Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial intelligence for caesarean surgical site infection surveillance: A scoping review of risk prediction, detection and clinical translation

Maiga Ayub Hussein
maigayub@gmail.com
Busitema University
Cahya Tri Purnami
Diponegoro University
Bagoes Widjanarko
Diponegoro University
Julius Nteziyaremye
Busitema University
Lintang Dian Saraswati
Diponegoro University

Systematic Review

Keywords: artificial intelligence, CFIR, caesarean section, surgical site infection, wound surveillance

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11515352/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Caesarean surgical site infection (cSSI) can prolong recovery, increase costs and require readmission. Surveillance is challenging after discharge, when wound changes may go unnoticed and travel, cost and limited clinical access delay assessment. Artificial intelligence (AI) could support risk stratification and infection detection, but its clinical value depends on performance and integration into maternal care. This review mapped AI applications for cSSI and examined clinical translation using the Consolidated Framework for Implementation Research (CFIR).

Methods

This scoping review followed Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews and was preregistered on the Open Science Framework (DOI: https://doi.org/10.17605/OSF.IO/JCXD2). Scopus, PubMed, Web of Science, IEEE Xplore, and Cochrane Library were searched from inception to 14 October 2026, supplemented by Google Scholar citation searching. Two independent reviewers screened records. English-language studies reporting cSSI and AI-assisted surveillance were eligible. Study characteristics, applications, performance, and validation were charted and synthesised descriptively. Implementation-relevant findings were mapped to the updated CFIR domains. Performance estimates were not pooled because of methodological and clinical heterogeneity.

Results

Eight studies published between 2019–2026 from Rwanda, the United States and Australia. Applications included clinical-data risk prediction, colour and thermal wound-image classification, smartphone-based image processing and language-model auditing of clinical notes. The largest caesarean-specific study included 5,272 deliveries. Its area under the receiver operating characteristic curve (AUC), declined from 0.78 during training to 0.69 [95% confidence interval 0.63–0.75] in external testing, indicating modest discrimination in a separate population. Colour-image models reported median AUC of 0.639–0.655, whereas thermal transfer learning achieved 0.90 using only 30 infected wounds and internal evaluation. A language-model audit achieved 93% patient-level accuracy with 100% sensitivity and 86% specificity in 28 mixed-surgery patients. CFIR mapping identified acquisition standardisation, adaptation and information infrastructure as translation considerations; leadership, sustained financing, recipient engagement and implementation outcomes were insufficiently examined. No study demonstrated improved maternal postoperative outcomes attributable to AI deployment.

Conclusions

AI offers potential support for cSSI surveillance, but technical performance does not establish clinical effectiveness. Clinical translation requires independent validation and prospective evaluation of human-supervised pathways, including timely assessment, equitable access, staff workload and sustainable intervention delivery.

1. Introduction

Surgical site infections (SSI) complicate 2–7% of all caesarean section (CS). They are psychological and physical burden to the mother contributing 3% to maternal mortality[1, 2]. Globally, the pooled cesarean surgical site infection (cSSI) incidence is at 5.63% (95% confidence interval [CI] 5.18–6.11%), and highest at 11.91% (95% CI 9.67–14.34%) in the African region [3]. The traditional surveillance confined to admission misses detecting 60.1% of SSI that occur after hospital discharge [4]. In rural Rwanda, post-discharge travel and household expenditure add to the burden of caesarean recovery [5]. Telemedicine can extend contact, but access and diagnostic reliability are separate questions. A prospective community health worker (CHW)-supported study of 787 women reported 97.6% specificity and 36.8% sensitivity for remote SSI assessment [6]. Thus, successful image transmission does not ensure that infection is recognised.

Artificial intelligence (AI) applications address different points in this pathway. Clinical-data models estimate subsequent infection risk; colour and thermal classifiers identify signs of infection present at imaging; language models search documentation for evidence of infection [7–9]. A capture application may standardise the input without itself establishing clinical benefit. However clinical translation cannot be inferred from the ability of the AI to discriminate between SSI and none SSI alone [10]. The World Health Organization (WHO) guidance emphasises autonomy, safety, accountability and equity, while subsequent regulatory and language-model guidance addresses lifecycle oversight and the risks of generated outputs [11]. Artificial intelligence implementation literature likewise identifies dependencies on data, workflows and human oversight [12].

Previous reviews have examined AI tools for surgical site infection surveillance; however, the evidence remains dispersed across broad surgical procedure and not explicitly cesarean section [13, 14]. A focused synthesis of AI use for cSSI surveillance is needed to clarify on prediction, detection and clinical translation. The updated Consolidated framework for implementation research (CFIR) provides a determinant framework for examining this objective through Innovation, Outer Setting, Inner Setting, Individuals and Implementation Process [15]. It is not a diagnostic appraisal tool, an effectiveness measure or a prescribed implementation sequence. Its value here is to organise what the included reports reveal about clinical translation and what remains unknown. The review questions were: (1) Which AI applications address cSSI risk, detection or surveillance? (2) How were their performance and transportability evaluated? (3) What implementation-relevant evidence and gaps can be mapped to CFIR?

2. Methods

2.1 Design

The scoping review followed JBI scoping-review guidance and was reported with reference to Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [16, 17]. An a priori protocol was registered on the Open Science Framework (DOI: https://doi.org/10.17605/OSF.IO/JCXD2)[18].

2.2 Study eligibility

Eligibility was organised using the population, concept and context framework [16]. Eligible English-language reports concerned women who underwent caesarean delivery and AI or machine-learning applications for SSI risk prediction, detection or surveillance in any country or care setting. Direct cSSI reports were distinguished from mixed-surgery reports with identifiable caesarean relevance. Non-AI telemedicine, reviews and studies of non-infectious wound outcomes were excluded from the AI evidence set, although relevant sources informed the introduction and discussion.

2.3 Information sources and search strategy

Five electronic databases were searched: Scopus, PubMed, Web of Science, IEEE Xplore and the Cochrane Library from inception to 14 October 2026, with forward and backward citation searching in Google Scholar. English-language restriction was applied during the search process. The search strategy combined three blocks of terms relating to caesarean birth, surgical site infection, and artificial intelligence or machine learning.

Block 1: (cesarean* OR caesarean* OR cesarian* OR caesarian* OR “c-section” OR “c-sections” OR “c section” OR “c sections” OR “abdominal delivery” OR “abdominal deliveries” OR “abdominal birth”). Block 2: (“surgical site infection*” OR “surgical wound infection*” OR “postoperative infection*” OR “post-operative infection*” OR “postcesarean infection*” OR “postcaesarean infection*” OR SSI OR SSIs OR cSSI OR wound* OR incision* OR “surgical site”). Block 3: (“artificial intelligence” OR “machine learning” OR “deep learning” OR “computer vision” OR “image recognition” OR “image classification” OR “image segmentation” OR “automated image analysis” OR “natural language processing” OR “large language model*” OR “neural network*” OR “convolutional network*” OR “transfer learning” OR “random forest*” OR “support vector machine*” OR “gradient boosting” OR “extreme gradient boosting” OR XGBoost OR LightGBM OR “decision tree*” OR “naive bayes” OR “naïve bayes” OR “ensemble learning” OR “supervised learning” OR “unsupervised learning” OR “AI-assisted” OR “AI-based” OR “AI-enabled” OR “AI-driven”). The three blocks were combined using the Boolean operator “AND” across all five databases. The search syntax “TITLE-ABS-KEY” was used as the field specification and adapted to the requirements of each database.

2.4 Study selection

Records from the databases were imported into EndNote version 21, and duplicates were removed automatically with manual confirmation. Two reviewers (MAH and NJ) independently screened titles/abstracts and full texts, with disagreements resolved by consensus or a third reviewer (CTP). The eight included reports are summarised in the PRISMA flowchart in results.

2.5 Data extraction

A structured predefined data extraction excel file was used to capture specific study characteristics by one independent reviewers (MAH) and later verified by the second reviewer (NJ). Charted items included source type, country, population, data modality, clinical purpose, timing, algorithm, intended user, reference standard, performance, validation and implementation-relevant reporting. Tables 1 and 2 separate study characteristics from task-specific performance and readiness. The same participants may appear in related Rwanda reports; sample sizes were therefore not summed. Heterogeneity in targets, reference standards, thresholds and validation precluded pooling. Formal risk-of-bias assessment was not undertaken, a permissible methodological choice since the did not aim at quality appraisal of literature. Design limitations were described without assigning formal quality scores [19].

