




# Systematic Review of Digital Innovations in Surgery: Machine Learning Applications and Implementation Guidelines

Naseralla Juma Elsaadi Suliman<sup>1</sup>  Ateia Hussain Ateia Gaber<sup>2</sup> Nagat Mohamed Abougila Milad<sup>1</sup>

<sup>1</sup>Benghazi Medical Center, University of Benghazi, Benghazi, Libya

<sup>2</sup>The 7th October Hospital, Faculty of Medicine, University of Benghazi, Benghazi, Libya

**Address for correspondence** Naseralla Juma Elsaadi Suliman, PhD, Benghazi Medical Center, University of Benghazi, Benghazi, Libya (e-mail: naseralla.elsaadi@uob.edu.ly).

Libyan Int Medical Univ J

## Abstract

Digital technologies, particularly machine learning (ML), are increasingly integrated into contemporary surgical practice, though implementation barriers remain. This systematic review examined the current evidence on digital innovations in surgery, focusing on ML applications, implementation challenges, and evidence-based implementation strategies. A comprehensive search of seven electronic databases identified 87 studies published between January 2018 and December 2024, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses framework. Evidence synthesis encompassed six domains: artificial intelligence and ML applications, extended reality (XR) technologies, clinician-led innovation, sustainable surgical practices, specialized training models, and nontechnical skills development. ML models demonstrated improved performance in preoperative risk stratification compared with conventional statistical approaches. Receiver operating characteristic analysis showed ML models, including deep neural networks (area under the curve [AUC]  $\approx$  0.82–0.96), random forests (AUC  $\approx$  0.86–0.93), and support vector machines (AUC  $\approx$  0.86), outperformed traditional logistic regression (AUC  $\approx$  0.68–0.74) for predicting postoperative complications. Computer vision algorithms improved procedural precision, while XR technologies (virtual reality/augmented reality) enhanced surgical training, showing skill acquisition comparable or superior to traditional methods. However, substantial implementation barriers were identified, including algorithmic bias, insufficient training in digital competencies, regulatory constraints, and documented concerns regarding bias in nonrandomized studies of XR technologies. Environmental impact assessments revealed that telemedicine applications reduced carbon emissions, whereas robotic surgical systems demonstrated higher resource consumption. The successful integration of digital innovations requires a phased implementation approach, multidisciplinary collaboration, comprehensive competency development, and systematic evaluation of both clinical and operational outcomes. This review provides recommendations for translating digital innovations, addressing technical, regulatory, and human factors.

## Keywords

- ▶ artificial intelligence
- ▶ machine learning
- ▶ extended reality
- ▶ clinician-led innovation
- ▶ implementation guidelines
- ▶ surgical innovation

DOI <https://doi.org/10.1055/s-0045-1814772>.  
ISSN 2519-139X.

© 2026. The Author(s).

This is an open access article published by Thieme under the terms of the Creative Commons Attribution License, permitting unrestricted use, distribution, and reproduction so long as the original work is properly cited. (<https://creativecommons.org/licenses/by/4.0/>)

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

## ملخص المقال باللغة العربية

## مراجعة منهجية للابتكارات الرقمية في الجراحة: تطبيقات التعلم الآلي وتوجهات التنفيذ

**المؤلفون:** نصرالله جمعة الساعدي، عطية حسين جابر، نجاة محمد بوعقبلة. مركز بنغازي الطبي، كلية الطب، جامعة بنغازي، ليبيا.

**المؤلف المسؤول:** نصرالله جمعة الساعدي، [يريد الإلكتروني: naseralla.elsaadi@uob.edu.ly](mailto:naseralla.elsaadi@uob.edu.ly)

يشهد المجال الجراحي المعاصر تزايداً ملحوظاً في دمج التقنيات الرقمية، وخاصة تقنيات التعلم الآلي، رغم استمرار وجود تحديات تعيق التطبيق العملي لهذه التقنية. تستعرض هذه المراجعة المنهجية أحدث الأدلة حول الابتكارات الرقمية في الجراحة، مع التركيز على تطبيقات التعلم الآلي، والعوائق المرتبطة بها، والاستراتيجيات المستندة إلى الأدلة لتطبيقها.

أجري بحث شامل في سبع قواعد بيانات إلكترونية، نتج عنه تحديد 87 دراسة نُشرت بين يناير 2018 وديسمبر 2024، وفقاً لإطار عمل برسما PRISMA. وقد جُمعت الأدلة ضمن ستة محاور رئيسية: تطبيقات الذكاء الاصطناعي والتعلم الآلي، وتقنيات الواقع الممتد (XR)، والابتكار بقيادة الأطباء، والممارسات الجراحية المستدامة، ونماذج التدريب المتخصصة، وكذلك تطوير المهارات غير التقنية.

أظهرت نتائج الدراسات أن نماذج التعلم الآلي حققت أداءً متفوقاً في تصنيف المخاطر قبل الجراحة مقارنة بالأساليب الإحصائية التقليدية. فقد بين تحليل منحنى خصائص تشغيل المستقبل (ROC) أن مساحة تحت المنحنى لشبكات العصبية العميقة (DNN AUC ≈ 0.82–0.96)، والغابات العشوائية (RF AUC ≈ 0.86–0.93)، وآلات المتجهات الداعمة (SVM AUC ≈ 0.86) تفوقت جميعها على الانحدار اللوجستي التقليدي (LR ≈ 0.68–0.74) في التنبؤ بمضاعفات ما بعد الجراحة.

كما ساهمت خوارزميات الرؤية الحاسوبية في رفع دقة الإجراءات الجراحية، بينما عززت تقنيات الواقع الافتراضي والمعزز (VR/AR) التدريب الجراحي، وأظهرت نتائج مماثلة أو أفضل من الطرق التقليدية في اكتساب المهارات.

ورغم هذه الإنجازات، برزت عدة عوائق أمام التطبيق، من أبرزها: التحيز الخوارزمي، ونقص التدريب على الكفاءات الرقمية، والقيود التنظيمية، والمخاوف المتعلقة بالتحيز في الدراسات غير العشوائية الخاصة بتقنيات الواقع المعزز. أما من الناحية البيئية، فقد أظهرت التقييمات أن تطبيقات الطب عن بُعد ساهمت في تقليل انبعاثات الكربون، في حين أن أنظمة الجراحة الروبوتية ارتبطت بزيادة استهلاك الموارد.

خلصت هذه المراجعة إلى أن التكامل الناجح لهذه الابتكارات الرقمية يتطلب: نهجاً تدريجياً في التنفيذ، وتعاوناً متعدد التخصصات، وتطويراً شاملاً للكفاءات الرقمية، وأخيراً تقييماً منهجياً للنتائج السريرية والتشغيلية. وفي الختام، تقدم هذه المراجعة توصيات عملية لترجمة الابتكارات الرقمية إلى واقع تطبيقي، مع مراعاة الأبعاد التقنية والتنظيمية والبشرية لضمان تحقيق أفضل النتائج.

