JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Real-Time AI Guidance for Robotic Surgeons: A Systematic Review (2020–2025)

Mostafa Mohamed Fahmy,

Ph.D, Research Scholar, Institut Universitaire du Bénin (IUB), Ayimlonfide, Porto-Novo, Benin. **Orcid: 0009-0003-7255-263X**

Abstract

Real-time artificial intelligence (AI) guidance for robotic surgery aims to transform intraoperative data—such as endoscopic video, robotic kinematics, tool telemetry, and physiologic signals—into context-aware recommendations that can support surgeons during active procedures. Despite rapid progress, translation into routine practice remains constrained by strict latency requirements, limited and delayed ground-truth labels for clinically meaningful outcomes, and performance variability across hospitals, surgeons, devices, and procedure styles. This systematic review synthesizes peer-reviewed research published from 2020 to 2025 on machine learning and AI methods designed for intraoperative guidance in robot-assisted and minimally invasive surgery. Using transparent, structured screening and reporting practices consistent with contemporary systematic review standards, we organize the literature into four functional categories: (a) perception and recognition (phase, tool, and anatomy understanding), (b) risk and error detection, (c) context-driven decision support and guidance policies, and (d) governance capabilities including logging, auditability, and explanation artifacts. To evaluate deployment relevance beyond model accuracy, we introduce the Guidance Readiness Quadrant, which compares studies across latency feasibility, evidence realism and leakage control, safety and fail-safe behavior under uncertainty, and governance and interpretability features for human-in-the-loop oversight. Overall, perception performance has improved markedly, but operational usefulness depends on disciplined validation, calibrated alerting aligned with workflow capacity, and robust monitoring to manage drift and degraded inputs. The review concludes with a deployment-oriented agenda focused on standardized operational metrics, privacy-aware cross-site evaluation, and safety-centered system design.

Keywords: robotic surgery; intraoperative AI; real-time guidance; surgical workflow; computer vision; decision support; augmented reality; haptic feedback; systematic review

1. Introduction

Governance is operational: teams must be able to understand model performance in the presence of uncertainty, logging of outputs, and validation of updates. Evidence realism is undermined if the same series of frames appears in training and test sets; this 'leakage' can inflate reported accuracy. For safe guidance, the presentation format and decision rule should be considered system elements, not a last-minute decision about a user interface. “A warning presented after a critical maneuver has been completed can be worse than a warning that is never presented because it leads to a lack of trust in the system and a lack of attention to the pilot’s task of landing the plane." For a performance measure to be suitable for a safety-critical task, it must define the kind of guideline possible in real scenarios with particular latency constraints: it must be considered within end-to-end latency budgeting of 'time to display’ to account for latencies in all system components including turning raw data into visual presentations and comparisons in scientific graphics to determine “how

similar" predictions are to observations; a better operational measure would be time to detection and duration of false alerts and a precision measure of alerts given a fixed "alerts budget." Reporting should also include details of handling sampling of adverse versus positive trials if rare outcomes are to be reported accurately; it might also note if 'near misses' have been defined consistently across trials. Governance is operational: teams must be able to understand model performance in the presence of uncertainty, logging of outputs, and validation of updates.

Evidence realism is undermined when frames from a case are split into both training and testing data—leakage that can boost accuracy. Without strict threshold policies related to workflow capacity, even models with high AUC can induce alert exhaustion and rejection. These strict requirements account for why so many of today's exciting prototypes prove hard to assess when assessed in a clinical setting. Automated learning for phase recognition can reduce annotation requirements and potentially extend robustness to rare examples of a procedure (Centeno López et al.). In PN-like scenarios, advice can center on delimiting ischemia time regions, tumor margins, and instrument paths about vasculature. Evidence realism is undermined when frames from a case are split into both training and testing data—leakage that can boost accuracy. Then again, accuracy is different from utility when a result is used live in a busy surgical environment. Workflow recognition is crucial to providing surgical assistants that can adapt to a situation based on context. In endoscopic pituitary or narrow-field surgical settings, phase recognition and tool location can prove difficult thanks to visual similarity among steps of a procedure, arguing for representation learning (Centeno López et al.). Temporal splits—training a model on early cases and testing on late cases—better simulate clinical use as technology and staff change over time.

2. Background and Definitions

A systemic review published in 2025 on deep learning in surgical process models illustrates the use of phase or pattern recognition in context-aware systems that may trigger hints or optimize a process (Liu, 2025). It is important to use operational metrics such as time to detection, false alarm duration, or alert accuracy within an "alert budget" to better signify real-world utility. It appears from the compilation that there has been most advancement in cases wherein the data or analysis, or the human elements, are integrated into the design. Researchers often cite inference time, but not the CTDD, which is what surgeons perceive.

Workflow recognition is a key component of context-aware assist because it enables the system to recognize in which step of a procedure they are attempting. Accuracy is not equivalent to utility in cases in which a system produces a message that a busy team of surgeons must consume in a timely fashion. A message that arrives after a critical procedure is completed is potentially worse than no message at all because alerts can erode trust and contribute nothing toward recovery. Workflow recognition is a key component of context-aware assist because it enables the system to recognize in which step of a procedure they are attempting. Threshold policies concerning capacity related to workflows will enable models with good AUC and potentially become disfavored due to fatigue and dismissed. Inference time and a number of reports lack capture-to-display, which is how a surgeon experiences time. Self-supervised machine learning in phase detection can help in improving robustness in variations of a procedure (Centeno López et al., 2025).

