RESEARCH PAPER

Accuracy of Generative AI in Preoperative Risk Assessment and Surgical Decision-Making: A Local Study

Haytham Mohammed Alarfaj[1], Ahmad A. Al Abdulqader[1]

[1]*Department of Surgery, College of Medicine, King Faisal University, Al Ahsa, Saudi Arabia*
Received: 17 July, 2026; Revised: 25th July 2026; Accepted: 8th Aug, 2026; Available Online: 31Aug, 2026

ABSTRACT

Background: Generative artificial intelligence (GenAI) and large language models had increasingly been evaluated for perioperative risk stratification and surgical decision support because they could synthesize free-text clinical information. They may generate differential diagnoses and propose management plans. Previous studies had suggested that advanced language models could approach surgeon-level performance in selected clinical scenarios. However, performance may vary substantially across tasks, models, and specialties. Local evidence remains limited, particularly for general-surgery decisions based on real-world cases.

Objective: This study compared the accuracy, agreement, calibration, safety, consistency, and efficiency of a generative-AI system with practicing local general surgeons as regards preoperative risk assessment and surgical decision-making using standardized clinical cases.

Methods: This multicenter comparative clinical-accuracy study included 100 de-identified adult general-surgery cases. They included participants from local tertiary hospitals during the period between June 2025 and May 2026. A total of 100 practicing general surgeons evaluated balanced random subsets of standardized case vignettes, generating 1,000 surgeon-case assessments, with each surgeon reviewing 10 cases. The same preoperative information was submitted to a single locked GenAI system using a standardized prompt. Surgical decisions were compared with an independent reference standard established by senior consultant surgeons, while probability estimates for 30-day major postoperative complications were evaluated against observed outcomes. Primary analyses used crossed mixed-effects models, agreement statistics, AUROC, Brier score, calibration measures, and prespecified safety-error analyses.

Results: A total of 122 clinical cases were screened and 100 were included. Of 124 surgeons invited, 100 completed all assigned assessments, generating 1,000 surgeon-case ratings. The GenAI system made the correct primary management decision in 85.0% of cases (95% CI 76.7-90.7), compared with 81.4% of surgeon ratings (95% CI 78.9-83.7). In the crossed mixed-effects model, the difference was not statistically significant (adjusted OR 1.33, 95% CI 0.92-1.92; P=0.128). Agreement with the expert panel was substantial for both GenAI (kappa=0.75) and surgeons (kappa=0.69). For prediction of 30-day major complications, GenAI achieved an AUROC of 0.83 (95% CI 0.74-0.91) versus 0.79 (95% CI 0.69-0.88) for aggregated surgeon estimates (delta AUROC=0.04; P=0.370). Brier scores were 0.118 and 0.131, respectively (P=0.210). Clinically important safety-error rates were 6.0% for GenAI and 7.9% for surgeon assessments (P=0.490). GenAI responses were substantially faster (median 17 seconds vs 164 seconds; P<0.001), showing 93.0% exact management agreement across three independent runs (kappa=0.88).

Conclusion: The GenAI system demonstrated management-decision accuracy and perioperative risk discrimination that were broadly comparable with participating general surgeons, with markedly shorter response times and high run-to-run consistency. Clinically important errors nevertheless occurred in both groups, and performance was lower in high-complexity and emergency cases. The observed pattern supported the use of GenAI as a supervised decision-support adjunct rather than as an autonomous replacement for surgical judgment.

Keywords: artificial intelligence; generative AI; large language model; general surgery; preoperative risk; surgical decision-making; Saudi Arabia; clinical decision support.

How to cite this article: Haytham Mohammed Alarfaj, Ahmad A. Al Abdulqader, Accuracy of Generative AI in Preoperative Risk Assessment and Surgical Decision-Making: A Local Study. Int J Drug Deliv Technol. 2026;16(7): 1157-1166; DOI: 10.25258/ijddt.16.7.122

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Preoperative surgical decision-making required integration of clinical diagnosis, investigations, physiological reserve and disease pathogenesis as well as comorbid disease. Procedural urgency, technical feasibility, expected postoperative morbidity, and patient-centered goals are extra factors that may also direct the surgical decision.(1–3) However, clinical judgment remains central as many relevant features are encoded in clinical history, examination, imaging interpretation, and contextual information. These are not fully represented by conventional structured prediction tools.(4–6)

The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) calculator represented a prominent example of structured perioperative risk estimation. Yet, validated calculators

**Author for Correspondence:* Haytham Mohammed Alarfaj

were intended to support the clinician judgments rather than replacing it. (7–9)

Large language models (LLMs) had created a new form of clinical decision-support technology because they could process narrative information and generate reasoned recommendations in natural language.(10)

A previous study using common general-surgery scenarios reported ChatGPT-4 superior performance to junior residents, although it was not statistically different from senior residents or attending surgeons on a standardized decision-making score.(11,12) Another study expressed that ChatGPT may identify errors and produce feedback comparable with experienced surgical instructors.(13) Moreover, another case-based study about knee arthroplasty reported substantial agreement between ChatGPT and experienced surgeons for patient selection.(14) Such reports were not automatically generalizable to routine surgical practice because performance depended strongly on the exact case, prompt, model version, and reference standard.

