JOURNAL OF THE EUROPEAN SOCIETY OF SURGERY
04/01/2026

ARTIFICIAL INTELLIGENCE IN HEALTHCARE WITH EMPHASIS ON SURGICAL DISCIPLINES: A MIXED-METHODS EVALUATION OF IMPLEMENTATION AND OUTCOMES

Vinod Kumar Singhal[1], Adil Mohammed Suleman[2], Faris Dawood Alaswad[3], Mr. Hemant. Sharma[4], Abhishek Panikar[5], Dr Ahmad Mokhtar Mohamed Ziad Rahima[6], Vidher V V Singhal[7], Nadim Mohammad Biswa[8], Minisha Singhal[9]

[1] Consultant General Surgeon, Prime Hospital, Dubai, UAE, drvinodsinghal@yahoo.co.in
[2] Specialist General Surgeon, Prime Hospital, Dubai, UAE, dradil.mohammed@primehospital.ae
[3] Consultant General Surgeon, Gladstone Hospital, Perth, Australia, faris_alaswad@yahoo.com
[4] MD MS FACS, FEBS, Diplomate Multi-Organ Transplant & HPB Surgery, (American Society of Transplant Surgeons), Transplant and Vascular Access Surgeon Royal Liverpool University Hospital.
[5] General Surgery Resident (R1), Email abhishek.panikar@gmail.com
[6] General Surgery Resident, Prime Hospital, Mokhtarrahima@gmail.com
[7] Postgraduate UCL, UK, vidher947@gmail.com
[8] Research Associate, Dhaka, Bangladesh
[9] BTech AI, BIT, Ranchi, India

Abstract

Background: Surgical outcomes depend on reliable perioperative coordination as much as operative skill, yet complications and inefficiencies remain common. In the Gulf region, high cardiometabolic burden, including diabetes, increases perioperative risk and strains discharge and infection prevention. AI tools may improve risk stratification and workflow, but real-world evidence on safe, effective implementation is limited, making implementation-focused mixed-methods evaluation essential.

Aim of the study: This study aimed to assess adoption, governance and safety performance, and early clinical and operational effects of AI-enabled perioperative decision-support and workflow tools across multi-centre Gulf-region surgical services.

Methods: A convergent mixed-methods, multi-centre pre-post evaluation (2024–2025) was conducted across three Gulf tertiary surgical hospitals, analyzing 320 consecutive adult surgical cases (pre n=160, post n=160) and 30 perioperative stakeholder interviews; AI tool use, staff training and 1–5 Likert perception domains, clinical outcomes (operative time, length of stay, complications, ICU admission, 30-day readmission, mortality), operational metrics (cancellations, turnover, time-to-incision, radiology and discharge turnaround, antibiotic stewardship, documentation burden), and AI-related safety events were extracted from routine systems and compared pre vs post using SPSS, while interviews were thematically analyzed in NVivo and integrated with quantitative findings via joint displays.

Results: Across 320 surgical cases (pre-160, post 160) across three Gulf tertiary centers, cohorts were similar overall, but emergency cases increased post-implementation (2.5% to 14.4%, $p=0.032$). AI-supported cases rose from 43.8% to 74.4% ($p<0.001$), with greater use in pre-op planning (17.5% to 32.5%, $p<0.001$), and clinician override rates among AI-exposed cases remained stable. Staff AI training increased markedly (31.2% to 76.9%), and all perception domains improved significantly. Length of stay decreased (4.57 to 3.62 days, $p=0.002$) and ICU admissions fell (16.2% to 6.2%, $p=0.008$), while operative time, complications, readmissions, and mortality were broadly unchanged. Service metrics improved, and 28 mostly minor AI-related safety events were reported, with declining rates over time; interviews (n=30) highlighted workflow fit, alert fatigue, governance visibility, training, and data integration as key drivers.

Conclusion: AI rollout increased uptake and improved staff readiness and operational performance, with lower length of stay and ICU admissions despite more emergency cases, while other clinical outcomes remained stable. Safe scaling requires strong governance, workflow fit, training, reliable data integration, alert control, and equity monitoring.

Keywords: Artificial intelligence, Perioperative decision support, Surgical workflow optimization, and Implementation science

JESS

RESEARCH ARTICLE

Volume *1*

Issue *1*

****Corresponding author***

Vinod Kumar Singhal

Submitted *05 March 2026*

Accepted *26 March 2026*

Published *01 April 2026*

Citation
Vinod Kumar Singhal. Et.al *Artificial Intelligence in Healthcare with Emphasis on Surgical Disciplines, A Mixed-Methods Evaluation of Implementation and Outcomes, Vol 1: Iss 1, 2026.*

The article was originally submitted in ENGLISH. Translations into other languages were reviewed and validated by the authors and the editorial team. Nevertheless, for the most accurate representation of the subject matter, readers are encouraged to consult the article in its original language.

Introduction

Surgical care is high-volume and high-risk; outcomes depend not only on operative skill but also on perioperative coordination, information flow, and team decisions across emergency and elective pathways. Despite advances, major gaps in access, timeliness, and safety still drive preventable morbidity, disability, and economic loss worldwide, reinforcing the Lancet Commission's call for surgical system strengthening as a development priority [1]. Even in well-resourced settings, postoperative complications and mortality vary widely between institutions and countries, underscoring that system reliability and care processes are as important as individual performance [2].

At the same time, the surgical case-mix is shifting toward higher baseline risk due to multimorbidity, aging, and the growing burden of non-communicable disease. In the Gulf region, cardiometabolic risk is especially prominent; high diabetes prevalence across GCC countries increases perioperative vulnerability and complicates discharge planning, readmissions, and infection prevention [3]. Surgical safety challenges such as healthcare-associated infections remain consequential, and pooled GCC estimates for surgical site infection suggest a substantial, procedure-dependent burden that warrants stronger surveillance and prevention tailored to local case-mix [4]. Together, these pressures require perioperative services to improve safety and operational performance, including theatre efficiency, timely diagnostics, and complete documentation, while maintaining governance standards for privacy and accountability.