**الكلمات المفتاحية:** الذكاء الاصطناعي؛ التعلم الآلي؛ الواقع المعزز؛ الابتكار بقيادة الأطباء؛ إرشادات التنفيذ؛ الابتكار الجراحي.

## Introduction

### Historical Context and Technological Evolution in Surgery

The evolution of surgical practice has long been closely tied to technological advancements, from the introduction of anesthesia and antisepsis to modern robotic techniques. The surgical field is currently experiencing integration of digital technologies, particularly artificial intelligence (AI) and machine learning (ML), across the entire surgical pathway. This digital integration has the potential to enhance diagnostic accuracy, refine surgical precision, improve resource management, and contribute to improved patient outcomes.<sup>1</sup> This acceleration is driven by the convergence of computational power, large data sets, and algorithmic advances.

### Machine Learning in Surgical Practice

ML, a critical subset of AI in which algorithms learn from data to make predictions or decisions, has demonstrated measurable potential in surgical applications.<sup>2,3</sup> ML applications span the entire surgical pathway, including preoperative planning, intraoperative guidance, and postoperative monitoring. This systematic application supports its utility in clinical decision-making and procedural optimization.

### Extended Reality Technologies in Surgery

Extended reality (XR) technologies represent an umbrella term for immersive technologies that create novel opportunities for surgical training, preoperative planning, and intraoperative guidance.<sup>4</sup> This spectrum includes virtual reality (VR), which creates a fully simulated environment; augmented reality (AR), which overlays digital information; and mixed reality (MR), enabling real-time physical-digital interaction. The integration of these technologies into surgical education and practice has demonstrated potential for enhancing procedural training and intraoperative support.<sup>5,6</sup>

### Implementation Challenges and Research Gaps

Despite documented potential, integrating digital innovations into surgical practice faces challenges spanning technical limitations, governance, ethical uncertainties, and requisite training.<sup>7</sup> Furthermore, the rapid pace of technological digital frequently outpaces the generation of robust, high-quality evidence needed to inform clinical implementation and evidence-based policymaking.

### Objectives and Research Questions

This systematic review assesses the current evidence on digital innovations in surgery by synthesizing findings across six key domains: AI/ML, XR technologies, clinician-led

innovation, sustainable surgical pathways, specialized training, and nontechnical skills (NTS). The review addresses three critical research questions: evaluating the current landscape of digital innovations, specifically ML, across the surgical pathway; identifying the prevailing implementation challenges; and developing evidence-based guidelines for effective, safe, and long-term sustainable integration.

## Methods

### Research Strategy

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines.<sup>8</sup> The focus is on digital innovations in surgery, specifically ML, to develop clinically actionable implementation guidelines.

### Search Strategy Resources

We searched seven electronic databases (PubMed/MEDLINE, Embase, Google Scholar, Cochrane Library, Web of Science, IEEE Xplore, and ACM Digital Library) to combine comprehensive biomedical coverage with targeted capture of engineering and computer-science literature on AI, ML, and XR. Scopus and ScienceDirect were excluded due to substantial content overlap with the selected sources, allowing for focused retrieval of interdisciplinary and technical studies most relevant to implementation.

The search strategy integrated three core conceptual domains: (1) surgery and surgical procedures, (2) digital innovations and technologies, and (3) applications and outcomes. The Boolean search string employed was: (*surgery OR surgical OR surgeon OR operative OR operation OR “minimally invasive surgery” OR “robotic surgery”*) AND (*digital OR technology OR innovation OR “artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network” OR “computer vision” OR “extended reality” OR “virtual reality” OR “augmented reality” OR “mixed reality” OR robotics OR “digital health” OR telemedicine OR “digital medicine”*) AND (*training OR education OR planning OR “decision support” OR outcome OR performance OR safety OR efficiency OR precision OR accuracy OR sustainability OR fellowship OR entrepreneurship OR “non-technical skills”*).

To reflect the rapidly evolving nature of this field, the search was restricted to English-language publications from January 2018 to December 2024.

### Inclusion and Exclusion Criteria

#### Inclusion Criteria

Studies were included if they met all of the following criteria:

Study type: Original research articles, systematic reviews, meta-analyses, and high-quality narrative reviews. Scope: Focus on digital innovations in surgical practice, training, or education. Technologies: Investigation of ML, AI, XR, or related digital technologies within a surgical context. Outcomes: Reporting of outcomes of surgical performance,

patient outcomes, training effectiveness, or implementation challenges.

#### Exclusion Criteria

Studies were excluded for the following reasons:

Publication type: Conference abstracts, letters, editorials, and opinion pieces without original data. Relevance: Focus on basic technical aspects without clear surgical application or clinical translation. Methodological rigor: Insufficient methodological detail for critical appraisal. Duplication: Duplicate publications or multiple reports of the same study. Scope: Primary focus on nonsurgical medical applications.

The eligibility criteria for the review were intentionally focused on high-volume regions, applied without geographical restriction, and based solely on methodological stringency and relevance to the research questions, thereby defining the scope of exclusion for other areas.

### Study Selection Process

A two-stage screening process was employed, wherein two independent reviewers first assessed all titles/abstracts and subsequently evaluated the full-text articles of potentially eligible studies. The PRISMA flow diagram illustrates this sequential process; following screening of 1,247 identified records, 87 studies met the eligibility criteria for the qualitative synthesis.

### Data Extraction

A standardized data extraction form was used to systematically collect information from the included studies. The area under the curve (AUC) values and other predictive performance metrics for various ML models (e.g., deep neural networks [DNNs], random forest [RF], support vector machines) were obtained through systematic extraction from the included primary research articles, and were not generated by code-based ML models, online servers, or other computational methods. The extracted data encompassed: Study characteristics [*author(s), year of publication, country, study design, sample size, and participant characteristics*]; technology characteristics [*type of digital innovation, technical specifications, and implementation details*]; application context [*surgical specialty, stage of surgical pathway, and purpose*]; outcomes [*primary and secondary outcomes, assessment methods, and statistical analyses*]; and implementation factors [*barriers and facilitators, cost implications, and training requirements*].

