Perspective | Open Access

Artificial Intelligence Surgery

Mastari. *Art Int Surg.* 2026;6:433-40 **DOI:**10.20517/ais.2026.49

Auditable medical AI in surgery: an architectural response to the ethics imperative

Fatima Azzahra Mastari

Citation: Mastari FA. Auditable medical AI in surgery: an architectural response to the ethics imperative. *Art Int Surg.* 2026;6:433-40. https://dx.doi.org/10.20517/ais.2026.49

Received: 2 Jun 2026 **First Decision:** 12 Aug 2026 **Revised:** 19 Aug 2026 **Accepted:** 26 Aug 2026 **Published:** 3 Sep 2026

Academic Editor: Andrew Gumbs **Copy Editor:** Tong Wang **Production Editor:** Tong Wang

INTRODUCTION: A FORENSIC QUESTION, ASKED FROM A CLINICAL CHAIR

In *Mata v. Avianca* (S.D.N.Y., 2023), counsel was sanctioned for filing generative-model citations that could not be verified when challenged[1]. The same forensic question - on what basis did the system say this, from which source, and when - will arise when a surgical artificial intelligence (AI) output is reviewed after an adverse event. The system either produces a reconstructible evidentiary artefact or it does not.

The journal's founding editorial placed surgeons at the centre of defining Artificial Intelligence Surgery[2]. Capelli *et al.* articulated trustworthy surgical AI as explainable, fair, accountable, robust, and safe[3]; World Health Organization (WHO) framed comparable commitments for health AI[4]; the surgical-training consensus extended them to pedagogy[5]; and the Artificial Intelligence Organization for the Next Generation of Surgeons (AIONS) supplied working definitions for AI Surgery, surgomics, and robotics[6].

These documents articulate what trustworthy systems should be; they do not specify the architectural artefacts by which those claims can be externally verified. This Perspective addresses that implementation gap.

The proposition is that high-trust surgical AI requires an assurance architecture spanning pre-operative support, training, and intra- and post-operative ambient assist, built around four externally inspectable properties. These are proposed as engineering determinants of defensibility, not regulatory mandates.

WHERE THE ETHICS FRAMEWORK REACHES ITS OPERATIONAL LIMIT

Ethical principles remain necessary but do not by themselves make system behaviour reconstructible. Transparency, for example, cannot be established from fluent outputs alone; it depends on what the system is structurally required to retrieve, refuse, and record. Architecture therefore supplies artefacts against which ethical claims can be tested.

CLINETHIX LLC, Sheridan, WY 82801, USA.

Correspondence to: Dr. Fatima Azzahra Mastari, CLINETHIX LLC, Sheridan, WY 82801, USA. E-mail: contact@clinethix.com

Evidence and adjacent professional failures illustrate the problem: a widely implemented sepsis model performed substantially worse on external validation than originally reported[7]; generative systems can fabricate citations[1]; and clinical language-model outputs may require verification against the health record[8]. Aggregate performance or fluent output alone therefore cannot establish defensibility at the point of use.

The task is to identify architectural properties that make ethical claims auditable rather than merely asserted. The four properties below are candidate choices for high-trust clinical AI environments and are treated as load-bearing commitments rather than features.

THE FIRST ARCHITECTURAL COMMITMENT: SOURCE-GROUNDED RETRIEVAL

A clinical AI system that generates medical content from model parameters without runtime retrieval against a curated, version-controlled source corpus is structurally weaker under audit: it cannot reconstruct the source pathway supporting a contested answer.

Source-grounded retrieval is not a novel form of retrieval-augmented generation (RAG). Conventional RAG combines parametric generation with retrieved non-parametric memory[9], and recent conceptual clinical work has proposed curated medical knowledge bases, provenance-aware RAG, and tamper-evident logging[10]. The contribution advanced here is their coupling to permission and authorisation gates: for evidence-dependent assertions, retrieved material must support the claim within identifiable provenance, version, and authorised scope; otherwise, the system enters a typed defer/refusal state and records it.