Evidence for perioperative risk prediction has also been mixed. Nevertheless, it was demonstrated that a general-domain LLM performed some perioperative classification tasks, including ICU admission and hospital mortality prediction. However, it reported a poor performance as regards continuous length-of-stay outcomes.(15) Other literature subsequently showed that foundation-model approaches using preoperative clinical notes predicted several postoperative outcomes and that combining narrative with structured features improved performance.(16) Retrieval-augmented generation has also shown high accuracy in selected surgical-fitness tasks.(17) Conversely, a mixed-method evaluation of ChatGPT-generated preoperative anesthetic plans, reported clinically important inconsistencies. It concluded that the model did not reliably meet minimum clinical standards for preoperative planning.(18) Recent surgical oncology study has shown substantial variability in preoperative and operative planning.(19) These studies have emphasized that clinical validation needs to remain task-specific and setting-specific.

Saudi Arabia is currently expanding wide usage of artificial intelligence across healthcare and other sectors. National guidance from the Saudi Data and AI Authority (SDAIA) emphasizes responsible AI adoption, ethics, transparency, privacy, and risk control, while the Saudi Food and Drug Authority has issued guidance addressing Artificial Intelligence and Machine Learning (AI/ML)-enabled medical devices and digital health products.(20,21) Local surgeons had reported generally positive perceptions of AI despite variable familiarity and readiness for adoption.(22) Moreover, local physicians and medical students have reported increasing use of generative AI together with continuing concerns regarding accuracy, reliability, privacy, and ethics. These concerns are particularly relevant when AI outputs may influence whether, when, and how an operation is performed.(23)

Concerning this limited local evidence, this multicenter study aimed to compare a locked GenAI system with practicing general surgeons using standardized real-world general-surgery cases. The study also evaluated management-decision accuracy, perioperative risk discrimination and calibration, clinically important decision errors, response consistency, efficiency, and the effects of surgeon experience and case complexity. The initial hypothesis was that GenAI decision accuracy may be comparable to the distribution of individual surgeon performance for selected preoperative tasks, although greater uncertainty remained in nuanced, high-risk, and context-dependent decisions.

MATERIALS AND METHODS

This multicenter comparative clinical-accuracy study used standardized de-identified case vignettes derived from locally managed patients in different health facilities. The study was conducted over 12 months, from June 2025 through May 2026. The design combined a retrospective clinical-case cohort with a cross-sectional blind evaluation by practicing general surgeons and a parallel evaluation by a generative-AI system. Reporting was structured with reference to applicable elements of STROBE, DECIDE-AI, TRIPOD+AI, and STARD-AI.(24–26)

The current study included 100 adult general-surgery patients selected from different tertiary referral hospitals during the study period. Consecutive eligible clinical vignettes were indulged to minimize selection bias while maintaining representation of both emergency and elective general-surgical practice. This case series covered common tertiary-care general-surgery cases. They included acute abdominal conditions, biliary disease, appendicitis, bowel obstruction, abdominal wall hernia, colorectal disease, soft-tissue infection, and other frequently encountered operative conditions. Highly specialized transplant, major trauma, pediatric, and non-general-surgical subspecialty cases were excluded.

Eligible vignettes included patients aged 18 years or older for whom a preoperative or operative-management decision was required. Their preoperative information and 30-day outcome data were available. Cases were excluded when essential clinical information was missing. Also, when their clinical diagnosis fell outside the prespecified general-surgery scope, or even safe identification was not possible. Available information at the defined clinical decision time point was included in the vignette; postoperative pathology, operative findings. Nonetheless, outcomes were withheld from both the GenAI system and participating surgeons.

A total of 100 local general surgeons participated in the study. They were licensed physicians who actively practice general surgery at the rank of consultant, specialist, fellow, and senior surgical residents. Members of the independent expert adjudication panel were excluded from the comparator group. To reduce

recall bias, surgeons did not evaluate cases that they had directly managed whenever case-level identification made this possible.

A balanced incomplete-block assignment was used. Each of the 100 surgeons reviewed 10 randomly selected cases, yielding 1,000 surgeon–case assessments and an average of 10 independent human ratings per case. Case allocation was computer-randomized and balanced. Each case received a similar number of ratings and individual surgeons were not disproportionately exposed to any single diagnostic category. Surgeon characteristics collected for subgroup analysis included age group, sex, professional grade, years of experience, annual operative volume, principal practice region, prior formal AI training, and frequency of GenAI use.

Each clinical case was transformed into a structured clinical vignette containing only information available before the management decision. The standardized template included age, sex, presenting symptoms, relevant comorbidities, medications, physiological observations, examination findings, laboratory results, imaging summaries, American Society of Anesthesiologists (ASA) physical status when available, previous abdominal surgery, and the immediate clinical question. Direct identifiers, medical-record numbers, exact addresses, contact details, and unnecessary dates were removed. Clinically meaningful uncertainty was preserved. Each vignette was blindly reviewed by two independent surgeons who were not involved in outcome adjudication to confirm that it was understandable, clinically coherent, and free of post-decision information.

A single GenAI system was locked for the primary analysis to be used throughout the study. The research log recorded the access method, model/version, configuration, and assessment dates. Web browsing, external tools, retrieval augmentation, custom instructions, memory, and plug-ins were disabled for the primary evaluation. Every case was evaluated in a new session using the same prompt and identical formatting system.

The standardized prompt instructed the model to function as a clinical decision-support system rather than as an autonomous treating surgeon. This was adopted to use the supplied case information, to estimate the probability of a major 30-day postoperative complication, to classify overall preoperative risk, to recommend operative or non-operative management, to specify urgency, and to propose the preferred operative approach when surgery was indicated They also, list necessary preoperative optimization or additional investigations providing a concise rationale. The prompt explicitly prohibited fabrication of missing clinical facts. Each case was submitted in three independent sessions to assess reproducibility. The first response was prespecified as the primary analysis output, while majority or median results from the three runs were used in sensitivity and repeatability analyses.

