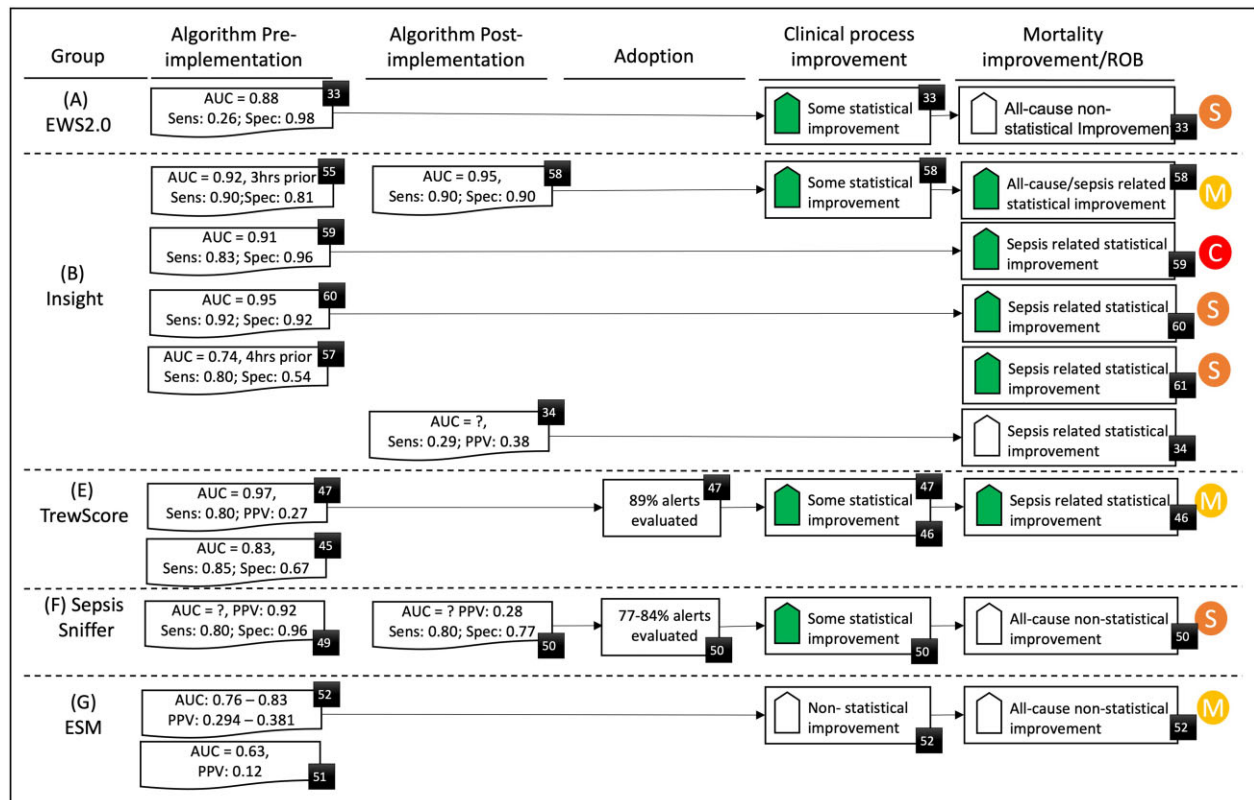


Figure 2. Evaluation results for algorithm performance, adoption, clinical process, and mortality improvement for each MLA group. Figure includes risk of bias (RoB) assessment for studies reporting mortality assessment (M = moderate [some concerns for RoB-2], S = serious, C = critical). Black numbered squares denote the cited paper for that result. AUC: area under the receiver operating curve; Sens: sensitivity; Spec: specificity; PPV/NPV: positive/negative predictive value. Solid shaded up arrow: significant improvement, whereas hollow up arrow: non-significant improvement.

MLA adoption, measured as the proportion of alerts clinicians responded to, was only reported in Group (E) at 89%, Group (F) at 77%–84% and Group (C) at 100%.

Clinical impact. Of 36 distinct clinical process outcomes reported across 9 papers in 5 groups, the most common were median lead time to first antibiotic use (5 papers, 4 groups), the 3-h sepsis care bundle compliance rate and the increases in antibiotic use (both 3 papers, 3 groups).

Ten different patient outcomes were reported, most commonly mortality and length of hospital stay (LOS) (both 5 groups, 9 papers). All 9 papers^{34,44,46,50,52,58–61} reported decreased mortality, be that all-cause, sepsis-related or both, although this was statistically significant only for 5 studies: Group (B), both all-cause⁵⁸ and sepsis-related,^{59–61} and Group (E), for sepsis-related only,⁴⁶ both involving a large samples of >13 500 septic patients. Only Group (E) adjusted their findings for differences between cohorts in patient characteristics. Only Group (B) performed more than one independent post-implementation mortality study, with all 5 studies showing improved mortality,^{58–61} although Topiwala et al³⁴ reported poor post-implementation MLA performance and no significant improvement in mortality.

Of the 2 groups reporting significantly improved sepsis-related mortality, only Group (E) reported strong MLA adoption (89%) and significant decreases in antibiotic lead time.⁴⁶ Group (B) reported no adoption data and only their smallest study reported improvement in a single process outcome: lead time to antibiotic

use.⁵⁸ Despite group (F) reporting high adoption rates (77% to 84%) and significantly improved rates of sepsis care bundle compliance, post-implementation the MLA specificity dropped markedly, from 96% to 80%, and there was no significant change in all-cause mortality.⁵⁰

Identification of implementation barriers, enablers, and decision points and mapping to SALIENT AI implementation framework

Barriers, enablers, and decision points provide real-world evidence of factors that are reported by practitioners and can impact MLA implementation success.