Support is assessed at claim level, not by exact-text matching. Multiple sources may be synthesised if each material claim remains traceable and disagreement is preserved. When no guideline directly addresses atypical anatomy or a rare intra-operative presentation, the system may provide a bounded synthesis of relevant primary literature, state the evidence gap, request context, or defer a prescriptive recommendation. Refusal applies to the unsupported or unauthorised claim, not to all potentially useful evidence.

Independent clinical work supports separating information supply from claim authorisation. In a controlled breast-cancer decision-snapshot evaluation, RAG without an authorisation gate did not reduce unjustified inference relative to an unconstrained large language model (LLM), whereas an evidence-graded authorisation framework reduced unsupported claims and increased appropriate refusal[11]. This preprint does not establish a standard, but supports the distinction: retrieval supplies evidence; governance determines what may be asserted.

The audit objective is reconstructibility: which source passage, corpus version, authorisation rule, and system state supported or prevented an assertion. Benchmarks such as MedHELM remain useful for task-level performance[12], but do not replace output-level provenance.

In surgery, the pre-operative guideline query is one text-based example, not the architecture's modality limit. Before laparoscopic cholecystectomy or liver resection, a query about updated society guidance should return an answer bound to an identifiable document section and date, or an evidence-insufficiency/defer state rather than an unsupported assertion. Intra-operative observations may instead originate from video or sensors; their governance is addressed below.

THE SECOND ARCHITECTURAL COMMITMENT: PERMISSION-FIRST GOVERNANCE

Medical AI often operates at the boundary between patient care and governed clinical content, including society guidelines, copyrighted literature, and institutional protocols. A permission-first architecture inverts ingestion: content is ingested, indexed, retrieved for synthesis, or used to generate an output only when an explicit upstream permission or licence covers that use. Content scopes are defined before deployment and enforced at retrieval time rather than moderated after the fact.

Machine-readable frameworks establish precedents for *ex ante* control. The World Wide Web Consortium (W3C) Open Digital Rights Language (ODRL) represents permissions, prohibitions, duties, and constraints over digital assets[13], while the Global Alliance for Genomics and Health (GA4GH) Data Use Ontology (DUO) encodes allowable uses of biomedical datasets to support authorisation decisions[14]. They govern different objects but show that permitted uses can be represented with content and evaluated computationally.

For guideline authorities, retrospective takedown cannot fully preserve source authority once content has been ingested and paraphrased. Under this proposal, issuing bodies define permitted uses ex ante and those constraints are enforced during retrieval and synthesis, preserving source authority, version, scope, and jurisdiction.

For surgical societies and guideline authorities, permission-first governance therefore offers a mechanism for AI-mediated distribution while preserving enforceable control over content, version, scope, and jurisdiction[13,14]. Figure 1 illustrates the evidence-governance path when an evidence-dependent question lacks an eligible, current source in the authorised corpus.

THE THIRD ARCHITECTURAL COMMITMENT: REFUSAL AS A FIRST-CLASS BEHAVIOUR

Abstention is established in selective prediction and learning-to-defer, where models withhold or defer predictions when uncertainty or expected error is unacceptable[15,16]; medical-LLM work also shows failures to abstain appropriately under clinical uncertainty[17]. Here, refusal is extended beyond statistical uncertainty as a typed governance event. The working taxonomy separates capability states (out-of-scope, insufficient retrieval evidence, internal source conflict) from authorisation states (out-of-licence, out-of-mandate)[18]. Each event is recorded with its trigger in the audit trail.

Refusal is an AI output state, not an instruction to halt care. Intra-operatively, the system withholds an unsupported or unauthorised assertion while the surgeon retains control and the standard clinical pathway. Fallbacks may request missing context - an interactive strategy evaluated in MediQ[19] - provide a traceable bounded synthesis, expose unresolved conflicts, defer to the clinician, or withhold the unsupported recommendation. Selective-prediction studies frame abstention against the costs of error and non-assistance[20], while ClinDet-Bench shows that incomplete information can produce both premature conclusions and excessive abstention[21]. Refusal frequency should therefore depend on task, case mix, evidentiary availability, and institutional mandate; the auditable requirement is the recorded trigger, resulting state, and human override.