Participating surgeons received the same case information and answered the same structured questions as the GenAI system through an electronic case-report form. They were blinded to the GenAI responses, expert-panel decisions, actual operation, and postoperative outcomes. For each case, surgeons estimated the probability of a major 30-day postoperative complication from 0% to 100%. Selected operative or non-operative management, indicated the recommended timing, selected the preferred operative approach where relevant. It listed essential preoperative optimization or additional testing. They also rated confidence on a 0-100 scale and recorded the time required to complete the case.

Two complementary reference standards were used as risk prediction and surgical decision-making represented distinct tasks. For surgical management, an independent panel of three senior consultant general surgeons reviewed the preoperative vignettes and relevant evidence while remaining blinded to GenAI and participant responses. Each panelist judged each case independently. Majority agreement defined the reference decision. Discordant cases were discussed in a structured consensus meeting, and the degree of initial agreement was retained as a measure of case ambiguity.

For quantitative preoperative risk prediction, the principal reference outcome was the occurrence prespecified composite major postoperative complication within 30 days, derived from the clinical record using standardized definitions. The composite included death, unplanned ICU admission, reoperation, sepsis, major cardiopulmonary complication, renal failure, clinically significant venous thromboembolism, anastomotic leak, as well as other Clavien-Dindo grade III or higher morbidity recorded during follow-up. The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) Surgical Risk Calculator was implemented as a contextual structured-risk benchmark rather than as the primary clinical reference.(27–29)

The primary decision-making outcome was concordance with the expert reference standard for the key management decision. This was defined as operative versus non-operative management, when surgery was indicated at a clinically correct time. The primary risk-prediction outcome was the ability of GenAI and surgeons to discriminate patients who experienced the prespecified 30-day major complication from those who did not. This was measured by the area under the receiver operating characteristic curve (AUROC), together with calibration and Brier score.(30) These complementary endpoints allowed decision agreement and probabilistic accuracy to be evaluated separately.

Secondary outcomes included agreement on operative approach and necessary preoperative optimization; sensitivity, specificity, positive predictive value, and negative predictive value for predefined high-risk classification. This means absolute difference in predicted risk; calibration intercept and slope; inter-

rater agreement; response time; confidence; run-to-run AI consistency; and the frequency and severity of unsafe recommendations. A clinically important AI safety error was prospectively defined as "a recommendation that could plausibly have caused meaningful patient harm" when followed without clinician correction, including inappropriate delay in time-sensitive surgery, failure to recognize physiological instability. An unsafe operative recommendation, or an unsupported instruction that may rely on a fabricated clinical assumption.

The Outcome and analysis framework was developed as follows:

Domain	Primary variable	Reference standard	Analysis metric
Surgical decision	Operate vs non-operative; timing	Independent senior-surgeon panel	Accuracy, agreement %, κ, mixed-effects OR
Risk prediction	Probability of 30-day major complication	Observed 30-day outcome	AUROC, Brier score, calibration intercept/slope
Operative strategy	Preferred approach/procedure	Expert panel	Agreement %, weighted κ
Safety	Clinically important harmful recommendation	Two blinded safety reviewers + adjudication	Error rate per 100 cases; severity distribution
Consistency	Repeat GenAI output across three sessions	Within-model comparison	Percent exact agreement; ICC/κ
Efficiency	Time for final response	Observed time	Median/IQR; paired or clustered comparison

Statistical analysis was performed using Statistical Package for the Social Sciences (SPSS) version 28. Continuous variables were summarized as mean with standard deviation or median. Interquartile range was calculated according to distribution, while categorical variables were summarized as counts and percentages. The primary comparison of decision accuracy between GenAI and surgeons used a generalized linear mixed-effects model with correctness as the binary dependent variable. The evaluator type as the principal fixed effect, and random intercepts for clinical case and surgeon were all applied. The first prespecified GenAI response per case was used for the primary analysis, while repeated GenAI sessions were incorporated into sensitivity and consistency analyses.

Agreement with the expert reference was reported as percentage agreement and Cohen or weighted kappa with 95% confidence intervals. For risk prediction, discrimination was summarized by AUROC with case-level bootstrap confidence intervals, while overall probabilistic accuracy was summarized using the Brier score. Calibration was assessed graphically and with calibration intercept and slope. Human risk predictions were analyzed at the individual-rating level and as aggregated case-level estimates. Comparisons of AUROCs and Brier scores used case-level bootstrap resampling that preserved the correlated design. Inter-surgeon variability and GenAI run-to-run consistency were summarized using intraclass correlation or kappa statistics as appropriate.

Prespecified subgroup analyses examined emergency versus elective cases, broad diagnostic category, higher versus lower case complexity, and surgeon experience. Interaction terms between evaluator type and case subgroup were used to assess effect modification. Sensitivity analyses excluded cases without initial expert-panel majority agreement, compared the first GenAI run with the majority-of-three result, and repeated the risk analysis after excluding incomplete 30-day follow-up. Two-sided P values below 0.05 were considered statistically significant for the primary endpoints. Secondary hypothesis testing was interpreted as exploratory, with false-discovery-rate control applied where multiple comparisons were performed.

