https://doi.org/10.1038/s44484-026-00017-3

# Guidance on AI-enhanced surgical practice: a Delphi consensus on ontology, data, implementation and evaluation

Check for updates

Justin W. Collins[1,2,3] ✉, Marco Montesi[4,5], Danail Stoyanov[3,6,7], Carla M. Pugh[8], Marcio C. Moschovas[4], Denny Yu[9], Shady Saikali[4], Jeffrey S. Levy[10,11], Zafer Tandogdu[1,2,3], Filippo Filicori[12], Dimitrios Stefanidis[13], Swaroop S. Vedula[14], Martin A. Martino[15] & Gretchen P. Jackson[16,17]

AI-enhanced surgical care lacks standardized frameworks for implementation, evaluation, and governance, limiting safe and scalable adoption. This study integrates a systematic review, expert consensus meeting, and accelerated Delphi process involving 50 expert stakeholders to address these gaps. Consensus was reached with high reliability in over 150 elements defining standards in ontology, data, implementation, and evaluation. The resulting guidance emphasizes rigorous study design, real-world evidence, and lifecycle oversight, providing a structured, consensus-based framework to support clinicians, leaders, and policymakers in responsibly implementing and evaluating AI in surgical practice. Across all stakeholders, success depends on maintaining a closed-loop system in which evaluation continuously informs both data strategy and implementation. Operational adoption may be approached in phases, beginning with institutional readiness and governance, followed by pilot deployments, scaled integration, and ultimately the development of continuous learning systems. This consensus on guidance is intended to serve as an operational blueprint, enabling coordinated, multidisciplinary implementation that ensures AI enhances surgical performance while maintaining safety, trust, and clinical accountability.

Artificial intelligence (AI) is increasingly transitioning from an experimental capability to clinical translation in modern surgical care[1]. The integration of AI technologies into surgery in the past 5 years has increased by more than 400%, rapidly evolving the way patients are treated and services are delivered globally[2]. Advances in machine learning, computer vision, and real-time data integration have enabled AI-enhanced systems to support surgeons across the full perioperative pathway, including preoperative planning, intraoperative guidance, and postoperative monitoring[3]. These technologies promise to augment human performance rather than replace it, with the aim of improving decision-making, procedural precision, workflow efficiency, and ultimately patient outcomes. As healthcare systems face increasing procedural complexity, workforce constraints, and expectations for safer, more consistent care, AI-enabled surgical solutions are being positioned as a key workable solution to deliver scalable, high-quality surgical practice[4].

The potential benefits of AI-enhanced surgical care are substantial. Predictive analytics can assist with patient selection and risk stratification[5], image-based algorithms can enhance anatomical visualization and surgical navigation[6], and automated monitoring systems can provide early detection of complications[7]. Collectively, these capabilities offer opportunities to reduce variability in surgical performance, shorten learning curves, and support evidence-based intraoperative decisions. When appropriately integrated, AI systems may also contribute to improved resource utilization, reduced operative times, and more personalized surgical care[8].

However, the implementation of AI in surgical settings introduces significant technical, clinical, ethical, and organizational risks. Algorithmic

[1]Division of Surgery & Interventional Science, University College London, London, UK. [2]Division of Uro-Oncology, University College London Hospital, London, UK. [3]University College London, London, UK. [4]AdventHealth Global Robotics Institute, Celebration, FL, USA. [5]Urology Unit of Fondazione Policlinico Gemelli, IRCCS, Rome, Italy. [6]UCL Hawkes Institute, London, UK. [7]Medtronic plc, Minneapolis, MN, USA. [8]Stanford University, Stanford, CA, USA. [9]Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN, USA. [10]Department of Surgery, Indiana University, Indianapolis, IN, USA. [11]Institute for Surgical Excellence, Washington, DC, USA. [12]Northwell Health, New York, NY, USA. [13]Department of Surgery, Indiana University School of Medicine, Indianapolis, IN, USA. [14]Malone Center for engineering in healthcare, Johns Hopkins University, Baltimore, MD, USA. [15]Ascension St Vincent's, Jacksonville, FL, USA. [16]Intuitive Surgical, Sunnyvale, CA, USA. [17]Department of Pediatric Surgery, Vanderbilt Health, Nashville, TN, USA. ✉e-mail: justin.collins@ucl.ac.uk

**Table 1 | Important hierarchical areas/levels to collect data to predict outcomes of surgery**

| Level | Description | Data to collect |
|---|---|---|
| Level 0 | Ground truth. Reference data to assess 'success'<br>Highest level temporal component of an operation: includes access, procedural specific surgical objectives, closure<br>(what was planned and what were identified risks, benefits – metadata) | 1. Pre-operative data: Patient characteristics, prehabilitation program, pre-op planning from imaging<br>2. Operative data: planned procedure (**agreed metrics – segmentation, each with defined start and end points**)<br>3. Post-operative data: planned recovery protocol, outcome data (oncological and functional) |
| Level 1<br>(OR data) | Procedure specific segmentations<br>(what actually happened) | Annotation of procedure for phases, steps, tasks completed, landmarks identified, errors and pre-errors |
| Level 2<br>(OR data) | Procedure specific tasks<br>('Bridge' between annotated data and automated data collection) | Tasks defined as triplet of 'instrument, verb, target' |
| Level 3<br>(OR data) | Machine collected data. Automated collection of surgical gestures and force<br>(tissue handling) | Kinematic data, haptic data etc |

**Table 2 | Recommended data to collect to aid AI enhanced surgical care (levels 0 to 3)**

| Level | Description | Generic data to collect (% agreement on importance to record) |
|---|---|---|
| Level 0 | **Patient specific** | BMI (92%)<br>Previous history of abdominal surgery- same site as planned surgery (94%)<br>Previous history of radiotherapy -same site as planned procedure (86%)<br>Size of tumor (88%)<br>Site of tumor (84%)<br>Anticoagulation (86%)<br>Anatomical/Vascular anomalies (84%)<br>Complexity of surgery related to vascularity (88%)<br>American Society of Anaesthesiologists - ASA score (82%) |
| Level 1<br>(OR data) | **Time specific:**<br>Recommended time points to record to aid efficiency assessments | Patient enters OR or anaesthetic room (86%)<br>Patient anaesthetized and in OR (86%)<br>Patient draped and positioned (78%)<br>Robot docked (88%)<br>Knife to skin (84%)<br>Console time starts (84%)<br>Console time ends (82%)<br>Transfer of patient from OR table to recovery (86%)<br>"Turnaround time" - Patient leaves OR till next patient enters OR or anaesthetic room (88%) |
| Level 2<br>(OR data) | **Procedure specific** | Tasks defined as triplet of 'instrument, verb, target' e.g., right needle holder, place suture at 6'o clock, right hand forehand under on bladder neck, going from outside to inside |
| Level 3<br>(OR data) | **Procedure specific** | Automated machine data such as kinematic data and computer vision[18,45] |

bias, data limitations in diversity, volume and quality, lack of model explainability, and challenges in validation across diverse patient populations raise concerns regarding safety and equity[9]. Dependence on AI-driven recommendations may alter clinical judgment, accountability, and team dynamics in the operating theatre[10]. In addition, regulatory uncertainty, cybersecurity risks, interoperability constraints, and the need for robust governance frameworks complicate real-world deployment. These challenges underscore the importance of a cautious, evidence-driven adoption rather than technology-led enthusiasm. Additionally, many AI-driven capabilities are medical devices or software utilized through a medical device and hence must adhere to medical device regulation and requirements for validation and verification.