### Risk of Bias Assessment

Bias risk across all included studies was assessed using validated, design-specific tools. The Cochrane Risk of Bias 2.0 (RoB 2) tool was applied to randomized controlled trials (RCTs).<sup>9</sup> Nonrandomized intervention studies were evaluated using ROBINS-I, which assesses internal validity across seven domains, including confounding and deviations from intended interventions.<sup>10</sup> Qualitative studies were systematically appraised with the Critical Appraisal Skills Program (CASP) Qualitative Checklist, evaluating

methodological rigor across 10 domains such as data analysis and ethical considerations. Systematic reviews underwent independent evaluation using AMSTAR-2, which examines methodological quality across 16 domains, including protocol registration and bias assessment.<sup>11</sup> This comprehensive multitool approach ensured rigorous quality evaluation of primary studies and secondary reviews, strengthening confidence in the synthesized evidence.

### Data Synthesis

Given the substantial heterogeneity in study designs, interventions, comparators, and outcome measures, a narrative synthesis approach was adopted as the primary analytical strategy. The synthesis was structured around six prespecified domains (*mentioned in Objectives and Research Questions subsection*). Within each domain, common themes, patterns, and trends were identified and mapped; contradictory findings were highlighted where present; and gaps in the evidence base were documented.

### Ethical Approval Statement

This systematic review, not involving primary human data collection, did not require institutional ethical approval. The review adhered to ethical principles of transparency, integrity, and methodological rigor; the ethical standards of all included studies were assessed. Author contributions conformed to the CRediT taxonomy, as described in the “References” section.

## Results

### Overview of Included Studies and Study Quality

The systematic search identified 1,247 records, from which 87 studies met the eligibility criteria for final synthesis (►Fig. 1). During initial screening, 800 records were excluded, primarily comprising conference abstracts, letters, or studies with a nonsurgical focus. Subsequently, 447 articles underwent full-text assessment, resulting in the exclusion of 113 records. These exclusions were primarily due to nonsurgical applications, insufficient methodological details, basic technical focus, or duplication.

The studies examined various digital innovations in surgery, with AI and ML being the most common (42%), followed by XR technologies (28%), robotic systems (18%), and other digital health tools (12%). Most research came from high-income countries, particularly the United States ( $n=34$ ), United Kingdom ( $n=19$ ), and Germany ( $n=11$ ). The work covered multiple surgical specialties—general surgery, orthopaedics, neurosurgery, urology, and cardiothoracic surgery—as detailed in ►Table 1.

### Digital Innovations Across the Surgical Pathway

#### Artificial Intelligence and Machine Learning Applications

ML applications, as a critical subset of AI, are integrated across the surgical pathway, with performance metrics extracted directly from the published results of the included primary studies. For diagnostic and prognostic models, the

discriminative ability was quantified using the AUC values, while model accuracy was reported via metrics such as sensitivity, specificity, and Dice scores. These metrics, derived from receiver operating characteristic analysis and segmentation comparisons, respectively, demonstrate the measurable capability of ML to augment surgical decision-making and performance (►Fig. 2).

### Extended Reality Technologies

XR technologies, encompassing VR and MR, show measurable efficacy and significant potential in surgical education and planning. Studies comparing XR platforms to traditional training methods indicated equivalent or improved skill acquisition, particularly for procedures requiring advanced three-dimensional (3D) visualization (►Table 2).<sup>12,13</sup> For preoperative planning, XR facilitates interactive exploration of patient-specific 3D anatomical models, which has been associated with reduced operative times and enhanced team communication. Intraoperatively, MR overlays imaging data onto the surgical field, contributing to improved navigational accuracy and a decrease in procedural errors.<sup>14</sup> XR technologies, including immersive VR simulation, have demonstrated established acceptance and effectiveness in enhancing surgical education and training, as confirmed by comprehensive systematic and umbrella reviews of the field.<sup>15–17</sup>

### Sustainable Surgical Pathways

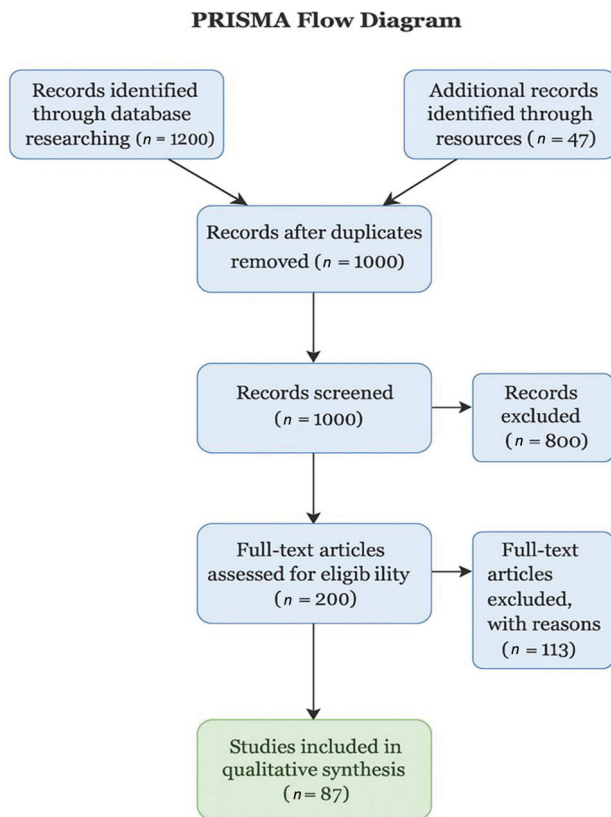
Quantitative analysis demonstrated that digital interventions significantly reduced the environmental impact of surgical care by targeting key pathway elements.<sup>18,19</sup> Specifically, telemedicine for preoperative assessments and postoperative follow-ups reduced travel-related carbon emissions. This was achieved while maintaining high patient satisfaction and comparable clinical outcomes to traditional care (►Fig. 3). Furthermore, digital therapeutics effectively managed conditions that might otherwise necessitate surgery, consequently lowering surgical demand and resource utilization.<sup>20</sup> Conversely, studies confirmed that the carbon footprint of robotic surgery is consistently more resource-intensive than laparoscopic approaches.<sup>21,22</sup>

### Implementation and Curriculum Development

Clinical entrepreneurship emerged as a driver for digital innovation, leveraging unique clinician insight into unmet needs.<sup>23</sup> These clinician-led initiatives demonstrated improvements in operational efficiency and cost reduction. They also enhanced patient satisfaction by streamlining preoperative assessment. Furthermore, these digital efforts augmented intraoperative decision-making and refined postoperative monitoring. Workforce development programs successfully equipped professionals with the necessary business and translational skills.

