Review article

Artificial intelligence in surgery

Chris Varghese [1], Ewen M. Harrison [2], Greg O'Grady[1,3] & Eric J. Topol [4] ✉

Artificial intelligence (AI) is rapidly emerging in healthcare, yet applications in surgery remain relatively nascent. Here we review the integration of AI in the field of surgery, centering our discussion on multifaceted improvements in surgical care in the preoperative, intraoperative and postoperative space. The emergence of foundation model architectures, wearable technologies and improving surgical data infrastructures is enabling rapid advances in AI interventions and utility. We discuss how maturing AI methods hold the potential to improve patient outcomes, facilitate surgical education and optimize surgical care. We review the current applications of deep learning approaches and outline a vision for future advances through multimodal foundation models.

Artificial intelligence (AI) tools are rapidly maturing for medical applications, with many studies determining that their performance can exceed or complement human experts for specific medical use cases[1–3]. Unimodal supervised learning AI tools have been assessed extensively for medical image interpretation, especially in the field of radiology, with some success in recognizing complex patterns in imaging data[3,4]. Surgery, however, remains a sector of medicine where the uptake of AI has been slower, but the potential is vast[5].

Over 330 million surgical procedures are performed annually, with increasing waiting lists[6,7] and growing demands on surgical capacity[8]. Substantial global inequities exist in terms of access to surgery, the burden of complications and failure to rescue (that is, post-complication mortality) after surgery[9–13]. A multifaceted approach to surgical system strengthening is required to improve patient outcomes, including better access to surgery, surgical education, the detection and management of postoperative complications and optimization of surgical system efficiencies. To date, minimally invasive surgery has been a dominant driver of improvements in surgical outcomes, reducing postoperative infections, length of stay and postoperative pain and improving long-term recovery and wound healing. Enhanced recovery programs[14], improved patient selection, broadening adjuvant approaches and organ-sparing treatments have also been important contributors. We are now entering an era where data-driven methods will become increasingly important to further improving surgical care and outcomes[15]. AI tools hold the potential to improve every aspect of surgical care; preoperatively, with regards to patient selection and preparation; intraoperatively, for improving procedural performance, operating room workflows and surgical team functioning;

and postoperatively, to reduce complications, reduce mortality from complications and improve follow-up.

Current AI applications in surgery have been mostly limited to unimodal deep learning (Box 1). Transformers are a particular recent breakthrough in neural network architectures that have been very effective empirically in several areas, owing to their improved computational efficiency through parallelizability[16] and their enhanced scalability (with models able to handle vast input parameters). Such transformer models have been pivotal in enabling multimodal AI and foundation models, with substantial potential in surgery.

Emerging applications of AI in surgery include clinical risk prediction[17,18], automation and computer vision in robotic surgery[19], intraoperative diagnostics[20,21], enhanced surgical training[22], postoperative monitoring through advanced sensors[23,24], resource management[25], discharge planning[26] and more. The aim of this state-of-the-art Review is to summarize the current state of AI in surgery and identify themes that will help to guide its future development.

Preoperative

There is much room for improvement of preoperative surgical care, encompassing areas of active surgical research such as diagnostics, risk prognostication, patient selection, operative optimization and patient counseling—all aspects of the preoperative pathway of patients receiving surgery where AI has emerging capabilities.

Preoperative diagnostics

Patient selection and surgical planning have become increasingly evidence based, but are still contingent on experiential intuitions

[1]Department of Surgery, University of Auckland, Auckland, New Zealand. [2]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK. [3]Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand. [4]Scripps Research Translational Institute, La Jolla, CA, USA. ✉e-mail: etopol@scripps.edu

Emergent AI models and terminology

Diverse terminologies are used in the literature and for clarity are defined here:

- Unimodal AI: models that operate and process data of a singular format, such as images, text or audio alone.
- Multimodal AI: models that simultaneously integrate diverse data formats provided as training and prompt inputs, including images, text, biosignals, -omics data and more.
- Generative AI: AI systems that can generate new content, such as texts, images (for example, DALL.E[170]) and other data formats. LLMs are an example of generative AI. These are trained on a vast corpus of language-based parameters.
- Transformer models: a type of model that uses self-attention rather than recurrence and convolutions in making predictions that offer advantages in understanding long-range dependencies in input parameters, scalability and improved performance[171].
- Foundation models: large pretrained models that serve a variety of purposes. These models are trained on diverse data sources and can be fine-tuned for specific tasks. Such models serve to decrease the need for recursive and extensive model training.
- Supervised learning: models trained on labeled datasets (the majority of applications to date).
- Self-supervised learning: models that attempt to create their own data labels on unlabeled input data.
- Unsupervised learning: models that evaluate unlabeled input data for inherent structures and relationships using clustering and dimensionality reduction methods.

(and biases), with profound individual and regional variabilities. The influence of AI may emerge most rapidly in the context of preoperative imaging for early diagnosis and surgical planning. As an example, a model-free reinforcement learning algorithm showed promise when applied to preoperative magnetic resonance images to identify and maximize tumor tissue removal while minimizing the impact on functional anatomical tissues during neurosurgery[27]. Technically challenging cases with high between-patient anatomical variations, such as in pulmonary segmentectomy, have been met with pioneering approaches to enhance preoperative planning with novel amalgamations of virtual reality and AI-based segmentation systems. In a pilot study by Sadeghi et al.[28], AI segmentation with virtual reality resulted in critical changes to surgical approaches in four out of ten patients.

Accurate preoperative diagnosis is an important area of surgical practice with substantial influence on clinical decision-making and therapeutic planning. For example, in the context of breast cancer diagnostics, the RadioLOGIC algorithm extracts unstructured radiological report data from electronic health records to enhance radiological diagnostics[29]. Extraction of unstructured reports using transfer learning (applying cross-domain knowledge to boost model performance on related tasks) showed high accuracy for the prediction of breast cancer (accuracy >0.9), and pathological outcome prediction was superior with transfer learning (area under the receiver operating characteristic curve (AUROC) = 0.945 versus 0.912). This report emphasizes the value of integrating natural language processing of unstructured text within existing infrastructures for promoting preoperative diagnostic accuracy. Another prominent example is a three-dimensional convolutional neural network that detected pancreatic ductal adenocarcinoma on

diagnostic computed tomography and visually occult pancreatic ductal adenocarcinoma on prediagnostic computed tomography with AUROC values of 0.97 and 0.90, respectively[30]. Streamlining preoperative diagnostics can optimize integrated multidisciplinary surgical treatment pathways and facilitate early detection and intervention where timely management is prognostically critical[31]. Progress with large language models and integration with electronic healthcare record systems— particularly the utility of foundation models empowering the analysis of unlabeled datasets—could be transformative in enabling earlier disease diagnosis and early treatment before disease progression[32,33].

AI-based diagnosis is one of the most mature areas of surgical AI where model accuracy and generalizability are seeing early clinical translation. Numerous in-depth and domain-specific explorations of the efficacy of AI in endoscopic[34,35], histological[36], radiological[37] and genomic[38] diagnostics have been outlined elsewhere (see ref. [39]). These task-specific advances are enabling more accurate diagnostics and disease staging in the oncology space, with substantial potential to optimize surgical planning. However, to date, all applications have been unimodal; novel transformer models that are able to integrate vastly more data, both in quantity and format, could spur further progress in the near future.