Table 1
Summary of characteristics, methods, and clinical applications for the eight included studies

Source	Setting	Design	Population and data	AI method	Outcome and application
Fletcher et al. (2019)[9]	Rwanda; Kirehe District Hospital	Model development; 100 random 75%/25% training/test splits	Follow-up: 572/729 eligible women at day 10 ± 3. Questionnaire: 572 women/61 infections; image analysis: 568/62.	Questionnaire L1/L2 logistic regression and SVM; engineered colour/texture features with L1 selection for image classification.	Classify SSI present at postoperative review using symptoms/signs and wound photographs; independent general-practitioner examination supplies labels.
Fletcher et al. (2021)[7]	Rwanda; Kirehe District Hospital	Colour-image CNN development with internal cross-validation	572 women/images; 62 infected and 510 non-infected; day 10 ± 3. Same recruitment period as the 2019 report.	Naïve CNN and pretrained ResNet50; augmentation and minority-class weighting; logistic-regression comparator.	Classify current wound infection from CHW-acquired RGB photographs against independent clinical examination; screening sensitivity prioritised.
Fletcher et al. (2021)[10]	Rwanda; Kirehe District Hospital	Prospective enrolment; thermal CNN development and internal cross-validation	530 women; 30 infected and 500 non-infected wounds; postoperative day 11 ± 3.	Naïve CNN and ResNet50 transfer learning; thermal temperature arrays, augmentation and class weighting.	Classify current SSI using phone-attached thermal imaging and general-practitioner reference assessment.
Alayande et al. (2023)[8]	Rwanda programme	Opinion and development lessons	No new independently recruited cohort.	Programme-level synthesis of visible/thermal image learning, computer-vision alignment, cropping and colour calibration.	Describe acquisition, labelling and generalisability requirements for CHW-supported home monitoring; quantitative examples summarise earlier work.
Wang (2022) [25]	MIT development; Rwanda data and intended CHW deployment	Master's thesis; platform development and technical testing	Stored-image tests: 32 infected + 32 non-infected; 32 + 550. Three Android phones; unique women and	Existing ResNet50 model converted to TensorFlow Lite; computer-vision capture,	Integrate standardised photography and SSI classification in Wound Screener; evaluate device runtime, pre-processing

AI = artificial intelligence; CHW = Community health worker; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; L1/L2 = Regularisation penalties; MIT = Massachusetts Institute of Technology; RGB = Red green blue; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy.

Source	Setting	Design	Population and data	AI method	Outcome and application
			test-set overlap unspecified.	quality checks, colour calibration; local/cloud inference.	consistency and classification delivery.
Larson et al. (2024)[22]	United States; Premier Healthcare Database, 2017–2021	Retrospective specialty-specific models and economic simulation	Caesarean training cohort: 937,875 (source Table 1); economic test cohort: 165,290. Broader study includes other procedures.	CatBoost selected after comparing tabular classifiers; predicted risk combined with prevention efficacy and cost inputs.	Predict preoperative composite SSC risk and simulate targeted sNPWT allocation. Endpoint includes SSI, dehiscence and haematoma/seroma.
Lobel et al. (2026)[23]	United States; two level IV academic centres	Retrospective development, internal validation and independent external testing	5,272 deliveries/181 cSSIs: development 2,336/109; external cohort 2,936/72.	XGBoost selected over random forest and regularised logistic regression; 15 discharge-available variables; Platt scaling.	Estimate subsequent cSSI risk at discharge and stratify follow-up. Outcome windows: six weeks in development and 30 days externally.
Satheakeerthy et al. (2025) [24]	Australia; two metropolitan hospitals	Retrospective case–control clinical-note audit	28 patients: 14 SSI/14 controls; 12/14 SSI cases followed caesarean birth. 379 case notes and 304 control notes.	Zero-shot, locally deployable 70-billion-parameter Llama model; criteria-based prompting of individual notes.	Identify documented SSI for audit; any positive note classifies the patient. Nurse labels confirmed by medical review; no separate cSSI accuracy.

AI = artificial intelligence; CHW = Community health worker; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; L1/L2 = Regularisation penalties; MIT = Massachusetts Institute of Technology; RGB = Red green blue; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy.

Table 2
Application-specific performance, validation limits, and clinical interpretation

Application	Inputs user and timing	Reported performance	Validation limits	Clinical interpretation
Symptom and image assessment [9]	Community health worker (CHW)-collected symptom questionnaires and red–green–blue (RGB) wound photographs at postoperative day 10 ± 3; physician-confirmed infection.	Questionnaire: Area under the receiver operating characteristic curve (AUC) 0.849–0.867; sensitivity 70.6–75.0%; specificity 98.5–100%; accuracy 95.8–96.5%. Image model: AUC 1.00; sensitivity and specificity 100%; accuracy 100%. Combined assessment: sensitivity and specificity 99%.	Repeated internal training–test splits without external validation. Differences in sample denominators and inconsistencies between reported image-model results.	Questionnaire AUC ≈ 0.85–0.87 indicates reasonably strong ranking within these splits, but sensitivity ≈ 71–75% corresponds to missing roughly 25–29 of every 100 infected women at those operating points. Near-perfect image separation is an internal, inconsistently reported result; it is not a dependable estimate for home screening.
Current wound classification [7]	CHW-acquired Red-Green-Blue images at day 10 ± 3; Convolutional neural network (CNN), ResNet50, and logistic regression compared against clinical examination.	Naïve CNN: AUC 0.655; sensitivity 75%; specificity 58%; Matthews correlation coefficient (MCC) 0.35. ResNet50: AUC 0.639; sensitivity 92–93%; specificity 18%; MCC 0.41. Logistic regression: AUC approximately 0.563.	Internal cross-validation involving 62 infected wounds; no external validation. Inconsistent specificity labels and performance summaries.	CNN discrimination was limited. Higher ResNet50 sensitivity was accompanied by approximately 82% false-positive classification among non-infected
Current wound classification [10]	Phone-attached thermal camera; postoperative day 11 ± 3; naïve CNN/ResNet50; doctor-assessed SSI.	Naïve CNN: median AUC 0.844 (IQR 0.34), sensitivity 71%, specificity 87%, MCC 0.37; maximum accuracy 86%. Transfer CNN: median AUC 0.900 (IQR 0.19), MCC 0.44; maximum accuracy 94%. Sensitivity/specificity reported as 95%/84% in the narrative but reversed in table.	Internal evaluation; five-fold transfer-model validation; 30 infected and 500 non-infected wounds. 95% CIs not reported. Conflicting sensitivity and specificity estimate for the transfer model.	Thermal imaging showed promising discrimination, but inconsistent operating-point estimates and limited infection cases restrict clinical interpretation

AUC = Area under the receiver operating characteristic curve; CHW = Community health worker; CI = Confidence interval; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; F1/F2 = Precision–recall harmonic scores (F2 gives more weight to recall); IQR = Interquartile range; L1/L2 = Regularisation penalties; MCC = Matthews correlation coefficient; Ms = milliseconds; RF = Random forest; RGB = Red, green blue; RLR = Regularised logistic regression; ROC = Receiver operating characteristic; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy;