This paper examines the implementation of AI-enhanced surgical care with a balanced focus on its potential benefits and associated risks. By analysing current data sources, applications, implementation considerations, and emerging governance approaches, our aim was to create consensus based expert recommendations that will provide structured guidance frameworks for clinicians, healthcare leaders, and policymakers seeking to optimize their current data resources and to responsibly implement and evaluate AI technologies into surgical practice[11].

## Results

### Formulation of guidance

We had 100% (50/50) response rate in all three rounds of the Delphi. There was high inter-rater reliability, which was >0.80. After three rounds of Delphi surveys, a consensus was obtained in more than 150 elements in the areas being considered: Ontology; Data; Implementation; and Evaluation.

### Guidance on ontology

Ontologies provide a consistent vocabulary and formal data collection structure. There was 100% agreement within the panel that that a common ontology will aid data interpretation and ability to action data and 98% agreement that a common ontology will be beneficial to data collection from different settings and the application of AI. To achieve this the panel agreed (98%) that it is important that the ontology is aligned with both engineering and clinical frameworks and thinking.

Agreement ranged from 88 to 100% on the following important elements to annotate in surgical video: generic segmentation (relevant to different procedures) to include patient positioning and closure, procedure specific segmentation, tasks (subcomponent), action (verb), errors and key anatomical landmarks (target). There was agreement that recording audio communications from surgeon (92% agreement) and bedside nurse (86% agreement) is important for nontechnical skills evaluation.

There were high levels of agreement on the important hierarchical areas/levels to collect data to predict outcomes of surgery (summarized in Table 1) and 96% agreement that detail collected in each level should be granular enough to be able to be aligned with data collection from the hierarchical rows above and below. Important generic data to be collected for 'Level 0' and 'Level 1' is summarized in Table 2 with the percentage agreement in brackets. The hierarchical nature of an individual surgical

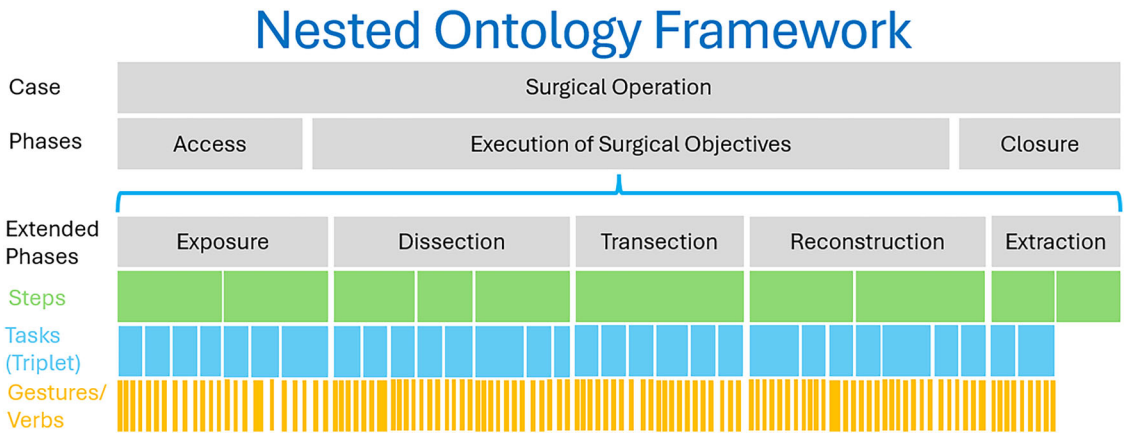


**Fig. 1** | Description of a nested ontology framework.

operation is summarized in Fig. 1, in the modified 'nested ontology framework'[12], There was agreement that elements of an operation can be generic such as the 'extended phases', or procedure specific such as the tasks (level 2).

When recording and to aid in avoiding errors, it was felt it was important to consider errors that are procedure specific errors (100%), errors that are shared across procedures i.e., generic errors that can be universally applied to all or subsets of multiple procedures (100%), the level of severity of errors for common procedures (90%), identify pre-error events e.g., moving instruments out of view; wrong instrument; cautery time; bleeding; dropping needle (94%), errors that impact patient outcomes versus not (94%) and to identify technical malfunction errors (92%).

There were high levels of consensus on multiple areas of peri-operative care that would benefit from predictive AI: patient selection for individual surgeon (90%), patient selection to optimize list utilization i.e., matching complexity with OR team expertise (92%), prehabilitation to optimize pre-op status (92%), improving patient turn around in the OR (98%), enhanced recovery in the hospital (88%), home monitoring with patient wearables (94%), follow-up planning (88%).

When we collect patient related data, this can be collected manually or automatically[13,14], the panel recommended that to ensure accuracy of data collection, this should be verified by two or more data sources when feasible. For home monitoring including patient wearables, recommended biometric data to be collected included: temperature (92%), heart rate (94%), BP (92%), Mobility (96%), exercise performance (88%), spirometry (80%), sleeping patterns (86%). For Patient reported outcome measure (PROMs) recommended data to be collected included: mobility (86%), drinking (80%), eating (82%), flatus and bowel function (84%), wound pain (86%), medications taken (88%), nausea (82%), spirometry (88%), signs of infection (82%).

### Guidance on data use

Identified key areas for data: (1) Technical infrastructure for data acquisition, storage and access, (2) data annotation and sharing, (3) data analytics, and (4) aspects related to clinical translation. Consensus findings are summarized in Table 3.

### Guidance on implementation

Identified key areas for improving the success of implementation of AI enhanced Intraoperative decision support were: (1) maximizing adaptability, (2) adequate training for safe implementation, (3) Design, (4) Support, (5) Fostering buy-in and (6) mitigating risk. Consensus findings are summarized in Table 4.

### Guidance on evaluation

Identified key areas for evaluation: (1) Clinical and Surgical Effectiveness, (2) Technical Reliability and Safety, (3) Human–AI Interaction and Workflow Integration, (4) Data Quality and Algorithmic Validity, (5) Ethical, Legal, and Regulatory Considerations, (6) Data Privacy, Security, and Governance, (7) Implementation, Cost, and Scalability and (8) Evaluation and Research Design. The consensus findings are summarized in Table 5.

## Discussion

Historically the safety and efficacy of surgery have depended on the skills of the surgeon, operating room staff and appropriate patient selection[15,16]. With the advent of objectively recorded outcome data available in electronic patient records and data rich operative devices such as robotic surgery, we have new opportunities to enhance decision making and skills performance with AI[17]. However, AI is only as good as the data it is derived from and the 'art of surgery' has always been susceptible to subjective analysis. The risk of "garbage in, garbage out" (GIGO) can be reduced through improved data annotation in robotic surgery using a domain-specific ontology established by consensus, thereby enabling better interoperability and long-term system reliability[12,18].