Clinical risk prediction and patient selection

High-accuracy risk prediction seeks to enable enhanced patient selection for operative management to improve outcomes, reduce futility[40], better inform patient consent and shared decision-making[41], triage resource allocation and enable pre-emptive intervention. It remains an elusive goal of surgical research[18]. Numerous critical reports have highlighted the high risk of bias[42] and overall inadequacy of the high majority of clinical risk scores in the literature, with few penetrating routine clinical practice[43]. It is important to remember that in the pursuit of enhanced predictive capabilities, the novelty of a tool (such as AI) should not supersede a tool's utility. Finlayson et al.[44] expertly summarize the often false dichotomy of machine learning and statistics; they argue that dichotomizing machine learning as separate to classical statistics neglects its underlying statistical principles and conflates innovation and technical sophistication with clinical utility.

The majority of current AI-based risk prediction tools offer sparse advances over existing tools, and few are used in clinical practice[45]. The COVIDSurg mortality score is one machine learning prediction score based on a generalized linear model (chosen for its superiority to random forest and decision tree alternatives) that shows a validation cohort AUROC of 0.80 (95% confidence interval = 0.77–0.83)[46]. Numerous other machine learning risk scores exist for preoperative prediction of postoperative morbidity and mortality[45,47–51]. One notable example is the smartphone app-based POTTER calculator, which uses optimal classification trees and outperforms most existing mortality predictors with an accuracy of 0.92 at internal validation[52], 0.93 in an external emergency surgery context[48] and 0.80 in an external validation cohort of patients >65 years of age receiving emergency surgery[47]. Notably, POTTER also showed improved predictive accuracy compared with surgeon gestalt[53]. Deep learning methods have also shown utility in neonatal cardiac transplantation outcomes, with high accuracy for predicting mortality and length of stay (AOROC values of 0.95 and 0.94, respectively)[54].

The use of AI in surgical risk prediction remains an emerging field that is lacking in randomized trials[55] and external validation, and has a high risk of bias[56]. Future work should move toward predictions of relevance to clinicians and patients[57] and prioritize compliance with the CONSORT-AI extension[58], TRIPOD (and its upcoming AI extension)[59,60], DECIDE-AI[61] and PRISMA AI[62] and other relevant reporting guidelines, to advance the field in a standardized, safe and efficient manner while minimizing research waste.

Preoperative optimization

Preoperative optimization is still an underdeveloped concept that is beginning to receive more attention in surgical research[63] and could

be leveraged with multimodal inputs. A multifaceted appreciation of patients' cardiovascular fitness, frailty, muscle function and optimizable biopsychosocial factors could be accurately characterized through multimodal AI approaches leveraging the full gamut of -omics data[64,65]. For example, research using AI to detect ventricular function using 12-lead electrocardiograms could rapidly streamline preoperative cardiovascular assessment[66–68]. While more information does not necessarily correlate with improved risk prediction, a more holistic understanding of patient factors in the preoperative setting could be leveraged to optimize characteristics such as sarcopenia, anemia, glycemic control and more, to facilitate improved surgical outcomes.

Patient-facing AI for consent and patient education

Large language models (LLMs)—a form of generative AI—are a generational breakthrough, with the emergence and adoption of ChatGPT occurring at an unprecedented pace and other LLMs emerging at an equally rapid pace. These models have attained high scores on medical entrance exams[69,70] and contextualized complex information as competently as surgeons[71], and there is the potential for patients to interact with them as an initial clinical contact point[72,73]. AI models can augment clinician empathy[74], contribute to reliable informed consent[41,71] and reduce documentation burdens. A recent report demonstrates promising readability, accuracy and context awareness of chatbot-derived material for informed consent compared with surgeons[71]. These advances offer a unique opportunity for tailored patient-facing interventions.

While clinical implementation of AI is a work in progress, there is great potential for superior patient-facing digital healthcare. A pilot clinical trial by the company Soul Machines (Auckland, New Zealand) highlights the potential power of amalgamating LLMs and avatar digital health assistants (or digital people)[75]. OpenAI's fine-tuned generative pretrained transformers and assistant application programming interfaces could be leveraged for such a purpose if solutions to trust and privacy concerns are found[76,77]. The COVID-19 pandemic highlighted the value of decentralized digital health strategies to enable wider access to healthcare, and as global healthcare demands rise, these promising reports offer a valuable augment to the delivery of healthcare. These concepts also offer a step toward a hospital-at-home future that aims to further democratize healthcare delivery. Such innovations have particular utility in surgical care, where preoperative counseling, surgical consent and postoperative recovery and follow-up could all be augmented by patient-facing AI models validated to show high reliability for target indications[41,71,78,79].

In current practice, informed consent and nuanced discussions about surgical care plans are frequently confined to time-limited clinic appointments. Chatbots powered by accurate LLMs offer an opportunity for patients to ask more questions, facilitating ongoing communication and better-informed care. Integrated with accurate deep learning-based risk prediction, such AI communication platforms could offer a personalized risk profile, answer questions about preoperative optimization and postoperative recovery and guide patients through the surgical journey, including postoperative follow-up consultations[80]. Early generative AI models are probably already primed for translation to such clinical education settings[81], with many more rapidly emerging (for example, Hippocratic AI, Sparrow[82] and Gemini (Google DeepMind), BlenderBot 3 (Meta Platforms), HuggingChat and more). Nuanced appreciations of real-world complexity[74] and the introduction of multi-agent conversational frameworks will be key for the testing and implementation of medical AIs[83]. At present, these models are yet to incorporate the vast and historic corpus of the medical literature; however, with specialized fine tuning and advances in unsupervised learning, the accuracy and generality of these tools is likely to improve. Nevertheless, further work to improve the transparency and reliability of such integrations is required[84], as evidenced by recent examples of inaccurate and unreliable information from LLMs in breast cancer screening[85].

In summary, multimodal approaches may transform the preoperative patient flow paradigm. The use of unstructured text from electronic health records, in conjunction with preoperative computed or positron emission tomography, genomics, microbiomics, laboratory results, environmental exposures, immune phenotypes, personal physiologies, sensor inputs and more will enable deep phenotyping at the individual patient level to optimize personalized risk prediction and operative planning. Such advances are highly sought after to improve shared decision-making, patient selection and offer individualized targeted therapy.

Intraoperative

The intraoperative period is a data-rich environment, with continuous monitoring of physiological parameters amid complex insults and alterations to anatomy and physiology. This time is the core of surgical practice. Advances in intraoperative computer vision have enabled preliminary progress in the analysis of anatomy, including assessment of tissue characteristics and dissection planes, as well as pathology identification. Likewise, progressing the reliable identification of instruments and stage of operation and the prediction of procedural next steps are important foundations for future autonomous systems and data-driven improvements in surgical techniques[19].

Events inside the operating theater have substantial impacts on recovery, postoperative complications and oncological outcomes[86]. Yet, despite their pivotal importance, minimal data are currently recorded, analyzed or collected in this setting. Valuable data streams from the intraoperative operative period should be harnessed to contribute to advances in surgical automation and to underpin the utility of AI in the theater space[15]. We envision a future operating room with real-time access to patient-specific anatomy, operative plans, personalized risks and dashboards that integrate information in real time throughout a case, updating based on surgeon and operating team prompts, actions and decisions (Fig. 1).

Intraoperative decision-making

Most pragmatically, enhanced pathological diagnostics from tissue specimens (as overviewed above in the section 'Preoperative diagnostics') could optimize surgical resection margins, reduce operative durations and optimize surgical efficiency[87,88]. One such example is a recent patient-agnostic transfer-learned neural network that used rapid nanopore sequencing to enable accurate intraoperative diagnostics within 40 minutes[88], enabling early information for operative decision-making. Multimodal AI interrogation of the surgical field could aid the determination of relevant and/or aberrant anatomy (with major strides toward such surgical vision already occurring in laparoscopic cholecystectomy[19,89]), augment the surgeon's visual reviews (for example, by employing a second pair of AI 'eyes' to run the bowel when looking for perforation), inform the need for biopsies and quantify the risk of malignancy[90].