Application	Inputs user and timing	Reported performance	Validation limits	Clinical interpretation
Programme lessons [8]	Visible/thermal imaging; computer-vision alignment, cropping and colour calibration for proposed home surveillance.	Visible-image classification: sensitivity improved from 81.3% to 97%, specificity from 65.3% to 87%, and AUC from 0.64 to 0.86 following workflow refinements. Thermal performance labels were inconsistent.	Opinion article; no new independent cohort, denominators or paired comparative analysis. Calibration estimates must not be counted as an additional validation.	Standardized image acquisition may improve classification performance, but clinical effectiveness and reproducibility remain unverified.
Technical delivery [25]	Android-based RGB wound photography, image-quality feedback, and ResNet50 classification using TensorFlow Lite or cloud inference.	Balanced dataset: sensitivity 53.1–93.8%; specificity 6.3–46.9%; accuracy 48.4–54.7%. Imbalanced dataset: sensitivity 37.5–87.5%; specificity 13.3–76.4%; accuracy 17.4–74.2%. Device processing: pre-processing 428–4,755 milliseconds; inference 317–1,504 milliseconds.	Stored-image technical testing without prospective clinical validation. Unspecified patient overlap, inconsistent platform outputs, and differences between tested and implemented thresholds. Android/Python outputs disagreed.	Model performance was highly threshold-dependent. The highest reported accuracy in the imbalanced dataset coincided with missing 62.5% of infected images, demonstrating why accuracy alone is insufficient. Device-dependent processing and inconsistent model outputs require resolution before clinic
Preoperative prevention support [22]	Preoperative demographic and claims-derived variables; CatBoost prediction of surgical site complications and simulated single-use negative-pressure wound therapy (sNPWT) allocation.	Composite complication prediction: incidence 1.7%; AUC 0.74; precision 69%; recall 15%; F1 score 0.24; F2 score 0.18. Economic simulation: sNPWT allocation was modelled as cost-saving for 61% of 165,290 caesarean patients.	Retrospective internal validation using a composite outcome; no infection-specific performance, external testing, or prospective evaluation of prevention benefits.	Moderate discrimination but low recall, with 85% of complications missed at the reported threshold. Simulated cost savings do not establish clinical effectiveness.
Discharge risk stratification [23]	Fifteen electronic medical record variables available at discharge; extreme gradient boosting (XGBoost) with	AUC: training 0.78 (95% confidence interval [CI] 0.73–0.82); internal validation 0.72 (95% CI 0.66–0.76); external testing 0.69 (95% CI 0.63–0.75). Risk	Independent external validation, but both cohorts came from specialized academic centres.	External discrimination was modest. Risk-targeted follow-up could prioritize high-risk women but would miss

AUC = Area under the receiver operating characteristic curve; CHW = Community health worker; CI = Confidence interval; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; F1/F2 = Precision–recall harmonic scores (F2 gives more weight to recall); IQR = Interquartile range; L1/L2 = Regularisation penalties; MCC = Matthews correlation coefficient; Ms = milliseconds; RF = Random forest; RGB = Red, green blue; RLR = Regularised logistic regression; ROC = Receiver operating characteristic; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy;

Application	Inputs user and timing	Reported performance	Validation limits	Clinical interpretation
	Platt calibration for subsequent caesarean surgical site infection (cSSI).	stratification: highest-risk 0.7% of women had 22.7% cSSI incidence. The two highest bands included 4.5% of women and captured 18.1% of infections.	Different outcome windows and infection prevalence; no prospective clinical impact	approximately 82% of infections if restricted to the two highest bands.
Clinical-note surveillance [24]	Inpatient and outpatient clinical notes analysed using a zero-shot 70-billion-parameter Llama model; nurse assessments confirmed by medical review.	Patient-level: sensitivity 100%; specificity 85.7%; accuracy 92.9%; F1 score 93%; AUC 0.91 (95% CI 0.73–1.00). Detection timing: AI flagged seven infections before, six on, and one after documented onset; median difference − 0.5 days.	Retrospective case–control study involving 28 patients, including 14 infections. No separate cSSI estimates, prospective validation, or confirmed improvement in clinical detection time.	All sampled infections were identified, with two false positives. Small sample size and inclusion of non-caesarean cases limit generalisability and clinical readiness.

AUC = Area under the receiver operating characteristic curve; CHW = Community health worker; CI = Confidence interval; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; F1/F2 = Precision–recall harmonic scores (F2 gives more weight to recall); IQR = Interquartile range; L1/L2 = Regularisation penalties; MCC = Matthews correlation coefficient; Ms = milliseconds; RF = Random forest; RGB = Red, green blue; RLR = Regularised logistic regression; ROC = Receiver operating characteristic; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy;

2.6 Data synthesis

A narrative synthesis was used guided by the updated Consolidated framework for implementation research (CFIR), for this synthesis, the innovation was defined as an AI-supported cSSI surveillance service, with risk prediction, image triage and clinical note review treated as alternative or complementary components rather than a single validated product. The Inner Setting was the maternity, community-follow-up or infection-prevention service responsible for delivery; the Outer Setting comprised the surrounding community, referral environment, policy and financing arrangements. Women were innovation recipients; community health workers (CHWs), maternity clinicians and infection-control staff were potential deliverers. Implementation-relevant observations were mapped deductively to the updated CFIR domains of Innovation, Outer Setting, Inner Setting, Individuals and Implementation Process [15]. In Table 3, the domain definitions were operationalized for an AI-supported cSSI surveillance service. Author-described constraints and proposed workflows were distinguished from measured technical results and participant-reported experience. Clinical translations are review interpretations, not observed intervention effects. Predictors of infection were not treated as implementation determinants and the mapping estimate are not frequencies or causal influences of the determinants on implementation outcomes [20].

Table 3

CFIR domain definitions, evidence mapping, and implications for clinical translation

CFIR domain	Operational definition	Evidence synthesis and supported constructs	Clinical translation
Innovation	Features of the AI-supported surveillance component itself that affect its use: evidentiary support, adaptability, design, complexity and cost. The component may be a wound classifier, risk model, capture platform or note-audit tool.	Innovation Evidence Base: independent cSSI risk-model testing yielded AUC 0.69, while thermal evidence remained internal and small [10, 23]. Innovation Design and Adaptability: capture targets, automatic cropping, colour calibration and recapture prompts replaced variable free-form images; offline/cloud options changed delivery conditions [8, 25]. Innovation Complexity and Cost: model conversion, phone-dependent runtime and a US$200 thermal module introduce device and maintenance requirements [10, 25].	Validate the exact model–pre-processing–device combination intended for service use, including poor-quality-image handling and threshold behaviour. Treat external risk discrimination as support for testing targeted contact, not excluding lower-risk women from care. Thermal equipment procurement requires a locally tested diagnostic pathway, rather than relying on the internally reported AUC alone [23, 25].
Outer setting	Community and environmental conditions outside the delivering maternity/community/infection-prevention service that influence women's ability to participate and the technology's ability to function.	Local Conditions: rural internet limitations restricted cloud inference; offline use shifts dependence towards phone capability and affordability (Wang, 2022). Physical and financial burdens of returning to facilities motivate home monitoring in the Rwanda programme (Alayande et al., 2023). These are author-described contextual conditions, not participant-derived estimates of AI acceptability.	For a home-monitoring pilot, test offline capture/inference and later synchronisation under actual network conditions. Measure whether flagged women can reach assessment, including travel time and out-of-pocket cost; otherwise, improved classification may not yield completed care. Maintain a contact/referral route for women who cannot use the digital pathway (Alayande et al., 2023; Wang, 2022).

AI = Artificial intelligence; AUC = Area under the receiver operating characteristic curve; CFIR = Consolidated Framework for Implementation Research; CHW = Community health worker; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; LLM = Large language model; US$ = United States dollars.

CFIR domain	Operational definition	Evidence synthesis and supported constructs	Clinical translation
Inner setting	Infrastructure, resources and routines within the responsible service that enable capture, computation, documentation, review and clinical follow-up.	Structural Characteristics Information Technology Infrastructure: LLM surveillance depends on complete, timely EMR notes and can operate locally; discharge risk modelling requires available EMR variables [23, 24]. Available Resources Materials and Equipment: phone capability and software/library compatibility affect delivery [25]. Compatibility: preoperative allocation is designed to allow device preparation before closure, whereas discharge risk estimation and note auditing enter different existing workflows; these are proposed fits, not demonstrated adoption [22, 23].	Check feature availability and note coverage at the intended decision time before integration. Assign responsibility for receiving image alerts or audit flags and confirming suspected infection. Test the complete software version on service phones and estimate review volume at local prevalence; a locally hosted model still requires a functioning documentation and response service [23–25].
Individual characteristics	Role-specific capability and opportunity of deliverers and recipients to perform required tasks: acquire usable images, interpret an output, document findings and obtain clinical assessment. This row corresponds to the updated CFIR Individuals domain.	Innovation Deliverers Capability: the thesis identifies limited CHW clinical/technical training as a design constraint and supplies immediate quality feedback; early symptom assessment depended on the administrator's clinical experience [9, 25]. Innovation Deliverers Opportunity: note-based surveillance depends on staff having the opportunity to create timely documentation, which the audit authors identify as a	Evaluate whether CHWs can capture a usable wound image and follow a recapture/escalation instruction during observed tasks. Test whether clinicians distinguish a risk score, current-wound classification and audit flag, and can correct false positives such as prophylaxis misclassification. These competency checks are proposed responses to the reported requirements; acceptance or confidence cannot be inferred from staff participation in data collection [9, 24, 25].