Barriers and enablers

We identified 14 unique barriers (Table 2) and 26 unique enablers (Table 3) from a total of 70 mentions across all studies. The most common barriers, identified by at least 3 groups, were lack of clinician trust (B1), alert fatigue (B4) and dismissal of alerts, mainly because clinicians perceived no clinical signs of deterioration (B3). However, 8 barriers were unique to a single group (D), and despite more enablers than barriers, just 2 groups (D, E) provided 80% of the group-level enabler instances.^{41–43,48} The most commonly reported enabler was frequent communications to raise awareness of the MLA during and after clinical trials (E4), with clinician involvement (E1), improvement cycles (E3), clinical champions (E5), and test versions for training (E6) reported by more than 2 groups. Overall, 90% of all barriers and enablers were AI task agnostic,

Table 2. Implementation barriers

Barriers (stage)		Component or element	A	B	C	D	E	F	G	H	Total
B1	Lack of clinician trust (IV+) ¹⁸	ICA	2		1	1	1				5/4
B2	MLA retraining concerns: Feedback loops arise when alerts lead to timely treatment. (IV+)	AI				1					1/1
B3	Alerts dismissed for wrong reasons, for example, patients with no sepsis symptoms or with higher acute complexity (IV+)	CW	2				2			1	5/3
B4	Alert fatigue (II+) ¹⁸	AI; CW; ICA	1			1		1			3/3
B5	Differential nurse/Doctor role, perceptions of role and value (IV+) ¹⁸	CW; ICA	1			1					2/2
B6	Inherent limitations of EHR data, which can be plagued by missingness, inaccuracies, and changes in practice patterns over time (II+)	DP	1			1					2/2
B7	Data entry delays, leading to delayed predictions (III+)	DP			1						1/1
B8	Inventors/company equity owners may have COI and inadvertently act in bias ways towards the evaluation of their system (III+)	ET				1					1/1
B9	Surveillance bias: important to monitor impact of Alerts on non-septic patients for over-prescription of antibiotics (IV+)	ET; EM					1				1/1
B10	Substantial cost involved for infrastructure, implementation personnel time and ongoing maintenance (All) ¹⁸	ICA; GOV				1					1/1
B11	Lack of individual proficiency of health professionals in the use of hardware and software (IV+)	CW; ICA			1						1/1
B12	Clinicians perceive they are better at diagnosing sepsis than the AI and the alert occurs after they already suspect (IV+) ¹⁸	CW; AI; ICA	1			1					2/2
B13	Lack of machine learning foundational knowledge and firsthand experience (II+) ⁶³	CW; ICA				1					1/1
B14	Clinician concern over reliance on system (IV+) ⁶³	GOV; QS; ICA					1				1/1
Total papers			8	0	3	9	5	1	0	1	27
Count of barriers identified by the group			6	0	3	9	4	1	0	1	24

For each barrier, the number of papers that identify the barrier within each group are noted in columns A to H. The totals column is in the format of: total number of papers/total number of groups. The associated element or component in the derived framework is also identified where ICA: Implementation, change management & adoption; AI: AI model; CW: clinical workflow; DP: data pipeline; GOV: governance; QS: Quality & safety; EM: Evaluation and monitoring. Beside each barrier is listed the stage in parentheses, that is associated with that barrier.

with just one barrier (B12) and 3 enablers (E2, E11, E26) specific to sepsis prediction.

All barriers and enablers could be mapped to the SALIENT AI implantation framework (see Figure 3). All barriers ($n=14$) were located between the silent trial stage (III) and the large trial or roll-out stage (V). Most enablers and barriers were identified for the clinical workflow solution component ($n=6$ and $n=8$, respectively in stages IV and V) and the cross-stage element, 'Implementation, change management and adoption' ($n=8$ and $n=14$, respectively). No barriers were identified that related to the regulatory and legal policy domain or the human computer interface solution component, whereas enablers were identified in all solution components and all cross-stage policy and organizational elements.

Decision points

Twenty-two decision points were identified in our review, with 17 identified by at least 2 groups; all were mapped to the SALIENT implementation framework (Table 4) and depicted in Figure 3.

Definition decision points (D1-D4). The target population and care locations were reported by 7 groups (D1); all included the ED, 5 added the ICU or general wards and 4 targeted all areas. No study reported use of different algorithms for the ICU and non-ICU wards, despite ICUs collecting more data elements at higher frequency.

Other decisions related to identifying all hospitalized patients with sepsis, including at ED presentation, or only those acquiring it whilst in hospital,³⁸ and whether to identify only patients at higher risk of mortality for prioritized clinical review, thus minimizing clinician workload.⁴⁶

Twenty-six different definitions of sepsis were used (see Supplementary Appendix SF, Table F2), ranging from sepsis to severe sepsis to septic shock (D2). The prime purpose for implementing sepsis MLAs varied which in turn determined how they were trained and evaluated (D3),⁴³ with evaluation metrics and success criteria varying depending on whether increasing sepsis care bundle compliance,⁵⁰ providing a sepsis detection and management system,⁴¹ reducing anti-microbial overuse⁵² or decreasing patient mortality and LOS were primary objectives.^{34,46,58,59,61} The algorithm objective also determined the minimum expected performance for the MLA (D4), in terms of sensitivity (proportion of septic cases detected) and false alarms (proportion of non-septic cases misidentified as sepsis). Different thresholds were chosen according to the anticipated impacts on clinical processes and clinician workload and adoption.^{60,61}