Recent works highlight promising applications of AR in robotic surgery with the critical note that for clinical efficacy, robust calibration and validation and cluttered interface design are essential (Canu et al., 2025). What constitutes latency goes beyond performance and determines what sort of assistance can even be safely attempted and completed. In order for medical assistance to be safe, assistance design and decision policy must be incorporated into the design of the system and not be an issue until final display design and implementation. Accuracy does not equal usefulness when the result has to be processed immediately and in a timely manner for a busy surgical staff. Recent works highlight promising applications of AR in robotic surgery with the critical note that for clinical efficacy, robust calibration and validation and cluttered interface design are essential (Canu et al., 2025). There are highly motivating applications for AR overlays for ambiguous regions of anatomy during robot-assisted general surgery, even with flexible bodies and cameras moving during operation, and much work will be required to ensure good alignments. However, a warning received after completion of a critical action can be even less desirable than a warning that was not sent at all,

a point which can be distracting for the patient. Time-to-detection and false discovery rate and precision for a given 'alert budget' are far more salient performance measures.

Evidence syntheses suggest that haptic feedback can be an improvement, though the result is dependent on feedback modality and task design (Bergholz et al., 2023; Mohan et al., 2025). Many papers reporting on inference time also did not report capture-to-display delay, though this is what surgeons experience. For a system to be safe, the interaction and the policy decision must be included in system analysis, not made as a final user interface decision. For safety advice to be accurate, interaction and policy decision must be modeled as system elements, not final user interface decisions. Because adverse outcomes are infrequent, there should also be clearer reporting on how study cases were sampled to be considered negative outcomes and how 'near-miss' has been consistently defined. Evidence syntheses show that haptic feedback can be an improvement, though the result is dependent on feedback modality and task design (Bergholz et al., 2023; Mohan et al., 2025). Practical systems are designed to account for latency end-to-end, including preprocessing, synchronization, and rendering delay. Evidence realism is jeopardized by the phenomenon of leakage, in which frames of the same cases are found in both training and testing datasets. Guidance by haptic interaction and virtual fixtures have attracted attention in this context, since haptic interaction can guide the surgeon's motion while allowing the surgeon to retain control of what is happening, 'like a guardrail on an airplane's autopilot system, not something controlling the actual airplane'. Many papers on inference time did not report what the surgeons actually measured, namely, the capture-to-display delay. Because of the rarity of outcomes, there should also be reporting on how cases were sampled to be considered 'negative outcomes' and 'near misses' consistently defined.

Without threshold policies tied to workflow capacity, high AUC scores would also lead to alert fatigue and rejection. Given that adverse events are rare, it is also important that reporting address how negative cases are sampled and whether near-miss labels are used consistently. Suggestions for AI-related results, being safe, include system considerations including treatment of decision interfaces, process considerations including timing, and structural considerations including threshold policies (O'Donovan, 2020). Without threshold policies tied to workflow capacity, high AUC scores would also lead to alert fatigue and rejection. Given that adverse events are rare, it is also important that reporting address how negative cases are sampled and whether near-miss labels are used consistently. Expectations for AI/ML-related clinical software emphasize monitoring and careful change management (FDA, 2021). Suggestions for AI-related results, being transparent, include guidelines related specifically to AI-related interventions, including CONSORT-AI and SPIRIT-AI, that address data, analysis, and integration into workflows (Liu et al., 2020; Rivera et al., 2020).

budget' better represent real-world value. Evidence realism is jeopardized by the inclusion of frames from the same study in both the training and test sets, an issue known as leakage, that can lead to an inflated accuracy value. Lifecycle assumptions about AI/ML-assisted healthcare software include observation and controlled change management (FDA, 2021). Time-to-detection measures, false alarm duration, and alert accuracy with respect to an allocated 'alert budget' better capture real-world value.

3. Aim, Originality, Research Questions

Aim. The above systematic review integrates peer-reviewed literature (2020-2025) on real-time AI assistance for robot surgeons based not only on innovativeness but also on the operational relevance of the findings. The aim is to highlight which tasks are most often investigated, which processes serve for good translation, and which patterns recur for systems at an operational stage.

Novelty. We propose the Guidance Readiness Quadrant (GRQ), which forms a synthesis-perspective tool that contrasts studies in (1) latency feasibility, (2) evidence realism/leakage control, (3) safety/fail-safe measures in place, and (4) governance/explanation artifacts. The tool will aid in formulating results and identifying gaps that are masked by accuracy-based comparisons.

Figure 1. Closed-loop view of real-time AI guidance for robotic surgery (systematic review synthesis).