### Specialized Training and Nontechnical Skills

The rapid integration of digital tools has created a notable gap in the skills provided by traditional surgical curricula,



**Fig. 1** Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram detailing the systematic search and selection process of studies included in the qualitative synthesis. Data were derived from the systematic search conducted across seven electronic databases (*PubMed/MEDLINE, Embase, Web of Science, IEEE Xplore Digital Library, ACM Digital Library, Cochrane Library, and Google Scholar*) and generated using Microsoft PowerPoint 365, adhering to the PRISMA 2020 guidelines. The interpretation: The diagram details the flow of information through the different phases of the systematic review. A total of 1,247 records were identified through database searches, leading to 87 studies included in the final qualitative synthesis. Following the removal of 247 duplicate records, 1,000 unique records underwent title and abstract screening. Of these, 800 records were excluded at the screening phase, primarily due to being nonpeer-reviewed formats (e.g., conference abstracts,  $n = 500$ ) or having a nonsurgical focus ( $n = 300$ ). The remaining 200 full-text articles were assessed for eligibility, resulting in the exclusion of 113 articles for the following reasons: nonsurgical applications ( $n = 40$ ), insufficient methodological detail ( $n = 35$ ), basic technical focus without clinical application ( $n = 25$ ), and duplicate publications ( $n = 13$ ). The final cohort of 87 included studies represents peer-reviewed, full-text publications with adequate methodological rigor and direct relevance to digital innovations in surgical practice.

which currently offer minimal instruction in AI, ML, and XR. Specialized fellowships in robotics, data science, and AI have emerged to address this deficit, yielding demonstrable improvements in technical ability and confidence among participants.<sup>22</sup>

Human factors and NTS remain important in digitally enhanced surgical settings.<sup>19</sup> Although digital tools offer benefits, they can elevate cognitive load. This increased load necessitates vigilance against new error pathways. Therefore, resilient team structures and robust escalation

protocols are necessary to ensure safe technology use. Synthesis of this evidence yielded key guidelines for optimal implementation, including specific considerations for low-resource settings (LRS) (►Table 3).

Assessment of methodological quality revealed distinct variations across different technology types. Studies focusing on robotics and ML/AI generally presented a lower overall risk of bias (RoB). Conversely, many XR studies were flagged with “some concerns,” primarily due to nonrandomized designs, the difficulty in blinding participants and assessors in intervention studies involving immersive technology.<sup>24</sup> While ►Fig. 4 demonstrates efficacy and methodological caveats in XR and ML for surgical training, a comprehensive breakdown of the RoB based on the specific assessment tool used is provided in ►Fig. 5.

## Discussion

This systematic review highlights an increasingly integrated approach in modern surgical practice, which is primarily driven by digital innovations and their core ML applications.

### Machine Learning Applications and Clinical Integration

#### Economic and Implementation Implications of ML

ML models demonstrate strong technical efficacy in surgical risk stratification, evidenced by AUC values frequently exceeding 0.90 for high-stakes tasks (►Fig. 2).<sup>25</sup> However, transitioning these prototypes to routine clinical use requires demonstrating long-term cost-effectiveness alongside accuracy. High-performance metrics suggest that economic benefits are primarily realized through the downstream effects of improved predictive capabilities. Specifically, robust discriminatory power translates to reduced preventable complications, shorter hospital stays, and more efficient management of high-cost resources like intensive care unit beds. This perspective aligns with health economics literature, suggesting the value of surgical AI lies in optimizing the entire surgical pathway. Ultimately, the focus must expand beyond technical performance to ensure sustainable financial viability.

#### Superior Predictive Performance and Technical Challenges

ML's dominance in the literature (42%) reflects research prioritizing technical development over implementation science.<sup>2,3</sup> Deep learning models demonstrate substantial gains in risk stratification, achieving AUC values of 0.96 versus approximately 0.68 for traditional methods.<sup>7</sup> For image-based tasks, convolutional neural networks (CNNs) consistently reach Dice similarity coefficients (DSCs) exceeding 0.85, confirming their effectiveness in tissue segmentation and phase recognition. These models excel by processing high-dimensional data through multilayered transformations, identifying subtle patterns that augment—not replace—surgical expertise in risk assessment and intraoperative guidance.

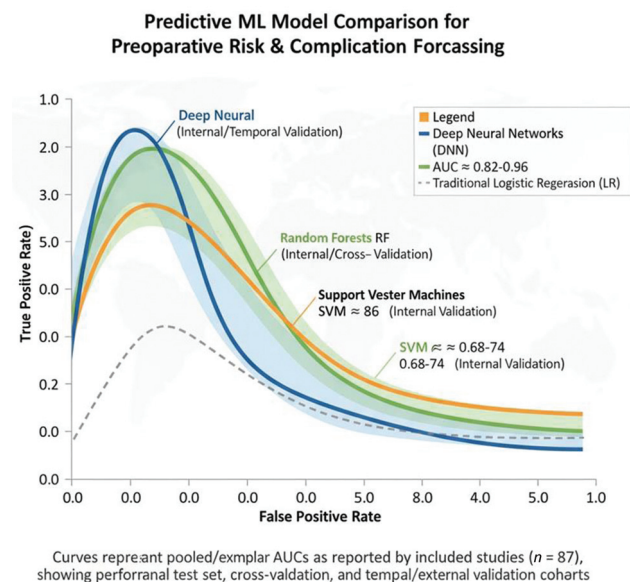
**Table 1** Overview of key characteristics of the 87 studies included in the systematic review (January 2018 and December 2024)

Surgical specialty	Number of studies (%)	Primary focus areas
General surgery	24 (27.6)	Risk prediction, operative video analysis, training
Orthopaedics	18 (20.7)	Preoperative planning, AR guidance, and rehabilitation
Neurosurgery	14 (16.1)	Imaging analysis, navigation, simulation
Urology	12 (13.8)	Robotic surgery, performance metrics, training
Cardiothoracic	9 (10.3)	Decision support, outcome prediction, planning
Other specialties	10 (11.5)	Various applications

Abbreviation: AR, augmented reality.

Note: The data represent the categorization of the 87 included studies based on the primary surgical specialty and the focus area of the digital innovation investigated. The percentages reflect the proportion of the total included studies.

Data source: Categorization and counts derived from the study characteristics extraction table in Excel.



**Fig. 2** (Machine learning [ML] model comparison) A comparison of predictive performance between machine learning models and traditional statistical approaches, showing receiver operating characteristic (ROC) curves with area under the curve (AUC) values. The ML model comparison (ROC curves

and AUC annotations for exemplar tasks: preoperative risk prediction, complication detection). Data source: Study-level AUCs and validation type extracted to an Excel spreadsheet from included articles.

Figure generation: ROC curves plotted and annotated in Python (matplotlib, seaborn); AUC summary table produced in Excel and imported to the plotting script. Note: This figure shows ROC curves per model family comparing the predictive performance of various ML algorithms with annotated AUC values (deep neural networks [DNN AUC  $\approx$  0.82–0.96], random forests [RF AUC  $\approx$  0.86–0.93], and support vector machines [SVM AUC  $\approx$  0.86]) against traditional logistic regression [LR AUC  $\approx$  0.68–0.74] for predicting postoperative complications. The ML models consistently demonstrate higher AUC values (0.82–0.89) compared with logistic regression (0.74).