The advantages of AI in hypothetico-deductive surgical decision-making are expertly overviewed by Loftus et al.[25]. Deep learning (in particular neural networks) is devised in an attempt to replicate human intuition—a key element of rapid decision-making among experienced surgeons[91]. One of the first machine learning tools developed for intraoperative decision-making[92], which has undergone validation and translation to clinical use, is the hypotension prediction index[93], which has shown proven benefit in two randomized trials[20,94]. This represents an early example of a supervised machine learning algorithm that has undergone external validation and the gold standard of randomized clinical testing to demonstrate benefit. Notably, since its advent, numerous advances in AI methods have emerged to strengthen algorithmic performances[95].

Such models could be improved in the future through continuous learning, ongoing iteration with constant refinement, and external validation. This requires collaboration with regulatory bodies to facilitate
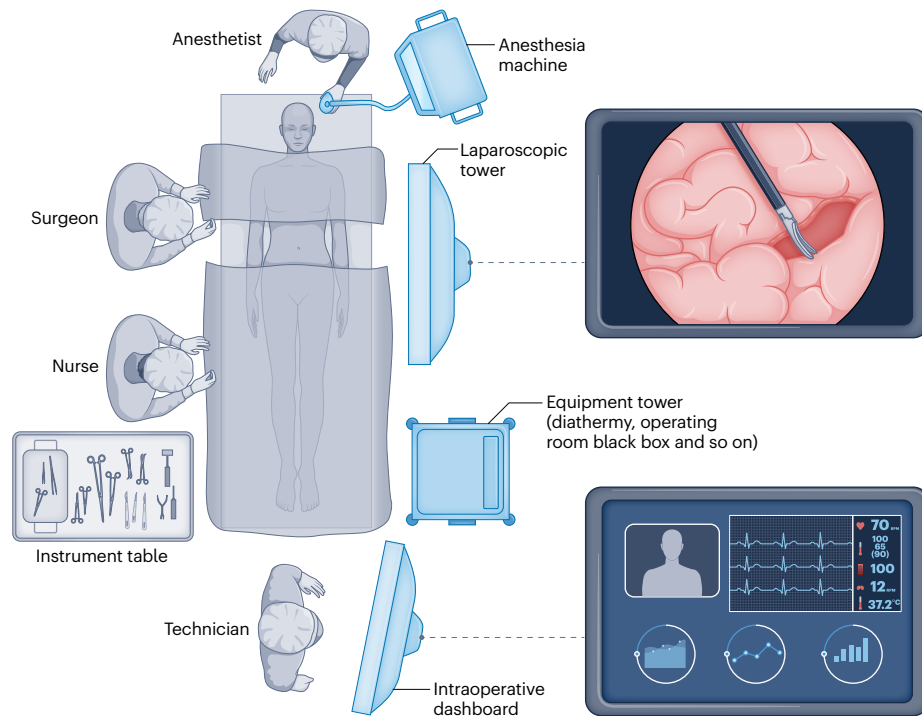
Fig. 1 | Integration of novel AI-powered digital interventions in the intraoperative setting. Operating room components with the potential for AI integration are shown in blue. Traditional laparoscopic towers could be integrated with virtual or augmented reality to facilitate improved three-dimensional views, adjustable overlaid annotations and warning systems for aberrant anatomy. They could also overlay the individual patient's imaging with AI diagnostics to improve R0 resections in oncological surgery, identify anatomical differences and better identify complex planes. Existing diathermy towers could incorporate voice assistants and black box-type systems for audit and quality control. An intraoperative dashboard aligned with the entire theater team could enable virtual consultations, virtual supervision for trainee surgeons and AI-powered access to the corpus of medical knowledge and surgical techniques, all contextualized to the operative plan. In addition, continuous vital signs, anesthetic inputs and patient-centered risks (for example, of hypotension) could be available as the operation progresses, to help the planning of pre-emptive actions and postoperative care. Personalized screens for scrub nurses, indicating stock location, phase detection[169] and the predicted next instrument needed could also improve efficiency.

safe and monitored development and maturation of algorithms as the field rapidly advances. The iterative nature of these models may pose challenges for clinical evidence requirements to keep pace with the rate of innovation.

The operative team

Early efforts to gather data in the operating room include the OR Black Box, the aim of which is to provide a reliable system for auditing and monitoring intraoperative events and practice variations[96,97]. A particularly novel advance toward optimizing surgical teamwork comes from preliminary work toward an AI coach to infer the alignment of mental models within a surgical team[98]. Shared mental models, whereby teams have a collective understanding of tasks and goals, have been identified as a critical component to decreasing errors and harm in safety-critical fields such as aviation and healthcare. Such approaches require further interrogation within a real-world operating room context, but highlight the breadth of opportunities for digital innovation in surgery.

On the theme of surgical teamwork, multimodal digital inputs, including physiological inputs (for example, skin conductance and heart rate variability) for the identification of operative stress, anesthetic inputs (continuous pharmacological and vitals outputs), nursing team staffing inputs and equipment stock and availability inputs, are all routine elements of the operating room experience that could be quantified digitally and integrated into a digital pathway suitable for automation and optimization. The expansion of multimodal inputs and use of generative AI models incorporating both patient and environmental inputs within the operating room present opportunities to augment nontechnical skills that are pivotal in surgery, including communication, situational awareness and operative team functioning[99,100]. Operative fatigue, anesthesiologist–surgeon miscommunication, staffing changeovers and shortages, and equipment unavailability are common causes for intraoperative mistakes and are all amenable to digital tracking. A digitized surgical platform can therefore be envisaged to facilitate an AI-enhanced future. The importance of investment toward the platform itself to leverage utility from digital innovations was, for example, embraced by Mayo Clinic in a recent CEO overview[101].

Surgical robotics and automation

While there has been much progress[19,102], early attempts at computer vision have been limited to specific tasks and have lacked external validation. AI has been applied to unicentric, unimodal video data to identify surgical activity[103], gestures[104], surgeon skill[105,106] and instrument actions[107]. A demonstrative advance has been made by Kiyasseh et al.[108], who have developed a unified surgical AI system that accurately identifies and assesses the quality of surgical steps and actions performed by the surgeon using unannotated videos (area under the curve >0.85 at external validation for needle withdrawal, handling and driving). This procedure-agnostic, multicentric approach, with a view to generalizability, facilitates integration into real-world practice[108]. Technical advances, such as through Meta's self-supervised (SEER) model currently offer particular promise in the realm of computer vision[109]. Similar efforts in the future that aim to improve the feedback available to surgeons, tactile responses from laparoscopic and robotic systems and the identification of optimal surgical actions in the intraoperative window could be advanced through multimodal inputs, including rich physiological monitoring, rapid histological diagnostics and virtual reality-based guidance (for example, toward

identification of aberrant anatomy and tissue planes, perfusion assessments and more). Low-risk opportunities for the integration of these emerging technologies include co-pilot technologies for operative note writing; with surgeon oversight, providing verification and the potential for iterative model improvement[110–112].

Computer vision, surgical robotics and autonomous robotic surgery are at the very early stages of development, with incremental but exciting strides occurring. Importantly, robust frameworks have recently been developed to progress the development of surgical robotics and complementary AI technologies, with guidelines for evaluation, comparative research and monitoring throughout clinical translational phases[113]. Several reviews offer more in-depth analysis of these emerging topics[108,114,115].