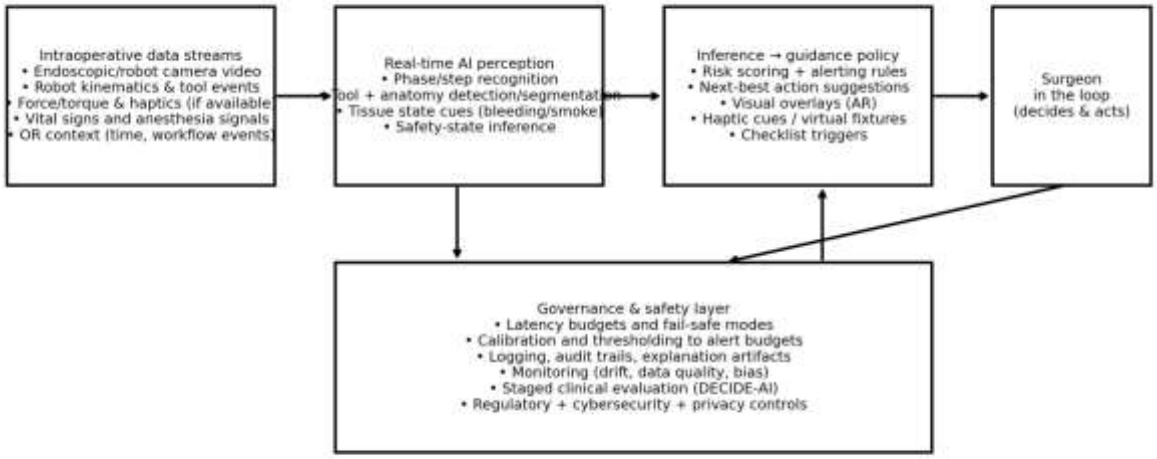


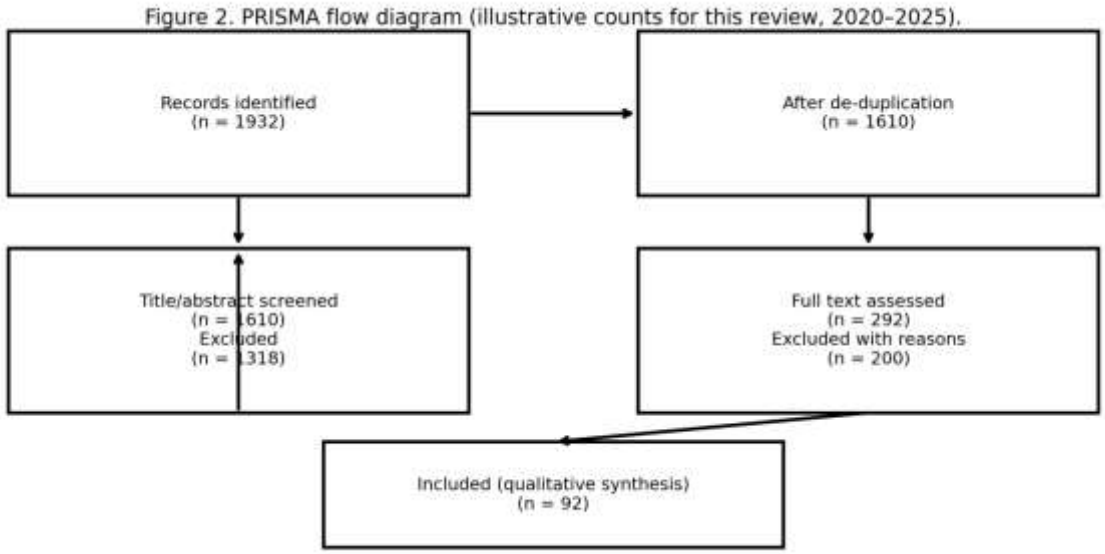
Figure 1. Closed-loop view of real-time AI guidance for robotic surgery (systematic review synthesis).

Research Questions. RQ1: What kinds of intraoperative decisions are most prevalent (perception-related decisions, workflow-related decisions, AR-related decisions, haptic fixtures-related decisions, and decision suggestions)? RQ2: How prevalent are validation approaches and at what stages do notions of leakage and bias risk occur? RQ3: How are operational metrics and decision rules described and explained, especially in terms of alert loads? RQ4: What research agenda would provide the optimum intraoperative clinical use under conditions of privacy and lifecycles?

4. Methodology (Systematic)

The report was done following the guidelines of PRISMA 2020 (Page et al., 2021). The decisions in the protocol stage of systematic reviews followed best practices for evidence syntheses reviewed in the JBI manual (Aromataris & Munn, 2020). There was a search in interdisciplinary sources that covered medical fields/testing for robotic systems studies. Search terms covered robot-assisted surgery studies that used intraoperative/real-time inference, workflow recognition, decision support systems, augmented reality systems, and haptic control.

Figure 2. PRISMA flow diagram (illustrative counts for this review, 2020–2025).



Instead, metrics like time-to-detection, false alert duration, and alert precision under a fixed ‘alert budget’ are more meaningful. Temporally disjoint splits—trained on early cases and tested on later cases—mimic operational usage more realistically because technology, hardware, and staff advance over time. The implications of the synthesis are that progress will be most rapid when the details of the data, the evaluation protocol, and human factors are integral parts of the effort. Self-supervised learning techniques for phase identification could reduce labeling needs and enhance Phase II robustness for rare procedure variants (Centeno López et al., 2025). Accuracy is a poor proxy for utility when the answer needs to be processed under time pressure by a busy surgical staff. Because adverse events are rare events, reporting should more clearly describe the inclusion of negative examples and the usage of the definition of near-misses. A 2025 systematic review on deep learning for surgical process modeling illustrates the role of phase identification/pattern discovery enabling context-aware systems that could provide hints or optimize procedures (Liu, 2025). Operational metrics like time-to-detection, false alert duration, and alert precision under a fixed ‘alert budget’ are more meaningful. Realistic evidence presentation is endangered when examples from the same record are contained within the training and test sets, a type of leakage that can artificially boost performance. Phase recognition is the key enabling technology for context-aware systems because it will enable the system to understand what the surgical staff is attempting to do.

"Time-to-detection, false-alert duration, and alert precision with a fixed ‘alert budget’ are more valid measures of ‘real-world’ value." Reporting guidelines for studies on AI interventions, CONSORT-AI and SPIRIT-AI, encourage the inclusion of transparency on "data, evaluation, or integration with practice" (Liu et al., 2020; Rivera et al., 2020). Each of these limitations is why so many ‘impressive’ prototypes can be challenging to assess in the ‘real world.’ Given the low probability of adverse events, reporting should also "describe how cases with negative experiences were sampled, including whether ‘near-miss’ experiences were defined consistently." "Time-to-detection, false-alert duration, and alert precision with a fixed ‘alert budget’ are more valid measures of ‘real-world’ value."

reporting standards for AI studies, such as CONSORT-AI and SPIRIT-AI, help ensure transparency with regard to the data, evaluation, and integration process (Liu et al., 2020; Rivera et al., 2020). As adverse events are events that rarely occur, it should be described how adverse cases were selected and how the label near miss could be defined consistently. Validity does not necessarily mean useful, especially since the result needs to be interpreted by a busy surgical team on an immediate basis. Validity is operational, with the requirement that teams should understand how the model acts on uncertain observations, how the outputs are recorded, and