Artificial intelligence, including machine learning and natural language processing, has been positioned as a pragmatic lever to address these dual pressures by augmenting clinical and operational decision-making across the surgical pathway. Contemporary surgical AI use cases increasingly extend beyond single-task diagnostics toward perioperative risk stratification, triage signaling for urgent pathways, prediction of complications, and workflow optimization, with the explicit intent to support clinician decisions rather than automate them [5]. However, the translational evidence base remains uneven. Systematic syntheses of AI in surgery highlight that many models are developed retrospectively, validated inconsistently, and reported with variable transparency, with limited emphasis on external validity, dataset representativeness, and real-world safety monitoring, all of which are prerequisites for trustworthy deployment [6]. Operational applications illustrate both promise and the evidence gap: for example, machine learning prediction of surgical case duration can outperform traditional scheduling heuristics in some contexts, yet reported efficiency gains and implementation impact are not consistently demonstrated, and context-sensitive workflow integration is often under-described [7].

This translational gap has prompted a shift in the field from "accuracy-first" narratives to implementation-aware evaluation that asks, does the tool fit the workflow, does it improve outcomes that matter, and what new risks does it introduce. Mixed-methods research has become increasingly important here because adoption, trust, and effective use are socially and operationally mediated, not purely technical. A large mixed-methods inventory of implementation barriers and strategies underscores recurring challenges, including workflow disruption, data quality and interoperability constraints, unclear accountability, alert burden, and limited organizational readiness, alongside actionable strategies such as local champions, training, and governance visibility [8]. Empirical implementation reviews using frameworks such as CFIR similarly show that successful deployment depends on sociotechnical alignment, including stakeholder engagement, iterative adaptation, and measurable enabling conditions, not simply model performance [9]. At the organizational level, systematic evidence from hospital-based implementations highlights the difficulty of moving from static, retrospective datasets to live clinical data streams, and emphasizes continuous monitoring, auditability, and learning health system principles as core requirements for sustainable AI use [10].

Accordingly, consensus guidance now emphasizes structured, stage-appropriate evaluation and transparent reporting. DECIDE-AI provides a reporting framework tailored to early-stage, real-world clinical evaluation of AI decision support, with explicit attention to human factors and safety considerations in live settings [11]. Complementary evaluation frameworks such as TEHAI propose multidomain assessment of capability, utility, and adoption, supporting appraisal across the AI lifecycle rather than at a single "go-live" point [12]. Broader consensus on trustworthy, deployable AI, articulated in the FUTURE-AI framework, further stresses fairness, universality, traceability, usability, robustness, and explainability as practical requirements for implementation and monitoring [13]. In parallel, reporting extensions such as CONSORT-AI and SPIRIT-AI clarify the minimum information needed to interpret trials and protocols involving AI interventions, particularly around human-AI interaction, input and output handling, and error analysis, reducing ambiguity for clinicians, reviewers, and regulators [14,15].

Despite these advances, there remains a clear gap in multi-center, implementation-focused surgical AI evaluations that jointly quantify early clinical and operational effects while explaining mechanisms of uptake, overrides, and safety signals in context, especially in Gulf-region surgical systems where rapid digital transformation, diverse workforces, and heterogeneous patient populations may magnify both the benefits and the risks of AI-supported workflows. Therefore, the aim of this study was to evaluate the implementation, adoption mechanisms, safety governance performance, and early clinical and operational impacts of AI-enabled perioperative decision-support and workflow tools in surgical disciplines across multi-centre Gulf-region surgical services.

METHODOLOGY & MATERIALS

This mixed-methods implementation evaluation examined how artificial intelligence was introduced into surgical services in the Middle East, and its early effects on workflow, safety, and clinical outcomes. A convergent parallel design was used: quantitative routinely collected operational and clinical data were analyzed in parallel with qualitative semi-structured interviews of perioperative stakeholders, then integrated to support an implementation- focused interpretation. The study was a multi-center pre-post evaluation conducted across tertiary surgical facilities in the Gulf region, covering mixed elective and emergency pathways, over an approximately 12-month window (2024-2025), with a defined pre-implementation period and a post-implementation period after deployment and stabilization. During the study period, 320 consecutive surgical cases and 30 completed staff interviews were analyzed.

The AI intervention comprised AI-enabled perioperative tools embedded into routine workflows to support, not replace, clinician decision-making. Capabilities varied by service line and included perioperative risk stratification with decision-support prompts, triage or prioritization signals for acute pathways, documentation support, and operational analytics for theatre readiness and coordination. A human-in-the-loop policy was applied throughout: AI outputs were advisory, clinicians retained full accountability for decisions, and overrides were permitted and logged. Governance measures included role-based access control, structured training with local champions, defined operational support and downtime escalation pathways, and periodic audits of safety signals, override patterns, and data-quality performance.