AI = Artificial intelligence; AUC = Area under the receiver operating characteristic curve; CFIR = Consolidated Framework for Implementation Research; CHW = Community health worker; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; LLM = Large language model; US$ = United States dollars.

CFIR domain	Operational definition	Evidence synthesis and supported constructs	Clinical translation
		condition of performance [24]. These are author-described requirements rather than measured capability or opportunity scores.	
Implementation process	Preparatory or delivery activities used to assess context, adapt the intervention and examine how the innovation functions in its intended workflow; technical preparation is distinguished from live service implementation.	Assessing Context: programme and platform reports identify, capture variability, connectivity and device limitations [8, 25]. Tailoring Strategies: targets, automatic calibration, image-quality feedback and local/cloud operation respond to these conditions [8, 25]. Reflecting and Evaluating Innovation: pre-processing, runtime and Android/Python comparisons identify unresolved delivery problems; external risk testing examines transportability [23, 25]. This is retrospective mapping of development activities, not evidence of a completed CFIR-guided implementation programme.	First resolve software equivalence and evaluate a fixed model/device/threshold prospectively without altering care. A subsequent supervised pilot can track alert acknowledgement, confirmatory review, referral completion, false alerts and missed infections. Compare these process measures with time to assessment and treatment, so adaptation addresses the specific point at which the surveillance pathway fails [23–25].

AI = Artificial intelligence; AUC = Area under the receiver operating characteristic curve; CFIR = Consolidated Framework for Implementation Research; CHW = Community health worker; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; LLM = Large language model; US$ = United States dollars.

3. Results

3.1 Study Selection

Eight reports were included from 80 identified records in total. After removal of 36 duplicates from the 77 records identified from the databases, 41 records underwent title and abstract screening; 34 were excluded. Seven full-text were retrieved and assessed for eligibility with 3 more records identified by citation from google scholar. Out of 10 eligible records, 2 were excluded (one did not evaluate artificial intelligence [6] and one was not specific on

surgical site infection [21]), leaving eight studies that were included for review. Figure-1 summarize the study selection processes.

3.2 Characteristics of included studies

As detailed below, the final selection comprised eight reports published between 2019 and 2026, evaluating distinct AI modalities, validation methods, and clinical targets.

[Insert Table-1. Here]

Table 1 shows the eight studies published between 2019–2026, four conducted in Rwanda [7–10], the United States (n = 2) [22, 23], and Australia (n = 1)[24], alongside a US-based technical development study using Rwandan data (n = 1)[25]. Three Rwandan studies developed questionnaire-based, colour-image, and thermal-image infection classifiers using postoperative clinical assessments as reference standards[7, 9, 10]. Alayande, Prasad [8] synthesized programme-level implementation lessons without recruiting a new cohort, while Wang [25] evaluated smartphone-based image acquisition, pre-processing, and model deployment. The two US retrospective studies developed machine-learning models for preoperative surgical site complication prediction and targeted prevention (937,875 caesarean procedures), and discharge-time cSSI risk stratification (5,272 deliveries), respectively [22, 23]. An Australian case–control study evaluated large language model classification of clinical notes from 28 patients, including non-caesarean cases[24]. AI methods included regularised logistic regression, support vector machines, convolutional neural networks, ResNet50 transfer learning, CatBoost, XGBoost, and Llama. Applications spanned postoperative wound classification, prospective risk prediction, prevention allocation, and clinical documentation surveillance. Validation approaches ranged from repeated random splits and internal cross-validation to independent external testing and technical performance assessment. Several Rwandan publications shared underlying data, while the US prevention model used a composite complication outcome and the Australian study did not report separate cSSI accuracy; therefore, populations and outcomes were not directly comparable

3.3 Performance, Validation and Clinical interpretation of AI Applications

Table 2 details the diagnostic accuracy, predictive capabilities, and structural validation constraints of AI implementations across the reviewed literature

[Insert Table-2. Here]

Across the eight studies summarize in Table 2, artificial intelligence (AI) performance varied by application, input modality, validation approach, and outcome definition. For postoperative surgical site infection (SSI) detection, Fletcher, Olubeko [9] reported questionnaire-based areas under the receiver operating characteristic curve (AUCs) of 0.853–0.867, with sensitivities of 70.6–75.0% and specificities of 98.5–100%, while image-based classification achieved an internally reported AUC of 1.00. However, subsequent red–green–blue (RGB) image classification using convolutional neural networks (CNNs) showed lower discrimination (AUC 0.639–0.655), with ResNet50 achieving 92–93% sensitivity but only 18% specificity[7]. Thermal imaging demonstrated stronger internal discrimination (AUC 0.844–0.900), although conflicting sensitivity and specificity estimates limited interpretation[10]. Alayande, Prasad [8] reported an AUC improvement from 0.64 to 0.86 following image-acquisition refinements, without independent validation. Wang [25] demonstrated smartphone-based deployment,

but classification performance varied by threshold, with sensitivity declining from 87.5% to 37.5% across tested thresholds in the imbalanced dataset, alongside device-dependent processing times.

For risk prediction, Larson, Roakes [22] reported an AUC of 0.74, precision of 0.69, and recall of 0.15 for composite surgical site complications. Economic simulation suggested cost-saving targeted single-use negative-pressure wound therapy for 61% of caesarean patients, without demonstrated clinical benefit. Lobel, Wagner [23] reported declining discrimination from training (AUC 0.78) to internal validation (0.72) and external testing (0.69). Targeting the highest two risk bands identified 18.1% of subsequent infections among 4.5% of women. For clinical-note surveillance, Satheakeerthy, Stretton [24] reported 100% sensitivity, 85.7% specificity, and an AUC of 0.91 in a 28-patient case–control sample, without separate caesarean surgical site infection (cSSI) estimates. Overall, evidence was predominantly retrospective or internally validated, with heterogeneous populations, outcomes, and performance measures. Only one study reported independent external model testing Lobel, Wagner [23], while none demonstrated prospective improvements in maternal outcomes.

3.4 Implementation relevant evidence mapped to CFIR

[Insert Table-3. Here]

Table 3, shows summarised evidence from the eight studies retrospectively mapped to the five updated CFIR domains[15]. Within the Innovation domain, external validation showed modest discrimination for caesarean surgical site infection (cSSI) risk prediction (area under the receiver operating characteristic curve [AUC] 0.69), while thermal-image classification achieved an internally validated AUC of 0.90. Image standardisation, offline functionality, device compatibility, and equipment costs, including a US$200 thermal module, emerged as relevant design considerations[8, 10, 23, 25]. The Outer Setting highlighted connectivity limitations and travel-related barriers to home surveillance, supporting the need for offline functionality and accessible referral pathways[8, 25]. The Inner Setting identified electronic medical record availability, documentation quality, equipment compatibility, and integration into existing clinical workflows as prerequisites for deployment [22–24]. The Individuals domain highlighted community health workers' image-acquisition capabilities, clinicians' interpretation of model outputs, and timely clinical documentation, although these competencies were not formally evaluated[9, 24, 25]. Within the Implementation Process, image calibration, quality feedback, technical testing, and external model validation represented preparatory activities rather than demonstrated clinical implementation[8, 23, 25]. Collectively, the findings support prospective evaluation of complete surveillance pathways, including technical reliability, staff competency, alert review, referral completion, and timely clinical assessment. However, no included study established successful routine implementation or improved maternal outcomes

4. Discussion

This scoping review included eight publications (2019–2026): five from the Rwandan imaging programme, two US retrospective studies, and one Australian study. AI applications showed promising internal performance (thermal AUC 0.90), but limited external validation and CFIR-identified technical, workforce, and referral challenges constrained clinical translation [10, 15, 23].