Ontology is fundamental to achieving true interoperability in AI because it provides a shared set of concepts and relationships, with formally specified vocabulary that ensures data, models, and software components carry consistent meaning across organizational and technical boundaries[19]. This is especially important due to the inherently interdisciplinary nature of AI-enhanced surgery. Without an explicit objectively defined ontology, different datasets may use the same terms to describe different concepts or different terms for the same concept, leading to ambiguity, brittle integrations, and unsafe downstream inferences. Formally specified and empirically validated ontology frameworks regarding surgical practice remain an evolving area. However, the 'Nested ontology framework'[12] is empirically validated by application to >13 K cases. While no universally adopted ontology currently exists the emerging "nested ontology" approaches developed in the context of surgical data science and annotation for computer vision have proposed hierarchical representations that link procedural phases, steps, tasks, gestures, within a unified structure[12].

A rigorously constructed ontology encodes domain knowledge, relationships, constraints, and hierarchical structures in a machine-interpretable format. This enables semantic alignment across heterogeneous data sources and supports reliable data integration, model portability, and reproducibility[20]. Formalizing surgical activities as triplets of the used instruments, actions performed, and target anatomies is becoming a gold standard approach for surgical activity modeling and recording[21]. In complex environments such as healthcare, where AI must interact with electronic records, imaging systems, registries, and clinical workflows, ontological clarity is not merely a technical convenience but a prerequisite for safety, regulatory transparency, and scalable deployment. By grounding AI development in shared semantic structures, ontology transforms interoperability from a fragile mapping exercise into a principled, extensible

**Table 3 | Consensus recommendations for guidance on data**

| Priority | |
|---|---|
| **Technical infrastructure** | **Average rating** |
| 1. Healthcare Institutions should enhance health information technology, including dedicated personnel resources, to ensure that relevant data can be prospectively, routinely and systematically acquired and stored | 4.64 |
| 2. Develop standards for data storage with respect to key aspects including data structure, format and longevity | 4.58 |
| 3. Health information technology in healthcare institutions should be enhanced so that data is relatable to clinical context and can be easily transferred between acquisition, storage and display systems | 4.54 |
| 4. Enhance health information technology to enable real-time or near-real-time AI inference in interventional settings | 4.44 |
| 5. Build local and international collaborations and partnerships involving clinical, research, public regulatory and industry stakeholders to implement the goals in accordance with internationally agreed standards | 4.42 |
| 6. Strategy for data acquisition should facilitate new intraoperative imaging methods for obtaining relevant information on tissue function, morphology and pathology | 4.12 |
| **Data annotation and sharing** | |
| 1. Establish annotation standards and protocols | 4.66 |
| 2. Establish best practices for assessing and assuring annotation quality | 4.60 |
| 3. Establish standardized ontologies for Surgical Data Science (SDS) | 4.50 |
| 4. Establish platforms and workflows for sharing data in a controlled manner | 4.46 |
| 5. Establish standards for identifying and addressing data biases using structured metadata | 4.46 |
| 6. Develop and disseminate openly accessible platforms to support adherence to annotation standards and protocols | 4.38 |
| 7. Incentivize adherence to best practices, standards and protocols relevant to data annotation and sharing | 4.38 |
| 8. Establish new methods for efficient data annotation | 4.34 |
| 9. Create more incentives for stakeholders to share and annotate data | 4.24 |
| **Data analytics** | |
| 1. Develop methods and validation frameworks focused on transparency and explainability | 4.50 |
| 2. Develop frameworks and workflows for facilitating regulatory approval of machine learning (ML) algorithms | 4.40 |
| 3. Develop methods and validation frameworks focused on the understanding of surgical processes, including variability resulting from technical-, personnel-, patient- and environment-specific factors | 4.38 |
| 4. Develop methods and validation frameworks focused on understanding, avoiding and dealing with biases | 4.36 |
| 5. Develop methods and validation frameworks focused on fast inference (AI models producing outputs quickly enough to enable real-time or near real-time clinical use) | 4.34 |
| 6. Develop methods and validation frameworks focused on robustness including generalization | 4.32 |
| 7. Develop methods and validation frameworks focused on learning efficiently from limited data | 3.90 |
| **Achieving clinical translation** | |
| 1. Develop widely accepted frameworks for obtaining hospital consent to data sharing | 4.58 |
| 2. Ensure high-quality external validation of AI-enhanced surgical practice applications | 4.44 |
| 3. Promote widely accepted frameworks to encourage industry to consent to data sharing | 4.42 |
| 4. Establish liability and medical insurance regulations for AI-enhanced surgical practice applications | 4.40 |
| 5. Develop widely accepted frameworks for obtaining patient consent to data sharing | 4.38 |
| 6. Improve transfer of knowledge among different stakeholders involved in developing AI-enhanced surgical practice | 4.36 |
| 7. Develop new performance metrics for AI algorithms that measure clinically relevant parameters not accounted for by existing metrics | 4.36 |
| 8. Allocate economic, infrastructural and personnel resources within healthcare institutions to support AI-enhanced surgical practice applications | 4.32 |
| 9. Build public trust in AI-enhanced surgical practice | 4.24 |
| 10. Catalyze clinical translation by generating visible and impactful success stories related to AI-enhanced surgical practice | 3.98 |
| 11. Establish AI-enhanced surgical practice, as a career path in healthcare institutions/universities | 3.76 |

foundation for trustworthy intelligent systems[12,22]. These frameworks enable alignment between technical performance, cognitive decision-making, and data annotation, supporting both AI model development and clinical evaluation.

Safe implementation of AI-enhanced intraoperative decision support in surgery depends not only on technical performance but on thoughtful integration into the surgical ecosystem[23]. Systems must be designed to maximize adaptability, accommodating variation in patient factors, operative techniques, and institutional workflows without introducing rigidity at critical moments[24]. Comprehensive standardized training for surgeons, anaesthesiologists, and operating room teams is essential to ensure appropriate interpretation of AI outputs, clear understanding of system limitations, and preservation of independent clinical judgment[25].