However, complex algorithms like DNNs raise significant interpretability and transparency concerns. Our synthesis connects technical success to persistent algorithmic bias, which reduces accuracy for underrepresented populations and complicates regulatory approval.<sup>6</sup> While technical efficacy is proven, the critical bottleneck remains bridging high-performing prototypes to equitable clinical deploy-

ment. This necessitates shifting research focus from efficacy to effectiveness, as the World Health Organization advocates.<sup>8</sup>

## Extended Reality in Surgical Training and Practice

### Efficacy and Methodological Caveats in XR Studies

XR technologies (VR, AR, and MR) are measurably influencing surgical education and practice.<sup>18,23</sup> Studies demonstrate XR's efficacy in surgical training, showing skill acquisition equivalent to traditional methods, with improved motion efficiency and reduced errors in VR simulation, supporting established construct validity ( $\rightarrow$  Fig. 4).<sup>16,17,23</sup> However, interpreting these findings requires caution, as bias assessments frequently identified performance and detection bias due to nonrandomized, proof-of-concept study designs.

Critical evaluation through RoB analysis reveals significant methodological quality variations ( $\rightarrow$  Figs. 4 and 5). A substantial proportion of XR studies using nonrandomized designs showed elevated concerns under ROBINS-I and CASP assessments.<sup>18,23</sup> This heightened risk stems from inherent design limitations: prevalence of small-scale feasibility studies, absence of control groups, and practical difficulties in blinding participants. These methodological constraints warrant careful interpretation when translating findings into clinical practice.

Furthermore, the stringent ROBINS-I tool highlights specific struggles in nonrandomized interventions regarding confounding variables and participant selection ( $\rightarrow$  Fig. 5). Consequently, to validate the preliminary promise of XR, there is an urgent need for higher-quality RCTs that address these methodological deficits.<sup>16,17</sup>

In distinct contrast, the reviewed ML literature consistently demonstrates superior predictive performance, particularly within preoperative risk stratification. Models such as DNNs and RFs exhibited high discriminatory power with AUC values up to 0.96, demonstrating improved performance over traditional statistical methods, which typically achieved AUCs around 0.68 ( $\rightarrow$  Fig. 2).<sup>24,25</sup> Furthermore, intraoperative computer vision, notably CNNs, supports real-time precision with DSC frequently exceeding

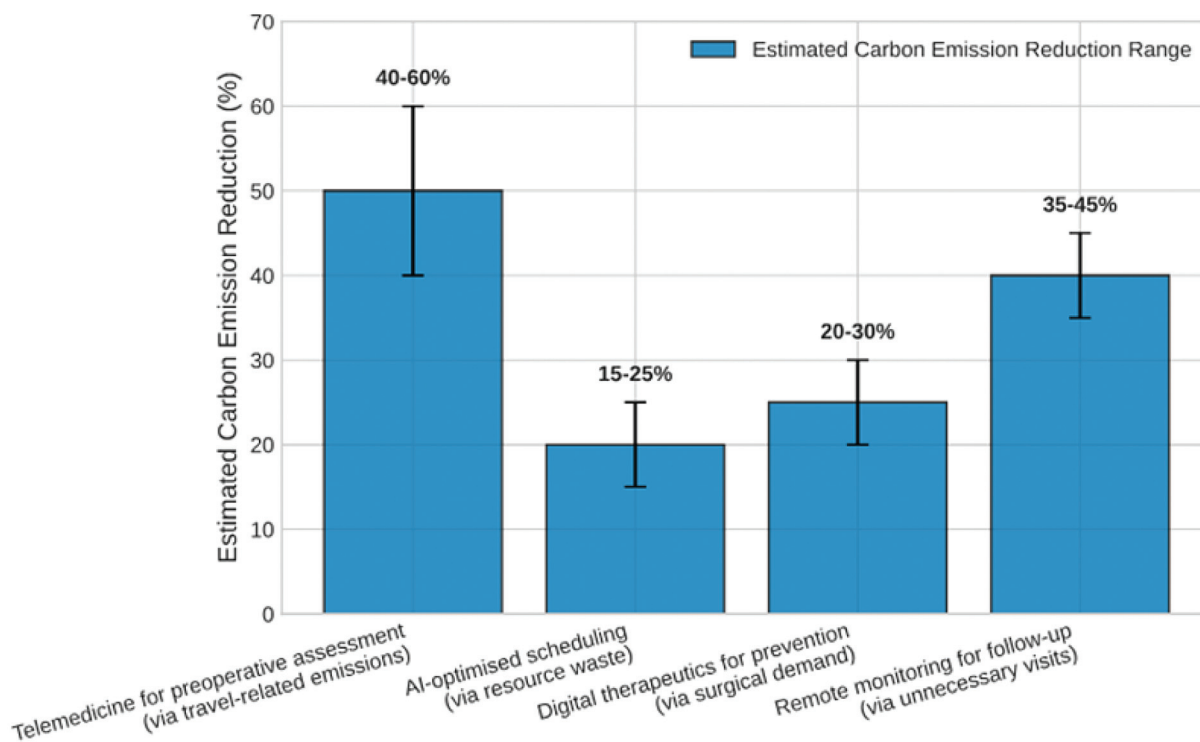
**Table 2** Comparative effectiveness of extended reality (XR)-based versus traditional surgical training methods.

Study	Procedure type	Training method	Performance metrics	Key findings
Huber et al <sup>15</sup>	Laparoscopic cholecystectomy	VR simulation vs. box trainer	Time, errors, motion efficiency	VR group showed 24% fewer errors and 18% better motion efficiency
Co et al <sup>16</sup>	Basic laparoscopic skills	VR simulation vs. standard training	OSATS scores, time to completion	VR group achieved proficiency in 30% less time
Toni et al <sup>17</sup>	Anatomical identification	AR models vs. textbook learning	Knowledge retention, spatial understanding	AR group showed 28% better retention at 6 weeks

Abbreviations: AR, augmented reality; OSATS, Objective Structured Assessment of Technical Skills; VR, virtual reality.

Notes: The table synthesizes key findings from selected studies comparing XR-based training modalities (VR, AR) against traditional training methods. The data highlights differences in performance metrics such as time, errors, and skill retention.

Data source: Extracted outcome measures and summary statistics from primary studies recorded in Excel.



**Fig. 3** (Carbon footprint reduction) A visualization of potential carbon emission reductions achievable through various digital interventions across the surgical pathway. The bar chart estimates carbon emission reductions from digital interventions across the surgical pathway.