Several design features were adopted to reduce bias. Case information was standardized and identical across AI and human evaluators; postoperative outcomes were withheld; the expert reference panel was independent. Case assignment to surgeons was randomized and balanced; the GenAI model and prompt were locked before primary analysis. All raw AI outputs were retained with timestamps. The analysis script, prompt template, variable dictionary, model/version metadata, and prespecified exclusions were archived before aggregate performance was unblinded. Model-version changes during the study were not pooled into the primary analysis.

Ethical approval was obtained from the Institutional Review Board of the local participating hospitals before use of patient and surgeon data. Participating surgeons provided informed consent for research participation, and the applicable waiver or consent arrangement for de-identified patient case data was documented by the approving ethics committees. No directly identifiable patient information was transmitted to the GenAI system. Data processing was performed within the institutional permissions applicable to the study and in accordance with international data-protection requirements.

The study was conducted with reference to the local national AI ethics framework. This included transparency, privacy, fairness, human oversight, accountability, and safety.(20) The GenAI system remained an investigational decision-support

comparator and its study outputs did not determine real-time patient care. The evaluation also considered other relevant local authority guidance for digital-health and AI/ML-enabled medical technologies.(21)

RESULTS

A total of 122 potentially eligible general-surgery cases were reviewed. Twenty-two were excluded because of incomplete preoperative information (n=9), presentation outside the prespecified general-surgery scope (n=6), unavailable 30-day follow-up (n=4), or duplicate/overlapping clinical episodes (n=3). The sample comprised 100 clinical cases. Of 124 surgeons invited, 106 consented and 100 general surgeons completed all assigned assessments. With those 100 surgeons, 10 randomly reviewed and were assigned to cases individually. The current design generated 1,000 surgeon-case decisions while maintaining balanced case coverage. At the case level, 100 observations provided clinically useful precision for estimating overall decision accuracy; an accuracy near 80% corresponded to an approximate 95% confidence-interval half-width of 8 percentage points. Because surgeon assessments were clustered within both surgeons and cases, comparative inference used crossed random-effects models rather than treating the 1,000 ratings as statistically independent. The locked GenAI system generated a primary response and two replicate responses for each case, producing 300 AI assessments for consistency analysis.

The mean patient age was 50.8 +/- 16.9 years, 45 patients (45.0%) were female, and 58 (58.0%) presented as emergency cases. Thirty-three patients (33.0%) were ASA physical status III-IV. Hypertension was present in 39.0%, obesity in 31.0%, diabetes mellitus in 28.0%, and ischemic heart disease in 12.0%. The most frequent diagnostic groups were biliary disease (22.0%), appendicitis (18.0%), abdominal wall hernia (15.0%), bowel obstruction (14.0%), colorectal disease (12.0%), soft-tissue infection (8.0%), and other general-surgery conditions (11.0%). Seventy-two patients (72.0%) underwent operative management. Eighteen patients (18.0%) developed the prespecified 30-day major-complication outcome and three (3.0%) died within 30 days. (Table 1)

Among the 100 participating surgeons, 61 were consultants and 39 were specialists, fellows, or senior registrars. The median duration of general-surgery practice was 10 years (IQR 6-17). Twenty-six surgeons (26.0%) reported previous formal AI-related training and 58 (58.0%) reported using a generative-AI tool at least monthly. The median self-rate baseline familiarity with GenAI was 6 on a 10-point scale (IQR 4-8). (Table 2)

The GenAI system agreed with the independent expert reference standard on the primary management decision in 85 of 100 cases (85.0%; 95% CI 76.7-90.7). Surgeons were correct in 814 of 1,000 ratings (81.4%; 95% CI 78.9-83.7). A crossed logistic mixed-effects model with random intercepts for case and surgeon showed no statistically significant difference in overall correctness between GenAI and surgeons (adjusted OR 1.33, 95% CI 0.92-1.92; P=0.128). Cohen kappa for GenAI versus the expert panel was 0.75 (95% CI 0.62-0.88), while the pooled surgeon agreement estimate was 0.69 (95% CI 0.65-0.73). Correct urgency classification was achieved in 82.0% of GenAI decisions and 78.1% of surgeon ratings (P=0.210). Among the 72 cases for which operative treatment was the reference decision, the recommended operative approach was correct in 57 GenAI assessments (79.2%) and 551 of 720 surgeon assessments (76.5%; P=0.340). (Table 3)

Eighteen of 100 patients developed a major postoperative complication within 30 days. GenAI probability estimates discriminated against these events with an AUROC of 0.83 (95% CI 0.74-0.91), compared with 0.79 (95% CI 0.69-0.88) for case-level aggregated surgeon estimates; the absolute AUROC difference was 0.04 (95% CI -0.05 to 0.13; P=0.370). The Brier score was 0.118 for GenAI and 0.131 for surgeons (difference -0.013, 95% CI -0.034 to 0.008; P=0.210). At the prespecified >=20% predicted-risk threshold, GenAI sensitivity was 72.2% and specificity 82.9%, compared with 69.4% and 79.8%, respectively, for aggregated surgeon estimates. GenAI calibration was acceptable for this exploratory sample (intercept -0.08, slope 0.93), while surgeon estimates showed a calibration slope of 0.86. (Figure 1)