Human-centered design principles should consider important issues such as clarity, minimal cognitive load, and intuitive interfaces to reduce distraction and prevent automation bias, while robust technical and clinical support structures ensure rapid troubleshooting and iterative refinement[26,27]. Equally important is fostering buy-in, engaging clinicians early in development and implementation to build trust, transparency, and shared ownership[28]. Finally, proactive risk mitigation including validation of the AI application in real-world settings with associated relevant metadata[22], contingency planning, cybersecurity safeguards, and continuous

**Table 4 | Consensus recommendations for guidance on implementation of AI**

| Priority | |
|---|---|
| **Maximizing adaptability** | **Average rating** |
| 1. Intraoperative AI decision support systems should be designed to function flexibly across diverse operating room environments and hardware setups | 4.34 |
| 2. AI systems should make explicit use cases (e.g., preoperative planning, intraoperative decision making, postoperative review) | 4.20 |
| 3. Integration with existing operating room infrastructure should be prioritized to minimize logistical challenges and setup complexity | 4.10 |
| 4. AI systems should support both manual and automatic toggling of overlays to accommodate varied use cases and comfort levels | 4.04 |
| 5. AI output should offer customizable interface options | 3.96 |
| 6. Adaptability in physical configuration e.g., display positioning is essential for institutional adoption | 3.86 |
| **Training** | |
| 1. Training should emphasize that AI decision support serves as an adjunct to—not a replacement for—clinical judgment | 4.74 |
| 2. Training for surgeons should include understanding of applicable populations, evidence of performance, and failure modes or limitations of the model | 4.66 |
| 3. Appropriate role-specific training should be mandatory prior to clinical implementation (e.g., surgeons, nurses, biomedical engineers, trainees) | 4.64 |
| 4. Physical or digital troubleshooting guides should be readily available for reference during surgical cases | 4.48 |
| 5. Hands-on simulations should be used to build familiarity and confidence before clinical deployment | 4.28 |
| 6. Ongoing refresher training and competency determination should be required to maintain safe and effective use when appropriate | 4.22 |
| **Design** | |
| 1. AI interfaces should be simple and intuitive to use | 4.74 |
| 2. Interface design should focus on reducing cognitive load and avoiding information overload during critical phases of surgery. | 4.66 |
| 3. Seamless integration of the AI tool into the surgical workflow is critical to successful adoption. | 4.66 |
| 4. Design should be refined through iterative testing involving multidisciplinary teams that understand the use case | 4.58 |
| 5. Visual cues (e.g., “Go,” “Slow,” or “Caution” zones) should be contextually clear | 4.46 |
| **Support** | |
| 1. Dedicated technical and educational support should be readily available during initial implementation phases as appropriate | 4.70 |
| 2. Vendors or technology providers should maintain open communication with surgical teams to address real-world challenges | 4.64 |
| 3. Continuous performance monitoring and appropriate software updates should be completed to maintain optimal performance | 4.62 |
| 4. Support structures should be established for maintenance, troubleshooting and data governance | 4.58 |
| 5. Feedback mechanisms should be built into the system to enable collection user experiences and feedback to guide iterative improvement | 4.42 |
| 6. Regular assessment of user satisfaction and outcomes should inform ongoing refinement of the technology | 4.26 |
| **Fostering buy-in** | |
| 1. Building and maintaining trust with clinicians is critical for widespread adoption and sustained utilization of intraoperative AI | 4.86 |
| 2. Early involvement of key stakeholders (e.g., surgeons, nurses, biomedical engineers) is essential for successful adoption | 4.78 |
| 3. A culture of safety and openness should be cultivated to normalize use of AI assistance without stigma or fear of judgment | 4.70 |
| 4. Building and maintaining trust between clinicians and patients is critical for widespread adoption and sustained utilization of intraoperative AI | 4.60 |
| 5. Institutional endorsement and leadership support are critical for widespread adoption and sustained utilization | 4.58 |
| 6. Transparent communication about the algorithm’s development, accuracy, and validation should be available to foster trust among users | 4.56 |
| 7. Local champions should be identified across professional roles to promote engagement and disseminate best practices | 4.48 |
| 8. Endorsement by relevant professional or specialty societies is critical for widespread adoption and sustained utilization | 4.42 |
| **Mitigating risk** | |
| 1. AI decision support must be explicitly introduced as an adjunct to clinical judgment to prevent overreliance | 4.50 |
| 2. AI interfaces should display uncertainty levels transparently to inform human oversight when appropriate | 4.48 |
| 3. Policy frameworks should inform roles and accountability in cases of AI-influenced surgical decisions when appropriate | 4.44 |
| 4. Overreliance and distraction risks should be intentionally mitigated when appropriate | 4.32 |

monitoring will be crucial to protect patient safety and maintain professional accountability[29]. Together, these considerations transform AI from a novel tool into a responsibly embedded component of high-reliability surgical care.

Rigorous evaluation of AI in surgery is essential to ensure that innovation translates into genuine clinical value rather than unintended harm[18]. Evaluation should include standardized objectives measures of successful utilization that delivers value[30]. For example, in a recent systematic review, existing evaluations of Large Language Models (LLMs) mostly focused on accuracy of question answering for medical examinations, without consideration of real patient care data. Dimensions such as fairness, bias, and toxicity and deployment considerations received limited attention. It was recommended that future evaluations should adopt standardized applications and metrics, use real world clinical data, and broaden focus to include a wider range of tasks and specialties[31].

Comprehensive evaluations should include studies of technical performance, usability, workflow integration, and health impact, and adaptive AI solutions undergo repeated evaluation as systems learn and evolve[32,33]. Assessment should establish clinical and surgical effectiveness,