4.1 Artificial intelligence applications for cSSI risk prediction, detection, and surveillance

The artificial intelligence (AI) applications reported included symptom screening, visible/thermal imaging, smartphone assessment, preoperative and discharge risk prediction, and clinical-note surveillance, using regression, support vector machines, convolutional neural networks, ResNet50, CatBoost, XGBoost, and Llama [9, 22–25]. The distinction between predictive, diagnostic, and surveillance applications is important because each supports a different clinical decision and requires different evidence of usefulness. Beyond caesarean care, van Boekel, van der Meijden [13] identified 85 machine-learning (ML) models from 24 studies used for postoperative surgical site infection (SSI) prediction. Furthermore, in this review, most ML models were most developed for abdominal surgeries including cesarean section. The most clinically relevant AI application for post-discharge cSSI care may be community-supported wound assessment, because it can connect women recovering at home with clinicians without relying exclusively on hospital return visits[7, 9]. In Rwanda, Kateera, Riviello [26] evaluated community health worker (CHW)-supported mobile-health follow-up, while Nkurunziza, Williams [6] demonstrated home-based wound photography transmitted to remote general practitioners for infection assessment. These studies provide evidence for community-based telemedicine delivery. The implication is that AI-assisted wound assessment should be evaluated as an enhancement to a CHW-to-clinician pathway, while preoperative prediction, discharge risk stratification, and clinical-note analysis should retain their distinct roles in prevention planning, additional follow-up, and documentation surveillance [7, 9, 27]

4.2 Performance, validation, and transportability of artificial intelligence applications

The performance of the included studies varied: the questionnaire area under the receiver operating characteristic curve (AUC) were 0.849–0.867, visible imaging 0.639–1.000, and thermal imaging 0.844–0.900; image standardization improved reported AUC from 0.64 to 0.86 [8–10]. Preoperative prediction achieved AUC 0.74 and 15% recall, with simulated cost-saving prevention for 61% of patients. Discharge prediction declined from AUC 0.78 to 0.69 externally, capturing 18.1% of infections in the two highest-risk groups [22, 23]. Clinical-note surveillance achieved 100% sensitivity and 85.7% specificity in 28 patients [24]. Evidence was predominantly retrospective or internally validated, with one external validation and no demonstrated maternal outcome improvements.

The principal methodological concern is whether reported performance reflects genuine predictive ability in new populations rather than favourable results within development datasets. A systematic review by van Boekel, van der Meijden [13], found that 81% of models underwent internal validation only, calibration was reported in just two studies, and overall risk of bias was high or unclear; importantly, machine-learning methods did not consistently outperform regression-based approaches.

Similarly, a 2025 systematic review of spinal surgery SSI prediction found that 77.8% of machine-learning studies used internal cross-validation, only 33.3% reported calibration information, and none underwent external validation[14]. These findings reinforce the concern that excellent internal discrimination cannot establish transportability, particularly when populations, infection prevalence, outcome definitions, and assessment procedures differ. Clinical usefulness also depends on the consequences of incorrect classifications, not simply the AUC or overall accuracy. The Rwandan prospective telemedicine study by Nkurunziza, Williams [6] achieved 97.6% specificity but only 36.8% sensitivity for physician interpretation of transmitted wound photographs, demonstrating that a feasible remote assessment pathway may still miss substantial numbers of infections.

This supports prioritizing clinically appropriate thresholds, independent validation, and assessment of false-negative and false-positive consequences before integrating AI outputs into referral decisions. TRIPOD + AI

provides standards for transparent model reporting, while PROBAST + AI supports appraisal of methodological quality, risk of bias, and applicability; together, they offer a basis for assessing whether cSSI models are sufficiently reliable for their intended populations and decision points [28, 29].

4.3 Clinical translation of artificial intelligence for cSSI surveillance using CFIR

The Consolidated Framework for Implementation Research (CFIR) mapping identified image quality, device compatibility, connectivity, documentation, workforce capability, and referral processes as untested implementation considerations [15, 25]. The updated CFIR provides a useful distinction between the technical characteristics of AI and the conditions required for its delivery within maternal health services. Within the Innovation domain, standardized photography, image-quality feedback, offline inference, and device compatibility are relevant because predictions may change when image capture, pre-processing, or software configurations differ [8, 25]. The Outer Setting is particularly consequential in rural maternity care, where connectivity and transport barriers may interrupt the pathway between detecting a suspected infection and obtaining treatment [8, 25]. Supporting evidence from a prospective Rwandan telemedicine study showed that 719 of 787 women (91.4%) were reached at home, but only 623 (79.0% of all enrolled women) completed the process with a clinician's diagnosis returned within one hour; diagnostic sensitivity was 36.8%, despite specificity of 97.6% [6]. Although this intervention did not use autonomous AI, it demonstrates that successful image collection, transmission, and clinical diagnosis are separate stages, each requiring evaluation before AI can be integrated into post-discharge surveillance.

The Inner Setting and Individuals domains further show why model deployment cannot be separated from existing clinical services [15]. Electronic medical record-based prediction and language-model surveillance require complete, timely clinical information, whereas home wound assessment requires compatible devices, trained CHWs, and clinicians available to review suspected infections [23–25]. A qualitative Rwandan study involving 14 women and 12 CHWs found that photo-based postoperative telemedicine was acceptable to participants, with perceived improvements in access and reduced costs, while trust in CHWs supported participation and concerns remained about the reliability of photographic diagnosis [30]. These findings support a CHW-mediated approach but do not establish patient acceptance of AI-generated decisions or the competence of staff to act on them.

Within the Implementation Process, image calibration, platform adaptation, and retrospective validation represent preparation rather than evidence of sustained clinical implementation [23–25]. DECIDE-AI emphasizes early live evaluation of clinical AI systems, including safety, human factors, workflow integration, and their interaction with healthcare professionals [27]. Accordingly, the next stages for cSSI surveillance should test a fixed model–device–threshold combination within a supervised CHW-to-clinician pathway, measuring usable image capture, alert acknowledgement, confirmatory examination, referral completion, missed infections, clinician workload, and time to treatment [6, 27]. These outcomes would allow CFIR constructs to be linked to observed service delivery and distinguish technical feasibility from clinical effectiveness.

4.4 Strength and limitations

This scoping review integrates evidence on artificial intelligence (AI) applications for caesarean surgical site infection (cSSI) prediction, detection, and surveillance, examining model performance, validation, and clinical translation using the updated Consolidated Framework for Implementation Research (CFIR). Including technical development reports alongside clinical studies captured emerging applications and practical deployment

challenges. However, only eight publications were included, with overlapping Rwandan datasets limiting independent evidence. Heterogeneous outcomes, small infection samples, inconsistent performance reporting, and predominantly internal validation prevented meaningful comparisons or quantitative pooling. Some studies assessed composite surgical complications or mixed surgical populations rather than cSSI exclusively. CFIR mapping was retrospective and relied on reported information, limiting assessment of actual implementation determinants. No included study demonstrated improved maternal outcomes following AI implementation.

5. Conclusion

AI shows potential for cSSI risk prediction, wound infection detection, and clinical-note surveillance, but promising model performance has not translated into demonstrated clinical benefit. Limited external validation, inconsistent diagnostic performance, and untested implementation conditions restrict clinical readiness. Future research should evaluate externally validated, community health worker-supported and clinician-linked surveillance pathways, assessing diagnostic safety, workforce capacity, referral completion, treatment timeliness, and maternal outcomes, particularly in resource-limited settings.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

The authors declare no funding was received

Author Contribution

MAH, Proposed the entire study and contributed to study registration, identifying, selecting relevant studies, extraction, reporting the results; drafted the initial manuscript; and took part in critical revisions as first reviewer. NJ played the second reviewer role; participated in screening, data extraction and supported data analysis with CTP, who doubled as consultant to resolve discrepancies. CTP, BW and LDS approved review protocol and supervised the entire study, guided on the methodology, analysis, critically reviewed the manuscript

Availability of data and materials

All data generated or analysed during this study are included in this published article