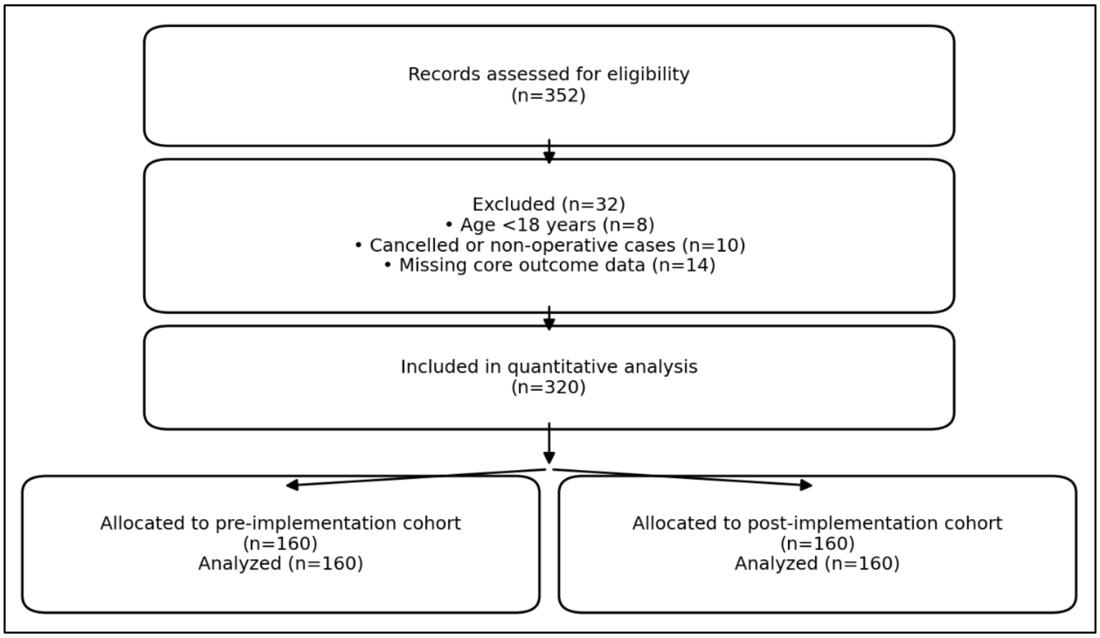


Figure 1. Quantitative sample flow diagram

Quantitative component, variables, and statistical analysis

The quantitative sample comprised 352 surgical records assessed for eligibility; 32 were excluded due to age <18 years (n=8), cancelled or non-operative encounters (n=10), or missing core outcome data (n=14), leaving 320 for analysis (Figure 1). Cases were evenly split into pre-implementation (n=160) and post-implementation (n=160) cohorts. AI utilization measures were derived from perioperative documentation and system use logs, and summarized as: AI-supported case status (yes, no), primary AI use point (pre-op planning, intra-op decision support, post-op monitoring, admin or scheduling, not used), and clinician action among AI-supported cases (followed, partially followed, overrode), reported as proportions within AI-exposed cases.

A brief structured staff response was collected for each included case (n=320), recording the respondent's role, years of experience, and whether they had completed formal AI training. Staff perceptions were rated on 1–5 Likert scales and summarized as mean ± SD across key implementation domains: perceived usefulness, ease of use, enabling conditions, trust and governance, privacy and ethics, intention to use, and perceived team impact. Quantitative outcomes included operative time, length of stay, postoperative complications, ICU admission, 30-day readmission, and in-hospital mortality. Key operational indicators were extracted from theatre and hospital systems and compared pre versus post, including cancellation rate, OR turnover time, time from decision to incision for urgent cases, radiology turnaround time, discharge summary completion time, antibiotic stewardship compliance, schedule overruns, and documentation burden. AI-related safety events were captured from incident and governance logs, classified by severity and type, and summarized as counts and as events per 100 AI-supported cases over baseline, 3 months, and 6 months.

Quantitative data were cleaned in spreadsheets and analyzed using SPSS (version 26.0). Categorical variables were summarized as n (%), continuous variables as mean ± SD or median (IQR) as appropriate; pre versus post comparisons used chi-square or Fisher's exact tests for categorical variables and t-tests or Mann–Whitney U tests for continuous variables. Where applicable, multivariable regression adjusted for case-mix, including site, age, urgency, and ASA class, reporting effect estimates with 95% confidence intervals; statistical significance was set at $p < 0.05$, two-sided.

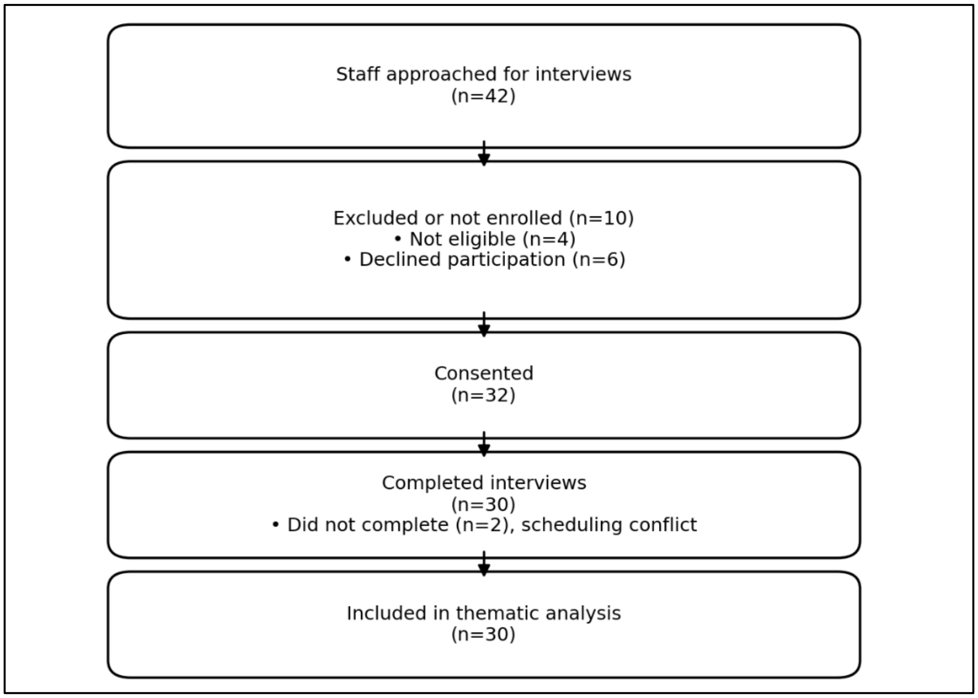


Figure 2. Qualitative interview sampling flow

Qualitative data collection and analysis

The qualitative component comprised semi-structured interviews with perioperative stakeholders involved in surgical pathways and AI implementation, including surgery, anesthesia, perioperative nursing, operating theatre management, quality and safety, informatics, and governance functions. Eligible participants were actively working in these pathways, had direct experience with AI outputs or with implementation, training, oversight, or safety monitoring, and could provide consent in English or Arabic, bilingual facilitation and translation were used when needed. Of 42 staff approached, 32 consented, 2 did not complete interviews due to scheduling conflicts, leaving 30 completed interviews included in thematic analysis. Interviews explored workflow fit, governance visibility and trust, training and local champions, alert burden and safety concerns, multilingual documentation issues, equity considerations in multinational populations, data quality and integration, perceived operational value, and accountability and override culture. Interviews were conducted in-person or virtually, audio-recorded when permitted, transcribed, de-identified, and securely stored. Transcripts were analyzed using implementation-oriented thematic analysis in NVivo, using a mixed deductive and inductive codebook; themes were synthesized as barriers and facilitators across individual, workflow, organizational, and system levels, and reported with theme frequencies as n (%) of interviews mentioning each theme.

Mixed-methods integration

Quantitative and qualitative findings were integrated using joint displays that aligned measured patterns, particularly AI uptake and staff perception shifts, clinical and operational outcomes, and safety event summaries, with stakeholder explanations, documenting convergence and complementarity. Integrated interpretation prioritized implementation mechanisms reflected in both datasets, especially workflow timing and fit, alert burden, governance visibility, multilingual documentation, and data quality and integration, as explanatory factors for observed adoption and outcome patterns.