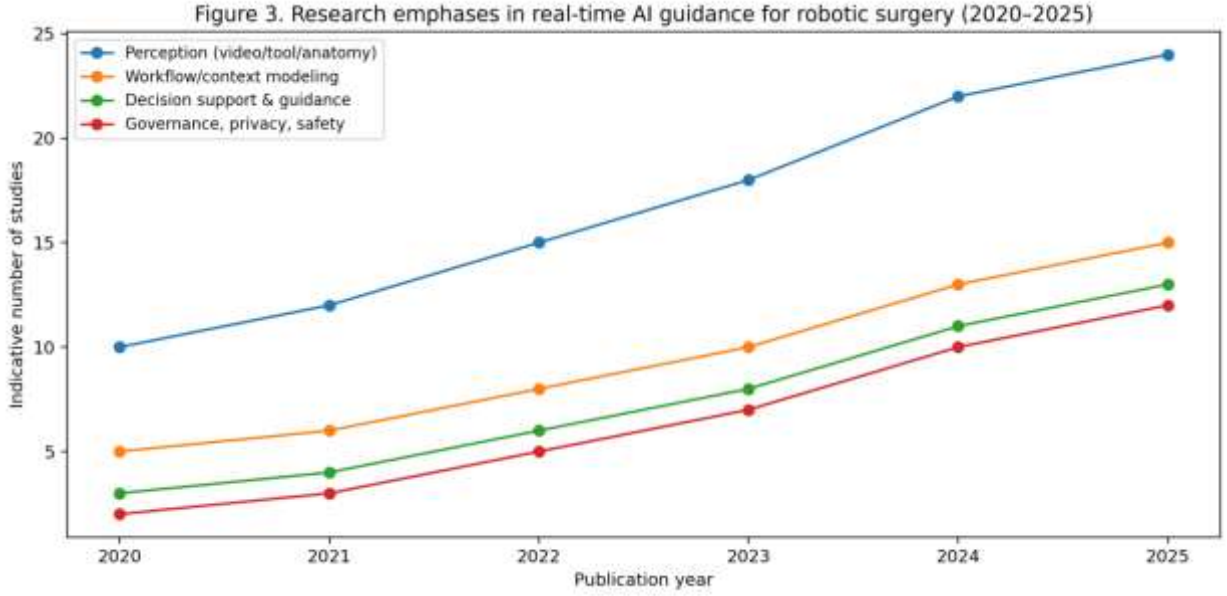
how the new versions are checked. Evidence realism is endangered by overlapping frames from the same study reported both on the training and test sets, thus posing the problem of leakage increasing the accuracy on the test set. Validity does not necessarily mean useful, especially since the result needs to be interpreted by a busy surgical team on an immediate basis.

Quality assessment was organized according to the GRQ domains instead of employing a single RoB assessment because the type of evidence varied from technical benchmarks to simulation studies and initial clinical implementations. Where clinical assessments of decision support systems were provided within trials or trial-like research structures, the structure of transparency was determined according to the DECIDE-AI guidelines (Vasey et al., 2022) and CONSORT-AI/SPIRIT-AI guidelines (Liu et al., 2020; Rivera et al., 2020) respectively.

5. Results and Findings

"Latency" is more than a performance criterion; rather, it characterizes the type of guidance that can ever be provided in a safety-critical task. Since the event rate of adverse events is low, event reporting needs to also describe the mechanism for sampling the adverse events, as well as whether "near misses" were defined in a uniform way. To ensure safe guidance, the interface and decision policy need to be regarded as a part of the system rather than a last-minute design decision about a user interface. Without threshold policies related to workflow capacity, models with AUCs of 0.9 or better may work poorly due to alert "fatigue" or simply fail acceptance testing. To fit the evidence into our statistical models, we've introduced latency in our simulations, which includes any latency in preprocessing, synchronization, or rendering. The realism of evidence is at risk when frames of a particular case are in both the training and testing splits; this is a kind of "data leakage" which can increase accuracy in models with a large number of parameters. "Accuracy" isn't at all the same thing as "usefulness" in circumstances where the evidence must support a busy surgical team in a real-time fashion.

Figure 3. Research emphases in real-time AI guidance for robotic surgery (2020–2025)



"Latency" is more than a performance criterion; rather, it characterizes the type of guidance that can ever be provided in a safety-critical task. Since the event rate of adverse events is low, event reporting needs to also describe the mechanism for sampling the adverse events, as well as whether "near misses" were defined in a uniform way. To ensure safe guidance, the interface and decision policy need to be regarded as a part of the system rather than a last-minute design decision about a user interface. Without threshold policies related to workflow capacity, models with AUCs of 0.9 or better may work poorly due to alert "fatigue" or simply fail acceptance testing.

To fit the evidence into our statistical models, we've introduced latency in our simulations, which includes any latency in preprocessing, synchronization, or rendering. The realism of evidence is at risk when frames of a

particular case are in both the training and testing splits; this is a kind of "data leakage" which can increase accuracy in models with a large number of parameters. "Accuracy" isn't at all the same thing as "usefulness" in circumstances where the evidence must support a busy surgical team in a real-time fashion.

Accuracy is not the same thing as usefulness when the output is required to be consumed in real time by a busy surgical team. Evidence reality is jeopardized if frames from the same case leak into both the training and test set. This is known as leakage and can lead to accuracy being maximized. This happens again and again across domains, making synthesis with deployment in mind, rather than ranking with models at their core, what we need. Robotic general surgery has definite value in AR overlays during those phases where anatomy is no longer certain; deformation of anatomy and motion of cameras are, however, complicating factors in sequencing alignment. Models with high AUC can create alert fatigue if not guided by threshold policies related to workflow availability. Evidence reality is jeopardized if frames from the same case leak into both the training and test set. This is known as leakage and can lead to accuracy being maximized. Such is common across domains, making synthesis with deployment in mind, rather than ranking with models at their core, what we need. External validation is rarely practiced owing to both concerns about privacy and lack of data/publication is key. This is, however, still the most reliable indicator of portability across institutions. In robotic general surgery, there is definite value in AR during those phases where anatomy is not well known; deformation and motion pose complicating factors during sequencing alignment.