Operative education

Surgical education has long been entrenched in apprenticeship models of learning, with little progress toward objective metrics and useful mechanisms for feedback to trainee operators. As discussed above, the operating room setting is a data-rich environment that could be leveraged toward automated, statistical approaches to tailored learning. Reliable feedback results in improved surgical performance[116–118], and data-driven optimization of surgical skill assessment has the potential to have a *trans*-generational impact on surgical practice[119,120]. Recently, the addition of human explanations to the supervised AI assessment of surgical videos improved reliability across different groups of surgeons at different stages of training, such that equitable and robust feedback could be generated through AI approaches[108]. This offers feedback to learners with mitigation against different quality feedback based on different surgeon sub-cohorts[121]. This is a promising example of the nuanced approaches to model development that will be key to the translation and implementation of AI models in real-world surgical education. An exploration of the potential biases of AI explanations in surgical video assessment[122]—namely under- and overskilling biases, based on the surgeon's level of training—highlights the importance of comparisons with current gold standards and utilizing AI outputs as data with which to iterate, learn and optimize toward real-world benefits. In another example, an AI coach had both positive and negative impacts on the proficiency of medical students performing neurosurgical simulation, including improved technical performance at the expense of reduced efficiency[123]. This example also shows the importance of expert guidance in the development and implementation of AI tools in specialized domains, as well as the need for ongoing assessment of such programs. The opportunity to harness AI in operative education is evidenced by the growing number of registered randomized trials evaluating this approach[22,124,125].

The intraoperative period is, therefore, a data-rich environment for surgical AI with early success seen with intraoperative diagnostic and surgical training models, as well as early emerging capabilities in computer vision and automation. Intraoperative applications are diverse and critical for the future of surgery, with vast potential to optimize nontechnical intraoperative functions such as communication, teamwork and skill assessment. Ongoing work toward computer vision systems will lay the groundwork for future autonomous surgical systems.

Postoperative

Postoperative monitoring

The aim of transforming hospital-based healthcare through hospital-at-home services is to liberalize and democratize healthcare and to improve equity and access while unburdening overloaded hospitals. Such a future will enable patients to recover in a familiar environment and will optimize patient recovery, convalescence and their return to functioning in society. Major strides have been made toward reducing postoperative lengths of stay, facilitating early discharge from hospital and improving functional recovery, largely through

minimally invasive surgical approaches, encouragement of earlier return to normal activities, enhanced postoperative monitoring, early warning systems and better appreciation of important contributors to recovery. The implementation of enhanced recovery after surgery programs has been pivotal toward this goal.

However, the postoperative period frequently remains devoid of data-driven innovations, crippling further progress. Many hospitals still rely on four-hourly nurse-led observations, unnecessarily prolonged postoperative stays driven by historic protocols and a 'one size fits all' approach to the immediate postoperative period. Ample opportunity exists for wearables to offer continuous patient monitoring, enabling multimodal inputs of physiological parameters that can contribute toward data-driven, patient-specific discharge planning. This would have the added benefit of unburdening nursing staff from cumbersome vital sign rounds, freeing up time and capacity for more patient-centered nursing care. Leveraging postoperative data can further guide discharge rehabilitation goals and interventions, inform analgesic prescriptions and prognosticate adverse outcomes.

One systematic review highlights 31 different wearable devices capable of monitoring vital signs, physiological parameters and physical activity[23], but further work is required to realize the potential of these data, including improving the quality of research and reporting[126,127]. We envision a future where continuous inputs can be integrated into predictive analytics and dashboard-style interfaces to enable rapid escalation, earlier prognostication of complications and reduced mortality from surgical complications[128,129]. Intensive care units are an example of a highly controlled, data-rich environment where such interventions are emerging, with the potential to modify the postoperative course[130]. Classical machine learning approaches, such as random forests, have been robustly applied in other heterogenous, multimodal time-series applications and stand to have particular value in the postoperative monitoring setting. For example, the explainable AI-based Prescience system monitors vital signs, predicts hypoxemic events five minutes before they happen and provides clinicians with real-time risk scores that continuously update with transparent visualization of considered risk factors[131,132].

To enable multimodal data-driven insights in postoperative sensors, a plethora of novel medical devices and sensors are being pursued (Fig. 2). Real-time physiological sensing of wound healing[133,134], remote identification of superficial skin infections[135,136] and cardiorespiratory sensors[137,138] are all putative technologies to enhance postoperative monitoring.

Complication prediction

The prediction of complications after surgery has been the goal of many academic studies[139,140] and presents a formidable challenge in a complex postoperative setting, with myriad variables affecting care and outcomes. However, the early detection of complications[138,141]—in particular, devastating outcomes such as anastomotic leaks after rectal cancer surgery and postoperative pancreatic fistulas after pancreatic surgery[142,143]—is likely to have a substantial impact on the ability of healthcare systems to reduce mortality following complications[144,145]. MySurgeryRisk represents one of the few advances in complication prediction, using a machine learning algorithm[50]. However, despite promising performance in single-center studies[68], there is little understanding of how to scale these algorithms to other health systems. The value of algorithmic approaches to complication prediction[18] and postoperative monitoring after pancreatic resections has been demonstrated in The Netherlands[146], serving as a reproducible model to aspire to. Wellcome Leap's US$50 million SAVE Program has identified failure to rescue from postoperative complications as a leading cause of death and the third most common cause of death globally[147], and has prioritized this as a target for innovation. Its goals focus on advanced sensing, monitoring and pattern recognition[148]. This remains a nascent field with numerous attempts but few breakthroughs, making complication prognostication
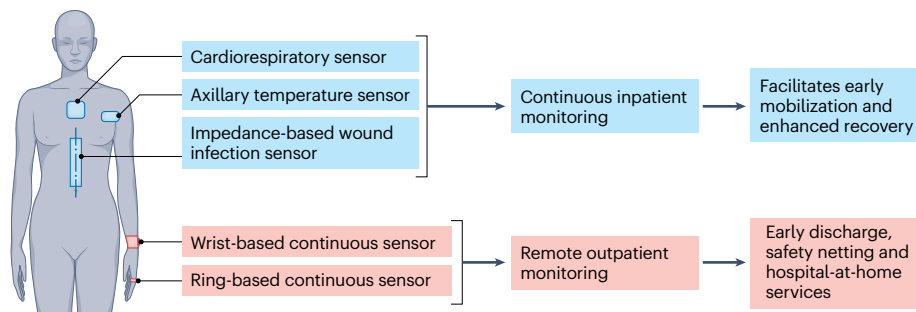
Fig. 2 | Sensor inputs for peri- and postoperative continuous monitoring. Examples of innovative sensors include chest- and axilla-based electrocardiogram, respiratory rate, tidal volume, temperature and skin impedance sensors. In the postoperative setting, when patients are mobilizing and discharged home, wrist- and finger-based sensors offer a safety netting system for the monitoring of sympathetic stress (via heart rate variability and skin impedance), postoperative arrhythmia and wound healing (for the early identification of superficial skin infection and/or wound dehiscence). Sensor-based technologies can be catergorized as continuous inpatient monitoring and early post-discharge monitoring to enable hospital-at-home services.

a high-value target for AI-based technologies, particularly as sensors[149], wearables[23] and devices capable of enabling multimodal, temporally rich inputs emerge.