Ethics, confidentiality, and data governance

Ethical approval was obtained from relevant institutional review and data governance bodies at participating Middle-East sites. Quantitative data were extracted and analyzed in de-identified form on secure institutional infrastructure with restricted access. Interview participants provided informed consent, transcripts were de-identified, and reporting avoided identifiable unit-level or individual-level attribution while maintaining transparency regarding implementation processes, safety monitoring, and governance practices.

RESULTS

The pre and post cohorts were broadly comparable, mean age 44.9 ± 12.9 vs 42.9 ± 13.9 years, BMI 28.0 ± 4.6 vs

28.5 ± 4.4 kg/m², with similar site distribution and gender mix, male 52.5% vs 51.9%. ASA class was also stable, class II was most common, 44.4% vs 45.0%. The main shift was case urgency, elective cases fell slightly, 69.4% to 66.2%, urgent cases decreased, 28.1% to 19.4%, while emergencies rose markedly, 2.5% to 14.4%, and this difference was statistically significant, p=0.032.

Table 1. Baseline characteristics of surgical cases (n=320)

Variable	**Category**	**Pre (n=160)**	**Post (n=160)**	**p-value**
		n (%)	**n (%)**	
Age (years), mean ± SD		44.9 ± 12.9	42.9 ± 13.9	0.175
BMI (kg/m²), mean ± SD		28.0 ± 4.6	28.5 ± 4.4	0.314
Site	Abu Dhabi tertiary center	69 (43.1%)	65 (40.6%)	0.304
	Dubai tertiary center	71 (44.4%)	65 (40.6%)	
	Sharjah tertiary center	20 (12.5%)	30 (18.8%)	
Gender	Male	84 (52.5%)	83 (51.9%)	1
Gender	Female	76 (47.5%)	77 (48.1%)	
	I	34 (21.2%)	34 (21.2%)	
	II	71 (44.4%)	72 (45.0%)	

ASA class	III	46 (28.8%)	50 (31.2%)	0.916
	IV–V	9 (5.6%)	4 (2.5%)	
Case urgency	Elective	111 (69.4%)	106 (66.2%)	0.032
	Urgent	45 (28.1%)	31 (19.4%)	
	Emergency	4 (2.5%)	23 (14.4%)	

Specialty case-mix did not differ significantly overall, p=0.234. General surgery and orthopedics remained the largest groups, 30.6% to 28.8% and 22.5% to 20.6%. The most noticeable directional change was urology increasing from 7.5% to 13.1%, while "other" declined modestly, 11.2% to 9.4%.

Table 2. Surgical case-mix by specialty (n=320)

Specialty	**Pre (n=160)**	**Post (n=160)**	**p-value**
	n (%)	**n (%)**	
General surgery	49 (30.6%)	46 (28.8%)	0.234
Orthopedics	36 (22.5%)	33 (20.6%)	
Urology	12 (7.5%)	21 (13.1%)	
Neurosurgery	16 (10.0%)	13 (8.1%)	
Cardiac/Thoracic	6 (3.8%)	9 (5.6%)	
ENT	10 (6.2%)	11 (6.9%)	
ObGyn	13 (8.1%)	12 (7.5%)	
Other	18 (11.2%)	15 (9.4%)	

AI-supported cases increased sharply from 43.8% (70/160) pre to 74.4% (119/160) post, p<0.001, with "not used" dropping from 56.2% to 25.6%. The primary use point shifted toward more perioperative embedding, pre-op planning rose from 17.5% to 32.5%, intra-op decision support from 7.5% to 11.9%, and post-op monitoring from 6.9% to 12.5%, all within an overall significant distribution change, p<0.001. Among AI-supported cases only, clinician adherence was similar, followed 68.6% to 73.9%, partially followed 15.7% to 15.1%, overrode 15.7% to 10.9%, with no significant difference, p=0.612.

Table 3. AI use patterns and clinician action (n=320)

Variable	**Category**	**Pre (n=160)**	**Post (n=160)**	**p-value**
		n (%)	**n (%)**	
AI-supported cases	Yes	70 (43.8%)	119 (74.4%)	<0.001
	No	90 (56.2%)	41 (25.6%)	
	Pre-op planning	28 (17.5%)	52 (32.5%)	<0.001
Primary AI use point	Intra-op decision support	12 (7.5%)	19 (11.9%)	
	Post-op monitoring	11 (6.9%)	20 (12.5%)	
	Admin/scheduling	19 (11.9%)	28 (17.5%)	
	Not used	90 (56.2%)	41 (25.6%)	
Clinician action	Followed (AI cases only)	48/70 (68.6%)	88/119 (73.9%)	0.612
	Partially followed (AI cases only)	11/70 (15.7%)	18/119 (15.1%)	
	Overrode (AI cases only)	11/70 (15.7%)	13/119 (10.9%)	

Respondents linked to cases had a mean of about 9 years in practice in both periods, 9.0 ± 5.9 pre vs 9.1 ± 6.1 post. Roles were balanced across time, surgeons were the largest group overall, 48.1%, followed by anesthesiologists, 20.0%, and OR nurses, 18.4%. The standout change was formal AI training, rising from 31.2% pre to 76.9% post, with "no training" dropping from 68.8% to 23.1%.

Table 4. Staff respondent profile linked to cases (n=320 responses)

Variable	Category	Overall (n=320)	Pre (n=160)	Post (n=160)
		n (%)	n (%)	n (%)
Years in practice, mean ± SD		9.0 ± 6.0	9.0 ± 5.9	9.1 ± 6.1
Role	Surgeon	154 (48.1%)	80 (50.0%)	74 (46.2%)
	Anesthesiologist	64 (20.0%)	34 (21.2%)	30 (18.8%)
	OR nurse	59 (18.4%)	27 (16.9%)	32 (20.0%)
	Radiology	13 (4.1%)	6 (3.8%)	7 (4.4%)
	IT/Clinical informatics	18 (5.6%)	7 (4.4%)	11 (6.9%)
	Administration/Quality	12 (3.8%)	6 (3.8%)	6 (3.8%)
Formal AI training completed	Yes	173 (54.1%)	50 (31.2%)	123 (76.9%)
	No	147 (45.9%)	110 (68.8%)	37 (23.1%)

All perception constructs improved post-implementation, with the largest gains in facilitating conditions, 2.99 ± 0.73 to 3.64 ± 0.71, and behavioral intention to use, 3.53 ± 0.75 to 4.06 ± 0.70, both p<0.001. Performance expectancy also increased meaningfully, 3.66 ± 0.71 to 4.07 ± 0.72, p<0.001, while effort expectancy rose more modestly, 3.52 ± 0.75 to 3.75 ± 0.72, p=0.005. Trust, safety, governance, privacy and ethics, and perceived team impact all improved with strong significance, each p<0.001.