THE FOURTH ARCHITECTURAL COMMITMENT: A CRYPTOGRAPHICALLY CHAINED AUDIT TRAIL

The three commitments converge on a fourth: retrievals, generations, refusals, and overrides are recorded in an integrity-protected audit trail. Secure logging mechanisms are well established: Schneier and Kelsey described constructions designed to make earlier entries resistant to undetectable modification or

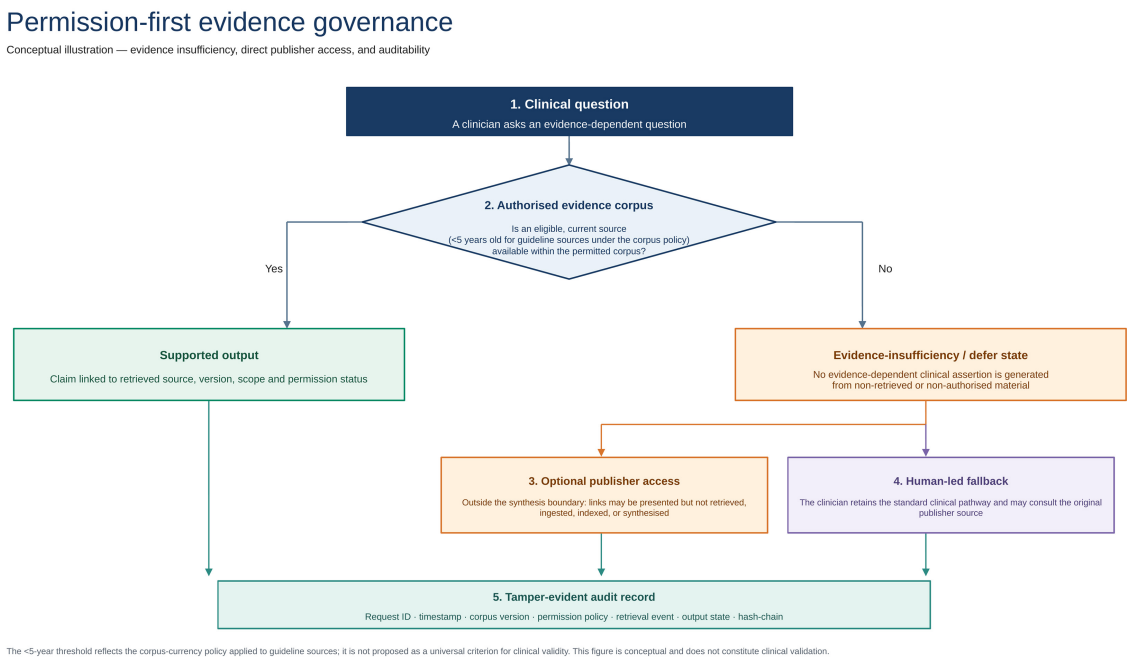

Figure 1. Permission-first evidence governance: supported output, defer state, human fallback, and auditable record. A clinician's evidence-dependent question is evaluated only against an authorised evidence corpus. When an eligible and current source is available within the permitted corpus, the system may produce a supported output linked to the retrieved source, its version, scope, and permission status. When no eligible source is available, the system enters an evidence-insufficiency/defer state and does not generate an evidence-dependent clinical assertion from material that has not been retrieved or is not authorised for synthesis. Official publisher links may optionally be presented outside the synthesis boundary for direct consultation, but are not retrieved, ingested, indexed, or synthesised. The clinician retains the standard clinical pathway. The request identifier, timestamp, corpus version, permission policy, retrieval event, output state, and hash-chain reference are retained in a tamper-evident audit record. The "less than five years" condition shown in the figure reflects the corpus-currency policy applied to guideline sources in the illustrated architecture; it is not proposed as a universal criterion for clinical validity. This figure is conceptual and does not constitute clinical validation.

destruction after compromise[22]. In the proposed architecture, records are hash-linked so each payload binds to the digest of the preceding record. Independent recomputation additionally requires deterministic serialisation or canonicalisation and an initial trust anchor; these are implementation specifications of this proposal, not requirements established by reference[22].