AI model decision points (D5-D8). We identified 1 statistical AI model (E), 5 ML models (A, B, F, G, H), 1 deep learning (DL) model (E), and 1 unknown (C), with no single model being utilized by more than one group. Selection of model type (D5) varied according

Table 3. Implementation enablers: for reach enabler, the number of papers that identify the enabler within each group are noted in columns A to H

	Enablers (stages)	Compo-nent	A	B	C	D	E	F	G	H	Tot.
E1	Clinician involvement is essential at all stages of model/HCI development and integration into clinical workflow (II+) ^{18,63}	AI; CW, HCI; GOV				2	2				4/2
E2	Better AI model training methods: for negative cases, use portion of patient journey when sick, not within 6h or discharge (II/III)	AI				1					1/1
E3	Conduct improvement initiatives (PDSA) cycles during implementation to quickly garner and act on clinical feedback (III+)	CW; QS; ICA								1	3/2
E4	Frequent communications to increase awareness during and after trial, for example, weekly meetings, emails, educational sessions giving progress and setting next goals and highlighting urgent need. (IV+) ¹⁸	ICA		1	1	1	2				5/4
E5	Appoint clinical champions to advocate for the tool (II+)	ICA				1	2				3/2
E6	Create a test version of the application to train clinicians and multi-channel training approaches incl. web (III)	CW				1	1				2/2
E7	Use a tablet with training loaded, plus feedback mechanism so training could occur on-the-job (IV+)	CW				1					1/1
E8	Implement alternative workflows during peak hours (IV+) ¹⁸	CW					1				1/1
E9	Clinicians were taught how to interpret risk scores (IV+)	CW; ET		1							1/1
E10	Iterative approach to design of clinical workflow, human-computer interface and AI model (II+)	AI; CW; HCI				1					1/1
E11	Visually delineating sepsis risk into colors (red cards as high risk, etc) and tracking patients across distinct tabs (patients to be triaged, screened out for sepsis, and those in the sepsis bundle) (III)	HCI				1					1/1
E12	Perform post-implementation interview (study) to identify improvements (IV+)	CW; HCI; QS; ICA				1					1/1
E13	HCI was augmented by completion and fallout indicators to visually guide the clinician to timely and appropriate care (III)	CW, HCI						1			1/1
E14	Report the number of cases the AI detects that clinicians miss: clinicians requested this—build trust (IV+)	EM; ICA				1	2				3/2
E15	Track and monitor data and model drift (III+)	QS; EM				1					1/1
E16	Work with regulatory officials to ensure the solution is qualified as CDS and not a diagnostic medical device (I+)	RL				2					2/1
E17	Establish a multi-disciplinary governance committee to promote usage, track compliance, provide training and plan for post-trial sustainability; and an external data safety board to oversee safety and AI efficacy (I+)	GOV				1					1/1
E18	A dedicated full-time role can work with frontline clinicians and stakeholders to integrate the tool (IV+)	ICA; CW			1	1					2/2
E19	Strong support from senior leadership (I+) ¹⁸	ICA; GOV				1					1/1
E20	Establish a transdisciplinary team of data scientists, statisticians, hospitalists, intensivists, ED clinicians, RRT nurses, and information technology leaders and develop capabilities across domains (I+)	ICA; GOV				1					1/1
E21	Staggered deployment across sites (V+)	ICA					1				1/1
E22	Although numbers and statistical trends were used as evidence, individual patient cases were important to frontline clinicians (IV+)	ICA				1					1/1
E23	Carefully navigate the lines of professional authority that physicians have toward the care of patients. Tool described as supporting physicians and nurses. The term AI was never used (I+)	ICA; ET				2					2/1
E24	Trust in the model increased as the clinician experienced the algorithm make correct predictions (IV+)	ICA				1					1/1
E25	A “Model Facts” sheet designed to convey relevant information about the model to clinical end users; (II)	ET; AI; ICA				1					1/1
E26	System sepsis monitoring was experienced as alleviating demands on attention and cognition (IV+)	ICA					1				1/1
	Total papers		0	4	2	23	12	1	0	1	43
	Count of group enablers		0	3	2	20	8	1	0	1	35

The totals column is in the format of: total number of papers/total number of groups. The associated components in the SALIENT framework are also identified where ICA: Implementation, change management and adoption; AI: AI model, HCI: human-computer interface, CW: clinical workflow, DP: data pipeline, GOV: governance; ET: Ethics; EM: evaluation and monitoring; RL: Regulatory& legal and QS: Quality & safety. Beside each enabler is listed the stage in parentheses, that is associated with that enabler.