Home-based recovery
In the United States, 50% of those who undergo a surgical procedure are over 65 years of age[150]. With advancing age, recovery can be prolonged and periods of return to baseline activities of daily living (ADLs) can extend beyond several months. Kim et al.[151] have proposed a multidimensional, AI-driven, home-based recovery model enabled by frequent, noninvasive assessments of ADLs. They centered their proposed paradigm shift on the basis of: (1) continuous real-time data collection; (2) nuanced assessment of relevant measures of activities of daily living; and (3) innovative assessments of ADLs to be leveraged in the postoperative, post-discharge and home-based setting. Again, these innovations would be driven by sensor technologies[152], including the continuous detection of video, location, audio, motion and temperature data in various home-based settings, integrated to provide a continuous assessment of activity patterns. These data contribute toward phenotyping recovery patterns and predicting adverse outcomes (for example, falls), informing care needs and personalized interventions in conjunction with multidisciplinary teams (such as occupational therapists). Systems-level implementation, data privacy and real-world prospective validation are awaited. ClinAIOps (clinical AI operations) is a recent framework for integrating AI into continuous therapeutic monitoring in a way that could be directly translated toward postoperative home-based monitoring[153].

As innovations proliferate toward the goal of remote postoperative monitoring, including mobile technologies[24], sensors[149], wearables[23] and hospital-at-home services, we have identified the key limitations to advancement to be the lack of routine large-scale implementation efforts, collaborations and comprehensive innovation evaluations in line with the IDEAL (idea, development, exploration, assessment and long-term follow-up) framework[154].

Building the evidence base
Emerging AI technologies need to be robustly evaluated in line with existing innovation frameworks[113,155] and, with the advent of multimodal and generative models, regulatory oversight and monitored implementation are pivotal. Complex intervention frameworks provide a robust tool to facilitate ongoing monitoring and rapid troubleshooting[156]. Engagement with all stakeholders, including patients, administrators, clinicians, industry and scientists, will be important to align visions and work concertedly toward improved surgical care.

While emerging models demonstrate promise, robust, prospective, randomized evidence is required to demonstrate improvements in patient care. To date, only six randomized trials of AI exist in surgery

(Table 1), all of which employed unimodal approaches, but the increasing number of trial registrations on the topic of assessing the efficacy of AI interventions is promising. AI offers diverse strengths and potential across many fields, but development alone is insufficient. Evaluation, validation, implementation and monitoring are required. The implementation of AI platforms at the pre-, intra- and postoperative phases should be guided by robust evidence of the benefits, such as a more accurate and timely diagnosis, reduced complications and improved systems efficiencies.

Future of surgical AI
Medicine is entering an exciting phase of digital innovation, with clinical evidence now beginning to accumulate behind advances in AI applications. Domain-specific excellence is emerging, with vast potential for translational progress in surgery. A sector of medical practice that once lagged behind in terms of evidence-based medicine[157], surgery has evolved to thrive on world-class research and evidence[158]. Surgery now equals other fields, such as cardiology, in terms of the quantity of randomized trials in AI applications, only lagging behind frontrunner fields such as gastroenterology and radiology, where task-specific applications are opportune, particularly around image processing[55]. In this Review, we have highlighted many of the most pragmatic and innovative emerging use cases of AI in surgery, with a particular focus on direct feasibility and preparedness for clinical translation, but there remain numerous additional examples and untapped avenues for further pursuit. As we pioneer surgical AI, the values of privacy, data security, accuracy, reproducibility, mitigation of biases, enhancement of equity, widening access and, above all, evidence-based care should guide our technological advances.

Reviews of AI in surgery frequently speculate toward autonomous robotic surgeons. In our view, this is the most distant of the realizable goals of surgical AI systems. While much attention has also been given to surgical automation[159], robotics and computer vision, these efforts should be contextualized in a time period where robotic surgery has yet to definitively demonstrate its advantage over other minimally invasive approaches[159–161]. In a resource-limited global surgical landscape, it remains to be seen whether AI-driven automation may offer the scalability to robotic surgical platforms that may help define its clinical value.

Surgery poses specific challenges for AI integration that are distinct from other areas of medicine. There is a paucity of digital infrastructure in most healthcare settings such that annotated datasets and digitized intraoperative records are rarely available[162]. In addition, procedural heterogeneity, acuity and rapidly changing clinical parameters represent a challenging and dynamic environment in which AI interventions will be required to deliver accurate and evolving output. Despite these known challenges, targeted work in these areas, including

Table 1 | Summary of the six published randomized controlled trials of AI in surgery

Application	Countries	Centre(s)	Sample size	Speciality	Domain	End points	Comparison	Results	Modality	AI Type
App-based, knowledge tailored AI-driven self-management strategies for patients[165]	Denmark and Norway	Multiple	461	Orthopedic	Lower back pain management	Questionnaire scores	AI versus routine care	Significant improvement: adjusted mean difference in Roland–Morris Disability Questionnaire=0.79 (95% CI=0.06–1.51)	Structured	Case Based Reasoning
Machine learning prediction of surgical resource utilization[166]	United States	Single	683	Colorectal and gynecology	Surgery length prediction	Mean absolute error time	AI versus routine care	Significant improvement in the prediction of case duration plus decreased wait time by 33.1 min per patient	Structured	Unspecified
AI-based intraoperative planning based on image segmentation[167]	France	Single	20	Orthopedic	Percutaneous vertebroplasty	Technical feasibility	Assisted versus unassisted	Significant improvement: decreased time to trocar deployment (642±210 versus 336±60 s; $P=0.001$)	Video	CNN
Machine learning predictor of intraoperative hypotension[168]	Greece	Single	99	Noncardiac	Hypotension management	Time-weighted hypotension	Assisted versus unassisted	Significant improvement: decreased time-weighted average mean arterial pressure intraoperatively (mean difference=−0.28; 95% CI = −0.48–0.09 mmHg; $P=0.0003$)	Waveform	Unspecified
Machine learning predictor of intraoperative hypotension[94]	Germany	Single	49	Orthopedic	Hypotension management	Time-weighted hypotension	Assisted versus unassisted	Significant improvement: decreased frequency of intraoperative hypotension ((n)/h) (48 versus 80%; $P<0.001$)	Waveform	Unspecified
Machine learning predictor of intraoperative hypotension[20]	The Netherlands	Single	60	Noncardiac	Hypotension management	Time-weighted hypotension	Assisted versus unassisted	Significant improvement: median difference of 16.7 min of intraoperative hypotension per patient (95% CI = 7.7–31.0; $P<0.001$)	Waveform	Unspecified

This table does not include AI trials for endoscopy. CI, confidence interval.

growing priority toward digital infrastructure, data security and privacy, as well as unsupervised AI paradigms, demonstrates substantial promise.

Transformer models are poised to enable real-time analytics of multi-layered data, including patient anatomy, biomarkers of physiology, sensor inputs, -omics data, environmental data and more. When leveraged by a fine-tuned understanding of the corpus of medical knowledge, such models stand to have a vast impact on surgical care[64]. At the time of writing, few examples exist for novel generative AI models in surgery. In the sections above, we present several opportunities for such generalizable AI models unburdened by labeling needs to be implemented in surgical care as the generalist AI surgeon augmenter. These approaches are common to AI in medicine, with the majority of approaches using decision trees, neural networks and reinforcement learning[55]. Early implementations of existing LLMs for text generation, data extraction and patient care are undoubtedly underway[163], with notable caveats such as model accuracy degradation, output overconfidence, lack of data privacy and regulatory approvals and a deficiency of prospective clinical trials yet to be overcome.

Numerous apprehensions remain with regard to the integration of AI into surgical practice, with many clinicians perceiving limited scope in a field dominated by experiential decision-making competency, apprentice model teaching structures and hands-on therapies. However, with the rapid development of AI in software, hardware and logistics, these perceived limitations in scope will be continuously tested. We envision a collaborative future between surgeons and AI technologies, with surgical innovation guided first and foremost by patient needs and outcomes.