**Table 5 | Consensus recommendations for guidance on evaluation of AI**

| Priority | |
|---|---|
| **Clinical and surgical effectiveness** | **Average rating** |
| 1. AI-driven guidance or decision-support tools must demonstrate clinically acceptable accuracy, sensitivity, and specificity before being implemented in surgical workflows | 4.74 |
| 2. AI systems used intraoperatively should clearly define their role—whether providing real-time decision support or functioning solely in a non-decisional, supportive capacity | 4.68 |
| 3. AI performance should be assessed across different procedures and patient populations to determine whether clinically significant degradation occurs under specific conditions | 4.62 |
| 4. There should be evidence-based consensus on what constitutes "ground truth" and surgical expertise for evaluating AI performance | 4.52 |
| 5. AI integration should demonstrate measurable improvements in key surgical outcomes (e.g., complication rates, operative time, procedural precision, and recovery time) | 4.44 |
| 6. External validation using independent real-world, modeled (synthetic) or combination of both datasets should be mandatory prior to clinical deployment | 4.40 |
| 7. AI effectiveness should be evaluated differently for preoperative, intraoperative, and postoperative decision support | 4.36 |
| 8. AI system performance should be benchmarked against expert surgeons and established standards of care using validated outcome metrics | 4.14 |
| **Technical reliability and safety** | |
| 1. Reliability and responsiveness of human override mechanisms when AI guidance is incorrect, delayed, or unsafe | 4.86 |
| 2. Assessment of potential AI system failure modes and the mechanisms for detecting and mitigating these failures | 4.82 |
| 3. Validation and safety testing of software updates before they are deployed in clinical settings | 4.72 |
| 4. Availability of redundancies or fallback processes if the AI system fails during a procedure | 4.70 |
| 5. Acceptable model performance thresholds guided by tangible benefits should be defined by developers and clinical stakeholders before clinical deployment | 4.56 |
| 6. The ability of the AI system to manage uncertainty or incomplete data (e.g., occlusion, poor imaging, hardware malfunction) without compromising safety | 4.54 |
| 7. Robustness of AI performance against adversarial attacks or unexpected data perturbations should be evaluated | 4.38 |
| 8. AI systems should provide real-time confidence, or uncertainty estimates with their outputs | 4.38 |
| 9. Defined expectations for how often the AI model require recalibration or retraining and who is responsible for oversight | 4.22 |
| **Human–AI interaction and workflow integration** | |
| 1. AI integration should minimize additional time, task burden, and workflow disruption | 4.58 |
| 2. Adequacy of training for surgeons and staff to interpret AI outputs and override them when necessary | 4.58 |
| 3. Influence of AI tools on enhancing or negatively distracting surgeon situational awareness during critical procedural steps | 4.54 |
| 4. Clarity and effectiveness of communication protocols governing AI alerts, recommendations, and team response | 4.38 |
| 5. Compatibility of the AI system with existing surgical platforms, imaging devices, OR workflows, and information systems | 4.28 |
| 6. Design of AI systems on the impact of real-time versus non-real-time feedback on performance, learning, and safety should be evaluated | 4.24 |
| 7. Alignment of AI task allocation with surgeon expertise to prevent under-reliance or over-reliance ("automation bias") | 4.18 |
| 8. Differential effects of AI integration on individual surgeon performance versus team-based performance should be evaluated | 4.16 |
| 9. Evaluation of how AI integration alters the roles and responsibilities of surgical team members | 4.16 |
| 10. Measurement and mitigation of cognitive load associated with AI-assisted surgery | 4.12 |
| **Data quality and algorithmic validity** | |
| 1. Processes for continuous post-market monitoring to detect performance drift or bias | 4.78 |
| 2. Quality, diversity, and representativeness of the data used to train and validate the AI system | 4.70 |
| 3. Feedback mechanisms that allow clinical outcomes to inform future model updates | 4.70 |
| 4. Use of standardized nomenclature, taxonomy, and definitions to support reproducibility and validation should be evaluated | 4.64 |
| 5. Transparency regarding the scope, limitations, and known biases of training, testing and validation datasets should be evaluated | 4.64 |
| 6. Performance consistency of the AI system across diverse patient populations, institutions, surgical environments, and equipment types | 4.58 |
| 7. Use of expert clinician-reviewed or annotated datasets during AI development | 4.22 |
| **Ethical, legal, and regulatory considerations** | |
| 1. Clarification of legal responsibility when an AI-assisted action results in patient harm (e.g., surgeon, institution, developer) | 4.80 |
| 2. Compliance of the AI system with relevant regulatory frameworks (FDA, EMA/MDR, EU AI Act, GMC, etc.) | 4.72 |
| 3. Application of ethical principles—autonomy, beneficence, non-maleficence, justice—in AI-supported surgical care | 4.60 |
| 4. Clarity of documented limitations, contraindications, and boundaries-of-use for AI-assisted surgical tools | 4.60 |
| 5. Processes to identify and manage conflicts of interest related to industry partnerships, data ownership, or commercial incentives | 4.50 |
| 6. There should be alignment of regulatory requirements with clinical workflow realities and risk stratification for practical implementation | 4.50 |
| 7. Ethical management of algorithmic and human bias in AI-supported surgical decision-making should be evaluated | 4.44 |

**Table 5 (continued) | Consensus recommendations for guidance on evaluation of AI**

| Priority | |
|---|---|
| **Clinical and surgical effectiveness** | **Average rating** |
| 8. Transparency and interpretability of AI decision-making processes | 4.42 |
| 9. Extent to which patients are informed and provide meaningful consent for AI use during their procedures | 3.98 |
| **Data privacy, security, and governance** | |
| 1. Cybersecurity protections against unauthorized access, tampering, or manipulation of AI-enabled surgical systems | 4.82 |
| 2. Security and privacy protocols governing the storage, processing, anonymization, and protection of intraoperative data (e.g., video, sensor streams) | 4.70 |
| 3. Providing accessible standardized labels describing AI system training, testing, and validation datasets, intended use, and known limitations | 4.58 |
| 4. Auditability of AI systems for traceability, accountability, and performance review | 4.56 |
| 5. Clarity of data ownership and control for information generated during AI-assisted surgery | 4.50 |
| 6. Transparency and enforceability of data-sharing agreements across hospitals, vendors, and developers | 4.44 |
| **Implementation, cost, and scalability** | |
| 1. Potential for AI-assisted surgery to reduce long-term healthcare costs through improved efficiency or clinical outcomes | 4.64 |
| 2. Barriers to adoption such as workflow disruption, surgeon acceptance, cost constraints, or technical limitations | 4.58 |
| 3. Scalability of AI solutions across institutions with varying infrastructure, staffing, or resource levels | 4.50 |
| 4. Level of surgeon confidence and trust in AI recommendations, including factors that influence adoption and consistent use | 4.48 |
| 5. Demonstrated return on investment (ROI) for health systems is important | 4.28 |
| 6. Costs associated with implementing and maintaining AI systems, including hardware, licensing, and staff training | 4.20 |
| 7. Presence of institutional leadership (e.g., Chief AI Officer) to guide AI implementation and governance | 3.98 |
| **Evaluation and research design** | |
| 1. Clarity and appropriateness of outcome metrics—including clinical, operational, and human factors—used to assess AI performance | 4.80 |
| 2. Inclusion of multiple stakeholders (clinicians, engineers, ethicists, regulators, and patients) in evaluation design | 4.62 |
| 3. The evaluation should include multidisciplinary stakeholders and include data quality, model performance, human factors, and clinical outcomes | 4.58 |
| 4. Incorporation of patient-reported outcomes where feasible | 4.58 |
| 5. Use of rigorous study designs (e.g., randomized trials, prospective cohorts, simulation studies) to evaluate AI in surgery | 4.46 |
| 6. Use of independent audits or third-party validation to confirm AI system safety, accuracy, and clinical effectiveness | 4.34 |
| 7. Have agreements in place to guide management of potential conflicts of interest (e.g., industry partnerships) | 4.32 |

demonstrating meaningful improvements in outcomes, precision, or efficiency, while concurrently verifying technical reliability and safety under real-world conditions[34,35]. Equally critical is understanding human–AI interaction and workflow integration, since even highly accurate systems can fail if they disrupt operative flow or degrade team cognition[36]. Evaluation must scrutinize data quality and algorithmic validity, including bias, generalizability, and performance across diverse patient populations, to avoid inequitable or unsafe deployment[37]. Broader considerations include ethical, legal, and regulatory compliance will frame accountability and transparency, while robust data privacy, security, and governance safeguards maintain public trust[9,24]. As AI continues to evolve, there is growing emphasis on transparency and interpretability in clinical contexts[38].

