OXFORD

Young BJS

Perioperative care

BJS, 2025, znaf217
https://doi.org/10.1093/bjs/znaf217
Young BJS

Bridging the gap: exposing the hidden challenges towards adoption of artificial intelligence in surgery

Ana Manzano Rodriguez[1,2,3,4,*], Cees G. M. Snoek[1,2,†] and Marlies P. Schijven[1,3,4,5,†]

[1]Data Science Center HAVA-Lab, University of Amsterdam, Amsterdam, the Netherlands
[2]Video & Image Sense Lab, Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands
[3]Amsterdam UMC Location University of Amsterdam, Surgery, Amsterdam, the Netherlands
[4]Amsterdam Public Health, Digital Health, Amsterdam, the Netherlands
[5]Amsterdam Gastroenterology and Metabolism, Amsterdam, the Netherlands

*Correspondence to: Ana Manzano Rodriguez, MSc, PhD Candidate, Data Science Center HAVA-Lab, University of Amsterdam, Vendelstraat 2, 1012 XX Amsterdam, the Netherlands, (e-mail: a.manzanorodriguez@amsterdamumc.nl)

†Equal advising.

Introduction

Artificial intelligence (AI) can be defined as the field of computer science focused on creating systems capable of performing tasks that normally require human intelligence, such as learning, reasoning, and decision-making. AI is advancing faster than ever before, driven by a new class of AI models known as foundation models[1] (see Supplementary material for definitions of key technical terms). These large-scale neural networks learn from extensive amounts of diverse types of data, including anything digital, be it text, video, or even protein structures, allowing them to capture complex patterns and relationships across domains. Traditionally, AI models relied on narrowly focused datasets, often much smaller in scope, where sets of data needed to be manually annotated to give proper meaning—a process that is costly, time-consuming, and dependent on domain experts. Foundation models (for example the large language models such as ChatGPT) overcome these limitations by first training on vast unlabelled data and then fine-tuning on smaller, labelled, task-specific datasets. This approach reduces the need for manual annotation and allows these general-purpose AI models to be applied across a wide range of applications. In surgery, where endoscopic and minimally invasive procedures generate enormous amounts of unlabelled video data, foundation models are particularly promising, as they could support multiple tasks—such as identifying procedural phases, recognizing instruments, or assessing surgical performance—within a single, adaptable model.

Despite its rapid progress in many other domains, the adoption of foundation models in the operating room remains limited to date. This leads to the question: why has AI already influenced and transformed so many aspects of our daily lives, yet struggles to gain proper traction in surgery? What factors are we overlooking as a community? Are we addressing the real needs of the surgical field, or is AI advancing without aligning with surgical clinical requirements?

Rather than focusing on issues that our community is already familiar with—such as the scarcity of properly annotated data, issues concerning data complexity, or the lack of data transparency[2]—we highlight less-discussed structural and collaborative obstacles. These include the fragmentation between surgical and AI research communities, which directly contributes to a lack of standardized data, tasks, and meaningful evaluation metrics. Consequently, it makes reproducibility difficult and ultimately slows progress. Drawing insights from how other fields have overcome similar challenges, we call for action to establish collaborative research standards to accelerate the integration of AI into surgical practice, addressing surgeons, data scientists, and related researchers.

Why is AI adoption in surgery so challenging?

Fragmentation of research communities

Adoption of AI in surgery requires bridging two distinct and previously disconnected fields: AI research and clinical surgical practice. Each of these disciplines comes with its own research culture, terminology, priorities, and methodologies. Experts from both sides have little interaction as they work in different buildings, attend different conferences and publish in separate fields, making true interdisciplinary collaboration difficult. Consequently, joint research questions are rarely addressed, limiting the utility of AI solutions and complicating cross-field adoption.

Practical challenges exist, as data scientists often lack access to clinical datasets, which most likely reside within secure hospital systems. This can lead to the development of AI models trained on limited, selective patient data, embedding bias. Researchers may have to rely on accessible datasets, such as Liu *et al.*[3] and Yuan *et al.*[4]. Public datasets, when available, may be poorly curated and unlikely to generalize across clinical settings, where variability extends beyond patient characteristics to institutional practices, surgical techniques, and technological infrastructure.