National Institute of Standards and Technology (NIST) Special Publication (SP) 800-92 addresses operational log management[23]; International Organization for Standardization (ISO) 27789:2021 defines electronic health record (EHR) audit trigger events and audit data[24]; and healthcare implementations have demonstrated immutable access logs on permissioned distributed ledgers[25]. These precedents ground durable, reviewable, tamper-evident records without prescribing this architecture.

The audit trail is the artefact available to regulators, review boards, or expert witnesses. Without it, *post-hoc* reconstruction of how unsupported content was generated or surfaced becomes substantially more difficult.

Three properties are central: integrity protection and tamper-evidence, content selectivity, and retention. Integrity protection makes unauthorised modification detectable; selectivity enables reconstruction of retrieved sources, refusal class, model and corpus versions, and applied thresholds without exposing clear-text patient data; retention keeps the artefact available on a medically meaningful time horizon.

Stronger implementations may add external timestamping, independent verification, or distributed-ledger anchoring; these are implementation options rather than requirements of the proposed architecture[22-25].

THE SURGICAL TRIAD IS ALREADY THERE: PREPARATION, TRAINING, DAILY COMPANION

Surgical practice already spans three operational phases: preparation, training, and daily clinical work. Source checking, institutional protocols, curricula, case records, and handoffs already embody expectations of justification and traceability; AI should preserve rather than erode them.

The question is therefore whether AI preserves the auditability expected of these established practices, not how the four commitments map onto a product line.

In preparation, a source-grounded retrieval system with a tamper-evident audit trail instruments the surgeon's existing source-checking reflex[9,10,23,24]. Assertions should remain traceable to retrieved passages, refusals classified, and interactions recorded for later reconstruction.

In training, the 2025 consensus supports simulation, AI-enabled objective feedback, automated skill assessment, error identification, and validation before widespread adoption[5]. The proposed architecture adds inspectability of evidentiary basis, system version, trainee interaction, and feedback, allowing training or accreditation bodies to review an auditable artefact.

In the daily companion layer, Article 14 of the EU AI Act makes human oversight salient for high-risk AI, while medical-device AI remains subject to the Medical Devices Regulation (MDR) and In Vitro Diagnostic Medical Devices Regulation (IVDR) and their interplay with the AI Act[26,27]. Mascagni *et al.* demonstrated a deep-learning computer-vision system that segments hepatocystic anatomy and assesses critical-view-of-safety criteria from laparoscopic images[28]. That perception layer is distinct from the proposed governance layer: provenance can identify video or frame context, model and version, and observation; permission and mandate constrain use; abstention or refusal governs outputs outside validated or authorised conditions; and audit records the resulting events. Source-grounded retrieval becomes relevant when an observation is converted into an evidence-dependent assertion or recommendation. The same governance pattern can surround future multimodal or vision-language components without requiring them.

Latency controls need not share one path. Pre- or post-operative consultation can perform retrieval, authorisation checks, and durable logging synchronously. In high-frequency intra-operative perception, the validated perception model remains on the real-time path while audit records are committed at event level asynchronously or in short batches; retrieval is invoked when an observation becomes an evidence-dependent assertion rather than on every frame. Event granularity, batching, and acceptable latency must be validated for intended use and risk; no universal latency budget is proposed. Auditable clinical RAG architectures likewise identify latency and usability as feasibility constraints[10].

Thus, the four commitments provide a coherent governance and assurance posture across phases, with implementation adapted to modality and workflow.

REGULATORY CONVERGENCE: OVERLAPPING OBLIGATIONS AND ARCHITECTURAL IMPLICATIONS

The EU AI Act, MDR/IVDR, and General Data Protection Regulation (GDPR) impose overlapping obligations rather than a single system architecture. The AI Act classifies systems as high risk in specified circumstances and Articles 9-15 address risk management, data governance, technical documentation,

record-keeping, transparency, human oversight, accuracy, robustness, and cybersecurity[26]. Where personal data are processed, GDPR adds accountability, data protection by design and default, records of processing where applicable, and security[29]. Medical Device Coordination Group (MDCG) 2025-6 addresses the interplay between MDR/IVDR and the AI Act for medical-device AI[27].