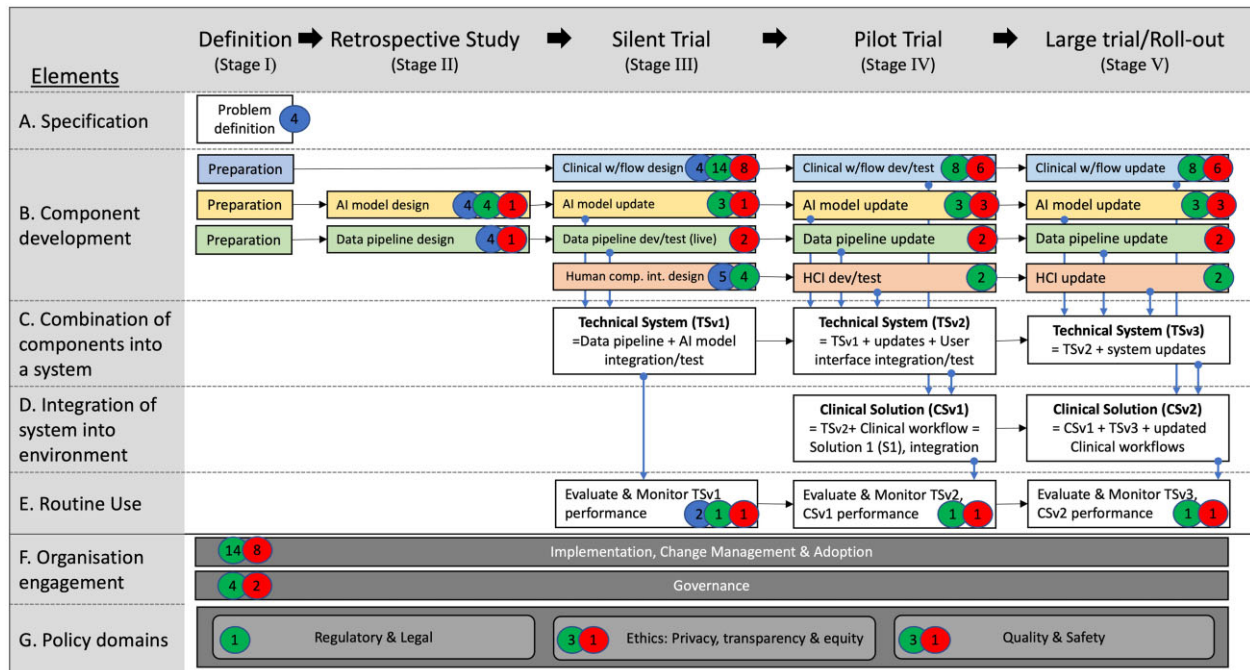


Figure 3. The number of barriers (red badge), enablers (green badge), and decision points (blue badge), denoted by the number within the badge, mapped to each stage and component of the SALIENT AI implementation framework. Implementation stages are labeled in the title row from left to right. The color-coded solution components are developed in row B, and consist of Clinical workflow (blue), AI model (yellow), data pipeline (green), and human-computer interface (red). The components are integrated in rows C and D, and rows F and G describe cross-stage elements required throughout the entire implementation process, such as governance and quality and safety assurance. w/flow: workflow; dev/test: development & test; HCI: human computer interface.

to perceived accuracy and adoption based on the level of model explainability,^{41,45,55} but with trade-offs according to the model’s ability to support time-series data,^{38,40} accommodate large, high-dimensional datasets,⁴⁸ and demonstrate better performance.⁴¹ Group (D) reported clinicians were willing to sacrifice explainability for more accurate predictions and better standardized treatment of all sepsis cases,⁴¹ while Ginestra et al found clinicians most wanted transparency regarding the predictive features generating the alerts.⁴⁴

Deciding which features to input and how simple (eg, vital signs only) or complex (eg, waveform data and laboratory results) they are was seen to influence model generalizability (D6) to different care locations.^{38,57,58,62} Group (B) supported different variables for different sites, claiming flexibility,⁶² although new models needed to be trained, validated and maintained at each site. Another decision was how quickly the MLA needed to make its first prediction after admission (D7), contingent upon the availability of the required data, with potential delays, for example, in obtaining laboratory investigation results.⁵³

Predicting onset of sepsis as early as possible involved trade-offs between: (1) alerts that were too early, where clinicians may not have known what to do, and therefore dismissed the alerts^{33,41,42,44,47,50,53}; and (2) alerts that were too late for patients for whom clinicians already suspected sepsis and had initiated appropriate care bundles (in one study up to half⁴⁴), thereby diminishing its clinical utility.^{44,50} The choice also had implications for MLA training and evaluation (D8).

Data pipeline decision points (D9-D12). Only group (D) contributed to data pipeline decisions for which only 2 barriers (B6, B7) and no enablers were reported. Decisions had to be made (D9)

about how to access the data: direct from the Electronic Health Record (EHR), which could entail partnering with the vendor, or indirectly from a real-time data warehouse or various feeder systems.⁴³ Similarly, whether to develop and implement the MLA in-house or use an external vendor (D10), which involved weighing up the capability to future-proof the organization for future AI solutions⁴¹ versus implementation and maintenance challenges arising from separate ownership of the input data and the AI model.⁴¹ The required level of data pipeline sophistication, including data imputation (D11) and transformation, also necessitated trade-offs between engineering effort versus model performance (D12),⁴¹ with group (D) having to remove a data imputation pipeline because of its complexity.⁴⁰

Clinical workflow decisions (D13-D15). Whether alerts were to be sent to and managed by dedicated clinical staff (centralized approach) or sent directly to clinicians responsible for individual patient care (distributed approach) varied across studies (D13). Five (A, C, E, F, G) groups chose the former, whereas Sandhu et al found physicians preferred the latter, having a nurse contact them directly, often in-person, rather than by means of EHR-generated alerts which imposed greater cognitive load and interruptions.⁴² However, the same physicians still saw nurse contacts as disruptive, while nurses found physicians often too busy to contact.⁴² Having a dedicated clinician receive calls minimized alarm fatigue,⁴¹ but group (E) found a distributed approach more scalable for monitoring multiple conditions, more feasible in small-staffed sites, and more able to provoke bedside reviews,⁴⁸ although clinicians often regarded the numbers of reviews as unmanageable.⁵⁹

The MLA alert threshold or setpoint determining the numbers of alerts was a key decision impacting clinician workload (D14).