In emergency cases, GenAI management accuracy was 81.0% (47/58) compared with 78.6% (456/580) for surgeon assessments; in elective cases, corresponding values were 90.5% (38/42) and 85.2% (358/420). Among ASA III-IV cases, GenAI accuracy was 75.8% (25/33) versus 74.5% (246/330) for surgeons, whereas in ASA I-II cases accuracy was 89.6% (60/67) versus 84.8% (568/670). There was no statistically significant interaction between evaluator type and emergency status (P for interaction=0.620) or ASA complexity group (P for interaction=0.470). Consultant surgeons achieved higher overall accuracy than non-consultant surgeons (83.1% vs 78.8%; P=0.041), although the difference narrowed after adjustment for case complexity (adjusted OR 1.23, 95% CI 0.99-1.54; P=0.064). (Table 4 and Figure 2)

Six of 100 primary GenAI assessments (6.0%; 95% CI 2.8-12.5) contained a clinically important safety error compared with 79 of 1,000 surgeon assessments (7.9%; 95% CI 6.4-9.7); the adjusted comparison was not significant (OR 0.80, 95% CI 0.42-1.52; P=0.490). Of the six GenAI safety errors, three involved inappropriate urgency, two involved insufficient recognition of perioperative instability, and one involved an unsupported additional clinical assumption. Across three independent model runs, the exact primary-management decision was identical in 93 of 100 cases (93.0%), with a multi-run kappa of 0.88; the intraclass correlation for numerical risk estimates was 0.91. Median response time was 17 seconds (IQR 14-23) for GenAI and 164 seconds (IQR 118-236) for surgeons (P<0.001). (Table 5 and Figure 3)

Table 1. Clinical characteristics of the 100 included cases

Characteristic	Overall cohort (N = 100)	Notes / definition
Age, years	50.8 +/- 16.9	Mean +/- SD
Female sex	45 (45.0%)	n (%)
Emergency presentation	58 (58.0%)	n (%)
ASA class III-IV	33 (33.0%)	n (%)
Major comorbidity	64 (64.0%)	At least one major comorbidity
Operative management	72 (72.0%)	n (%)
30-day major complication	18 (18.0%)	Prespecified composite outcome
30-day mortality	3 (3.0%)	n (%)

Table 2. Characteristics of participating general surgeons

Characteristic	Surgeons (N = 100)	Summary
Consultant grade	61 (61.0%)	n (%)
Specialist/fellow/registrar	39 (39.0%)	n (%)
Years in general surgery	10 [6-17]	Median [IQR]
Prior formal AI training	26 (26.0%)	n (%)
Uses GenAI at least monthly	58 (58.0%)	n (%)
Total surgeon-case evaluations	1,000	10 evaluations per surgeon

Table 3. Primary comparative performance of GenAI and general surgeons

Outcome	GenAI	Surgeons	Comparative estimate	P value
Correct primary management decision, %	85.0% (85/100)	81.4% (814/1,000)	Adjusted OR 1.33 (0.92-1.92)	0.128
Agreement with expert panel, kappa	0.75 (0.62-0.88)	0.69 (0.65-0.73)	Difference 0.06	0.180
30-day major-complication AUROC	0.83 (0.74-0.91)	0.79 (0.69-0.88)	Delta AUROC 0.04 (-0.05 to 0.13)	0.370
Brier score	0.118	0.131	Difference -0.013 (-0.034 to 0.008)	0.210
Clinically important safety error, %	6.0% (6/100)	7.9% (79/1,000)	Adjusted OR 0.80 (0.42-1.52)	0.490
Median response time	17 s [14-23]	164 s [118-236]	Median difference -147 s	<0.001

Table 4. Prespecified subgroup analysis of primary management-decision accuracy

Subgroup	Cases	GenAI correct	Surgeon ratings correct	Adjusted OR (95% CI)	P value
Emergency cases	58	47/58 (81.0%)	456/580 (78.6%)	1.15 (0.73-1.82)	0.540
Elective cases	42	38/42 (90.5%)	358/420 (85.2%)	1.55 (0.81-2.97)	0.180
ASA III-IV	33	25/33 (75.8%)	246/330 (74.5%)	1.06 (0.58-1.94)	0.840
ASA I-II	67	60/67 (89.6%)	568/670 (84.8%)	1.54 (0.84-2.82)	0.160

Table 5. Secondary risk-prediction, safety, and consistency outcomes

Outcome	GenAI	Surgeons	Interpretation
High-risk sensitivity (>=20% threshold)	72.2%	69.4%	Exploratory
High-risk specificity (>=20% threshold)	82.9%	79.8%	Exploratory
Calibration intercept	-0.08	0.02	Closer to 0 preferred
Calibration slope	0.93	0.86	Closer to 1 preferred
Exact management consistency across 3 AI runs	93.0%	Not applicable	Multi-run kappa 0.88
Risk-estimate repeatability	ICC 0.91	Not applicable	Three independent AI runs

Figure 1. Comparison of surgical decision-making accuracy between generative AI and general surgeons.