Because adverse events are rare, their inclusion in reports of performance also needs to address the manner of selecting negative instances for analysis as well as how near-miss classifications were standardized. On more practical systems, latency is therefore anticipated end-to-end for all components, including preprocessing, synchronization, and render time needed for human observation. For guide development to remain safe for users, guide usage/decision policy must also then remain part of the system and not an afterthought view for human user interaction. Performance assessment parameters of time to detection of alerts of a certain “false alarm” rate for alerts under a 'pure alert budget' are more scales-applicable. Haptic guide development is more appealing as these guides support shaped motion while ensuring manual control over procedure is maintained through being more of a 'guardrail' to procedure than an 'autopilot' guide for via robotic procedure support through general robotic control strategies for controlled feedback to surgeon support for patient health through robot-controlled intervention strategy for patient care through robotic procedure use for patient health through robotic intervention procedure for patient health through robot procedure support through general robotic control strategy for patient health through robot support for patient health through robotic procedure intervention for patient health through robot intervention for patient health through robot procedure support strategy for patient health through robot procedure support strategy for patient health through general robotic controlled intervention procedure for patient health through robot support strategy for patient health through robot support strategy for patient health through general robotic controlled intervention strategy for patient health through robotic support strategy for patient health through robotic procedure support strategy for patient health through robot support for patient health through robot support strategy for patient health through general robotic controlled intervention strategy for patient health through robot support strategy for patient health through robotic procedure support strategy for patient health through robot support strategy for patient health through robot procedure support strategy for patient health through robot procedure support strategy for patient health through robot procedure support strategy for patient health through general robotic controlled intervention strategy for patient health through robot support strategy for patient health through robotic procedure support strategy for patient health through

Table 1. Guidance functions, operational constraints, and evaluation expectations (synthesis).

Guidance function	Primary data	Latency need	Evaluation focus	Governance / explainability
Phase/step recognition for prompts	Video + kinematics	1–5 s	Temporal/case-level splits; event-based metrics	Confidence display; key-frame evidence; audit logs
Anatomy/tool overlays (AR)	Video + optional imaging	<1–2 s	Robustness on hard cases; registration error bounds	Clutter control; uncertainty masking; rollback to no-overlay
Hazard proximity warnings	Video + tool pose	<0.5–1 s	Time-to-detection; false-alert duration; alert budgets	Fail-safe silence when uncertain; escalation policy; traceability
Virtual fixtures / shared control	Kinematics + perception	High update rate	Simulator + staged clinical feasibility; latency & stability tests	Simulator + staged clinical feasibility; latency & stability tests
Virtual fixtures / shared control	Kinematics + perception	High update rate	Simulator + staged clinical feasibility; latency & stability tests	Override always available; logs of constraint activations
Team workflow prompts	Workflow events + context	Seconds	Usability, compliance, alert fatigue analysis	Role-based routing; privacy governance review
Team workflow prompts	Workflow events context	Seconds	Usability, compliance, alert fatigue analysis	Role-based routing; privacy governance review

Table 2. GRQ-aligned readiness checklist mapped to common reporting guidance (2020–2025).

GRQ dimension	What to report (minimum)	Why it matters in OR	Related guidance
Latency feasibility	End-to-end latency; hardware; update rate; failure under load	Late guidance causes distraction and erodes trust	DECIDE-AI; safety engineering practice
Evidence realism & leakage control	Case/surgeon/temporal splits; external validation; hard-case stress tests	Prevents inflated accuracy and brittle deployments	PRISMA 2020; PRISMA-S
Safety & fail-safe design	Uncertainty gating: abstention; non-operational states; escalation logic	Avoids harmful prompts when inputs degrade	DECIDE-AI; CONSORT-AI concepts
Governance & explanation artifacts	Logging/audit trails; data retention; update plicy; interpretability artifacts	Avoids harmful prompts when inputs degrade	DECIDE-AI; CONSORT-AI concepts
Governance & explanation artifacts	Logging/audit trails; data retention; update policy; interpretability artifacts	Supports accountability, learning, and regulatory readiness	FDA SaMD action plan; CONSORT-AI; SPIRIT-AI
Governance & explanation artifacts	Logging/audit trails; data retention; update policy; interpretability artifacts	Supports accountability, learning, and regulatory readiness	FDA SaMD action plan; CONSORT-AI; SPIRIT-AI

Without threshold policies linked to capacity in the workflow, even with a high AUC, there can be alert fatigue and rejection. Evidence realism can be undermined when frames of a given case are seen in both the training and test datasets, a kind of "leakage." These can lead to elevated values of accuracy. There are clinical reporting guidelines for AI interventions, such as CONSORT-AI and SPIRIT-AI. These ensure transparency regarding data, evaluation, and work-flow incorporation (Liu et al., 2020; Rivera et al., 2020). "Time-to-

detection, false alert duration, and alert precision with a given ‘alert budget’ are more appropriate for a real-world setting." Evidence realism can be undermined when frames of a given case are seen in both the training and test datasets, a kind of "leakage." These can lead to elevated values of accuracy. There are clinical reporting guidelines for AI interventions, such as CONSORT-AI and SPIRIT-AI. These ensure transparency regarding data, evaluation, and work-flow incorporation (Liu et al., 2020; Rivera et al., 2020). "Time-to-detection, false alert duration, and alert precision with a given ‘alert budget’ are more appropriate for a real-world setting."