Table 4. Decision points identified for each component in the SALIENT framework

Decision Points		A	B	C	D	E	F	G	H	Tot	
Definition											
D1	Which patients? Age; location: ICU, ED, all non-ICU				Identified by differences across papers						
D2	What to predict? sepsis, severe, shock? Should you prioritize on mortality? Patients admitted with sepsis and/or hospital acquired sepsis?				1	1	1			3/3	
D3	What objective/bundle compliance, early identification, mortality/LOS—primary and secondary outcomes; anti-microbial mis-use (flow on to model)	1			1				1	3/3	
D4	What is the minimum expected performance for alarms? precision v sensitivity?		1			1		1		3/3	
AI model											
D5	Which model: ML vs DL (explainable, earliness of prediction) and where trained	1			2	1				4/3	
D6	Which features: simple vs complex, set-in-stone or changeable. Noting this will impact earliest first prediction: immediately at ED or later?	1	3						1	5/3	
D7	How early to target alerts? (too early—no symptoms/signs, too late, no clinical utility)	2			1	2	1		1	7/5	
D8	What outcome basis for Train/Evaluate?	1			3		1			5/3	
Data pipeline											
D9	What data access approach to use: direct or separate				2					2/1	
D10	Whether inhouse vs external platform/product/solution				2					2/1	
D11	What methods of data imputation to use				2					2/1	
D12	What level of pipeline sophistication can be supported: model performance vs engineering effort				1					1/1	
Clinical workflow											
D13	Whether dedicated vs distributed model of alert handling		1		2	1				4/3	
D14	What determines the setpoint decision		1			1				2/2	
D15	How to deal with ambiguity over alerted patients that have NOT decompensated	2				2			1	5/3	
Human-computer interface											
D16	Whether integrated with EMR or not and if not—are tablets/phones allowed				3	1				4/2	
D17	Whether individual notification (hard alert) or aggregated dashboard (soft alert)				2	2	1			5/3	
D18	Which alert timing: suppression of alerts after first alert; one-time or repeat	1	2							3/2	
D19	Whether to provide clinician feedback or not										
D20	Whether prediction is explained or not	2	1		3	2				8/4	
Evaluation and monitoring											
D21	Which metrics to use				Identified by differences across papers						
D22	What process to follow: Silent trial or not and which trial method				1	2		1		4/3	
Count of papers		11	9	0	26	16	4	2	4	72	
Count of group decisions		8	6	0	14	11	4	2	4	49	

The numerals refer to the number of papers by group (A -> H) that discuss a particular decision. The totals column is in the format of: total number of papers/total number of groups.

EHR: electronic health record; ML: machine learning; DL: deep learning; ICU: intensive care unit; ED: emergency department.

Group (E) utilized an improvement cycle to decide on the alert threshold at each local implementation site in improving clinician adoption.^{47,59}

Related to the timing of alerts (D7), decisions about what actions clinicians should take for alerts involving patients showing no symptoms or signs of sepsis proved problematic (D15), as unclear roles and responsibilities constituted potential barriers to adoption (B3, B5).^{33,44}

Human computer interface (HCI) decisions (D16–D20). How algorithm predictions were presented to clinicians and whether they were accompanied by additional information or even

recommendations varied between groups. The HCI options comprised: (1) an alert only (Groups B, H), or with optional attached information (Group A) sent directly to clinicians via messaging systems (phones, e-mails, personal tablets) (D16); (2) content integrated within existing EHRs (Groups E, G); or (3) an external dashboard or application (Groups C, D, F) (D17). Integration into an EHR relied on organisations having a single EHR, otherwise multiple HCIs were required. Also, many EHRs did not have in-built capacity to support complex MLAs,⁴¹ whereas external dashboards conferred flexibility to design a bespoke solution that could also support mobile devices,⁴¹ although requiring clinicians to switch between applications interrupting workflows.⁴⁸ The type of alert

(D17) varied between hard alerts (such as a pop-up directive) requiring clinicians to immediately respond, and soft alerts (such as colored icons) that were more easily managed.^{41,42,46,48} No group indicated which method prompted more appropriate clinical actions and conferred better clinical outcomes.⁵⁰

Whether alerts were allowed to fire once or repeatedly until deactivated (D18) also varied between groups. The EWS2.0 (Group A) used a one-time alert, but found clinician evaluation of patients often occurred some hours after the alert fired.³³ Group (F) implemented completion and fall-out indicators for single alerts to visually guide clinicians to more timely review.⁵⁰ Group (B) supported multiple alerts for the same patient, but incorporated a snooze feature to suppress alerts within 6 h of the first alert.^{59,61} Whether to include more information about what caused alerts, versus just firing alerts alone (D20) had implications as to how the algorithm was trained. The decision by groups (E, F) to enable clinicians to feedback whether they thought the alert represented sepsis or something else (D19) enabled implementation teams to evaluate clinical utility, and provide feedback to clinicians about missed sepsis cases, which incentivized greater adoption.^{42,48}

Evaluation decisions (D21, D22). Evaluation decisions (D21) proved challenging as most groups omitted pre- and post-implementation evaluations of MLA performance using the same metrics. If done, it would have enabled linking of MLA performance with changes in clinical care or outcomes (Figure 2). Pre-implementation studies reported AUROC ranging from 0.63⁵¹ to 0.97⁴⁷ but only one post-implementation group (B) study⁵⁸ reported AUROC of 0.95, which was similar to pre-implementation studies.^{55,59,60}