Data source: Emission estimates and intervention effect sizes abstracted from included studies and supplemented by published lifecycle assessments; raw values compiled in Excel. Note: Telemedicine for preoperative assessment results in a 40 to 60% reduction in travel-related emissions; artificial intelligence (AI)-optimized scheduling demonstrates a 15 to 25% reduction in resource waste; digital therapeutics for prevention indicate a 20 to 30% reduction in surgical demand; and remote monitoring for follow-up shows a 35 to 45% reduction in unnecessary visits.

0.85.<sup>26,27</sup> However, clinical deployment faces significant constraints related to algorithmic bias, data privacy, and regulatory compliance.<sup>28</sup> Moreover, although ML-driven postoperative monitoring detects complications earlier than standard care, results are complicated by variable data sources and outcome definitions.<sup>29</sup>

#### Clinical Integration and Implementation Barriers

Clinically, XR facilitates the interactive exploration of patient-specific 3D models, a practice associated with reduced operative times and enhanced team communication. Intra-

operative MR overlays further contribute to improved navigational accuracy and a measurable decrease in procedural errors.<sup>14</sup> However, widespread implementation across both ML and XR domains faces multifaceted barriers, including interoperability gaps with electronic health records and limited clinician AI literacy. These challenges are compounded by significant infrastructure constraints, regulatory uncertainty, and a scarcity of large-scale randomized trials. Finally, the concentration of evidence in high-income nations limits global generalizability and raises distinct equity concerns regarding these technologies.

**Table 3** Evidence-based guidelines for implementing digital innovations in surgery

Implementation domain	Key recommendations	Evidence strength
Strategic approach*	Adopt phased implementation beginning with low-risk applications	Strong
Stakeholder engagement	Ensure multidisciplinary collaboration, including surgeons, technologists, and administrators	Strong
Training	Develop comprehensive programs covering technical operation and underlying principles	Moderate
Evaluation	Implement systematic assessment of clinical and operational impacts	Moderate
Workflow integration	Design technologies to seamlessly integrate with existing processes	Strong
Ethical considerations	Proactively address data privacy, informed consent, and algorithmic transparency	Strong
Sustainability	Balance technological advancement with environmental impact	Emerging

\*Prioritize low-cost, high-impact digital tools (e.g., telemedicine, diagnostic aids on mobile devices) over capital-intensive technologies like surgical robotics to ensure sustainability and equitable access. Data source: Recommendations synthesized from narrative synthesis and recorded in an Excel evidence matrix linking recommendations to source studies.

### Efficacy and Methodological Caveats in XR and ML for Surgical Training



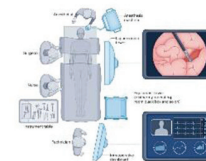
#### Extended Reality: Efficacy with Methodological Limitations

XR technologies demonstrate equivalent or superior skill acquisition compared to traditional methods<sup>[22][23][20][19]</sup>, with measurable improvements in motion efficiency and reduced errors in VR simulation.

**Evidence Base Maturity:** Less mature than ML; significant portion of studies show higher ROBINS-I/CASP concerns<sup>[22][23]</sup>

**Methodological Weakness:** High Risk of Bias stems from non-randomized, proof-of-concept, or small-scale feasibility studies and difficulty in blinding participants<sup>[19][20]</sup>

**Recommendation:** Higher-quality, randomized controlled trials are necessary to justify widespread, evidence-based curriculum changes.



#### Machine Learning: Superior Performance with Implementation Barriers

ML models demonstrate superior predictive performance across the surgical pathway, particularly in preoperative risk stratification<sup>[10][1]</sup>.

**Discriminatory Power:** AUC up to 0.96 vs. traditional methods (AUC ~0.68)

**Intraoperative Precision:** CNNs achieve high performance in tissue segmentation and surgical phase recognition, with DSC frequently exceeding 0.85<sup>[2][13]</sup>.

**Constraints:** Algorithmic bias reduces accuracy for underrepresented populations; complex data privacy and regulatory hurdles<sup>[15]</sup>

**Fig. 4** Efficacy and methodological caveats in extended reality (XR) and machine learning (ML) for surgical training. XR technologies offer equivalent or superior surgical skill acquisition, though evidence maturity is limited by methodological bias in nonrandomized studies. Conversely, ML models demonstrate superior predictive performance, with risk stratification reaching area under the curve (AUC) 0.96 (vs. logistic regression [LR] AUC 0.68) and segmentation accuracy at Dice similarity coefficient (DSC) > 0.85. However, ML implementation is hindered by algorithmic bias and complex regulatory hurdles. These findings underscore the need for higher-quality randomized controlled trials for XR technologies and robust governance frameworks for ML implementation in surgical practice.

## Implementation and Workforce Development

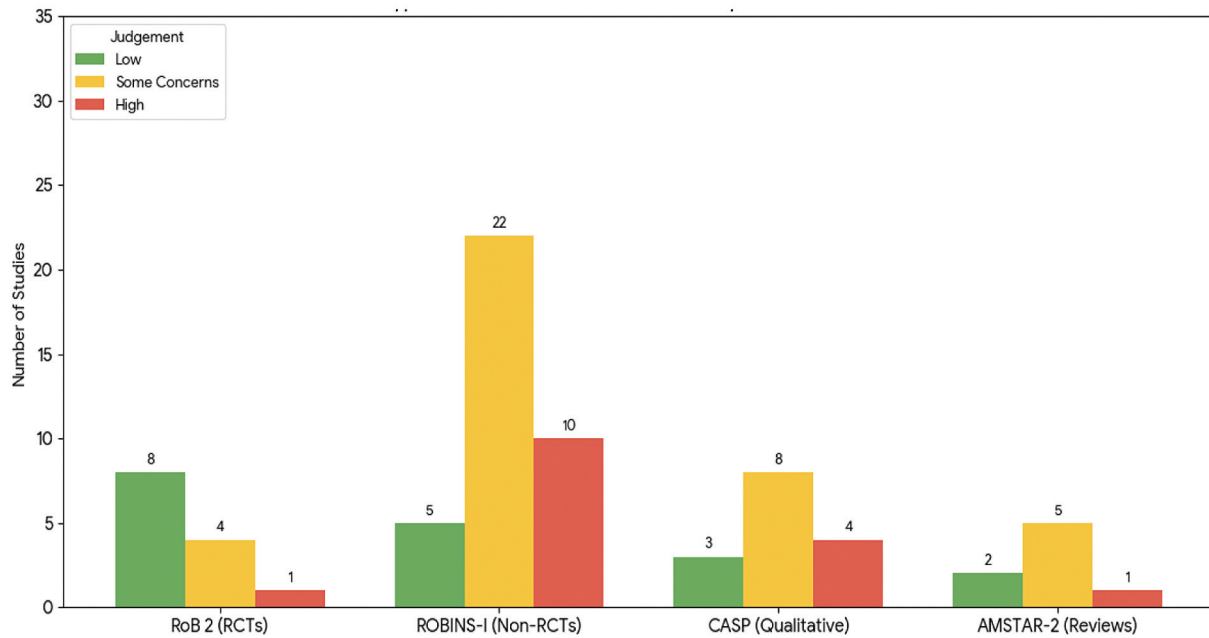
### Clinical Entrepreneurship as a Driver for Innovation