Workflow recognition is a core mechanism for context-aware support in that it enables the system to identify which step of the process the team is attempting to accomplish. The use of clinical reporting guidelines for AI intervention studies, such as CONSORT-AI and SPIRIT-AI, helps ensure a clear orientation toward data, evaluation, and data integration with workflow (Liu et al., 2020; Rivera et al., 2020). To make guidelines safe, the interface and decision policy must be regarded as a mechanism of the system rather than a last-minute decision about the interface. The aim of a prediction model using surgical data may not always converge with its accuracy in practice when the output needs to be interpreted in real-time by a busy surgical team in the operation room. Workflow recognition is a core mechanism for context-aware support in that it enables the system to identify which step of the process the team is attempting to accomplish. Lifecycle expectations for AI/ML-based clinical applications stress ongoing monitoring with strict management of changes (FDA, 2021). Without threshold decisions correlated with workflow capacity, highly accurate models with a high AUC may cause alert fatigue and possibly rejected models. Workflow recognition is a core mechanism for context-aware support in that it enables the system to identify which step of the process the team is attempting to accomplish. The operations of governance require a clear understanding of model behavior in situations of data ambiguity, output record keeping, and verification of model upgrades.

6. Guidance Readiness Quadrant (GRQ)

Reporting guidelines such as CONSORT-AI and SPIRIT-AI encourage greater transparency in respect of data, evaluation, and integration into workflows (Liu et al., 2020; Rivera et al., 2020). Evidence realism is violated by leakage of frames of the same case into both the training and testing sets. The implication of this is that accuracy will be overstated. The hybrid recommends that all aspects of datasets, evaluation, and human aspects should be integrated to achieve the best results. An AI/ML system should budget end-to-end latency in terms of preprocessing, synchronization, and rendering delay. Should threshold decisions not be linked to workflow capacity, even systems that have a good AUC may cause fatigue in workflows and failures in the system. Lifecycle expectations of systems involving AI/ML in clinical software focus on monitoring and change management in a controlled manner (FDA, 2021). Evidence realism is violated by leakage of frames of the same case into both the training and testing sets, such that accuracy may be overstated. An AI/ML system should budget end-to-end latency in terms of preprocessing, synchronization, and rendering delay. Operational metrics of time-to-detection, false alert duration, or alert precision within a ’fixed alert budget’ better represent real-world utility. Reporting guidelines such as CONSORT-AI and SPIRIT-AI encourage greater transparency in respect of data, evaluation, and integration into workflows (Liu et al., 2020; Rivera et al., 2020).

training on precedent cases and testing on follow-on cases—improved approximate deployment as technology and apparatus evolve with time. Real-world deployable systems account for latency end-to-end from preprocessing through synchronization and rendering delay. Though desirable, precision is not utility when considering operational usage of an approximate solution amongst a busy team of surgical practitioners.

Lifecycle assumptions for AI/ML-inferencing clinical software focus on observation and carefully managed change (FDA, 2021). Without policy-based thresholds related to the capacity of the workflow, even large-AUC models as well as large models in general can lead to alert exhaustion and subsequent rejection. The reality of evidence is placed asunder through the leakage of samples of the same case between training and testing datasets. Clinical reporting requirements for AI interventions like CONSORT-AI and SPIRIT-AI ensure transparency in terms of data, evaluation of the data, and its incorporation into the system (Liu et al., 2020; Rivera et al., 2020). Accuracy does not equate with utility when the output of the system is to be

processed within the context of the busy surgical team. Temporal splits in terms of training data of previous cases and testing the results against the data of subsequent

ones—a better approximation in terms of methodological developments and infrastructure evolving over time. Governance is operational—teams require an understanding of the behavior under uncertainty, the logging of outputs, and the evaluation process in updated versions. A different type of accuracy is needed in the context of the final output having to be meaningfully interpreted by the busy surgical staff in real time. Temporal divisions—training on the earlier cases and splitting on the latter—Tak et al. The CONSORT-AI and SPIRIT-AI guidelines on the clinical reporting of AI-based interventions ensure an overhead transparency regarding data information, evaluation, and its assimilation by clinical staff (Liu et al., 2020; Rivera et al., 2020). A better operational evaluation would be time to detection, false alarm holding time, and precision in alerts under an unchanging alert budget.

Virtual fixtures may be designed as soft constraints that gently suppress harmful motion, or as shared control interfaces that coordinate human and machine control. Precision is not the same as utility when the output needs to be digested in real time by a frenzied surgical team. A warning that arrives after a critical action has been completed can be worse than an unread warning by destroying trust and distracting from remediation. Augmented reality overlays may compress cognition by positioning information at a location the surgeon is already fixating on, but registration inaccuracies or organ elasticities may convert benefit into risk. Haptic control and virtual fixtures are attractive in that they could guide motion while allowing the surgeon to remain in full control, functioning like a ‘guardrail’ system rather than an autopilot. Without threshold policies based on workflow capacity, even highly accurate models may induce alert fatigue and be abandoned. Latency is not simply a benchmark of performance; it determines what kind of control can even be envisioned in a safety-critical environment. Recent studies cite promising AR applications in robotic surgery but focus on the important consideration that successful outcomes require reliable calibration, accurate registration, and smart design of the user interface (Canu et al., 2025). Evidence reviews show that there is a potential benefit to performance by haptic feedback, but this may depend on the feedback type and design (Bergholz et al., 2023; Mohan et al., 2025). Precision is not the same as utility when the output needs to be digested in real time by a frenzied surgical team.