Received: March 28, 2025. **Revised:** September 01, 2025. **Accepted:** September 09, 2025

Table 1 Cultural and structural differences between artificial intelligence (AI) and surgical research practices

Aspect	Surgical research conventions	AI research conventions
Data sharing	Clinical data are often siloed due to privacy, institutional and/or legal constraints	Public datasets are widely available (for example Imagenet, COCO)
Modelling and code availability	Code and annotation of data are rarely shared in methodology of publications where more emphasis is placed on study model, sample size collection, use of statistics and clinical outcomes	Open-source code is expected (for example GitHub repositories) to promote transparency and reuse
Preprint culture	Rare use of preprints; preference for peer-reviewed and PubMed/Medline-indexed journals with impact factor	Preprints (for example arXiv) are commonly used to share research quickly, followed by peer-reviewed publication in conferences and journals
Benchmarking and leaderboards	Lack of standardized benchmarks: performance of AI models is often evaluated in isolated studies without disclosure of modelling methods	Centralized benchmarks (for example GLUE, MedQA) and public leaderboards drive progress
Reproducibility standards	Clinical care follows strict protocols (for example safety checklists), but research reproducibility (such as code sharing or standardized reporting) is less formalized	Reproducibility is increasingly prioritized, with shared code and standardized documentation
Collaboration practices	Collaborations often occur within institutions or regional networks linked to conditions, surgical domain, and/or disease patterns	Cross-institutional and international collaborations are common across sectors

The pace of research also differs. AI advances rapidly, with thousands of papers published monthly across peer-reviewed conferences (for example NeurIPS, CVPR, AAAI) and on preprint servers, often unnoticed by the surgical community, which relies on peer-reviewed journals with inherent delay. Literature searches vary as well: medical research uses databases like PubMed, Medline, CINAHL, or Embase, whereas AI research is accessed via preprints, conference proceedings, and platforms like Google Scholar.

This raises the question: where can these communities meaningfully engage? Even the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), the leading conference at the intersection of AI and medicine, remains largely unknown to many surgical professionals. According to data provided by MICCAI's Submission Platform Manager (Wong K, personal communication, March 2025), among all attendees at the 2024 conference, only 14.9% were associated with hospitals, medical institutes, or medical universities, not necessarily meaning they were practising healthcare professionals. In contrast, 44.9% were part of technical, engineering, or computer science institutions, 2.3% had mixed expertise, and for 38%, these data were not specified (see Supplementary material). Efforts to create interdisciplinary venues for Surgical Data Science exist, but awareness and engagement remain low[5].

Lack of benchmark standardization

Just as surgical guidelines guarantee consistency and safety in the operating room, benchmarks in AI provide a standard way to measure and compare results. Benchmarking refers to the process of establishing standardized datasets, tasks, and evaluation metrics to ensure reproducibility, enable fair comparisons and track community progress.

The AI research community operates on principles of open science and collaboration, building quickly upon shared prior work. By refining and expanding existing models using shared datasets and standardized evaluation methods, researchers can focus on meaningful improvements instead of reinventing the wheel. Collaboration in surgical AI research is more challenging —Eckhoff *et al.* reported that 'it is difficult to perform multi-institutional studies involving surgical video due to the lack of well-defined data structure standards'[6]. Surgical professionals, who are generally not AI experts, require robust guidance and well-defined references to develop benchmarks that are useful for research.

Next, tasks and evaluation metrics studied for surgical AI applications lack consistency. Surgical AI research tends to focus on solving highly specific tasks—such as segmentation of anatomical structures, tool tracking, or skill assessment—developing datasets and models tailored to a single procedure or problem, which limits broader applicability. This approach misses the opportunity to benefit from the latest technologies such as foundation models, which could serve as versatile assistants supporting surgeons throughout entire procedures rather than solving isolated tasks. Evaluation also lacks uniformity, as studies rely on different metrics, making direct comparisons difficult. Moreover, these metrics are often purely technical and do not capture clinically meaningful outcomes, limiting their impact in real-world scenarios.

Taken together, these issues highlight a fundamental challenge: surgical AI lacks proper benchmarking when adopting modern AI. This lack of structure, whilst being in a hyperregulated situation, creates a reproducibility crisis, making previous studies difficult to replicate, limiting the maturation of AI models and further adaptation to other procedures or applications. If surgical AI continues to rely on isolated, task-specific research, it risks falling behind in its development and integration into the broader AI landscape.