To translate this consensus into practice, the recommendations should be applied as an integrated, iterative framework across the four domains of Ontology, Data, Implementation, and Evaluation. These domains are designed to function as a continuous cycle of define, deploy, measure, and refine to support safe, scalable, and evidence-based adoption of AI in surgical care. A defining strength of this consensus is the deliberate inclusion of a multidisciplinary group of stakeholders, including robotic and minimally invasive surgeons, surgical society representatives, AI experts, researchers, industry leaders, and regulatory authorities. This breadth of representation is essential, as effective implementation requires alignment across clinical reality, technical validity, operational feasibility, and regulatory oversight. Accordingly, the recommendations should be interpreted as cross-domain guidance grounded in real-world feasibility, and this multidisciplinary model should be replicated at the institutional level through structured governance and oversight committees.

In practice, clinicians should adopt AI tools within clearly defined workflows, maintain active clinical judgment, and participate in structured feedback and safety reporting. Researchers should prioritize prospective and real-world evaluation, incorporating human–AI interaction metrics and standardized methodologies. Healthcare institutions should establish multidisciplinary governance structures, invest in interoperable data infrastructure, integrate AI into existing workflows, and implement continuous monitoring for safety and performance. Developers should focus on human-centered design, transparency, robust validation across diverse environments, and ongoing lifecycle management, including drift detection and recalibration. Across all stakeholders, success depends on maintaining a closed-loop system in which evaluation continuously informs both data strategy and implementation.

Operational adoption may be approached in phases, beginning with institutional readiness and governance, followed by pilot deployments, scaled integration, and ultimately the development of continuous learning systems. Overall, this consensus document is intended to serve as an operational blueprint, enabling coordinated, multidisciplinary implementation that ensures AI enhances surgical performance while maintaining safety, trust, and clinical accountability. Importantly, the strong expert consensus across domains underscores both the urgency and feasibility of establishing shared standards for AI integration in surgery. Together, the 4 domains of ontology, data, implementation and evaluation with governance, form a cohesive framework that supports the safe, effective, and trustworthy integration of AI into surgical care.

One important aspect that we have not accounted for is the cost, pricing or business considerations for developing, deploying and maintaining such systems and devices. This requires further studies and understanding of the medical device market in different regions to understand the fiscal viability. Future AI-enhanced surgical practice applications will

require prospective trails to ensure wide-spread adoption is safe, reproducible and advantageous. This consensus guidance will require further validation.

In conclusion the practical adoption of AI-enhanced surgical practice requires careful attention to implementation strategy, cost, and scalability to ensure that benefits justify investment and can be sustained across diverse institutions. Rigorous evaluation and research design, including standardized ontology, prospective trials, real world evidence generation, and continuous post-market monitoring, are essential to produce credible and reproducible evidence.

## Methods

### Evidence acquisition

The study was carried out in three phases:

1. A steering group was formed to review the literature and summarize the current evidence for AI-Enhanced Surgical Practice.
2. A larger expert panel convened and discussed the important aspects of AI-Enhanced Surgical Practice related to Ontology, Data, Implementation and Evaluation, based on the current evidence. Following presentations and open group discussions, a survey was created, with input from the panel members.
3. Thirdly, panel-based consensus findings were determined using an online Delphi process to formulate guidance and provide recommendations for future research.

### Review of the literature

Our systematic review was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA)[39].

A comprehensive literature search was conducted up to December 5, 2025, using the Medline and Scopus databases. Supplementary material provides a detail research string and filters used for the research.

The study protocol was registered in the International Prospective Register of Systematic Reviews database [PROSPERO registration for this study, record number **CRD420261279890**].

Two independent reviewers (M.M. and J.C.) conducted an initial screening of the studies. Relevant articles identified from the reference lists of selected papers were also included. Each study underwent a full-text review to determine eligibility. Discrepancies were resolved by consensus among the reviewers.

### Inclusion and exclusion criteria

The inclusion criteria were chosen according to PICOS (Population, Intervention, Comparator, Outcome, Study design) framework[40].

Studies were included if they involved surgical teams, or surgical simulation models (Population) in which artificial intelligence was used to enhance any aspect of surgical practice, including intraoperative guidance, workflow analysis, automation, decision support, or training (Intervention), with or without a comparison to standard non-AI surgical practice (Comparator), and reported technical, workflow, implementation, performance, or surgical outcomes (Outcomes). Peer-reviewed original research articles written in English and indexed in MEDLINE and Scopus or listed in ClinicalTrials.gov were included.

### Data extraction

A dedicated data sheet built for the study was used to collect the data of interest. The following data were extracted for analysis: Study characteristics (year, country, design), surgical specialty and procedure, AI application domain, type of AI system, dataset characteristics, outcome measures. There were no disagreements on the data collection.

### Results of review

The initial search identified 241 records, of which 3 were duplicates. After title and abstract screening, 120 studies were assessed for eligibility and 59 were excluded, most commonly because the AI application was not related to surgery (see Fig. 2 below).

### Evidence synthesis

The systematic review of 61 studies reveals that artificial intelligence is increasingly applied to enhance surgical practice, particularly in the domains of: workflow recognition, intraoperative guidance, surgical skill assessment and training and simulation.

Most studies focused on implementation and performance evaluation, while fewer addressed data infrastructure, annotation methodologies, and standardization. Across specialties, AI demonstrated promising accuracy in recognizing surgical phases, assessing technical performance, and supporting intraoperative decision making. However, heterogeneity in study design, outcome measures, and datasets limited direct comparison across studies.

### Risk of bias assessment

Given the methodological heterogeneity of the included studies, design-specific risk of bias tools were applied according to study type. Diagnostic accuracy studies were evaluated using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool[41].

Randomized clinical trials were assessed using the Cochrane Risk of Bias tool for randomized trials 2 (RoB 2), and non-randomized interventional studies were evaluated using the Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool[42,43].

Overall, the risk of bias across studies ranged from low to moderate, with most concerns related to small sample sizes, limited external validation, and potential confounding.

### Expert panel

An advisory panel was formed, which comprised societal leads, global key opinion and industry leaders from the fields of surgery, regulation, engineering and computer science with a specialist interest in AI-enhanced surgical practice. In total, 50 experts were brought together to discuss and develop guidance on AI-enhanced surgical practice related to ontology, data, implementation and evaluation. The panel breakdown included the 12 co-authors and the other 38 contributors mentioned in the acknowledgements. This included 37 from the US from 14 different states, 12 from Europe from 4 different countries, and 1 from Israel. Professions included surgeon (52%), engineer (34%), researcher in AI (36%), involved in education and training (30%), there was also a qualified lawyer with expertise in this field.10 surgical specialties were represented. Of the academics listed as co-authors the median H-index and i10-index were 31 (range 3–78) and 72 (range 1–342) respectively. During the panel hybrid meeting, that was chaired by J.L. and J.C., the evidence identified in the literature was presented to inform panel opinion on the current state of AI use in surgery.