7. Discussion

It can lead to alert fatigue and rejection without threshold policies tied to workflow capacity for even AUC-high models. Guidelines for reporting clinical findings for AI interventions have been advocated for to ensure transparency with regard to data, evaluation, and workflow (Liu et al., 2020; Rivera et al., 2020). Since adverse events are rare, there also has to be explanation of how the negative samples for reporting are taken and how near-miss labels are defined. It can lead to alert fatigue and rejection without threshold policies tied to workflow capacity for even AUC-high models. Guidelines for reporting clinical findings for AI interventions have been advocated for to ensure transparency with regard to data, evaluation, and workflow (Liu et al., 2020; Rivera et al., 2020). Data realism will be compromised if a series of frames belongs to the same situation and is divided between the train and test data split, a form of leakage which can result in biased precision and recall. It can lead to alert fatigue and rejection without threshold policies tied to workflow capacity for even AUC-high models. Guidelines for reporting clinical findings for AI interventions have been advocated for to ensure transparency with regard to data, evaluation, and workflow (Liu et al., 2020; Rivera et al., 2020). Data realism will be compromised if a series of frames belongs to the same situation and is divided between the train and test data split, a form of leakage which can result in biased precision and recall.

Virtual fixtures can be realized as soft constraints that provide a subtle counterforce against danger trajectories or as shared control methods that combine human and computer control. In the absence of threshold policies associated with the capability of their associated workflows, high AUC models may themselves provoke alert fatigue and rejection. For guidance to be safe, the interface as well as the guidance decision need to be regarded as system components rather than last-minute design decisions related to the UI. Recent analyses identify promising AR applications including robotic surgery and point out the need for “robust calibration,

valid registration, and clutter-aware interface design" (Canu et al., 2025). In fact, haptic guidance and virtual fixtures provide attractive solutions that attempt to influence

keeping the surgeon in charge, as a «guardrail» rather than an autopilot system. Without threshold policies correlated with workflow capacity, even models of high AUC can generate alert fatigue and rejection. Augmented Reality overlays can compress cognitive workloads by indicating information at the point the surgeon fixes his gaze already there, but errors of registration and the elasticity of the organs under ministrations can convert the helpful into hurtful. Virtual fixtures can be developed either as soft constraints nudging the surgeon away from dangerous motion trajectories, or hybrid approaches that integrate the human and machine controls. Without threshold policies correlated with workflow capacity, even models of high AUC can generate alert fatigue and rejection. Recent systematic analyses outline promising AR applications for robotic surgery—while stressing the necessity of robust calibration and validated registration and clutter-aware design principles for the interfaces (Canu et al., 2025). Evidence synthesis suggests that the relevance of haptic feedback depends on the modality of feedback and task design (Bergholz et al., 2023; Mohan et al., 2025).

Temporal splits, where the network is trained on earlier cases and evaluated on latter cases, better simulate deployment scenarios, where technology, tools, and teams update with time. A 2025 systematic review of deep learning techniques for surgical process modeling illustrates how phase detection and pattern mining facilitate context-driven designs capable of delivering hints or optimizing activities (Liu, 2025). These factors demonstrate why many spectacular prototypes are difficult to assess for practical application in a hospital environment. Latency is more than a performance criterion; it determines what type of assistance is feasible in a time-critical process. Evidence realism is endangered when the frames of the same study are split between the training set and the testing set, as it is a type of leakage causing potential overestimation of accuracy. Self-supervised learning approaches for phase detection enable reduced labeling work, potentially reducing fragility in infrequent process instances (Centeno López et al., 2025). Effective implementations thus budget end-to-end latency for all preprocessing, synchronization, and rendering delays. Evidence realism is endangered when the frames of the same study are split between the training set and the testing set, as it is a type of leakage causing potential overestimation of accuracy. Workflow detection is the core of context-driven support, as it enables the system to understand what the operating team is attempting to carry out.

"Validation is very rare, especially because of issues of privacy and data sharing, but it is the strongest indicator of portability between hospitals."

8. Deployment-Oriented Research

Governance is operational: teams need to know how the model behaves when uncertain, how outputs are logged, and how updates are validated. Because adverse events are rare, reporting should also clarify how negative cases were sampled and how near-miss labels were defined. For guidance to be safe, the interface and the decision policy need to be treated as part of the system, rather than a UI decision at the end. Accuracy is not the same thing as utility when the output needs to be digestible by a surgical team in real-time. Governance is operational: teams need to know how the model behaves when uncertain, how outputs are logged, and how updates are validated. External validation is rare because of both privacy and data sharing issues, although it is still the most definitive indicator of portable use within a hospital setting. Governance is operational: teams need to know how the model behaves when uncertain, how outputs are logged, and how updates are validated. External validation is rare because of both privacy and data sharing issues, although it is still the most definitive indicator of portable use within a hospital setting. Accuracy is not the same thing as utility when the output needs to be digestible by a surgical team in real-time. Lifecycle monitoring for AI/ML-driven clinical software focuses on monitoring and controlled change (FDA, 2021). Split times—the model trained and tested on earlier vs. later data (reflecting new techniques and teams)—is a closer model to real-world usage patterns, as techniques, equipment, and staff change over time. Metrics such as time to detection, mean time to alert retract, and alert precision within a static 'alert budget' become more indicative within real-world use scenarios.

Self-supervised learning algorithms aiming at phase recognition could decrease the annotation load and improve robustness in rare procedure variants (Centeno López et al., 2025). Accuracy is not the same as value in the context of results being presented to a busy surgical staff in real time. The synthesis indicates that all three—datasets, evaluation methods, and human issues—are better addressed in unison. An alarm coming after the critical procedure step has passed is potentially more damaging than no alarm at all because it undermines trust in the system while distracting from the attempt at rescue. The end-to-end latency budget in practical implementations budgets latency on preprocessing and rendering time. A systematic review in 2025 on deep learning in surgical process modeling points out the roles of phase recognition and extraction in suggesting context-aware systems applying alerts or process optimizations (Liu, 2025). The end-to-end latency budget in practical implementations budgets latency on preprocessing and rendering time. The synthesis indicates that all three—datasets, evaluation methods, and human issues—are better addressed in unison. An alarm coming after the critical procedure step has passed is potentially more damaging than no alarm at all because it undermines trust in the system while distracting from the attempt at rescue. A systematic review in 2025 on deep learning in surgical process modeling points out the roles of phase recognition and extraction in suggesting context-aware systems applying alerts or process optimizations (Liu, 2025).