Clinical entrepreneurship emerges as a significant driver for digital innovation, underscoring the importance of clinician leadership. Surgeons contribute unique insights into clinical needs and workflow constraints, ensuring that technologies address genuine problems. This clinician-led innovation is particularly critical for the sustainable adoption of digital surgery in diverse global settings, where context-appropriate, are often more effective than high-cost commercial systems. The codesign of tools by clinicians, engineers, and informaticians accelerates translation from prototype to practice and mitigates unintended consequences.<sup>30</sup> For example, a surgeon in a resource-constrained environment

facing high rates of surgical site infections due to inconsistent postoperative wound care documentation could develop a simple, secure, and locally hosted Web form or a shared cloud-based spreadsheet. This low-cost digital tool, accessible via existing hospital Wi-Fi, standardizes the collection of wound-check data, automates the calculation of a risk score, and triggers an alert for high-risk patients.

### Training, Nontechnical Skills, and Digital Readiness

The rapid integration of digital tools has exposed a persistent gap between traditional surgical training and the competencies required in the digital era.<sup>22,31</sup> Embedding digital literacy, data ethics, and human factors into core curricula is essential, while specialized fellowships and AI-robotics pairing effectively address this deficit through automated,



**Fig. 5** (Bar chart) Risk of bias (RoB) distribution by assessment tool. The graph breaks down the risk of bias by the specific tool used for assessment, which is crucial for understanding the context of the judgments. Data source: Counts aggregated from the same Excel extraction sheet. The chart displays the number of studies categorized as low (green), some concerns (yellow), or high (red) risk of bias for each research tool: 1. RoB 2 (randomized controlled trials [RCTs]): Assesses randomized controlled trials. 2. ROBINS-I (non-RCTs): Assesses nonrandomized studies. 3. Critical Appraisal Skills Program (CASP) (qualitative): Assesses qualitative research. 4. AMSTAR-2 (reviews): Assesses systematic reviews. Interpretation: This shows that the highest risk of bias and “concerns” are associated with nonrandomized studies (assessed by ROBINS-I), which is a common finding. The reviews (AMSTAR-2) and RCTs (RoB 2) generally fared better, though the number of these high-quality studies was lower.

objective feedback that enhances skill acquisition beyond subjective traditional assessment.<sup>31</sup>

Crucially, NTSs—communication and situational awareness—remain vital in digitally enhanced settings. However, complex tools risk increasing cognitive load and introducing new error pathways. Mitigating these risks requires resilient team structures, clear role delineation, and robust escalation protocols as operational necessities for advanced surgical technologies. Finally, successful ML integration faces two critical nontechnical challenges: algorithmic bias and data privacy/regulatory hurdles.<sup>19</sup> Persistent concerns regarding data protection and information governance mandate secure, encrypted systems, underscoring that technical advancement alone cannot ensure safe, equitable deployment without addressing these foundational barriers.

## Implementation Guidelines and Sustainability

### Implementation Framework

The long-term viability of digital innovations requires robust ethical and regulatory frameworks.<sup>32</sup> The synthesis of our evidence yields a pragmatic, safety-oriented framework for implementation, which begins with the staged implementation of low-risk applications to enable iterative refinement and local evidence generation. Success is predicated on multidisciplinary collaboration among surgeons, technologists, administrators, and patients, ensuring governance aligns with clinical and operational goals. Comprehensive training is essential to ensure user readiness, moving beyond technical operation to include conceptual understanding and

critical output interpretation. Furthermore, robust evaluation via balanced scorecards tracking safety, efficacy, and equity supports continuous improvement, while workflow integration prioritizes interoperable systems to avoid duplicative standalone technologies. Proactive ethical and regulatory compliance, addressing data privacy and algorithmic transparency, is nonnegotiable for building trust.

### Sustainability and Low-Resource Settings

Sustainability must be embedded in adoption strategies, as digital interventions can measurably reduce the surgical pathway’s environmental impact.<sup>20</sup> Telemedicine for preoperative assessment offers the largest travel-related carbon emission reduction (40–60%), with remote monitoring providing 35 to 45% reduction (► Fig. 3). Digital therapeutics and AI-optimized scheduling further contribute through reduced surgical demand and resource waste (20–30% and 15–25% reduction, respectively). Conversely, capital-intensive technologies like robotic surgery have higher carbon footprints, necessitating careful balancing of technological advancement with environmental impact.<sup>18,21</sup>

Given that current evidence predominantly originates from high-income settings, specific adaptation strategies for LRS are essential. Implementation in LRS faces critical infrastructural constraints: inconsistent power, limited high-speed Internet, and absent data infrastructure.<sup>1</sup> Adaptation must therefore prioritize resilience and local capacity through low-bandwidth telemedicine solutions, open-source or locally developed AI models requiring minimal computational power, and phased, modular technology integration approaches.

### Limitations of the Review

This review's strengths include its comprehensive scope across six fields of digital innovation and its emphasis on practical implementation guidelines. However, several limitations must be acknowledged. First, the rapid pace of technological evolution means that some evidence may become outdated quickly, necessitating ongoing surveillance. Second, significant heterogeneity in study designs, interventions, comparators, and outcome measures constrained opportunities for quantitative synthesis and meta-analysis, limiting the ability to estimate pooled effect sizes. Third, the predominance of studies from high-income countries may restrict generalizability to resource-constrained settings, where differences in infrastructure and regulatory frameworks substantially impact the feasibility and equity of implementation. Finally, future reviews should aim to include non-English databases to create a more globally representative evidence base.

### Implications for Practice

The successful translation of digital innovation requires targeted action from all stakeholders. For surgeons, maintaining clinical efficacy necessitates cultivating digital literacy, adapting to redefined workflows, and retaining strong critical appraisal skills for patient-centered care. Health care organizations must pair investments in secure infrastructure and interoperability with governance structures that integrate clinical and technical perspectives to ensure safety and equity. Technology developers are required to adopt user-centered design principles and ensure transparency regarding performance boundaries, failure modes, and data usage to foster trust. Finally, policymakers and regulators must urgently establish adaptive frameworks to balance patient safety with innovation, supporting robust evaluation and clear guidance on data governance.