AI in surgery is a rapidly developing and promising avenue for innovation; the realization of this potential will be underpinned by increased collaboration[154], robust randomized trial evidence[55], the exploration of novel use cases[164] and the development of a digitally minded surgical infrastructure to enable this technological transformation. The role of AI in surgery is set to expand dramatically and, with correct oversight, its ultimate promise is to effectively improve both patient and operator outcomes, reduce patient morbidity and mortality and enhance the delivery of surgery globally.

References

1. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
2. Wallace, M. B. et al. Impact of artificial intelligence on miss rate of colorectal neoplasia. *Gastroenterology* **163**, 295–304.e5 (2022).
3. Sharma, P. & Hassan, C. Artificial intelligence and deep learning for upper gastrointestinal neoplasia. *Gastroenterology* **162**, 1056–1066 (2022).
4. Aerts, H. J. W. L. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA Oncol.* **2**, 1636–1642 (2016).
5. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
6. COVIDSurg Collaborative. Projecting COVID-19 disruption to elective surgery. *Lancet* **399**, 233–234 (2022).
7. Weiser, T. G. et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *Lancet* **385**, S11 (2015).
8. Meara, J. G. & Greenberg, S. L. M. The Lancet Commission on Global Surgery Global surgery 2030: evidence and solutions for achieving health, welfare and economic development. *Surgery* **157**, 834–835 (2015).
9. Alkire, B. C. et al. Global access to surgical care: a modelling study. *Lancet Glob. Health* **3**, e316–e323 (2015).
10. Grönroos-Korhonen, M. T. et al. Failure to rescue after reoperation for major complications of elective and emergency colorectal surgery: a population-based multicenter cohort study. *Surgery* **172**, 1076–1084 (2022).
11. GlobalSurg Collaborative. Mortality of emergency abdominal surgery in high-, middle- and low-income countries. *Br. J. Surg.* **103**, 971–988 (2016).
12. GlobalSurg Collaborative & National Institute for Health Research Global Health Research Unit on Global Surgery. Global variation in postoperative mortality and complications after cancer surgery: a multicentre, prospective cohort study in 82 countries. *Lancet* **397**, 387–397 (2021).
13. GlobalSurg Collaborative. Surgical site infection after gastrointestinal surgery in high-income, middle-income, and low-income countries: a prospective, international, multicentre cohort study. *Lancet Infect Dis.* **18**, 516–525 (2018).
14. Paton, F. et al. Effectiveness and implementation of enhanced recovery after surgery programmes: a rapid evidence synthesis. *BMJ Open* **4**, e005015 (2014).
15. Vedula, S. S. & Hager, G. D. Surgical data science: the new knowledge domain. *Innov. Surg. Sci.* **2**, 109–121 (2017).
16. Kaddour J. et al. Challenges and applications of large language models. Preprint at https://arxiv.org/abs/2307.10169 (2023).
17. Bonde, A. et al. Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study. *Lancet Digit. Health* **3**, e471–e485 (2021).
18. Gögenur, I. Introducing machine learning-based prediction models in the perioperative setting. *Br. J. Surg.* **110**, 533–535 (2023).
19. Mascagni, P. et al. Computer vision in surgery: from potential to clinical value. *NPJ Digit. Med.* **5**, 163 (2022).
20. Wijnberge, M. et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *J. Am. Med. Assoc.* **323**, 1052–1060 (2020).
21. Kalidasan, V. et al. Wirelessly operated bioelectronic sutures for the monitoring of deep surgical wounds. *Nat. Biomed. Eng.* **5**, 1217–1227 (2021).
22. Fazlollahi, A. M. et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw. Open* **5**, e2149008 (2022).
23. Wells, C. I. et al. Wearable devices to monitor recovery after abdominal surgery: scoping review. *BJS Open* **6**, zrac031 (2022).
24. Dawes, A. J., Lin, A. Y., Varghese, C., Russell, M. M. & Lin, A. Y. Mobile health technology for remote home monitoring after surgery: a meta-analysis. *Br. J. Surg.* **108**, 1304–1314 (2021).
25. Loftus, T. J. et al. Artificial intelligence and surgical decision-making. *JAMA Surg.* **155**, 148–158 (2020).
26. Safavi, K. C. et al. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. *JAMA Netw. Open* **2**, e1917221 (2019).
27. Dundar, T. T. et al. Machine learning-based surgical planning for neurosurgery: artificial intelligent approaches to the cranium. *Front. Surg.* **9**, 863633 (2022).
28. Sadeghi, A. H. et al. Virtual reality and artificial intelligence for 3-dimensional planning of lung segmentectomies. *JTCVS Tech.* **7**, 309–321 (2021).
29. Zhang, T. et al. RadioLOGIC, a healthcare model for processing electronic health records and decision-making in breast disease. *Cell Rep. Med.* **4**, 101131 (2023).
30. Korfiatis, P. et al. Automated artificial intelligence model trained on a large data set can detect pancreas cancer on diagnostic computed tomography scans as well as visually occult preinvasive cancer on prediagnostic computed tomography scans. *Gastroenterology* **165**, 1533–1546.e4 (2023).
31. Maier-Hein, L. et al. Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* **1**, 691–696 (2017).

32. Verma, A., Agarwal, G., Gupta, A. K. & Sain, M. Novel hybrid intelligent secure cloud Internet of things based disease prediction and diagnosis. *Electronics* **10**, 3013 (2021).

33. Ghaffar Nia, N., Kaplanoglu, E. & Nasab, A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discov. Artif. Intell.* **3**, 5 (2023).

34. Gong, D. et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol. Hepatol.* **5**, 352–361 (2020).

35. Wang, P. et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol. Hepatol.* **5**, 343–351 (2020).

36. Kiani, A. et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit. Med.* **3**, 23 (2020).

37. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).

38. Dias, R. & Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* **11**, 70 (2019).

39. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).

40. Javanmard-Emamghissi, H. & Moug, S. J. The virtual uncertainty of futility in emergency surgery. *Br. J. Surg.* **109**, 1184–1185 (2022).

41. Ali, R. et al. Bridging the literacy gap for surgical consents: an AI–human expert collaborative approach. *NPJ Digit. Med.* **7**, 63 (2024).

42. Vernooij, J. E. M. et al. Performance and usability of pre-operative prediction models for 30-day peri-operative mortality risk: a systematic review. *Anaesthesia* **78**, 607–619 (2023).

43. Wynants, L. et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *Br. Med. J.* **369**, m1328 (2020).

44. Finlayson, S. G., Beam, A. L. & van Smeden, M. Machine learning and statistics in clinical research articles—moving past the false dichotomy. *JAMA Pediatr.* **177**, 448–450 (2023).

45. Chiew, C. J., Liu, N., Wong, T. H., Sim, Y. E. & Abdullah, H. R. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Ann. Surg.* **272**, 1133–1139 (2020).

46. COVIDSurg Collaborative. Machine learning risk prediction of mortality for patients undergoing surgery with perioperative SARS-CoV-2: the COVIDSurg mortality score. *Br. J. Surg.* **108**, 1274–1292 (2021).

47. Maurer, L. R. et al. Validation of the AI-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator in patients 65 years and older. *Ann. Surg.* **277**, e8–e15 (2023).

48. El Hechi, M. W. et al. Validation of the artificial intelligence-based Predictive Optimal Trees in Emergency Surgery Risk (POTTER) calculator in emergency general surgery and emergency laparotomy patients. *J. Am. Coll. Surg.* **232**, 912–919.e1 (2021).