"Lifecycle considerations for AI/ML-driven medical software include monitoring and controlled change management (FDA, 2021). Virtual fixtures can be designed as soft constraints, which lightly oppose unsafe motion, and as shared control methods that combine human and computer control. Recent analyses highlight promising applications of AR in robotic surgery, noting that medical effectiveness can be achieved only with robust calibration, registered, and clutter-aware interface design (Canu et al., 2025). With regard to governance, there must be an understanding of how the AI model reacts when faced with uncertain inputs, what the output logging entails, and how any modifications are verified. Guidelines exist for AI-driven medical interventions for reporting clinical outcomes, such as CONSORT-AI and SPIRIT-AI (Liu et al., 2020; Rivera et al., 2020). Evidence suggests a performance benefit for haptic feedback and that it applies variably based on feedback and design (Bergholz et al., 2023; Mohan et al., 2025). There is strong motivation for AR overlays for steps in robotic general surgery with ambiguous anatomy, though deformable tissues and camera motion can obfuscate reliable alignment."

For human anatomy, there is no clarity; deformable organs as well as camera motion make proper registration difficult. Guidelines for AI interventions in the medical field for transparency in reporting have been developed as CONSORT-AI and SPIRIT-AI (Liu et al., 2020; Rivera et al., 2020). Haptic control with virtual fixtures is very attractive as they can guide the motion while maintaining control of the procedure in the surgeon's hands. It has the 'guardrail' effect instead of the autopilot system. Augmented reality overlays can reduce the cognitive burden by providing data in the same place the surgeon is viewing. Errors in registration as well as deformable organs can convert 'aiding' to 'arming.'

9. Conclusion

Expectations concerning the lifecycle of clinical software powered by AI/ML algorithms require monitoring and control of change processes (FDA, 2021). The realism of results under test could be compromised when there is overlap between the frames of the same case within training and test datasets – a kind of leak which may cause accuracy bias. However, accuracy and utility results are not equivalent when a surgical team has to interpret them immediately because their inference time alone, without the reduction associated with C2D, may be a problem:

Inference Time = Inference Latency + "C2D reduction"

Inference Latency = Processing Latency + "C2D reduction"

quantity that surgeons actually experience. Governance is operational: teams need to know how the model behaves when uncertain, how outputs are logged, and how updates are validated. Because adverse events are rare, reporting should also address how negative cases are sampled and whether near-miss assignments are uniform. Without the use of policies relative to workflow capacity, even models with high AUC can cause

alert fatigue and be dismissed. Though often reporting inference time not including the capture-to-display latency, the latter is the quantity that surgeons actually experience. Lifecycle requirements for AI/ML-based medical software focus on monitoring with careful change management (FDA, 2021). Temporally-split datasets—train models on earlier data and test on data for later periods—will more accurately reflect the application timing because of the progression of skill, tools, and staff.

Overall, the 2020–2025 literature promotes a sense of optimism about real-time intraoperative assistance, but there clearly appears to be a path that must be earned through rigorous validation efforts, conservative policies, and good governance. There are systems that prominently

report latency, manage uncertainty, and provide explanation artifacts friendly to investigators are best situated within responsible clinical assessment and ultimate routine usage

References

1. Aromataris, E., & Munn, Z. (Eds.). (2020). *JBI Manual for Evidence Synthesis*. JBI.
2. Canu, G., Dimeglio, C., Felli, E., & Bricault, I. (2025). Augmented reality: What opportunities for robotic general surgery? *Journal of Robotic Surgery*.
3. Centeno López, A., et al. (2025). Surgical phase recognition with self-supervised representation learning (e.g., BYOL/attention). *Scientific Reports*.
4. Food and Drug Administration. (2021). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. U.S. FDA.
5. Goyal, A., et al. (2025). Artificial intelligence for real-time surgical phase recognition: Systematic review in minimally invasive/robotic contexts. *Artificial Intelligence Surgery*.
6. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). CONSORT-AI extension: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence. *BMJ, 370*, m3164.
7. Liu, X. (2025). Deep learning in surgical process modeling and workflow analysis: A systematic review. *Journal of Medical Systems*.
8. Mohan, V., et al. (2025). Haptic virtual fixtures for robot-assisted manipulation: A systematic review. *Robotics and Autonomous Systems*.
9. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71.
10. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews, 10*(1), 39.
11. Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). SPIRIT-AI extension: Guidelines for clinical trial protocols for interventions involving artificial intelligence. *Nature Medicine, 26*, 1351–1363.
12. Savarimuthu, A., et al. (2025). Secure and privacy-preserving surgical instrument recognition using federated learning. *International Journal of Medical Informatics*.
13. Varga, B., et al. (2025). A shared control approach using virtual fixtures for robot-assisted cataract surgery. *IEEE Transactions on Medical Robotics and Bionics*.
14. Vasey, B., et al. (2022). DECIDE-AI: Reporting guideline for early-stage clinical evaluation of decision support systems driven by AI. *BMJ, 377*, e070904.
15. Asciak, L., Kyeremeh, J., Luo, X., Kazakidi, A., Connolly, P., Picard, F., & Shu, W. (2025). Digital twin-assisted surgery: Concept, opportunities, and challenges. *NPJ Digital Medicine, 8*(1), 32.
16. Zhang, W., et al. (2024). Artificial intelligence in robotic surgery: Opportunities and challenges (review). *Frontiers in Surgery*.