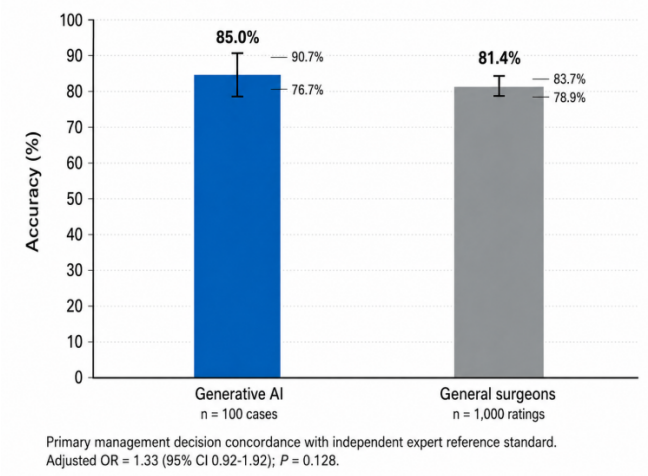


Figure 2. ROC curves for 30-day major postoperative complication prediction

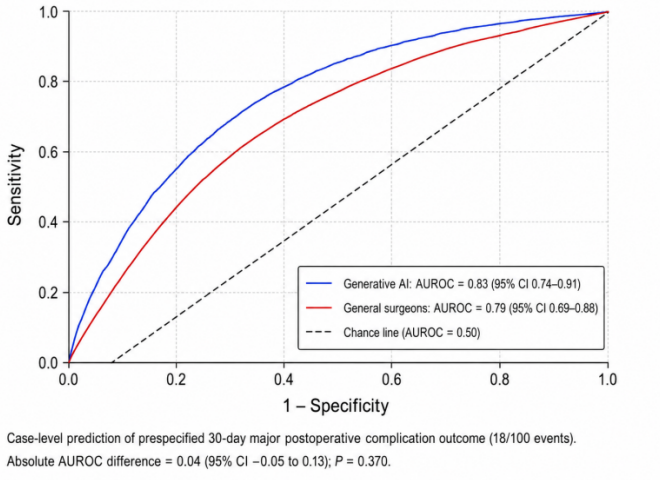
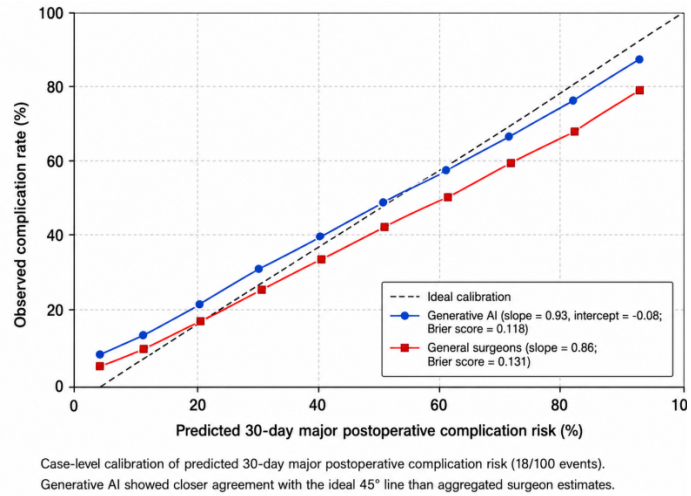


Figure 3. Calibration plot for preoperative risk prediction



DISCUSSION

The GenAI system correctly identified the expert-panel management decision in 85.0% of cases, compared with 81.4% of 1,000 surgeon-case assessments. Although the point estimate favored GenAI, the crossed mixed-effects analysis showed no statistically significant difference (adjusted OR 1.33, 95% CI 0.92-1.92; P=0.128). The observed performance was therefore broadly comparable within this dataset, although the study was not designed as a formal equivalence or non-inferiority trial. GenAI also demonstrated substantial agreement with the expert reference standard (kappa=0.75) and high repeatability across independent runs (93.0% exact decision agreement; kappa=0.88).

The results were directionally consistent with a previous report in which advanced LLMs approached experienced clinicians on selected surgical reasoning tasks.(10) The AUROC of 0.83 for 30-day major-complication prediction was numerically higher than the aggregated surgeon AUROC of 0.79, but the difference was small and non-significant (P=0.370). Likewise, the lower Brier score for GenAI (0.118 vs 0.131) did not reach statistical significance. These findings indicated that favorable discrimination alone did not establish clinical superiority when calibration, case complexity, and safety were considered alongside AUROC.

The subgroup pattern was clinically informative. GenAI accuracy was highest in elective and lower-complexity cases (90.5% and 89.6%, respectively) and lower in emergency and ASA III-IV cases (81.0% and 75.8%). Surgeons showed a similar decline with complexity, and no significant evaluator-by-subgroup interaction was detected. This pattern suggested that clinical complexity was an important source of uncertainty for both human and AI decision-makers. The findings did not support autonomous replacement of surgeons because bedside examination, operative judgment, patient preferences, and professional accountability remained essential components of surgical care.

Safety findings qualified the overall accuracy results. Clinically important errors occurred in 6.0% of GenAI assessments and 7.9% of surgeon assessments, with no significant difference. The GenAI errors included inappropriate urgency classification, incomplete recognition of physiological instability, and one unsupported clinical assumption. Thus, high average accuracy coexisted with recommendations that required expert correction. The large efficiency difference - a median of 17 seconds for GenAI versus 164 seconds for surgeons - suggested value as a rapid second reader, but the clinical usefulness of this speed depended on continued surgeon oversight and transparent handling of uncertainty.

The local setting added practical relevance. In the surgeon cohort, only 26.0% reported formal AI training although 58.0% used GenAI at least monthly, suggesting that adoption had occurred faster than structured education. The findings therefore reinforced the importance of combining technical validation with surgeon education in prompt limitations, hallucination recognition, privacy, documentation, and human accountability, consistent with Saudi AI ethics

principles and applicable digital-health requirements.(20,21)

Strength of the current study is that it involved multicenter participation with 100 real-world case vignettes, 100 surgeons contributing 1,000 independent case ratings, identical information for AI and human evaluators, and independent expert adjudication. It observed 30-day outcomes for risk validation, repeat AI runs, and statistical methods that accounted for clustering by both surgeon and case. The design also captured dimensions omitted by simple examination benchmarks, including calibration, safety errors, timing decisions, repeatability, and efficiency.