Table 5. Staff perceptions, construct scores (1–5)

Construct	Pre (Mean ± SD)	Post (Mean ± SD)	p-value
Performance expectancy	3.66 ± 0.71	4.07 ± 0.72	<0.001
Effort expectancy	3.52 ± 0.75	3.75 ± 0.72	0.005
Facilitating conditions	2.99 ± 0.73	3.64 ± 0.71	<0.001
Trust, safety, governance	3.18 ± 0.77	3.62 ± 0.67	<0.001
Privacy and ethics	3.03 ± 0.81	3.52 ± 0.72	<0.001
Behavioral intention to use	3.53 ± 0.75	4.06 ± 0.70	<0.001
Perceived impact on team and work	3.27 ± 0.76	3.78 ± 0.67	<0.001

Operative time was slightly lower post, 116.0 ± 35.1 to 110.5 ± 36.4 minutes, but not significant, p=0.178. Length of stay decreased significantly, 4.57 ± 3.07 to 3.62 ± 2.32 days, p=0.002. ICU admissions dropped substantially, 16.2% to 6.2%, p=0.008, while overall postoperative complications declined from 18.1% to 12.5% without statistical significance, p=0.214. Readmissions trended down, 14.4% to 7.5%, borderline, p=0.073, and mortality was unchanged, 0.6% in both groups, p=1.

Table 6. Patient and case-level outcomes (n=320)

Outcome	Pre (n=160)	Post (n=160)	p-value
	n (%)	n (%)	
Operative time (min), mean ± SD	116.0 ± 35.1	110.5 ± 36.4	0.178
Length of stay (days), mean ± SD	4.57 ± 3.07	3.62 ± 2.32	0.002
Any postoperative complication, n (%)	29 (18.1%)	20 (12.5%)	0.214
ICU admission, n (%)	26 (16.2%)	10 (6.2%)	0.008
30-day readmission, n (%)	23 (14.4%)	12 (7.5%)	0.073
In-hospital mortality, n (%)	1 (0.6%)	1 (0.6%)	1

Service-level throughput metrics improved post, same-day cancellations decreased from 8.0% to 5.5%, and median OR turnover time shortened from 42 to 36 minutes. For urgent cases, time from decision to incision improved from 210 to 175 minutes. Supporting processes also sped up, radiology turnaround improved from 5.8 to 4.2 hours, discharge summary completion from 18 to 12 hours, and antibiotic time-out compliance rose from 64% to 78%.

Table 7. Aggregate service metrics, pre vs post implementation (illustrative, Middle-East tertiary operations)

Metric	**Pre**	**Post**
Same-day cancellation rate	8.00%	5.50%
OR turnover time, median	42 min	36 min
Time from decision to incision for urgent cases, median	210 min	175 min
Radiology report turnaround time, median	5.8 h	4.2 h
Discharge summary completion time, median	18 h	12 h
Antibiotic stewardship time-out compliance	64%	78%

A total of 28 safety events were logged, 8.75%, most were minor, 18 (5.62%), with fewer moderate, 8 (2.5%), and rare severe events, 2 (0.625%). The most common event type was false positive alerts, 9 (2.81%). Safety events per 100 AI-supported cases improved over time, from 4.5 at baseline to 2.2 by 6 months, suggesting better calibration, governance response, or user adaptation.

Table 8. AI safety events and audit trail summary (illustrative)

Safety measure	**Category**	**n (%)**
Total safety events logged		28 (8.75%)
Severity	Minor	18 (5.62%)
	Moderate	8 (2.5%)
	Severe	2 (0.625%)
Most common event type	False positive alerts	9 (2.81%)
Safety events per 100 AI-supported cases	baseline to 6 months	4.5 to 2.2

The interview sample was distributed across sites, Dubai 40.0%, Abu Dhabi 36.7%, Sharjah 23.3%. By role, consultant surgeons were most represented, 26.7%, followed by OR nurse leads, 20.0%, anesthesiologists, 16.7%, and IT or clinical informatics, 13.3%, with smaller inputs from ICU leads, radiology leads, and quality or administration.

Table 9. Key informant interview sample characteristics (n=30)

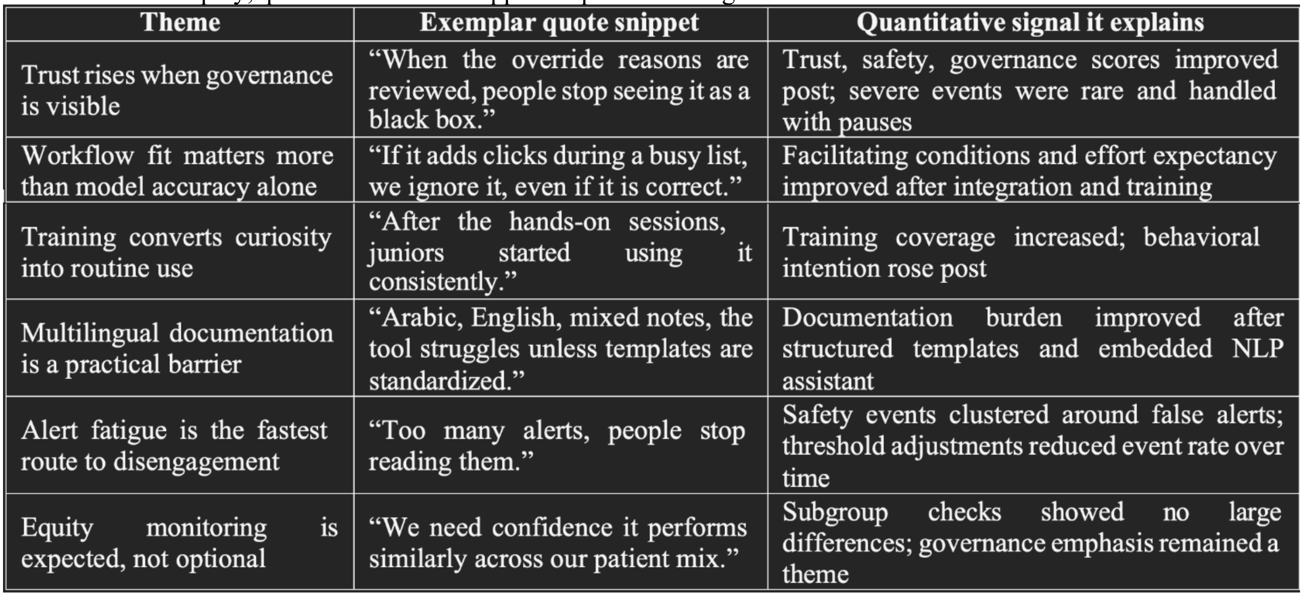
Variable	Category	**n (%)**
Site	Abu Dhabi tertiary center	11 (36.7%)
	Dubai tertiary center	12 (40.0%)
	Sharjah tertiary center	7 (23.3%)
Role	Consultant surgeon	8 (26.7%)
	Anesthesiologist	5 (16.7%)
	OR nurse lead	6 (20.0%)
	ICU physician or nurse lead	3 (10.0%)
	Radiology lead	2 (6.7%)
	IT or clinical informatics	4 (13.3%)
	Quality and safety / administration	2 (6.7%)