These regimes do not mandate source-grounded retrieval, typed refusal, or cryptographic chaining[26,27,29]. The four commitments are candidate engineering patterns that may help operationalise and evidence requirements for documentation, record-keeping, transparency, oversight, governance, accountability, and security. Their records can contribute to broader conformity and accountability artefacts but do not establish regulatory compliance by themselves.

EXTENDING THE SAME COMMITMENTS TO ADJACENT LAYERS

The same governance principles extend beyond the immediate clinical decision to institutional governance and scholarly knowledge, with implementation adapted to modality, content type, and workflow.

At the institutional layer, hospitals, academic centres, and guideline authorities define which AI behaviour is permitted, against which protocols, and under which oversight. Applying the same provenance, authorisation, refusal, and audit principles can connect a clinical event to the rule governing it.

At the scholarly knowledge layer, the relevant objects are the corpus, its version discipline, and its relationship to source authorities. Permission-first, machine-readable policies can support AI-mediated distribution while preserving explicit control over scope, version, and jurisdiction[13,14].

These extensions locate the argument at the structural layer where clinical assertions, institutional rules, and published guidance must remain reconstructible and attributable.

CONCLUSION: THE NEXT COLLECTIVE TASK

Surgery has advanced by embedding ethical duties into practices that alter the conditions under which harm occurs. Auditable medical AI requires the same move from principle to inspectable architecture.

Source-grounding, permission-first governance, typed refusal, and cryptographically chained audit trails are proposed as architectural conditions supporting defensibility in surgical AI.

The author-developed Refusal Stack specification underlying the taxonomy is publicly deposited[18] as a citable starting point, not a final standard. The surgical community should now subject the architectural layer to the same collective scrutiny applied to definitions, training standards, and clinical practice.

DECLARATIONS

Authors' contributions

The author contributed solely to the article.

Availability of data and materials

No datasets were generated or analysed. The author-developed technical specifications informing this Perspective are publicly deposited on Zenodo: Sovereignty by Design - Why Guideline Authorities and Regulators Need Their Own Distribution Rail for Clinical AI, DOI:10.5281/zenodo.20258141; and The Refusal Stack - Engineering Medical AI for Adversarial Audit, DOI:10.5281/zenodo.20257894.

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tool ChatGPT (version GPT-5.6 Sol, released 2026-07-09) was used solely for language editing. The tool did not influence the study design, data collection, analysis,

interpretation, or the scientific content of the work. The author takes full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

None.

Conflicts of interest

Mastari FA is founder and CEO of CLINETHIX LLC, which developed the author-deposited technical specifications described above. The Perspective reports no original clinical data, author-generated benchmark results, patient-specific recommendations, or commercial offer. The specifications are disclosed as citable technical materials and are not used as independent evidence where established external literature is available. No commercial promotion of CLINETHIX systems is intended.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). Available from: https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1%3A2022cv01461/575368/54/. [Last accessed on 2 September 2026].
2. Gumbs AA, Perretta S, d'Allemagne B, Chouillard E. What is Artificial Intelligence Surgery? *Art Int Surg.* 2021;1:1-10. DOI
3. Capelli G, Verdi D, Frigerio I, et al. ; Artificial Intelligence Surgery Editorial Board Study Group on Ethics. White paper: ethics and trustworthiness of artificial intelligence in clinical surgery. *Art Int Surg.* 2023;3:111-22. DOI
4. World Health Organization. Ethics and governance of artificial intelligence for health. Available from: https://www.who.int/publications/i/item/9789240029200. [Last accessed on 2 September 2026].
5. Grasso SV, Spolverato G, Capelli G, et al. The role of advanced technologies and artificial intelligence (AI) in surgical training: a consensus report. *Cureus.* 2025;17:e98371. DOI PubMed PMC
6. Gumbs A, Diana M, Rawicz-Prusyński K, et al. AIONS consensus conference on definitions of artificial intelligence surgery, surgomics and robotics. *Art Int Surg.* 2026;6:98-113. DOI
7. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* 2021;181:1065-70. DOI PubMed PMC
8. Chung P, Swaminathan A, Goodell AJ, et al. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI.* 2026;3. DOI
9. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst.* 2020;33:9459-74. Available from: https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html [Last accessed on 2 September 2026].
10. Alu FF, Oluwadare S. An auditable and source-verified framework for clinical AI decision support: integrating retrieval-augmented generation with data provenance. *Front Artif Intell.* 2026;9:1737532. DOI PubMed PMC
11. Lin C, Lin JY, Lin YS. Evidence-graded decision authorization for safe clinical AI: a constrained reasoning framework. *medRxiv* 2026; Epub ahead of print. DOI
12. Bedi S, Cui H, Fuentes M, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med.* 2026;32:943-51. DOI PubMed PMC
13. World Wide Web Consortium. ODRL Information Model 2.2. W3C Recommendation, 15 February 2018. Available from: https://www.w3.org/TR/odrl-model/. [Last accessed on 2 September 2026].
14. Lawson J, Cabili MN, Kerry G, et al. The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genom.* 2021;1:100028. DOI PubMed PMC
15. Mozannar H, Sontag D. Consistent estimators for learning to defer to an expert. *Proc Mach Learn Res.* 2020;119:7076-87. Available from: https://proceedings.mlr.press/v119/mozannar20b.html [Last accessed on 2 September 2026].

16. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med.* 2021;4:4. DOI PubMed PMC
17. Machcha S, Yerra S, Gupta S, et al. Knowing when to abstain: medical LLMs under clinical uncertainty. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); 2026 Mar; Rabat, Morocco. Stroudsburg, PA, USA: Association for Computational Linguistics, 2026; pp. 6153-82. DOI
18. Mastari FA. The refusal stack - engineering medical AI for adversarial audit. Zenodo, 17 May 2026. Available from: https://zenodo.org/records/20257894 [Last accessed on 2 September 2026].
19. Li S, Balachandran V, Feng S, et al. MediQ: question-asking LLMs and a benchmark for reliable interactive clinical reasoning. Advances in Neural Information Processing Systems 37; 2024 Dec 10-15; Vancouver, BC, Canada. San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024; pp. 28858-88. DOI
20. Swaminathan A, Lopez I, Wang W, et al. Selective prediction for extracting unstructured clinical data. *J Am Med Inform Assoc.* 2024;31:188-97. DOI PubMed PMC
21. Watanabe Y, Kobashi Y, Kojima T, Iwasawa Y, Okuno Y, Matsuo Y. ClinDet-Bench: beyond abstention, evaluating judgment determinability of LLMs in clinical decision-making. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track); 2026 Jul; San Diego, California, USA. Stroudsburg, PA, USA: Association for Computational Linguistics, 2026; pp. 681-703. DOI
22. Schneier B, Kelsey J. Cryptographic support for secure logs on untrusted machines. Proc 7th USENIX Security Symposium, San Antonio, TX, 26-29 January 1998; pp. 53-62. Available from: https://www.usenix.org/conference/7th-usenix-security-symposium/cryptographic-support-secure-logs-untrusted-machines. [Last accessed on 2 September 2026].
23. Kent K, Souppaya M. Guide to Computer Security Log Management. NIST Special Publication 800-92. Gaithersburg, MD: National Institute of Standards and Technology; 2006. DOI
24. International Organization for Standardization. ISO 27789:2021. Health informatics - Audit trails for electronic health records. ISO, 2021. Available from: https://www.iso.org/standard/75313.html. [Last accessed on 2 September 2026].
25. Tith D, Lee JS, Suzuki H, et al. Application of blockchain to maintaining patient records in electronic health record for enhanced privacy, scalability, and availability. *Healthc Inform Res.* 2020;26:3-12. DOI PubMed PMC
26. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). EUR-Lex, 2024. Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. [Last accessed on 2 September 2026].
27. Medical Device Coordination Group. MDCG 2025-6: FAQ on Interplay between the Medical Devices Regulation (MDR) & In Vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). European Commission, June 2025. Available from: https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en. [Last accessed on 2 September 2026].
28. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Ann Surg.* 2022;275:955-61. DOI PubMed
29. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). OJ L 119, 4 May 2016; pp. 1-88. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj. [Last accessed on 2 September 2026].