To better illustrate these cultural differences, *Table 1* compares common practices in surgical and AI research. These comparisons are not meant to imply shortcomings but rather reflect distinct historical contexts, priorities, and constraints in each field.

Those fields demonstrating significant impacts from AI have already overcome benchmarking challenges. For example, the TREC Video Retrieval Evaluation (TRECVid) benchmark organized by the U.S. National Institute of Standards and Technology advanced the field of video retrieval by providing digitized video datasets, challenging retrieval tasks, developing an evaluation protocol and criteria, and mandating open presentation and publication of methodologies[7] . Similarly, the ImageNet competition transformed computer vision by providing a large, standardized dataset, clear evaluation metrics, and a yearly workshop to highlight the best approaches, enabling fair comparisons and driving

breakthroughs in deep learning architectures[8]. In the medical domain, radiology's DICOM standard improved data-sharing and reproducibility, with EU-backed initiatives such as the DICOM Library promoting its use[9,10].

The challenges already overcome by other domains should inform the development of surgical AI through adapting and customizing those strategies while respecting the unique legal, ethical, and technical issues of surgical research and practice.

An initial roadmap for surgical AI could include: developing shared task definitions and data formats to promote compatibility and reproducibility; organizing open surgical AI challenges with ethically approved datasets and standardized evaluation metrics; and hosting regular interdisciplinary events where surgeons and AI researchers discuss common benchmarks, challenges, and results.

Although institutional initiatives are key, individual surgeons can contribute by engaging in discussions, contributing annotated data, or offering clinical perspectives ensuring these technologies truly support surgical practice. It is also crucial to promote openness and transparency—encouraging authors to share data, code, and experimental setups—to strengthen reproducibility, build confidence in results, and reduce duplicated efforts. These initiatives serve a dual purpose: short-term initiatives catalyse progress, whereas long-term educational programmes maintain continuity and evolution.

A call for action

Bridging the gap between AI research and surgery is essential for reaping the benefits AI can bring to surgical practice. The path forward is clear: fostering better collaboration between these very different fields of expertise. Only through collective action can surgical AI move beyond isolated studies towards meaningful advancements creating a true ecosystem. With well-defined standards, the field can evolve faster, achieving the significant advances we are all expecting. The potential is immense, but without structured cooperation, it will remain unrealized. Now is the time for our disciplines to unite, plan and deliver.

Funding

The authors have no funding to declare.

Acknowledgements

We thank MICCAI for sharing attendee statistics. This work was supported by the University of Amsterdam's Data Science Centre, as part of the HAVA Lab.

Author contributions

Ana Manzano Rodriguez (Investigation, Writing—original draft), Cees G. M. Snoek (Conceptualization, Supervision, Writing—review & editing), and Marlies P. Schijven (Conceptualization, Supervision, Writing—review & editing).

Disclosure

The authors declare no conflict of interest.

Supplementary material

Supplementary material is available at *BJS* online.

References

1. Lam K, Qiu J. Foundation models: the future of surgical artificial intelligence? *Br J Surg* 2024;**111**:znae090
2. Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A *et al.* Surgical data science—from concepts toward clinical translation. *Med Image Anal* 2022;**76**:102306
3. Liu D, Li Q, Jiang T, Wang Y, Miao R, Shan F *et al.* Towards unified surgical skill assessment. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) Virtual, USA*. IEEE, 2021, 9517–9526.
4. Yuan K, Navab N, Padoy N. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Adv Neural Inf Process Syst* 2024;**37**: 122952–122983
5. Surgical Data Science Summer School. https://www.edu4sds.org (accessed 20 August 2025)
6. Eckhoff JA, Rosman G, Altieri MS, Speidel S, Stoyanov D, Anvari M *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* 2023;**37**:8690–8707
7. Smeaton AF. Large scale evaluations of multimedia information retrieval: the TRECVid experience. In: *International Conference on Image and Video Retrieval, Singapore*. Springer, 2005, 11–17.
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S *et al.* Imagenet large scale visual recognition challenge. *Int J Comput Vis* 2015;**115**:211–252
9. Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J Am Med Inform Assoc* 1997;**4**: 199–212
10. DICOM Library. https://www.dicomlibrary.com (accessed 1 March 2025)