49. Gebran, A. et al. POTTER-ICU: an artificial intelligence smartphone-accessible tool to predict the need for intensive care after emergency surgery. *Surgery* **172**, 470–475 (2022).

50. Bihorac, A. et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann. Surg.* **269**, 652–662 (2019).

51. Ren, Y. et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. *JAMA Netw. Open* **5**, e2211973 (2022).

52. Bertsimas, D., Dunn, J., Velmahos, G. C. & Kaafarani, H. M. A. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator. *Ann. Surg.* **268**, 574–583 (2018).

53. El Moheb, M. et al. Artificial intelligence versus surgeon gestalt in predicting risk of emergency general surgery. *J. Trauma Acute Care Surg.* **95**, 565–572 (2023).

54. Jalali, A. et al. Deep learning for improved risk prediction in surgical outcomes. *Sci. Rep.* **10**, 9289 (2020).

55. Han R. et al. Randomized controlled trials evaluating AI in clinical practice: a scoping evaluation. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.12.23295381 (2023).

56. Li, B. et al. Machine learning in vascular surgery: a systematic review and critical appraisal. *NPJ Digit. Med.* **5**, 7 (2022).

57. Artificial intelligence assists surgeons' decision making. *ClinicalTrials.gov* https://classic.clinicaltrials.gov/ct2/show/record/NCT04999007 (2024).

58. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K., SPIRIT-AI & CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).

59. Collins, G. S. et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).

60. Heus, P. et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open* **9**, e025611 (2019).

61. Vasey, B. et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Br. Med. J.* **377**, e070904 (2022).

62. Cacciamani, G. E. et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nat. Med.* **29**, 14–15 (2023).

63. Wynter-Blyth, V. & Moorthy, K. Prehabilitation: preparing patients for surgery. *Br. Med. J.* **358**, j3702 (2017).

64. Topol, E. J. As artificial intelligence goes multimodal, medical applications multiply. *Science* **381**, adk6139 (2023).

65. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).

66. Yao, X. et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat. Med.* **27**, 815–819 (2021).

67. Attia, Z. I. et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med.* **25**, 70–74 (2019).

68. Ouyang, D. et al. Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study. *Lancet Digit. Health* **6**, e70–e78 (2024).

69. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).

70. Matias Y. & Corrado, G. Our latest health AI research updates. *Google Blog* https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/ (2023).

71. Decker, H. et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw. Open* **6**, e2336997 (2023).

72. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).

73. Ke, Y. et al. Development and testing of retrieval augmented generation in large language models—a case study report. Preprint at https://arxiv.org/abs/2402.01733 (2024).

74. Perry, A. AI will never convey the essence of human empathy. *Nat. Hum. Behav.* **7**, 1808–1809 (2023).

75. Loveys, K., Sagar, M., Pickering, I. & Broadbent, E. A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: randomized pilot trial. *JMIR Ment. Health* **8**, e31586 (2021).

76. Introducing GPTs. *OpenAI* https://openai.com/blog/introducing-gpts (2023).

77. How assistants work. *OpenAI* https://platform.openai.com/docs/assistants/how-it-works (2023).

78. Ouyang, H. Your next hospital bed might be at home. *The New York Times* https://www.nytimes.com/2023/01/26/magazine/hospital-at-home.html (2023).

79. Temple-Oberle, C., Yakaback, S., Webb, C., Assadzadeh, G. E. & Nelson, G. Effect of smartphone app postoperative home monitoring after oncologic surgery on quality of recovery: a randomized clinical trial. *JAMA Surg.* **158**, 693–699 (2023).

80. März, K. et al. Toward knowledge-based liver surgery: holistic information processing for surgical decision support. *Int. J. Comput. Assist. Radiol. Surg.* **10**, 749–759 (2015).

81. Thirunavukarasu, A. J. et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med. Educ.* **9**, e46599 (2023).

82. Glaese, A. et al. Improving alignment of dialogue agents via targeted human judgements. Preprint at https://arxiv.org/abs/2209.14375 (2022).

83. Johri, S. et al. Testing the limits of language models: a conversational framework for medical AI assessment. Preprint at https://www.medrxiv.org/content/10.1101/2023.09.12.23295399v1 (2023).

84. Cacciamani, G. E., Collins, G. S. & Gill, I. S. ChatGPT: standard reporting guidelines for responsible use. *Nature* **618**, 238 (2023).

85. Haver, H. L. et al. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* **307**, e230424 (2023).

86. Birkmeyer, J. D. et al. Surgical skill and complication rates after bariatric surgery. *N. Engl. J. Med.* **369**, 1434–1442 (2013).

87. Zhang, J. et al. Rapid and accurate intraoperative pathological diagnosis by artificial intelligence with deep learning technology. *Med. Hypotheses* **107**, 98–99 (2017).

88. Vermeulen, C. et al. Ultra-fast deep-learned CNS tumour classification during surgery. *Nature* **622**, 842–849 (2023).

89. Mascagni, P. et al. Early-stage clinical evaluation of real-time artificial intelligence assistance for laparoscopic cholecystectomy. *Br. J. Surg.* **111**, znad353 (2023).

90. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

91. Bechara, A., Damasio, H., Tranel, D. & Damasio, A. R. Deciding advantageously before knowing the advantageous strategy. *Science* **275**, 1293–1295 (1997).

92. Van der Ven, W. H. et al. One of the first validations of an artificial intelligence algorithm for clinical use: the impact on intraoperative hypotension prediction and clinical decision-making. *Surgery* **169**, 1300–1303 (2021).

93. Hatib, F. et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology* **129**, 663–674 (2018).

94. Schneck, E. et al. Hypotension prediction index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: a single centre feasibility randomised blinded prospective interventional trial. *J. Clin. Monit. Comput.* **34**, 1149–1158 (2020).

95. Aklilu, J. G. et al. Artificial intelligence identifies factors associated with blood loss and surgical experience in cholecystectomy. *NEJM AI* **1**, AIoa2300088 (2024).

96. Jung, J. J., Jüni, P., Lebovic, G. & Grantcharov, T. First-year analysis of the operating room black box study. *Ann. Surg.* **271**, 122–127 (2020).

97. Al Abbas, A. I. et al. The operating room black box: understanding adherence to surgical checklists. *Ann. Surg.* **276**, 995–1001 (2022).

98. Seo, S. et al. Towards an AI coach to infer team mental model alignment in healthcare. *IEEE Conf. Cogn. Comput. Asp. Situat. Manag.* **2021**, 39–44 (2021).

99. Agha, R. A., Fowler, A. J. & Sevdalis, N. The role of non-technical skills in surgery. *Ann. Med Surg.* **4**, 422–427 (2015).

100. Gillespie, B. M. et al. Correlates of non-technical skills in surgery: a prospective study. *BMJ Open* **7**, e014480 (2017).

101. Farrugia, G. Transforming health care through platforms. *LinkedIn* https://www.linkedin.com/pulse/transforming-health-care-through-platforms-gianrico-farrugia-m-d-/ (2023).

102. Feasibility and utility of artificial intelligence (AI)/machine learning (ML)—driven advanced intraoperative visualization and identification of critical anatomic structures and procedural phases in laparoscopic cholecystectomy. *ClinicalTrials.gov* https://classic.clinicaltrials.gov/ct2/show/NCT05775133?term=artificial+intelligence,+surgery&type=Intr&draw=2&rank=4 (2024).

103. Zia, A., Hung, A., Essa, I. & Jarc, A. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018* 273–280 (Springer International Publishing, 2018).