References

1. Zuarez-Easton S, et al. Postcesarean wound infection: prevalence, impact, prevention, and management challenges. Int J Womens Health. 2017;9:81–8. https://doi.org/10.2147/ijwh.S98876.
2. Kawakita T, Landy HJ. Surgical site infections after cesarean delivery: epidemiology, prevention and treatment. Matern Health Neonatol Perinatol. 2017;3:12. https://doi.org/10.1186/s40748-017-0051-3.
3. Mojtahedi MF, et al. Global incidence of surgical site infections following caesarean section: a systematic review and meta-analysis. J Hosp Infect. 2023;139:82–92. https://doi.org/10.1016/j.jhin.2023.05.019.
4. Woelber E, et al. Proportion of surgical site infections occurring after hospital discharge: a systematic review. Surg Infect. 2016;17(5):510–9. https://doi.org/10.1089/sur.2015.241.
5. Niyigena A, et al. The true costs of cesarean delivery for patients in rural Rwanda: Accounting for post-discharge expenses in estimated health expenditures. Int J Equity Health. 2022;21(1):62. https://doi.org/10.1186/s12939-022-01664-x.
6. Nkurunziza T, et al. mHealth-community health worker telemedicine intervention for surgical site infection diagnosis: a prospective study among women delivering via caesarean section in rural Rwanda. BMJ GLOBAL HEALTH. 2022;7(7). https://doi.org/10.1136/bmjgh-2022-009365. Article e009365.
7. Fletcher RR et al. *Use of Convolutional Neural Nets and Transfer Learning for Prediction of Surgical Site Infection from Color Images*. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2021. Institute of Electrical and Electronics Engineers Inc.
8. Alayande BT, et al. Image-based surgical site infection algorithms to support home-based post-cesarean monitoring: Lessons from Rwanda. PLOS global public health. 2023;3(2):e0001584.
9. Fletcher RR, et al. Application of Machine Learning to Prediction of Surgical Site Infection. Annu Int Conf IEEE Eng Med Biol Soc. 2019;2019:2234–7. https://doi.org/10.1109/embc.2019.8857942.
10. Fletcher RR et al. *The Use of Mobile Thermal Imaging and Deep Learning for Prediction of Surgical Site Infection*. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2021. Institute of Electrical and Electronics Engineers Inc.
11. WHO, Ethics and governance of artificial intelligence for health. 2021: World Health Organization (WHO). Available from: https://www.who.int/publications/i/item/9789240029200
12. Gama F, et al. Implementation Frameworks for Artificial Intelligence Translation Into Health Care Practice: Scoping Review. J Med Internet Res. 2022;24(1):e32215. https://doi.org/10.2196/32215.
13. van Boekel AM, et al. Systematic evaluation of machine learning models for postoperative surgical site infection prediction. PLoS ONE. 2024;19(12):e0312968. https://doi.org/10.1371/journal.pone.0312968.
14. Ndjonko LC, et al. Evaluating predictive performance and generalizability of traditional and artificial intelligence models in predicting surgical site infections postspinal surgery: a systematic review. Spine J. 2026;26(2):280–91. https://doi.org/10.1016/j.spinee.2025.07.032.
15. Damschroder LJ, et al. The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci. 2022;17(1):75. https://doi.org/10.1186/s13012-022-01245-0.
16. Peters MD, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Implement. 2021;19(1):3–10. https://doi.org/10.1097/XEB.0000000000000277.

17. Tricco AC, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467–73. https://doi.org/10.7326/M18-0850.
18. Maiga AH, Translation C et al. https://doi.org/10.17605/OSF.IO/JCXD2
19. Pollock D, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid synthesis. 2023;21(3):520–32. https://doi.org/10.11124/JBIES-22-00123.
20. Reardon CM, et al. The Consolidated Framework for Implementation Research (CFIR) User Guide: a five-step guide for conducting implementation research using the framework. Implement Sci. 2025;20(1):39. https://doi.org/10.1186/s13012-025-01450-7.
21. Zhao L et al. *A Precision Early Warning Model for Fat Liquefaction After Cesarean Section Based on Supervised Learning*. in. 2025 *7th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*. 2025.
22. Larson BJ, et al. Precision in Prevention: Tailoring Single-Use Negative Pressure Wound Therapy Utilization Through Artificial Intelligence-Based Surgical Site Complications Risk and Cost Modeling [Article]. Surg Infect. 2024;25(4):315–21. https://doi.org/10.1089/sur.2023.274.
23. Lobel MM, et al. An externally validated machine learning model to predict cesarean surgical site infections. Pregnancy (Hoboken). 2026;2(1):e70231. https://doi.org/10.1002/pmf2.70231.
24. Satheakeerthy S, et al. Zero-shot large language model application for surgical site infection auditing [Article]. Infect Disease Health. 2025;30(4):337–42. https://doi.org/10.1016/j.idh.2025.05.001.
25. Wang L. Integrated Smartphone-Based Computer Vision and Machine Learning Platform for Identification of Surgical Site Infections. Massachusetts Institute of Technology; 2022.
26. Kateera F, et al. The Effect and Feasibility of mHealth-Supported Surgical Site Infection Diagnosis by Community Health Workers After Cesarean Section in Rural Rwanda: randomized Controlled Trial [Journal article]. JMIR mHealth uHealth. 2022;10(6):e35155. https://doi.org/10.2196/35155.
27. Vasey B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924–33. https://doi.org/10.1038/s41591-022-01772-9.
28. Moons KG et al. (2025). PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *bmj*, *388*. https://doi.org/10.1136/bmj-2024-082505
29. Collins GS et al. (2024). TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, *385*. https://doi.org/10.1136/bmj-2023-078378
30. Bikorimana L, et al. Acceptability of telemedicine for early surgical site infection diagnosis after cesarean delivery in rural Rwanda: a qualitative study. Maternal Health Neonatology Perinatol. 2025;11(1):3. https://doi.org/10.1186/s40748-024-00200-9.

Tables

Table 1. Summary of characteristics, methods, and clinical applications for the eight included studies

Source	Setting	Design	Population and data	AI method	Outcome and application
Fletcher et al. (2019)[9]	Rwanda; Kirehe District Hospital	Model development; 100 random 75%/25% training/test splits	Follow-up: 572/729 eligible women at day 10 ± 3. Questionnaire: 572 women/61 infections; image analysis: 568/62.	Questionnaire L1/L2 logistic regression and SVM; engineered colour/texture features with L1 selection for image classification.	Classify SSI present at postoperative review using symptoms/signs and wound photographs; independent general-practitioner examination supplies labels.
Fletcher et al. (2021)[7]	Rwanda; Kirehe District Hospital	Colour-image CNN development with internal cross-validation	572 women/images; 62 infected and 510 non-infected; day 10 ± 3. Same recruitment period as the 2019 report.	Naïve CNN and pretrained ResNet50; augmentation and minority-class weighting; logistic-regression comparator.	Classify current wound infection from CHW-acquired RGB photographs against independent clinical examination; screening sensitivity prioritised.
Fletcher et al. (2021)[10]	Rwanda; Kirehe District Hospital	Prospective enrolment; thermal CNN development and internal cross-validation	530 women; 30 infected and 500 non-infected wounds; postoperative day 11 ± 3.	Naïve CNN and ResNet50 transfer learning; thermal temperature arrays, augmentation and class weighting.	Classify current SSI using phone-attached thermal imaging and general-practitioner reference assessment.
Alayande et al. (2023)[8]	Rwanda programme	Opinion and development lessons	No new independently recruited cohort.	Programme-level synthesis of visible/thermal image learning, computer-vision alignment, cropping and colour calibration.	Describe acquisition, labelling and generalisability requirements for CHW-supported home monitoring; quantitative examples summarise earlier work.
Wang (2022) [25]	MIT development; Rwanda data and intended CHW deployment	Master's thesis; platform development and technical testing	Stored-image tests: 32 infected + 32 non-infected; 32 + 550. Three Android phones; unique women and test-set overlap unspecified.	Existing ResNet50 model converted to TensorFlow Lite; computer-vision capture, quality checks, colour calibration; local/cloud inference.	Integrate standardised photography and SSI classification in Wound Screener; evaluate device runtime, pre-processing consistency and classification delivery.
Larson et al. (2024)[22]	United States; Premier	Retrospective specialty-specific	Caesarean training cohort: 937,875 (source	CatBoost selected after comparing	Predict preoperative composite SSC risk and simulate

Source	Setting	Design	Population and data	AI method	Outcome and application
	Healthcare Database, 2017–2021	models and economic simulation	Table 1); economic test cohort: 165,290. Broader study includes other procedures.	tabular classifiers; predicted risk combined with prevention efficacy and cost inputs.	targeted sNPWT allocation. Endpoint includes SSI, dehiscence and haematoma/seroma.
Lobel et al. (2026)[23]	United States; two level IV academic centres	Retrospective development, internal validation and independent external testing	5,272 deliveries/181 cSSIs: development 2,336/109; external cohort 2,936/72.	XGBoost selected over random forest and regularised logistic regression; 15 discharge-available variables; Platt scaling.	Estimate subsequent cSSI risk at discharge and stratify follow-up. Outcome windows: six weeks in development and 30 days externally.
Satheakeerthy et al. (2025) [24]	Australia; two metropolitan hospitals	Retrospective case–control clinical-note audit	28 patients: 14 SSI/14 controls; 12/14 SSI cases followed caesarean birth. 379 case notes and 304 control notes.	Zero-shot, locally deployable 70-billion-parameter Llama model; criteria-based prompting of individual notes.	Identify documented SSI for audit; any positive note classifies the patient. Nurse labels confirmed by medical review; no separate cSSI accuracy.