However, several limitations were identified. This includes 100 cases and only 18 major-complication events limited precision for rare outcomes and calibration analyses. The second limitation is the vignette-based evaluation which did not reproduce physical examination, evolving physiology, multidisciplinary discussion, operative findings, or patient preferences. Moreover, the expert panel represented a judgment-based reference standard for management decisions. The voluntary surgeon recruitment could have introduced participation bias. Additionally, results from a single locked GenAI version were vulnerable to rapid technological obsolescence as models changed. Finally, the study evaluated performance in offline research setting and therefore did not measure the effect of AI-assisted decisions on actual patient outcomes.

The findings indicated the need for subsequent prospective silent evaluation and carefully governed clinician-in-the-loop trials before any autonomous clinical deployment was considered. Important next steps included evaluating whether surgeon access to GenAI changed management decisions, reduced omissions, improved risk communication, or affected patient outcomes, and validating performance across local regions, non-English language and bilingual prompts, locally grounded retrieval-augmented systems, and changing model versions.

CONCLUSION

Generative AI achieved 85.0% primary management-decision accuracy compared with 81.4% for surgeon ratings, while 30-day complication discrimination was 0.83 versus 0.79 by AUROC. Neither primary comparative difference was statistically significant, whereas GenAI was substantially faster and showed high run-to-run consistency. Clinically important errors nevertheless occurred in 6.0% of GenAI assessments, and performance was lower in emergency and high-complexity cases. Overall, the findings supported GenAI as a supervised adjunct for preoperative risk assessment and surgical decision support rather than as an autonomous substitute for general surgeons.

Declarations

Conflicts of interest: None.

Data availability: De-identified study data were handled in accordance with institutional, ethical, and local data-protection requirements. The approved sharing statement was to reflect the conditions governing the original dataset.

REFERENCES

1. Wan YI, Savonitto S. Improving decision-making for timing of surgery for high-risk comorbid patients. Br J Anaesth. 2025 Jan;134(1):8–10. doi:10.1016/j.bja.2024.10.008
2. Shi G, Xu H, Tang X, Ding Q, Tong Y, Xu Y, et al. Preoperative prehabilitation in adults: A concept analysis. Int J Nurs Stud Adv. 2026 Jun;10:100565. doi:10.1016/j.ijnsa.2026.100565
3. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS): III. Accurate Preoperative Prediction of 8 Adverse Outcomes Using 8 Predictor Variables. Ann Surg. 2016 Jul;264(1):23–31. doi:10.1097/SLA.0000000000001678
4. Plaza NA, Salcedo Echeverry GE, Arenas-Soto AF. Comparative Analysis of Machine Learning Models vs. Traditional Clinical Calculators for Cardiovascular Risk Prediction [Internet]. Cardiovascular Medicine; 2026 [cited 2026 Aug 29]. Available from: http://medrxiv.org/lookup/doi/10.64898/2026.06.11.26355407 doi:10.64898/2026.06.11.26355407
5. Fakhari H, Scherr CL, Moe S, Hoell C, Smith ME, Rasmussen-Torvik LJ, et al. From Calculation to Communication: Using Risk Score Calculators to Inform Clinical Decision Making and Facilitate Patient Engagement. Med Decis Making. 2024 Nov;44(8):900–13. doi:10.1177/0272989X241285036
6. Brennan M, Puri S, Ozrazgat-Baslanti T, Feng Z, Ruppert M, Hashemighouchani H, et al. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. Surgery. 2019 May;165(5):1035–45. doi:10.1016/j.surg.2019.01.002
7. Chudgar N, Yan S, Hsu M, Tan KS, Gray KD, Molena D, et al. The American College of Surgeons Surgical Risk Calculator performs well for pulmonary resection: A validation study. J Thorac Cardiovasc Surg. 2022 Apr;163(4):1509-1516.e1. doi:10.1016/j.jtcvs.2021.01.036 PubMed PMID: 33610360; PubMed Central PMCID: PMC8292429.
8. Wang X, Hu Y, Zhao B, Su Y. Predictive validity of the ACS-NSQIP surgical risk calculator in geriatric patients undergoing