In regard to pre-deployment silent or shadow trials evaluating algorithm performance against conventional clinical judgment in a live-data environment (D22),³³ 3 groups (A, D, E) conducted such trials for 6, 3 and an unknown number of months, respectively, during which algorithm validation was undertaken as well as end-to-end testing of the model, the data pipeline, the HCI and the clinical workflow.^{41,48}

DISCUSSION

Systematic review of sepsis MLA implementation studies

The systematic review served to learn how MLA performance, adoption, and different implementation modes were measured and how they impacted clinical care processes and patient outcomes. We found MLAs have potential to reduce mortality, but no definitive causal relationship has been demonstrated. At a minimum, the causal chain requires a high performing (high sensitivity/low false alarm) implemented MLA, clinician adoption and resulting positive changes to clinical processes (see Figure 2). Two groups (B, E) could demonstrate at least 2 of these, together with a significant reduction in mortality but only Group E reported definitive evidence of MLA adoption.

Demonstrating a causal link was limited by: (1) Non-randomized study designs being subject to confounding bias, such as sepsis awareness programs accompanying MLA implementation; and (2) Infrequently reported and non-standardized MLA performance metrics post-implementation, which, when they were reported, often showed decreased accuracy. Given these limitations, it remains unclear whether MLAs were responsible or needed for improved

mortality. In a meta-review of 55 observational studies of sepsis reduction programs using guideline-based care bundles,⁶⁴ a significant 34% overall reduction in mortality was achieved despite the absence of digitally embedded sepsis screening or alert tools in most studies (43/55, 78%).

Other important study findings were, firstly, clinical process improvements after MLA implementation did not always result in better patient outcomes, likely due to different clinical process improvement metrics ($N=36$). However, significant reductions in just one metric, median lead time from alert to first antibiotic, did coincide with significant reductions in mortality,^{46,47,58} suggesting this as an important indicator of MLA implementation success.

Second, it remains unclear whether MLA model choice impacts implementation success. Seven different algorithms were implemented with 5 reporting improved clinical indicators and mortality outcomes. The level of MLA performance post-implementation appears to be more important than choice of algorithm in predicting effectiveness. Across 2 different MLAs (Groups B and F), only the algorithm with high post-implementation performance was associated with significant mortality improvement.^{50,58} Similar results were seen for 2 independent implementations of the same algorithm (Group B).^{34,58}

Third, the choice of outcome definition, in this case sepsis, is critical as it can directly influence algorithm performance measures, particularly specificity.⁶ Definitions of sepsis varied from initial systemic inflammatory process (eg, Sepsis-1 definition)³ to multi-organ dysfunction (eg, Sepsis-3 definition)⁶⁵ reflecting a later, more advanced state of the illness. Importantly, the concern here is diagnosing sepsis (ie, using a diagnostic predictive algorithm) rather than predicting the likelihood of sepsis occurring in a patient before the inflammatory process begins (ie, a prognostic predictive algorithm).⁶⁶

Fourth, how algorithm predictions are presented to clinicians, and the extent to which they are accompanied by additional information or even recommendations are key determinants of clinician acceptance.⁶⁷

Mapping to SALIENT AI implementation framework

The second study objective was to map the review findings to the SALIENT framework to validate its coverage of important real-world implementation factors. Unlike in similar reviews,^{6-8,11,17,68,69} we conducted a novel 2-stage review wherein the second stage we identified related studies before or after the principal deployment study, which provided studies across the end-to-end MLA implementation process. We found the findings of each study could be mapped to one or more stages within SALIENT and that all SALIENT stages were utilized across all studies, indicating that SALIENT's implementation stages are both necessary and sufficient for real-world sepsis MLA implementation.

Secondly, every barrier, enabler, and decision identified in the review could be located to a stage (I-V) and either components (AI model, data pipeline, clinical workflow, HCI) or elements (A-G) within SALIENT. Knowing in advance what decisions are required (for example as a checklist), when they need to be made and in relation to which part of the implementation process is novel and could be informative to those engaged in AI implementation planning. We also found that most of the decision points, barriers and enablers identified were not specific to sepsis prediction, but were AI-task agnostic, suggesting SALIENT may have application for non-sepsis MLA implementation projects.

Strengths and limitations

As far as we know, our study is the first attempt to undertake a systematic review of sepsis prediction algorithms deployed in clinical settings, to identify barriers, enablers, and key decision points, and to map these to a single, inclusive, end-to-end implementation framework. The resulting framework and mapped items render these key decisions and contextual factors explicit, ordered and transparent, address gaps in current implementation guidance and offers a pragmatic staged approach for use by clinicians, informatics personnel and managers. Limitations relate to the small number of empirical studies of deployed sepsis-prediction algorithms, under-reporting of post-implementation performance metrics, focus on adult hospital settings, and potential publication bias from under-reporting of other sepsis MLA implementation studies.¹⁸ Although risk of bias for mortality reporting studies was moderate to high, all studies, including the 3 lowest bias papers,^{46,52,58} reported numerical reductions in mortality, with 5 being significant.^{46,58–61}

CONCLUSIONS

Our systematic review indicates that implementing MLAs within adult hospital care settings to predict sepsis has potential to reduce mortality, but no definitive causal link has been demonstrated. Implemented MLAs were few and only 2 provided some evidence of causation. The types of MLA models employed mattered less than their implementation accuracies and ability to alert clinicians to order antibiotics earlier.