### Internet survey and Delphi process

At the in-person meeting all panel members consented to contribute to all 3 rounds. The Delphi process was conducted to drive consensus among the experts. An Internet survey (Google forms) was generated based on the current literature and expert opinion and sent to the 50 committee members. The Delphi was divided into 3 main sections. The Supplementary material shows a full list of the survey questions. An accelerated e-consensus reaching exercise, over 3 consecutive days, using the Delphi methodology was then applied. The Delphi method organizes group communication in a structured manner to enable participants to effectively address complex problems. Questions in which there was ≥80% consensus were removed from the next round of the survey. Repeated iterations of anonymous voting continued over three rounds, where an individual's vote in the next round was informed by knowledge of the entire group's results in the previous round. To be included in the final recommendations, each survey item had to have reached group consensus (≥80% agreement) by the end of the three survey rounds. Ranking/prioritization questions (Tables 3–5) were asked only once to the group. In the Delphi process, the finding of "consensus" is more relevant than the level of consensus. Reliability of the formulated guidance was evaluated using Cronbach alpha to assess internal consistency among experts[44].

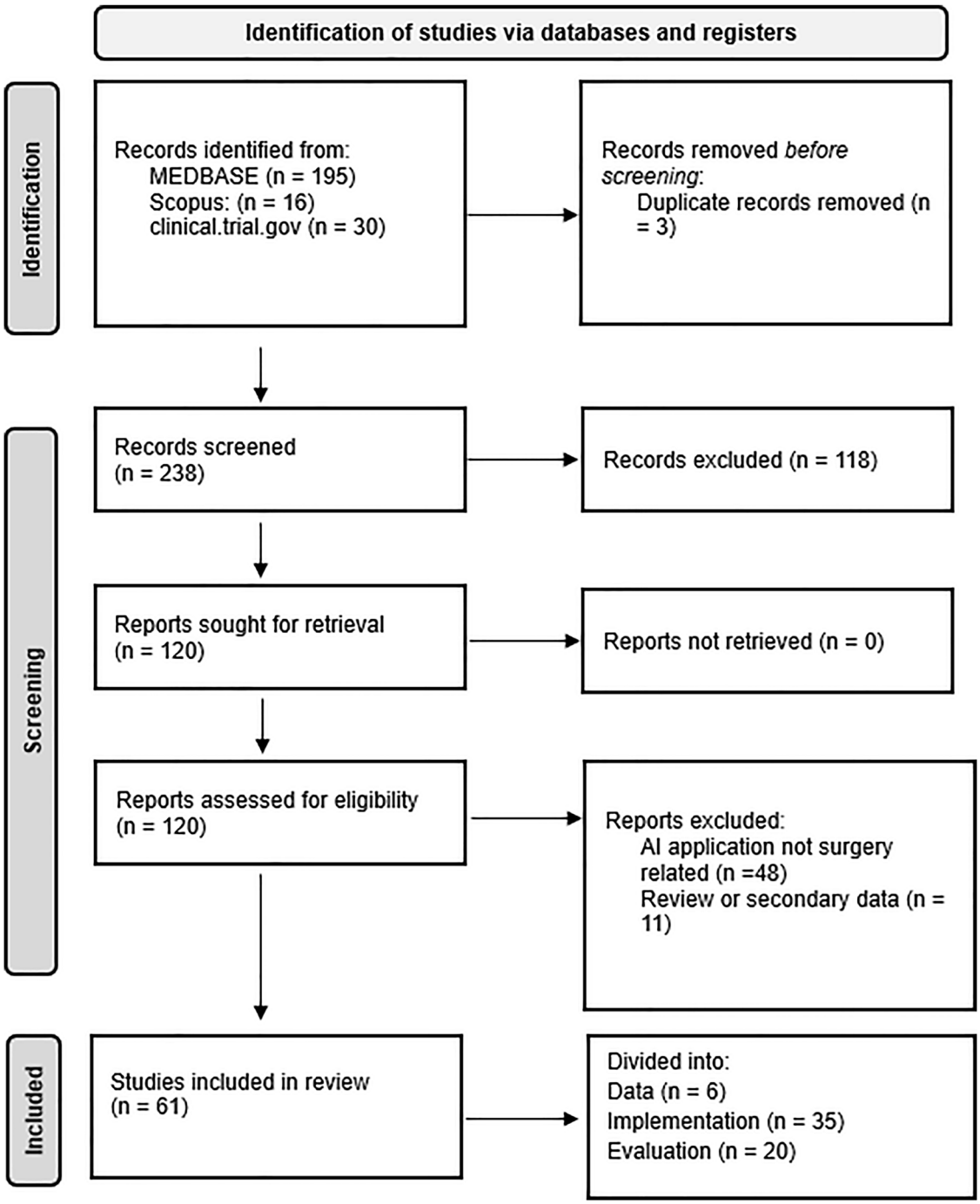


**Fig. 2** | Flow chart of the selection of papers for the literature review.

## Data availability

As supplementary materials, we have included the search strategy, and the resulting list of publications identified from the PRISMA systematic review. The 61 papers are listed with the primary domains (data, implementation and evaluation), which informed the panel to produce the initial questions for the Delphi process. For a summary of the PRISMA review see Figure 2: Flow chart of the selection of papers for the literature review. A full list of the Delphi questions is included in the supplementary materials.

## Code availability

'Google forms' was used for the 3 rounds of the Delphi Process and to collate the results of each round.

Received: 1 April 2026; Accepted: 22 May 2026;
Published online: 07 August 2026

## References

1. Li, H. et al. Artificial intelligence in surgery: evolution, trends, and future directions. *Int. J. Surg.* **111**, 2101–2111 (2025).
2. Malik, A. P., Ahmad, W. & Iqbal, J. The AI surge in surgery: evolution, challenges, and ethical considerations. *Ann. Biomed. Eng.* **53**, 1989–1992 (2025).
3. Maier-Hein, L. et al. Surgical data science - from concepts toward clinical translation. *Med. Image Anal.* **76**, 102306 (2022).