lumbar surgery. Medicine (Baltimore). 2017 Oct;96(43):e8416. doi:10.1097/MD.0000000000008416 PubMed PMID: 29069040; PubMed Central PMCID: PMC5671873.
9. Liu Y, Cohen ME, Hall BL, Ko CY, Bilimoria KY. Evaluation and Enhancement of Calibration in the American College of Surgeons NSQIP Surgical Risk Calculator. J Am Coll Surg. 2016 Aug;223(2):231–9. doi:10.1016/j.jamcollsurg.2016.03.040
10. Livieratos A, Lin J, Chasani P, Gaga M, Fousekis FS, Gogos C, et al. Large Language Models for Clinical Narrative Processing: Methods, Applications, and Challenges. Methods Protoc. 2026 May 1;9(3):69. doi:10.3390/mps9030069 PubMed PMID: 42201162; PubMed Central PMCID: PMC13214763.
11. Palenzuela DL, Mullen JT, Phitayakorn R. AI Versus MD: Evaluating the surgical decision-making accuracy of ChatGPT-4. Surgery. 2024 Aug;176(2):241–5. doi:10.1016/j.surg.2024.04.003 PubMed PMID: 38769038.
12. Vavekanand R, Karttunen P, Xu Y, Milani S, Li H. Withdrawn: Large Language Models in Healthcare Decision Support: A Review [Internet]. Computer Science and Mathematics; 2024 [cited 2026 Aug 29]. Available from: https://www.preprints.org/manuscript/202407.1842/v1 doi:10.20944/preprints202407.1842.v1
13. Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, Grasset Escobar E, Villagrán Gutiérrez I, Achurra Tirado P, et al. Surgeons vs ChatGPT: Assessment and Feedback Performance Based on Real Surgical Scenarios. J Surg Educ. 2024 Jul;81(7):960–6. doi:10.1016/j.jsurg.2024.03.012 PubMed PMID: 38749814.
14. Musbahi O, Nurek M, Pouris K, Vella-Baldacchino M, Bottle A, Hing C, et al. Can ChatGPT make surgical decisions with confidence similar to experienced knee surgeons? The Knee. 2024 Dec;51:120–9. doi:10.1016/j.knee.2024.08.015 PubMed PMID: 39255525.
15. Chinn LW, Nemeh I, Chinn NR. Artificial intelligence-enabled clinical decision support systems in preadmission testing: a scoping review of risk prediction, triage, and perioperative workflows (2020–2025). J Clin Monit Comput. 2026 Jan 31;40(2):525–45. doi:10.1007/s10877-025-01404-w
16. Alba C, Xue B, Abraham J, Kannampallil T, Lu C. The foundational capabilities of large language models in predicting postoperative risks using clinical notes. Npj Digit Med. 2025 Feb 11;8(1):95. doi:10.1038/s41746-025-01489-2
17. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025 Apr 5;8(1):187. doi:10.1038/s41746-025-01519-z PubMed PMID: 40185842; PubMed Central PMCID: PMC11971376.
18. Abdel Malek M, van Velzen M, Dahan A, Martini C, Sitsen E, Sarton E, et al. Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study. Br J Anaesth. 2025 May;134(5):1333–40. doi:10.1016/j.bja.2024.08.038 PubMed PMID: 39547871.
19. Colmenares O, Gómez-Hernández MT, Rivas CE, Fuentes MG, Manama M, Gómez F, et al. Preoperative Factors Associated with Surgical Complexity and Postoperative Outcomes in Patients Undergoing Robotic Anatomical Segmentectomy. Interdiscip Cardiovasc Thorac Surg. 2025 Dec 1;40(12):ivaf294. doi:10.1093/icvts/ivaf294 PubMed PMID: 41347837; PubMed Central PMCID: PMC12782741.
20. Saudi Data and Artificial Intelligence Authority (SDAIA). AI Ethics Principles [Guidelines / Regulatory document] [Internet]. Riyadh, Saudi Arabia: Saudi Data and Artificial Intelligence Authority (SDAIA); 2023 [cited 2026 Aug 29]. Available from: https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf?utm_source=chatgpt.com
21. Saudi Food and Drug Authority (SFDA). Guidance for Artificial Intelligence (AI) and Machine Learning (ML) Technologies Based Medical Devices. Riyadh, Saudi Arabia: Saudi Food and Drug Authority (SFDA). Report MDS-G010.
22. Altowijri AA, Alhemaidi SSM, Alali OA, Alali AA, Alwadai TA, Abubaker NJ, et al. Artificial Intelligence Awareness And Perceptions Among Surgeons In Saudi Arabia :A Cross-Sectional Observational Study. Afr J Biomed Res. 2025. doi:10.53555/AJBR.v28i4S.8421
23. Alhowaish TS, Alshafi NI, Aldekhyyel RN, Mesallam TA, Farahat M, Temsah MH, et al. Generative artificial intelligence use in medical research: A medical student vs. physician survey from Saudi Arabia. Int J Med Inf. 2026 Jul;215:106458. doi:10.1016/j.ijmedinf.2026.106458
24. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022

May;28(5):924–33. doi:10.1038/s41591-022-01772-9

25. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 16;385:e078378. doi:10.1136/bmj-2023-078378 PubMed PMID: 38626948; PubMed Central PMCID: PMC11019967.
26. Sounderajah V, Guni A, Liu X, Collins GS, Karthikesalingam A, Markar SR, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025 Oct;31(10):3283–9. doi:10.1038/s41591-025-03953-8
27. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Ann Surg. 2004 Aug;240(2):205–13. doi:10.1097/01.sla.0000133083.54934.ae PubMed PMID: 15273542; PubMed Central PMCID: PMC1360123.
28. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg. 2013 Nov;217(5):833-842.e1-3. doi:10.1016/j.jamcollsurg.2013.07.385 PubMed PMID: 24055383; PubMed Central PMCID: PMC3805776.
29. Haller G, Bampoe S, Cook T, Fleisher LA, Grocott MPW, Neuman M, et al. Systematic review and consensus definitions for the Standardised Endpoints in Perioperative Medicine initiative: clinical indicators. Br J Anaesth. 2019 Aug;123(2):228–37. doi:10.1016/j.bja.2019.04.041 PubMed PMID: 31128879; PubMed Central PMCID: PMC6676244.
30. Li L, Gu L, Kang B, Yang J, Wu Y, Liu H, et al. Evaluation of the Efficiency of MRI-Based Radiomics Classifiers in the Diagnosis of Prostate Lesions. Front Oncol. 2022 Jul 5;12:934108. doi:10.3389/fonc.2022.934108