AI = artificial intelligence; CHW = Community health worker; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; L1/L2 = Regularisation penalties; MIT = Massachusetts Institute of Technology; RGB = Red green blue; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy.

Table 2. Application-specific performance, validation limits, and clinical interpretation

Application	Inputs user and timing	Reported performance	Validation limits	Clinical interpretation
Symptom and image assessment [9]	Community health worker (CHW)-collected symptom questionnaires and red–green–blue (RGB) wound photographs at postoperative day 10 ± 3; physician-confirmed infection.	Questionnaire: Area under the receiver operating characteristic curve (AUC) 0.849–0.867; sensitivity 70.6–75.0%; specificity 98.5–100%; accuracy 95.8–96.5%. Image model: AUC 1.00; sensitivity and specificity 100%; accuracy 100%. Combined assessment: sensitivity and specificity 99%.	Repeated internal training–test splits without external validation. Differences in sample denominators and inconsistencies between reported image-model results.	Questionnaire AUC ≈0.85–0.87 indicates reasonably strong ranking within these splits, but sensitivity ≈71–75% corresponds to missing roughly 25–29 of every 100 infected women at those operating points. Near-perfect image separation is an internal, inconsistently reported result; it is not a dependable estimate for home screening.
Current wound classification [7]	CHW-acquired Red-Green-Blue images at day 10 ± 3; Convolutional neural network (CNN), ResNet50, and logistic regression compared against clinical examination.	Naïve CNN: AUC 0.655; sensitivity 75%; specificity 58%; Matthews correlation coefficient (MCC) 0.35. ResNet50: AUC 0.639; sensitivity 92–93%; specificity 18%; MCC 0.41. Logistic regression: AUC approximately 0.563.	Internal cross-validation involving 62 infected wounds; no external validation. Inconsistent specificity labels and performance summaries.	CNN discrimination was limited. Higher ResNet50 sensitivity was accompanied by approximately 82% false-positive classification among non-infected
Current wound classification [10]	Phone-attached thermal camera; postoperative day 11 ± 3; naïve CNN/ResNet50; doctor-assessed SSI.	Naïve CNN: median AUC 0.844 (IQR 0.34), sensitivity 71%, specificity 87%, MCC 0.37; maximum accuracy 86%. Transfer CNN: median AUC 0.900 (IQR 0.19), MCC 0.44; maximum accuracy 94%. Sensitivity/specificity reported as 95%/84% in the narrative but reversed in table.	Internal evaluation; five-fold transfer-model validation; 30 infected and 500 non-infected wounds. 95% CIs not reported. Conflicting sensitivity and specificity estimate for the transfer model.	Thermal imaging showed promising discrimination, but inconsistent operating-point estimates and limited infection cases restrict clinical interpretation
Programme lessons [8]	Visible/thermal imaging; computer-vision alignment, cropping and colour calibration for proposed	Visible-image classification: sensitivity improved from 81.3% to 97%, specificity from 65.3% to 87%, and AUC from 0.64 to 0.86 following workflow	Opinion article; no new independent cohort, denominators or paired comparative analysis. Calibration estimates must not	Standardized image acquisition may improve classification performance, but clinical effectiveness and

	home surveillance.	refinements. Thermal performance labels were inconsistent.	be counted as an additional validation.	reproducibility remain unverified.
Technical delivery [25]	Android-based RGB wound photography, image-quality feedback, and ResNet50 classification using TensorFlow Lite or cloud inference.	Balanced dataset: sensitivity 53.1–93.8%; specificity 6.3–46.9%; accuracy 48.4–54.7%. Imbalanced dataset: sensitivity 37.5–87.5%; specificity 13.3–76.4%; accuracy 17.4–74.2%. Device processing: pre-processing 428–4,755 milliseconds; inference 317–1,504 milliseconds.	Stored-image technical testing without prospective clinical validation. Unspecified patient overlap, inconsistent platform outputs, and differences between tested and implemented thresholds. Android/Python outputs disagreed.	Model performance was highly threshold-dependent. The highest reported accuracy in the imbalanced dataset coincided with missing 62.5% of infected images, demonstrating why accuracy alone is insufficient. Device-dependent processing and inconsistent model outputs require resolution before clinic
Preoperative prevention support [22]	Preoperative demographic and claims-derived variables; CatBoost prediction of surgical site complications and simulated single-use negative-pressure wound therapy (sNPWT) allocation.	Composite complication prediction: incidence 1.7%; AUC 0.74; precision 69%; recall 15%; F1 score 0.24; F2 score 0.18. Economic simulation: sNPWT allocation was modelled as cost-saving for 61% of 165,290 caesarean patients.	Retrospective internal validation using a composite outcome; no infection-specific performance, external testing, or prospective evaluation of prevention benefits.	Moderate discrimination but low recall, with 85% of complications missed at the reported threshold. Simulated cost savings do not establish clinical effectiveness.
Discharge risk stratification [23]	Fifteen electronic medical record variables available at discharge; extreme gradient boosting (XGBoost) with Platt calibration for subsequent caesarean surgical site infection (cSSI).	AUC: training 0.78 (95% confidence interval [CI] 0.73–0.82); internal validation 0.72 (95% CI 0.66–0.76); external testing 0.69 (95% CI 0.63–0.75). Risk stratification: highest-risk 0.7% of women had 22.7% cSSI incidence. The two highest bands included 4.5% of women and captured 18.1% of infections.	Independent external validation, but both cohorts came from specialized academic centres. Different outcome windows and infection prevalence; no prospective clinical impact	External discrimination was modest. Risk-targeted follow-up could prioritize high-risk women but would miss approximately 82% of infections if restricted to the two highest bands.

Clinical-note surveillance [24]	Inpatient and outpatient clinical notes analysed using a zero-shot 70-billion-parameter Llama model; nurse assessments confirmed by medical review.	Patient-level: sensitivity 100%; specificity 85.7%; accuracy 92.9%; F1 score 93%; AUC 0.91 (95% CI 0.73–1.00). Detection timing: AI flagged seven infections before, six on, and one after documented onset; median difference −0.5 days.	Retrospective case–control study involving 28 patients, including 14 infections. No separate cSSI estimates, prospective validation, or confirmed improvement in clinical detection time.	All sampled infections were identified, with two false positives. Small sample size and inclusion of non-caesarean cases limit generalisability and clinical readiness.