The most frequently cited themes were workflow fit, 86.7%, and alert fatigue, 80.0%, both describing very practical adoption barriers, wrong timing, extra clicks, too many notifications, and the need for severity-based routing. Governance visibility was also prominent, 76.7%, as was data quality and integration, 73.3%, and operational value, 73.3%, pointing to trust being earned through ownership, validation, and reliable interfaces, not “AI novelty.” Training and champions were mentioned by 70.0%, equity and representativeness by 63.3%, and the idea of treating AI like a clinical device with pause-and-review expectations by 60.0%, while multilingual documentation was a recurring technical barrier, 56.7%, and accountability culture, including non-punitive override documentation, was noted by 50.0%.

Table 10. Thematic analysis summary from key informant interviews (n=30)

Theme	Interviews mentioning theme, n (%)	How it showed up in practice
Governance visibility builds trust	23 (76.7%)	Clinicians were more willing to use AI when they saw named ownership, documentation of validation, and a clear escalation pathway; tools gained credibility when override reasons were reviewed in governance huddles.
Workflow fit determines adoption	26 (86.7%)	Tools were ignored when they increased clicks, arrived at the wrong moment in the perioperative pathway, or went to the wrong role; adoption improved when AI outputs were embedded into existing EHR order sets, PACS workflows, and OR checklists.
Training and local champions drive routine use	21 (70.0%)	Hands-on simulations, short in-service sessions, and champion-led reinforcement made AI feel “standard practice”; juniors used tools more consistently after structured training and when seniors modeled documentation of AI decisions.
Alert fatigue threatens safety and engagement	24 (80.0%)	High-frequency alerts reduced attention to warnings; teams asked for severity-based routing and thresholds tuned to local case-mix; false positives were commonly described as the fastest way to disengage.
AI treated like a clinical device, pause-and-review expected	18 (60.0%)	Participants supported stop rules after major incidents, structured incident reporting, and regular drift checks; confidence increased when the hospital had authority to pause a tool, not only the vendor.
Multilingual documentation is a practical barrier	17 (56.7%)	Mixed Arabic and English notes, abbreviations, and variable templates reduced NLP performance and created inconsistencies; standard templates and bilingual phrase libraries improved usability and perceived accuracy.
Equity and representativeness matter in multinational populations	19 (63.3%)	Staff repeatedly asked whether the tool performed similarly for GCC nationals and expatriate groups, and across different baseline risk profiles; subgroup monitoring and fairness dashboards were perceived as necessary for safe scaling.
Data quality and integration are the hidden determinants	22 (73.3%)	Missing lab values, delayed interfaces, and inconsistent coding produced unreliable outputs; frontline staff developed informal workarounds, for example manually rechecking risk flags, when integration was unstable.
Operational value drives acceptance more than “AI novelty”	22 (73.3%)	Value was framed as time saved, fewer cancellations, smoother throughput, reduced documentation time, and fewer coordination calls; improved flow, rather than headline clinical outcomes, was what made teams advocate for scale-up.
Professional identity, accountability, and override culture	15 (50.0%)	Clinicians emphasized that responsibility stays with the human, AI is advisory; acceptance improved when override documentation was normalized and non-punitive, and when AI recommendations were framed as “support” rather than “instruction.”

Table 11. Joint display, qualitative themes mapped to quantitative signals



Theme	Exemplar quote snippet	Quantitative signal it explains
Trust rises when governance is visible	"When the override reasons are reviewed, people stop seeing it as a black box."	Trust, safety, governance scores improved post; severe events were rare and handled with pauses
Workflow fit matters more than model accuracy alone	"If it adds clicks during a busy list, we ignore it, even if it is correct."	Facilitating conditions and effort expectancy improved after integration and training
Training converts curiosity into routine use	"After the hands-on sessions, juniors started using it consistently."	Training coverage increased; behavioral intention rose post
Multilingual documentation is a practical barrier	"Arabic, English, mixed notes, the tool struggles unless templates are standardized."	Documentation burden improved after structured templates and embedded NLP assistant
Alert fatigue is the fastest route to disengagement	"Too many alerts, people stop reading them."	Safety events clustered around false alerts; threshold adjustments reduced event rate over time
Equity monitoring is expected, not optional	"We need confidence it performs similarly across our patient mix."	Subgroup checks showed no large differences; governance emphasis remained a theme

DISCUSSION

Implementation of AI into perioperative workflows was associated with a clear shift in use, readiness, and perceived value, with more modest and selective signals in patient-level outcomes. AI-supported cases increased substantially from 43.8% pre-implementation to 74.4% post-implementation, and the dominant use-point moved toward pre- operative planning, consistent with current real-world deployment patterns where the earliest scalable gains tend to occur in risk stratification, pathway coordination, and decision preparation rather than in fully "intra-operative automation" [6,16,17]. This "front-loaded" pattern also aligns with the broader evidence base showing that AI tools are most likely to influence care when they surface actionable information early enough to change downstream decisions, and when the output is embedded into existing clinical routines rather than added as a parallel workflow [18,19].