This study also validated the SALIENT framework demonstrating real-world MLA implementation barriers, enablers, and decisions could be mapped to its constituent stages and components. Our findings highlight that AI implementation success has many more dimensions than the types of MLA employed, including evaluation methods and stages and the many decisions required throughout the multi-stage process. SALIENT may provide a roadmap for stakeholders to identify these stages, components and decisions which, with more robust studies, may be shown to conclusively link MLA implementation with significant improvement in patient outcomes. The SALIENT framework also has potential application to other MLA algorithms seeking to identify patients at risk of other acute hospital acquired conditions.

FUNDING

AHvdV was funded through a Queensland Government, Advanced Queensland Industry Research Fellowship grant. The Queensland Government had no role within this research.

AUTHOR CONTRIBUTIONS

AHvdV and IAS conceptualized the review. AHvdV, KD, and RJS conducted the title/abstract screening and full text review. VRK, RJS, and AHvdV performed the risk-of-bias assessments. AHvdV and KD performed all data extraction and tabular data collation. AHvdV derived the proposed framework. AHvdV, IAS, and KD drafted the manuscript with revisions and feedback from PJJ, VRK, and RJS.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

External review was conducted with thanks by Dr. Amith Shetty, Clinical Director at the NSW Ministry of Health.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests with respect to this publication.

DATA AVAILABILITY

There are no new data associated with this article.

REFERENCES

- Rudd KE, Johnson SC, Agesa KM, *et al.* Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* 2020; 395 (10219): 200–11.
- Fernando SM, Rochweg B, Seely AJE. Clinical implications of the third international consensus definitions for sepsis and septic shock (Sepsis-3). *CMAJ* 2018; 190 (36): E1058–9.
- Bone RC, Balk RA, Cerra FB, *et al.* Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest* 1992; 101 (6): 1644–55.
- Dugani S, Veillard J, Kissonoo N. Reducing the global burden of sepsis. *CMAJ* 2017; 189 (1): E2–3.
- Seymour CW, Gesten F, Prescott HC, *et al.* Time to treatment and mortality during mandated emergency care for sepsis. *N Engl J Med* 2017; 376 (23): 2235–44.
- Hassan N, Slight R, Weiland D, *et al.* Preventing sepsis; how can artificial intelligence inform the clinical decision-making process? A systematic review. *Int J Med Inform* 2021; 150: 104457.
- Ackermann K, Baker J, Green M, *et al.* Computerized clinical decision support systems for the early detection of sepsis among adult inpatients: scoping review. *J Med Internet Res* 2022; 24 (2): e31083.
- Wulff A, Montag S, Marscholke M, Jack T. Clinical decision-support systems for detection of systemic inflammatory response syndrome, sepsis, and septic shock in critically ill patients: a systematic review. *Methods Inf Med* 2019; 58 (S 02): E43–57.
- Schinkel M, Paranjape K, Nannan Panday RS, Skyttberg N, Nanayakkara PWB. Clinical applications of artificial intelligence in sepsis: a narrative review. *Comput Biol Med* 2019; 115: 103488.
- Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. Prediction of sepsis patients using machine learning approach: a meta-analysis. *Comput Methods Programs Biomed* 2019; 170: 1–9.
- Fleuren LM, Klausch TLT, Zwager CL, *et al.* Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020; 46 (3): 383–400.
- Stead WW, Haynes RB, Fuller S, *et al.* Designing medical informatics resource projects to increase what is learned. *J Am Med Inform Assoc* 1994; 1 (1): 28–33.
- Greenhalgh T, Wherton J, Papoutsis C, *et al.* Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res* 2017; 19 (11): e367.
- Reddy S, Rogers W, Makinen VP, *et al.* Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform* 2021; 28 (1): 1–7.

15. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. *J Med Internet Res* 2022; 24 (1): e32215.
16. Bakken S, Ruland CM. Translating clinical informatics interventions into routine clinical care: how can the RE-AIM framework help? *J Am Med Inform Assoc* 2009; 16 (6): 889–97.
17. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Front Med (Lausanne)* 2021; 8: 607952.
18. Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. *JAMIA Open* 2022; 5 (2): 1–11.
19. Schwartz JM, George M, Rossetti SC, et al. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: qualitative descriptive study. *JMIR Hum Factors* 2022; 9 (2): e33960.
20. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMS (Wash DC)* 2019; 7 (1): 1.
21. van der Vegt A, Scott I, Dermawan K, Schnetler R, Kalke V, Lane P. Implementation frameworks for end-to-end clinical AI: derivation of the SALIENT framework. 2023. Manuscript submitted for publication.
22. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022; 377: e070904. doi:10.1136/BMJ-2022-070904.
23. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009; 4: 50.
24. Reed JE, Howe C, Doyle C, Bell D. Successful Healthcare Improvements From Translating Evidence in complex systems (SHIFT-Evidence): simple rules to guide practice and research. *Int J Qual Health Care* 2019; 31 (3): 238–44.
25. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur J Clin Invest* 2015; 45 (2): 204–14.
26. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162 (1): W1–73.
27. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020; 370: m3164.
28. Moher D, Shamseer L, Clarke M, et al.; PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; 4 (1): 1.
29. Veritas Health Innovation. Covidence. 2022. www.covidence.org. Accessed August 4, 2022.
30. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016; 355: i4919.
31. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; 366: l4898.
32. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23 (3): 269–78.
33. Giannini HM, Ginestra JC, Chivers C, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med* 2019; 47 (11): 1485–92.
34. Topiwala R, Patel K, Twigg J, Rhule J, Meisenberg B. Retrospective observational study of the clinical performance characteristics of a machine learning approach to early sepsis identification. *Crit Care Explor* 2019; 1 (9): e0046.
35. Gonçalves LS, de Amaro ML, de Romero M, et al. Implementation of an artificial intelligence algorithm for sepsis detection. *Rev Bras Enferm* 2020; 73 (3): e20180421.
36. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018; 8 (1): e017833.
37. Scherer J de S, Pereira JS, Debastiani MS, Bica CG. Beyond technology: can artificial intelligence support clinical decisions in the prediction of sepsis? *Rev Bras Enferm* 2022; 75 (5): e20210586.
38. Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multi-task Gaussian process RNN classifier. In: *34th International Conference on Machine Learning, ICML 2017*. Vol 3; Sydney; 2017:1914–1922.
39. Futoma J, Hariharan S, Sendak M, et al. An improved multi-output Gaussian process RNN with real-time validation for early sepsis detection. *Proc Mach Learn Healthc* 2017; 68: 2017. <http://arxiv.org/abs/1708.05894>.
40. Bedoya AD, Futoma J, Clement ME, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open* 2020; 3 (2): 252–60.
41. Sendak MP, Ratliff W, Sarro D, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform* 2020; 8 (7): e15182.
42. Sandhu S, Lin AL, Brajer N, et al. Integrating a machine learning system into clinical workflows: qualitative study. *J Med Internet Res* 2020; 22 (11): e22421.
43. Sendak M, Elish MC, Gao M, et al. “The human body is a black box”: supporting clinical decision-making with deep learning. In: *FAT 2020—Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020:99–109. doi:10.1145/3351095.3372827.
44. Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit Care Med* 2019; 47 (11): 1477–84.
45. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015; 7 (299): 299ra122–299ra122. doi:10.1126/SCITRANSLMED.AAB3719/SUPPL_FILE/7-299RA122_SM.PDF.
46. 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): 1455–60.
47. Henry KE, Kornfield R, Sridharan A, et al. Human-machine teaming is key to AI adoption: clinicians’ experiences with a deployed machine learning system. *NPJ Digit Med* 2022; 5 (1): 97.
48. 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): 1447–54.
49. Harrison AM, Thongprayoon C, Kashyap R, et al. Developing the surveillance algorithm for detection of failure to recognize and treat severe sepsis. *Mayo Clin Proc* 2015; 90 (2): 166–75.
50. Lipatov K, Daniels CE, Park JG, et al. Implementation and evaluation of sepsis surveillance and decision support in medical ICU and emergency department. *Am J Emerg Med* 2022; 51: 378–83.
51. Wong A, Orles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021; 181 (8): 1065–70.
52. Schootman M, Wiskow C, Loux T, et al. Evaluation of the effectiveness of an automated sepsis predictive tool on patient outcomes. *J Crit Care* 2022; 71: 154061.
53. Brown SM, Jones J, Kuttler KG, Keddington RK, Allen TL, Haug P. Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department. *BMC Emerg Med* 2016; 16 (1): 31.
54. Burdick H, Pino E, Gabel-Comeau D, et al. Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. *BMC Med Inform Decis Mak* 2020; 20 (1): 1–10.
55. Calvert JS, Price DA, Chettipally UK, et al. A computational approach to early sepsis detection. *Comput Biol Med* 2016; 74: 69–73.

56. Calvert J, Desautels T, Chettipally U, *et al.* High-performance detection and early prediction of septic shock for alcohol-use disorder patients. *Ann Med Surg (Lond)* 2016; 8 (2016): 50–5.
57. Desautels T, Calvert J, Hoffman J, *et al.* Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 2016; 4 (3): e28.
58. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Resp Res* 2017; 4 (1): e000234.
59. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual* 2017; 6 (2): e000158.
60. Burdick H, Pino E, Gabel-Comeau D, *et al.* Evaluating a sepsis prediction machine learning algorithm in the emergency department and intensive care unit: a before and after comparative study. *bioRxiv* 2018; (872): 224014.
61. Burdick H, Pino E, Gabel-Comeau D, *et al.* Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Heal Care Informatics* 2020; 27 (1): 1–8.
62. Calvert J, Hoffman J, Barton C, *et al.* Cost and mortality impact of an algorithm-driven sepsis prediction system. *J Med Econ* 2017; 20 (6): 646–51.
63. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA - J Am Med Assoc* 2018; 320 (11): 1107–8.
64. Damiani E, Donati A, Serafini G, *et al.* Effect of performance improvement programs on compliance with sepsis bundles and mortality: a systematic review and meta-analysis of observational studies. *PLoS One* 2015; 10 (5): e0125827.
65. Seymour CW, Liu VX, Iwashyna TJ, *et al.* Assessment of clinical criteria for sepsis for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA - J Am Med Assoc* 2016; 315 (8): 762–74.
66. Faisal M, Scally A, Richardson D, *et al.* Development and external validation of an automated computer-aided risk score for predicting sepsis in emergency medical admissions using the patient's first electronically recorded vital signs and blood test results. *Crit Care Med* 2018; 46 (4): 612–8.
67. Kennedy G, Gallego B. Clinical prediction rules: a systematic review of healthcare provider opinions and preferences. *Int J Med Inform* 2019; 123 (November 2017): 1–10.
68. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
69. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021; 23 (4): e25759.