AUC = Area under the receiver operating characteristic curve; CHW = Community health worker; CI = Confidence interval; CNN = Convolutional neural network; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; F1/F2 = Precision–recall harmonic scores (F2 gives more weight to recall); IQR = Interquartile range; L1/L2 = Regularisation penalties; MCC = Matthews correlation coefficient; Ms = milliseconds; RF = Random forest; RGB = Red, green blue; RLR = Regularised logistic regression; ROC = Receiver operating characteristic; SSC = Surgical-site complication; SSI = Surgical site infection; SVM = Support vector machine; sNPWT = Single-use negative-pressure wound therapy;

Table 3. CFIR domain definitions, evidence mapping, and implications for clinical translation

CFIR domain	Operational definition	Evidence synthesis and supported constructs	Clinical translation
Innovation	Features of the AI-supported surveillance component itself that affect its use: evidentiary support, adaptability, design, complexity and cost. The component may be a wound classifier, risk model, capture platform or note-audit tool.	Innovation Evidence Base: independent cSSI risk-model testing yielded AUC 0.69, while thermal evidence remained internal and small [10, 23]. Innovation Design and Adaptability: capture targets, automatic cropping, colour calibration and recapture prompts replaced variable free-form images; offline/cloud options changed delivery conditions [8, 25]. Innovation Complexity and Cost: model conversion, phone-dependent runtime and a US$200 thermal module introduce device and maintenance requirements [10, 25].	Validate the exact model–pre-processing–device combination intended for service use, including poor-quality-image handling and threshold behaviour. Treat external risk discrimination as support for testing targeted contact, not excluding lower-risk women from care. Thermal equipment procurement requires a locally tested diagnostic pathway, rather than relying on the internally reported AUC alone [23, 25].
Outer setting	Community and environmental conditions outside the delivering maternity/community/infection-prevention service that influence women's ability to participate and the technology's ability to function.	Local Conditions: rural internet limitations restricted cloud inference; offline use shifts dependence towards phone capability and affordability (Wang, 2022). Physical and financial burdens of returning to facilities motivate home monitoring in the Rwanda programme (Alayande et al., 2023). These are author-described contextual conditions, not participant-derived estimates of AI acceptability.	For a home-monitoring pilot, test offline capture/inference and later synchronisation under actual network conditions. Measure whether flagged women can reach assessment, including travel time and out-of-pocket cost; otherwise, improved classification may not yield completed care. Maintain a contact/referral route for women who cannot use the digital pathway (Alayande et al., 2023; Wang, 2022).
Inner setting	Infrastructure, resources and routines within the responsible service that enable capture, computation, documentation, review and clinical follow-up.	Structural Characteristics Information Technology Infrastructure: LLM surveillance depends on complete, timely EMR notes and can operate locally; discharge risk modelling requires	Check feature availability and note coverage at the intended decision time before integration. Assign responsibility for receiving image alerts or audit flags and confirming suspected infection. Test the complete software version on service phones and estimate review volume at local prevalence;

		available EMR variables [23, 24]. Available Resources Materials and Equipment: phone capability and software/library compatibility affect delivery [25]. Compatibility: preoperative allocation is designed to allow device preparation before closure, whereas discharge risk estimation and note auditing enter different existing workflows; these are proposed fits, not demonstrated adoption [22, 23].	a locally hosted model still requires a functioning documentation and response service [23-25].
Individual characteristics	Role-specific capability and opportunity of deliverers and recipients to perform required tasks: acquire usable images, interpret an output, document findings and obtain clinical assessment. This row corresponds to the updated CFIR Individuals domain.	Innovation Deliverers Capability: the thesis identifies limited CHW clinical/technical training as a design constraint and supplies immediate quality feedback; early symptom assessment depended on the administrator's clinical experience [9, 25]. Innovation Deliverers Opportunity: note-based surveillance depends on staff having the opportunity to create timely documentation, which the audit authors identify as a condition of performance [24]. These are author-described requirements rather than measured capability or opportunity scores.	Evaluate whether CHWs can capture a usable wound image and follow a recapture/escalation instruction during observed tasks. Test whether clinicians distinguish a risk score, current-wound classification and audit flag, and can correct false positives such as prophylaxis misclassification. These competency checks are proposed responses to the reported requirements; acceptance or confidence cannot be inferred from staff participation in data collection [9, 24, 25].
Implementation process	Preparatory or delivery activities used to assess context, adapt the intervention and examine how the innovation functions in its intended workflow; technical preparation is distinguished from live service implementation.	Assessing Context: programme and platform reports identify, capture variability, connectivity and device limitations [8, 25]. Tailoring Strategies: targets, automatic calibration, image-quality	First resolve software equivalence and evaluate a fixed model/device/threshold prospectively without altering care. A subsequent supervised pilot can track alert acknowledgement, confirmatory review, referral completion, false alerts and missed infections. Compare

		feedback and local/cloud operation respond to these conditions [8, 25]. Reflecting and Evaluating Innovation: pre-processing, runtime and Android/Python comparisons identify unresolved delivery problems; external risk testing examines transportability [23, 25]. This is retrospective mapping of development activities, not evidence of a completed CFIR-guided implementation programme.	these process measures with time to assessment and treatment, so adaptation addresses the specific point at which the surveillance pathway fails [23-25].

AI = Artificial intelligence; AUC = Area under the receiver operating characteristic curve; CFIR = Consolidated Framework for Implementation Research; CHW = Community health worker; cSSI = Caesarean surgical site infection; EMR = Electronic medical record; LLM = Large language model; US$ = United States dollars.

Figures

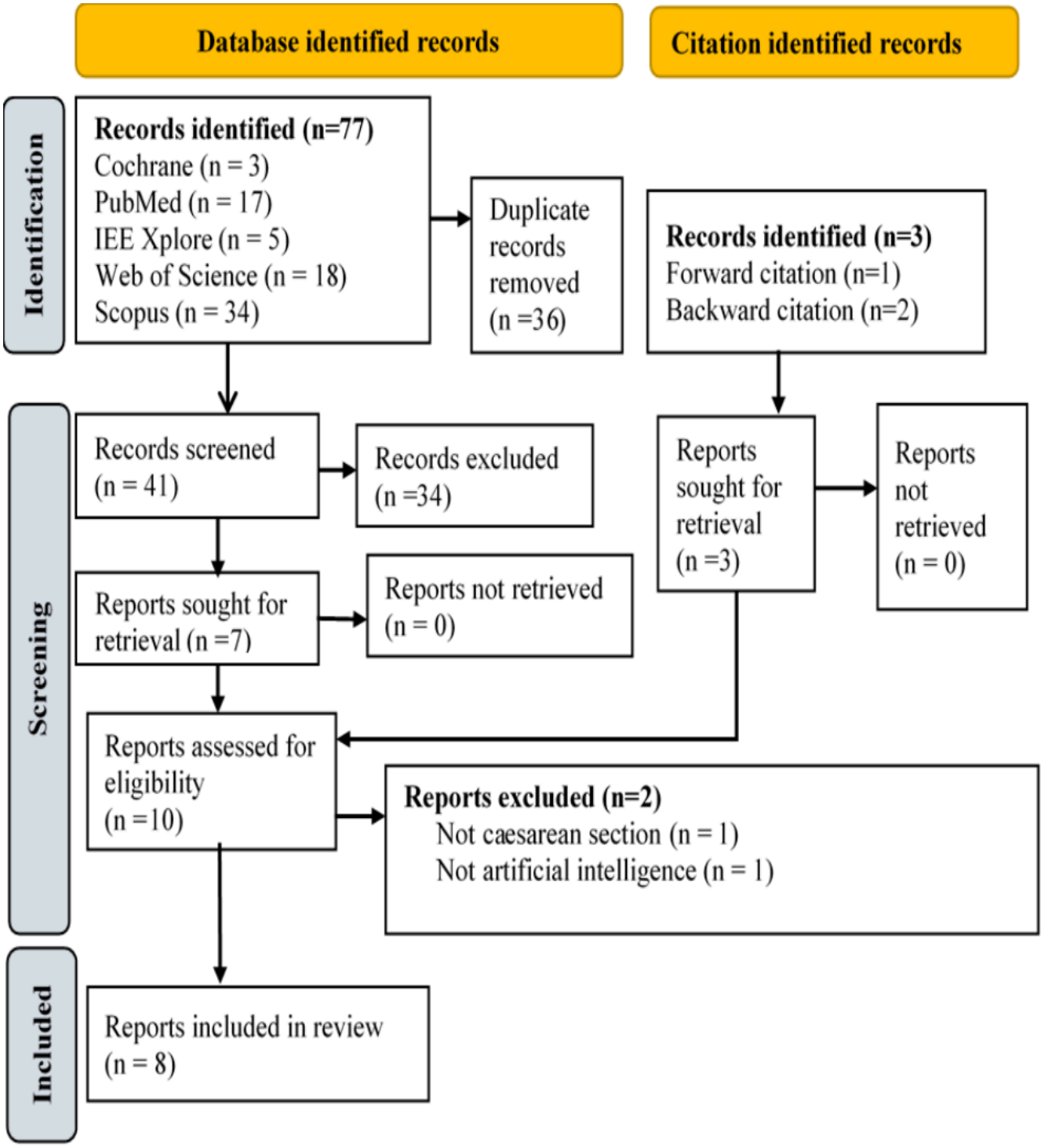

Figure 1. PRISMA flow diagram of study selection process

Figure 1

See image above for figure legend.