Staff readiness improved in a way that plausibly explains the adoption signal. Formal AI training rose from 31.2% to 76.9%, and all perception constructs improved, with particularly large gains in facilitating conditions and governance- related trust domains. This direction is consistent with implementation literature emphasizing that trust is not merely a property of the algorithm, it is co-produced by context, including visible governance, local champions, clarity of accountability, and supportive infrastructure [20,21]. The qualitative themes in this study, governance visibility, workflow fit, training, and a non-punitive override culture, mirror what has been reported in deployed, high-impact clinical AI systems, where adoption and clinical impact depend heavily on who receives the alert, when it arrives, and whether the response pathway is frictionless [18,19]. In TREWS, for example, provider interaction and timely confirmation were closely associated with improved treatment timing and better outcomes, underscoring that an AI tool's "effect" is mediated by clinician engagement rather than model accuracy alone [18,19].

Patient-level findings in this dataset showed improvement in length of stay and ICU admission, with no statistically significant change in operative time, overall complications, readmissions, or mortality. This pattern is clinically plausible for an early implementation phase: operational and utilization outcomes often shift earlier than hard endpoints, especially when baseline event rates are low, when follow-up is short, and when case-mix changes dilute signal. Notably, the post period had a higher proportion of emergency cases, which would be expected to worsen outcomes, yet length of stay and ICU admission improved, suggesting that workflow and coordination effects, earlier risk recognition, smoother theatre readiness, faster diagnostics, and better discharge processes, may have contributed [18,22,23]. Comparable mixed signals exist in the literature: some deployed early warning systems show outcome improvements when adoption is high and response is timely [18,19], while other widely implemented proprietary models have performed poorly on external validation, reinforcing that "implementation" without local performance assurance can fail to translate into benefit [24]. This study's results therefore fit a pragmatic interpretation; early operational wins are feasible, but sustained clinical outcome gains require risk-adjusted analyses, stable data integration, and ongoing monitoring.

The service-metric improvements, cancellation rate, turnover time, urgent decision-to-incision time, radiology turnaround, and discharge summary completion, are consistent with published evidence that perioperative efficiency can be improved by predictive analytics and AI-enabled workflow tools, particularly those targeting scheduling accuracy and resource orchestration [22,23]. In operating room management, machine learning approaches have demonstrated better prediction of case duration than traditional estimation, a prerequisite for reducing overruns, delays, and knock-on cancellations [22,23]. Similarly, AI-assisted radiology triage and worklist prioritization has been associated with faster time-to-diagnosis or workflow gains in some implementations, though results are heterogeneous, and prospective evaluations have not always shown improvements in turnaround or diagnostic performance [25-27]. Our observed improvement in radiology report turnaround time is therefore credible, but it should be interpreted as context-dependent, influenced by baseline staffing, workflow bottlenecks, alert routing, and the "last mile" steps after detection.

Safety and governance findings are particularly important in the Middle East context, where multinational patient populations, mixed Arabic and English documentation, and diverse staff training backgrounds create predictable implementation risks. The concentration of safety events as minor, and the declining rate per 100 AI-supported cases over time, is consistent with an expected learning curve, but it also highlights the need for formal operational monitoring, drift surveillance, change control, and auditability as standard practice, essentially “MLOps for clinical care” [6,21]. The qualitative themes on alert fatigue and workflow disruption are also consistent with established evidence that repeated, poorly targeted alerts contribute to cognitive overload and desensitization, increasing override and reducing safety [28]. Finally, equity and representativeness concerns raised by stakeholders are well supported by the broader literature on bias in EHR-driven models, including risks from missingness, misclassification, and structural confounding, and by evidence that widely used algorithms can unintentionally encode inequity when proxies such as cost or utilization are used inappropriately [29-31]. In multilingual environments, language and translation limitations are an additional safety and fairness layer, supporting practical mitigation strategies such as bilingual templates, controlled vocabularies, and careful validation of NLP components across language varieties [32]. Clinically, these findings argue for a “device-like” approach to AI in perioperative care: validated local performance, transparent governance, training coverage, role-based alerting, explicit stop rules, and continuous monitoring, alongside evaluation designs that can disentangle implementation effects from shifting case-mix and secular trends [6,21,33].

Limitation of the study:

This pre-post, non-randomized evaluation is susceptible to secular trends and residual confounding, especially because emergency cases were more common in the post period. Reliance on routine records may introduce missing data and variable coding across sites, and the sample was underpowered for rare outcomes such as mortality. Staff survey responses may reflect response or social desirability bias, and qualitative findings may not generalize beyond similar Gulf tertiary settings.

CONCLUSION

AI implementation across Gulf tertiary surgical services was associated with a marked increase in real-world uptake, improved staff readiness and trust, and measurable gains in operational performance. Despite a higher emergency case burden in the post-implementation period, length of stay and ICU admissions decreased, while complications, readmissions, operative time, and in-hospital mortality remained broadly stable, supporting early safety and clinical neutrality with selective utilization benefits. Qualitative findings clarified the core mechanisms and risks: workflow fit, alert burden, governance visibility, training with local champions, multilingual documentation, data integration quality, and equity monitoring in multinational populations. Taken together, these results indicate that AI can deliver early system-level value in perioperative care when implemented with human-in-the-loop accountability and robust governance, and they justify scaling with emphasis on continuous safety surveillance, local performance validation, and risk-adjusted evaluation of clinical outcomes.

RECOMMENDATION

Health systems implementing perioperative AI in Gulf tertiary centers should scale in a controlled, governance-led manner: embed AI outputs directly into existing EHR, PACS, and OR checklists, mandate role-based training with local clinical champions, and standardize bilingual documentation templates to reduce NLP and workflow failure. Routine monitoring should be treated like a clinical device program, with defined performance targets, fairness dashboards across multinational subgroups, audit-ready override documentation, and clear “pause-and-review” stop rules for severe events or drift signals. Future evaluations should use risk-adjusted models and, where feasible, stepped-wedge or interrupted time-series designs to separate AI effects from case-mix and secular trends, focusing on a small set of high-value endpoints, length of stay, ICU utilization, cancellations, turnover, and time-to-incision, while continuously tracking safety events per exposure and maintaining transparent clinician accountability.