4. Perivolaris, A. et al. Quality of interaction between clinicians and artificial intelligence systems. A systematic review. *Future Health. J.* **11**, 100172 (2024).
5. Tragaris, T., Benetos, I. S., Vlamis, J. & Pneumaticos, S. Machine learning applications in spine surgery. *Cureus* **15**, e48078. https://doi.org/10.7759/cureus.48078 (2023).
6. Knudsen, J. E., Ghaffar, U., Ma, R. & Hung, A. J. Clinical applications of artificial intelligence in robotic surgery. *J. Robot Surg.* **18**, 102 (2024).
7. Stein, M. J. & Rohrich, R. Artificial Intelligence and postoperative monitoring in plastic surgery. *Plast. Surg.* **33**, 312–317 (2025).
8. Wah, J. N. K. Revolutionizing surgery: AI and robotics for precision, risk reduction, and innovation. *J. Robot Surg.* **19**, 47 (2025).
9. Collins, J. W. et al. Ethical implications of AI in robotic surgical training: a Delphi consensus statement. *Eur. Urol. Focus* **8**, 613–622 (2022).
10. Kader, R. et al. A novel cloud-based artificial intelligence for real-time detection of colorectal neoplasia – a randomized controlled trial (EAGLE). *npj Digit. Med.* **9**, 84 (2026).
11. Pressman, S. M. et al. AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. *Healthcare* **12**, 825 (2024).
12. Mlambo, B. et al. Standardized temporal segmentation framework and annotation resource library in robotic surgery. *Mayo Clin. Proc. Digit Health* **3**, 100257 (2025).
13. Protserov, S. et al. Development, deployment and scaling of operating room-ready artificial intelligence for real-time surgical decision support. *NPJ Digit Med.* **7**, 231 (2024).
14. Ren, Y. et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. *JAMA Netw. Open* **5**, e2211973 (2022).
15. Palagonia, E. et al. The safety of urologic robotic surgery depends on the skills of the surgeon. *World J. Urol.* **38**, 1373–1383 (2020).
16. Vincent, C., Moorthy, K., Sarker, S. K., Chang, A. & Darzi, A. W. Systems approaches to surgical quality and safety: from concept to measurement. *Ann. Surg.* **239**, 475–82 (2004).
17. Nakanuma, H. et al. An intraoperative artificial intelligence system identifying anatomical landmarks for laparoscopic cholecystectomy: a prospective clinical feasibility trial (J-SUMMIT-C-01). *Surg. Endosc.* **37**, 1933–1942 (2023).
18. van Amsterdam, B. et al. Gesture recognition in robotic surgery with multimodal attention. *IEEE Trans. Med. Imaging* **41**, 1677–1687 (2022).
19. Neumann, J. et al. Ontology-based surgical workflow recognition and prediction. *J. Biomed. Inf.* **136**, 104240 (2022).
20. Parsanasab, E., Ahmadipour, A. & Mehraeen, E. Utilization of ontology to develop artificial intelligence systems in the healthcare industry. *Health. Inf. Res.* **31**, 320–330 (2025).
21. Nwoye, C. I. et al. CholecTriplet2022: show me a tool and tell me the triplet - An endoscopic vision challenge for surgical action triplet detection. *Med. Image Anal.* **89**, 102888 (2023).
22. Eckhoff, J. A. et al. SAGES consensus recommendations on surgical video data use, structure, and exploration (for research in artificial intelligence, clinical quality improvement, and surgical education). *Surg. Endosc.* **37**, 8690–8707 (2023).
23. Carayon, P. et al. Work system design for patient safety: the SEIPS model. *Qual. Saf. Health Care* **15**, i50–8 (2006).
24. Boal, M. W. E. et al. Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review. *Br. J. Surg.* **111**, znad331 (2024).
25. Hassan A. M. et al. Governance framework for safe and ethical implementation of artificial intelligence in surgery: a modified-Delphi consensus. *J. Am. Coll. Surg*. https://doi.org/10.1097/XCS.0000000000001834 (2026).
26. Yang, J. et al. An adaptive human-robotic interaction architecture for augmenting surgery performance using real-time workload sensing-demonstration of a semi-autonomous suction tool. *Hum. Factors* **66**, 1081–1102 (2024).
27. Wheelock, A. et al. The impact of operating room distractions on stress, workload, and teamwork. *Ann. Surg.* **261**, 1079–84 (2015).
28. Khashu, K. Operationalizing trustworthy artificial intelligence in clinical and operational workflows. *Front. Digit Health* **8**, 1779041 (2026).
29. Radanliev, P. et al. AI security and cyber risk in IoT systems. *Front. Big Data* **7**, 1402745 (2024).
30. Norton, J. et al. Video assessment of CaRdiac Surgery QualITY (VARSITY) Surgery Group. Impact of surgical non-technical skills on clinical outcomes: systematic review. *Br. J. Surg.* **113**, znaf271 (2025).
31. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333**, 319–328 (2025).
32. Petersen, C. et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. *JAMIA* **28**, 677–684 (2021).
33. Jackson, G. P. & Shortliffe, E. H. Understanding the evidence for artificial intelligence in healthcare. *BMJ Qual. Saf*. https://doi.org/10.1136/bmjqs-2025-018559 (2025).
34. Ward, T. M., Hashimoto, D. A., Ban, Y., Rosman, G. & Meireles, O. R. Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. *Surg. Endosc.* **36**, 6832–6840 (2022).
35. Aoyama, Y. et al. Artificial intelligence for surgical safety during laparoscopic gastrectomy for gastric cancer: Indication of anatomical landmarks related to postoperative pancreatic fistula using deep learning. *Surg. Endosc.* **38**, 5601–5612 (2024).
36. Vasey, B. et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. *JAMA Netw. Open* **4**, e211276 (2021).
37. Knop, M., Weber, S., Mueller, M. & Niehaves, B. Human factors and technological characteristics influencing the interaction of medical professionals with Artificial intelligence-enabled clinical decision support systems: literature review. *JMIR Hum. Factors* **9**, e28639 (2022).
38. Ali, I. S. et al. Artificial intelligence in planning for spine surgery. *Curr. Rev. Musculoskelet. Med.* **18**, 627–634 (2025).
39. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
40. Amir-Behghadami, M. & Janati, A. Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. *Emerg. Med. J.* **37**, 387–387 (2020).
41. Whiting, P. F. et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529–536 (2011).
42. Sterne, J. A. C. et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* **366**, l4898 (2019).
43. Sterne, J. A. et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4919 (2016).
44. Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O. & Alberti, C. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. *PLoS ONE* **6**, e20476 (2011).
45. Ma, R. et al. Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *npj Digit. Med.* **5**, 187 (2022).


## Acknowledgements

We would like to acknowledge the other contributors to the Delphi process who completed 3 rounds of the Delphi: Binita Ashar, Prem Batchu-Green, Alona Bilik, Sanket Chauhan, Prashant Chodagiri, Moise Danielpour, Andrew Gumbs, Marcel Hohl, Robert Holloway, William Kennedy, Peter Kim, Stephen Knych, Christiam Kriebel, Shlomi Laufer, Edward Lin, Rick Lombardi, Imanol Luengo, Ann Majewicz Fey, Ben Mueller, Keith Nahigian,

Yu Ozawa, Calvin Perumalla, Poya Pourhaderi, Jay Redan, Hugang Ren, Sharona Ross, Ganesh Sankaranarayanan, Rick Satava, Michael Sazbo, Rohan Sharma, Raimundo Sierra, Patricia Sylla, Meghan Williams, Yichuan Yan, John Yeh, Bryan Zediker.

## Author contributions

J.C. and M.M. wrote the main article initial draft. J.C. prepared Tables 1–5 and Figure 1. M.M. prepared Figure 2. D.S., C.P., M.C.M., D.Y., S.S., J.L., Z.T., F.F., D.S., S.V., M.M., and G.P. reviewed and contributed to the final manuscript. All co-authors contributed to the 3 rounds of the Delphi.

## Competing interests

D.S. is an employee of Medtronic PLC. G.J. is an employee of Intuitive Surgical. The other authors declare no competing financial or non-financial interests. F.F. is an Associate Editor of Nature Digital Surgery. F.F. was not involved in the journal's review of, or decisions related to, this manuscript.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s44484-026-00017-3.

**Correspondence** and requests for materials should be addressed to Justin W. Collins.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026