### Future Directions

The future direction of digital surgery research requires four key priorities to ensure responsible and effective integration.<sup>1</sup> First, rigorous, high-quality effectiveness studies, including RCTs, are urgently needed to evaluate clinical and cost-effectiveness in real-world settings. Second, scalability and sustainability research must focus on comparative implementation studies across diverse resource settings to inform adaptation strategies. Third, longitudinal impact assessments are necessary to assess the long-term effects on the surgical workforce, patient outcomes, and team dynamics, with an emphasis on equity and resilience. Finally, ethical inquiry must prioritize accountability, transparency, and equity by actively reducing algorithmic bias and safeguarding patient data to prevent deepening existing disparities.

### Conclusion

The review establishes that ML and digital innovations augment surgical practice across critical domains, enhancing training, operative planning, performance, and patient out-

comes. Successful integration, however, is contingent upon addressing critical challenges in technical interoperability, regulatory clarity, and generating robust clinical evidence. The guidelines derived from this analysis emphasize a strategic, staged implementation approach, supported by multidisciplinary collaboration, comprehensive training, and rigorous, ethical evaluation.

The practical application of these findings necessitates a dual focus on policy and education. Institutionally, efforts must formalize digital readiness and establish sustainable infrastructure; simultaneously, the surgical curriculum requires integration of AI literacy and data science principles. Policy-level frameworks are essential for the evidence-based validation and ethical deployment of AI tools, prioritizing patient safety and data governance. Responsible integration of digital technologies will define the next era of surgical care. Thoughtful adoption can enhance quality and efficiency while preserving the clinical judgment essential to surgical excellence.

### Funding

None.

### Conflict of Interest

None declared.

### Acknowledgments

The authors extend their sincere gratitude to colleagues for their invaluable support in helping with data collection and constructive critical review of the manuscript. The language corrections and improvements in this article were assisted by Google Gemini and QuillBot, AI language models, to ensure clarity, grammatical accuracy, and readability. The AI was used solely for language polishing and did not contribute to scientific content, data analysis, or conclusions.

### References

- Morris MX, Fiocco D, Caneva T, Yiapanis P, Orgill DP. Current and future applications of artificial intelligence in surgery: implications for clinical practice and research. *Front Surg* 2024;11:1393898
- Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg* 2018;268(01):70–76
- Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science - from concepts toward clinical translation. *Med Image Anal* 2022;76:102306
- Mao RQ, Lan L, Kay J, et al. Immersive virtual reality for surgical training: a systematic review. *J Surg Res* 2021;268:40–58
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(01):44–56
- Hung AJ, Chen J, Gill IS. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg* 2018;153(08):770–771
- Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689
- Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372(71):n71

- 9 Flemyng E, Moore TH, Boutron I, et al. Using Risk of Bias 2 to assess results from randomised controlled trials: guidance from Cochrane. *BMJ Evid Based Med* 2023;28(04):260–266
- 10 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:14898
- 11 Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 2017;358:j4008
- 12 Bunogerane GJ, Taylor K, Lin Y, Costas-Chavarri A. Using touch surgery to improve surgical education in low- and middle-income settings: a randomized control trial. *J Surg Educ* 2018;75(01):231–237
- 13 Meling TR, Meling TR. The impact of surgical simulation on patient outcomes: a systematic review and meta-analysis. *Neurosurg Rev* 2021;44(02):843–854
- 14 Aguilar-Salinas P, Gutierrez-Aguirre SF, Avila MJ, Nakaji P. Current status of augmented reality in cerebrovascular surgery: a systematic review. *Neurosurg Rev* 2022;45(03):1951–1964
- 15 Huber T, Paschold M, Hansen C, Wunderling T, Lang H, Kneist W. New dimensions in surgical training: immersive virtual reality laparoscopic simulation exhilarates surgical staff. *Surg Endosc* 2017;31(11):4472–4477
- 16 Co M, Chiu S, Billy Cheung HH. Extended reality in surgical education: a systematic review. *Surgery* 2023;174(05):1175–1183
- 17 Toni E, Toni E, Fereidooni M, Ayatollahi H. Acceptance and use of extended reality in surgical training: an umbrella review. *Syst Rev* 2024;13(01):299
- 18 Parsons D, MacCallum K. Current perspectives on augmented reality in medical education: applications, affordances and limitations. *Adv Med Educ Pract* 2021;12:77–91
- 19 Shoham MA, Baker NM, Peterson ME, Fox P. The environmental impact of surgery: a systematic review. *Surgery* 2022;172(03):897–905
- 20 Dang A, Arora D, Rane P. Role of digital therapeutics and the changing future of healthcare. *J Family Med Prim Care* 2020;9(05):2207–2213
- 21 de'Angelis N, Conso C, Bianchi G, et al; CERES (Collectif Eco-REsponsabilité en Santé) Systematic review of carbon footprint of surgical procedures. *J Visc Surg* 2024;161(2S):7–14
- 22 Rizan C, Steinbach I, Nicholson R, Lillywhite R, Reed M, Bhutta MF. The carbon footprint of surgical operations: a systematic review. *Ann Surg* 2020;272(06):986–995
- 23 Hudise JY, Mojiri ME, Shawish AM, et al. The role of virtual reality in advancing surgical training in otolaryngology: a systematic review. *Cureus* 2024;16(10):e71222
- 24 Cekic E, Pinar E, Pinar M, Dacinar A. Deep learning-assisted segmentation and classification of brain tumor types on magnetic resonance and surgical microscope images. *World Neurosurg* 2024;182:e196–e204
- 25 Wang J, Tozzi F, Ashraf Ganjouei A, et al. Machine learning improves prediction of postoperative outcomes after gastrointestinal surgery: a systematic review and meta-analysis. *J Gastrointest Surg* 2024;28(06):956–965
- 26 Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018;15(11):e1002686
- 27 Wijnberge M, Geerts BF, Hol L, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020;323(11):1052–1060
- 28 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447–453
- 29 Lam K, Abramoff MD, Balibrea JM, et al. A Delphi consensus statement for digital surgery. *NPJ Digit Med* 2022;5(01):100
- 30 Konda NN, Lewis TL, Furness HN, Miller GW, Metcalfe AJ, Ellard DR. Surgeon views regarding the adoption of a novel surgical innovation into clinical practice: systematic review. *BJS Open* 2024;8(01):zrad141
- 31 Dovramadjiev T, Dimova R, Dimov D, Manolova P. Human-Centered Innovations in Healthcare Education: Digital Skills for Ergonomics and Bioengineering Advancement. In: *International Conference on Computer and Communication Engineering*. Cham: Springer Nature Switzerland May 24, 2024:209–216. Doi: 10.1007/978-3-031-71079-7\_17
- 32 Chau M. Ethical, legal, and regulatory landscape of artificial intelligence in Australian healthcare and ethical integration in radiography: a narrative review. *J Med Imaging Radiat Sci* 2024;55(04):101733