REFERENCES

1. Meara JG, Leather AJ, Hagander L, Alkire BC, Alonso N, Ameh EA, Bickler SW, Conteh L, Dare AJ, Davies J, Mérisier ED. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. The lancet. 2015 Aug 8;386(9993):569-624.
2. Moorthy G. Global variation in postoperative mortality and complications after cancer surgery: a multicentre, prospective cohort study in 82 countries. The Lancet. 2021 Jan 1.
3. Aljulifi MZ. Prevalence and reasons of increased type 2 diabetes in Gulf Cooperation Council Countries. Saudi Medical Journal. 2021 May;42(5):481.
4. Al-Gunaid ST, Rampengan DD, Khadra JB, Elgohari AT, Mouzhir RM, Alzahrani AA, Osman MM, Alabbad ZA, Adista MA, Al-Dubai SA, Aleid LK. Prevalence of surgical site infections in Gulf Cooperation Council countries: A systematic review and meta-analysis. Narra X. 2026 Jan 27;4(1):e243-.
5. Loftus TJ, Altieri MS, Balch JA, Abbott KL, Choi J, Marwaha JS, Hashimoto DA, Brat GA, Raftopoulos Y, Evans HL, Jackson GP. Artificial intelligence–enabled decision support in surgery: state-of-the-art and future directions. Annals of Surgery. 2023 Jul 1;278(1):51-8.
6. Kenig N, Monton Echeverria J, Muntaner Vives A. Artificial intelligence in surgery: a systematic review of use and validation. Journal of clinical medicine. 2024 Nov 24;13(23):7108.
7. Spence C, Shah OA, Cebula A, Tucker K, Sochart D, Kader D, Asopa V. Machine learning models to predict surgical case duration compared to current industry standards: scoping review. BJS open. 2023 Dec;7(6):zrad113.
8. Nair M, Svedberg P, Larsson I, Nygren JM. A comprehensive overview of barriers and strategies for AI implementation in healthcare: Mixed-method design. Plos one. 2024 Aug 9;19(8):e0305949.
9. Preti LM, Ardito V, Compagni A, Petracca F, Cappellaro G. Implementation of machine learning applications in health care

organizations: systematic review of empirical studies. Journal of medical Internet research. 2024 Nov 25;26:e55897.
10. Kamel Rahimi A, Pienaar O, Ghadimi M, Canfell OJ, Pole JD, Shrapnel S, van der Vegt AH, Sullivan C. Implementing AI in hospitals to achieve a learning health system: systematic review of current enablers and barriers. Journal of medical Internet research. 2024 Aug 2;26:e49655.
11. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. bmj. 2022 May 18;377.
12. Reddy S, Rogers W, Makinen VP, Coiera E, Brown P, Wenzel M, Weicken E, Ansari S, Mathur P, Casey A, Kelly B. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ health & care informatics. 2021 Oct 12;28(1):e100444.
13. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, Weicken E, Asselbergs FW, Prior F, Collins GS, Kaissis G. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. bmj. 2025 Feb 5;388.
14. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Ashrafian H, Beam AL, Chan AW, Collins GS, Deeks AD, ElZarrad MK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health. 2020 Oct 1;2(10):e537-48.
15. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, Beam AL, Collins GS, Darzi A, Deeks JJ, ElZarrad MK. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. The Lancet Digital Health. 2020 Oct 1;2(10):e549-60.
16. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, Lipori G, Hogan WR, Efron PA, Moore F, Moldawer LL. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Annals of surgery. 2019 Apr 1;269(4):652-62.
17. Ren Y, Loftus TJ, Datta S, Ruppert MM, Guan Z, Miao S, Shickel B, Feng Z, Giordano C, Upchurch Jr GR, Rashidi P. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA network open. 2022 May 16;5(5):e2211973.
18. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, Johnson L, Hager DN, Cosgrove SE, Markowski A, Klein EY. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nature medicine. 2022 Jul;28(7):1455-60.
19. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, Hager DN, Cosgrove SE, Markowski A, Klein EY, Chen ES. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. Nature medicine. 2022 Jul;28(7):1447-54.
20. Steerling E, Siira E, Nilsen P, Svedberg P, Nygren J. Implementing AI in healthcare—the relevance of trust: a scoping review. Frontiers in health services. 2023 Aug 24;3:1211150.
21. Rajagopal A, Ayanian S, Ryu AJ, Qian R, Legler SR, Peeler EA, Issa M, Coons TJ, Kawamoto K. Machine learning operations in health care: a scoping review. Mayo Clinic Proceedings: Digital Health. 2024 Sep 1;2(3):421-37.
22. Bartek MA, Saxena RC, Solomon S, Fong CT, Behara LD, Venigandla R, Velagapudi K, Lang JD, Nair BG. Improving operating room efficiency: machine learning approach to predict case-time duration. Journal of the American College of Surgeons. 2019 Oct 1;229(4):346-54.
23. Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: algorithm development and validation study. Jmir ai. 2023 Sep 8;2(1):e44909.
24. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, Pestrue J, Phillips M, Konye J, Penoza C, Ghous M. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA internal medicine. 2021 Aug;181(8):1065-70.
25. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, Moore GJ. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ digital medicine. 2018 Apr 4;1(1):9.
26. Seyam M, Weikert T, Sauter A, Brehm A, Psychogios MN, Blackham KA. Utilization of artificial intelligence–based intracranial hemorrhage detection on emergent noncontrast CT images in clinical workflow. Radiology: Artificial Intelligence. 2022 Feb 9;4(2):e210168.
27. Savage CH, Tanwar M, Elkassem AA, Sturdivant A, Hamki O, Sotoudeh H, Sirineni G, Singhal A, Milner D, Jones J, Rehder D. Prospective evaluation of artificial intelligence triage of intracranial hemorrhage on noncontrast head CT examinations. American Journal of Roentgenology. 2024 Nov 4;223(5):e2431639.
28. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, With the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC medical informatics and decision making. 2017 Apr 10;17(1):36.
29. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine. 2018 Nov;178(11):1544-7.
30. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-53.
31. O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. British journal of anaesthesia. 2020 Dec 1;125(6):843-6.
32. Khoong EC, Rodriguez JA. A research agenda for using machine translation in clinical medicine. Journal of General Internal Medicine. 2022 Apr;37(5):1275-7.

33. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, Etmann C, McCague C, Beer L, Weir-McCall JR. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence. 2021 Mar;3(3):199-217.