104. Luongo, F., Hakim, R., Nguyen, J. H., Anandkumar, A. & Hung, A. J. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery* **169**, 1240–1244 (2021).

105. Funke, I. et al. Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* 467–475 (Springer International Publishing, 2019).

106. Lavanchy, J. L. et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci. Rep.* **11**, 5197 (2021).

107. Goodman E. D. et al. A real-time spatiotemporal AI model analyzes skill in open surgical videos. Preprint at https://arxiv.org/abs/2112.07219 (2021).

108. Kiyasseh, D. et al. A vision transformer for decoding surgeon activity from surgical videos. *Nat. Biomed. Eng.* **7**, 780–796 (2023).

109. Goyal, P. et al. Vision models are more robust and fair when pretrained on uncurated images without supervision. Preprint at https://arxiv.org/abs/2202.08360 (2022).

110. Chatterjee, S., Bhattacharya, M., Pal, S., Lee, S.-S. & Chakraborty, C. ChatGPT and large language models in orthopedics: from education and surgery to research. *J. Exp. Orthop.* **10**, 128 (2023).

111. Kim, J. S. et al. Can natural language processing and artificial intelligence automate the generation of billing codes from operative note dictations? *Glob. Spine J.* **13**, 1946–1955 (2023).

112. Takeuchi, M. et al. Automated surgical-phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence. *Ann. Surg. Oncol.* **29**, 6847–6855 (2022).

113. Marcus, H. J. et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. *Nat. Med.* **30**, 61–75 (2024).

114. Yip, M. et al. Artificial intelligence meets medical robotics. *Science* **381**, 141–146 (2023).

115. Shademan, A. et al. Supervised autonomous robotic soft tissue surgery. *Sci. Transl. Med.* **8**, 337ra64 (2016).

116. Hu, Y.-Y. et al. Complementing operating room teaching with video-based coaching. *JAMA Surg.* **152**, 318–325 (2017).

117. Bonrath, E. M., Dedy, N. J., Gordon, L. E. & Grantcharov, T. P. Comprehensive surgical coaching enhances surgical skill in the operating room: a randomized controlled trial. *Ann. Surg.* **262**, 205–212 (2015).

118. Goodman, E. D. et al. Analyzing surgical technique in diverse open surgical videos with multitask machine learning.*JAMA Surg.* **159**, 185–192 (2024).

119. Levin, M., McKechnie, T., Khalid, S., Grantcharov, T. P. & Goldenberg, M. Automated methods of technical skill assessment in surgery: a systematic review. *J. Surg. Educ.* **76**, 1629–1639 (2019).

120. Lam, K. et al. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digit. Med.* **5**, 24 (2022).

121. Kiyasseh, D. et al. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Commun. Med.* **3**, 42 (2023).

122. Kiyasseh, D. et al. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *NPJ Digit. Med.* **6**, 54 (2023).

123. Fazlollahi, A. M. et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. *JAMA Netw. Open* **6**, e2334658 (2023).

124. Artificial intelligence augmented training in skin cancer diagnostics for general practitioners (AISC-GP). *ClinicalTrials.gov* https://classic.clinicaltrials.gov/ct2/show/NCT04576416 (2024).

125. Testing the efficacy of an artificial intelligence real-time coaching system in comparison to in person expert instruction in surgical simulation training of medical students—a randomized controlled trial. *ClinicalTrials.gov* https://clinicaltrials.gov/study/NCT05168150 (2024).

126. Knight, S. R. et al. Mobile devices and wearable technology for measuring patient outcomes after surgery: a systematic review. *NPJ Digit. Med.* **4**, 157 (2021).

127. A clinical trial of the use of remote heart rhythm monitoring with a smartphone after cardiac surgery (SURGICAL-AF 2) *ClinicalTrials.gov* https://classic.clinicaltrials.gov/ct2/show/NCT05509517?term=artificial+intelligence,+surgery&recrs=e&type=Intr&draw=2&rank=14 (2024).

128. Sessler, D. I. & Saugel, B. Beyond 'failure to rescue': the time has come for continuous ward monitoring. *Br. J. Anaesth.* **122**, 304–306 (2019).

129. Xu, W. et al. Continuous wireless postoperative monitoring using wearable devices: further device innovation is needed. *Crit. Care* **25**, 394 (2021).

130. Bellini, V. et al. Machine learning in perioperative medicine: a systematic review. *J. Anesth. Analg. Crit. Care* **2**, 2 (2022).

131. Junaid, M., Ali, S., Eid, F., El-Sappagh, S. & Abuhmed, T. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease. *Comput. Methods Prog. Biomed.* **234**, 107495 (2023).

132. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).

133. Lu, S.-H. et al. Multimodal sensing and therapeutic systems for wound healing and management: a review. *Sens. Actuators Rep.* **4**, 100075 (2022).

134. Anthis, A. H. C. et al. Modular stimuli-responsive hydrogel sealants for early gastrointestinal leak detection and containment. *Nat. Commun.* **13**, 7311 (2022).

135. NIHR Global Health Research Unit on Global Surgery & GlobalSurg Collaborative Use of telemedicine for postdischarge assessment of the surgical wound: international cohort study, and systematic review with meta-analysis. *Ann. Surg.* 277, e1331–e1347 (2023).

136. McLean, K. A. et al. Evaluation of remote digital postoperative wound monitoring in routine surgical practice. *NPJ Digit. Med.* **6**, 85 (2023).

137. Savage, N. Sibel Health: designing vital-sign sensors for delicate skin. *Nature* https://doi.org/10.1038/d41586-020-01806-7 (2020).

138. Callcut, R. A. et al. External validation of a novel signature of illness in continuous cardiorespiratory monitoring to detect early respiratory deterioration of ICU patients. *Physiol. Meas.* **42**, 095006 (2021).

139. Shickel, B. et al. Dynamic predictions of postoperative complications from explainable, uncertainty-aware, and multi-task deep neural networks. *Sci. Rep.* **13**, 1224 (2023).

140. Zhu, Y. et al. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. Preprint at https://arxiv.org/abs/2402.01713 (2024).

141. Xu, X. et al. A deep learning model for prediction of post hepatectomy liver failure after hemihepatectomy using preoperative contrast-enhanced computed tomography: a retrospective study. *Front. Med.* **10**, 1154314 (2023).

142. Bhasker, N. et al. Prediction of clinically relevant postoperative pancreatic fistula using radiomic features and preoperative data. *Sci. Rep.* **13**, 7506 (2023).

143. Ingwersen, E. W. et al. Radiomics preoperative-Fistula Risk Score (RAD-FRS) for pancreatoduodenectomy: development and external validation. *BJS Open* **7**, zrad100 (2023).

144. Greijdanus, N. G. et al. Stoma-free survival after rectal cancer resection with anastomotic leakage: development and validation of a prediction model in a large international cohort. *Ann. Surg.* **278**, 772–780 (2023).

145. Wells, C. I. et al. "Failure to rescue" following colorectal cancer resection: variation and improvements in a national study of postoperative mortality. *Ann. Surg.* **278**, 87–95 (2022).

146. Smits, F. J. et al. Algorithm-based care versus usual care for the early recognition and management of complications after pancreatic resection in the Netherlands: an open-label, nationwide, stepped-wedge cluster-randomised trial.*Lancet* **399**, 1867–1875 (2022).

147. Nepogodiev, D., Martin, J., Biccard, B., Makupe, A. & Bhangu, A. Global burden of postoperative death. *Lancet* **393**, 401 (2019).

148. Dugan, R. E. & Gabriel, K. J. Changing the business of breakthroughs. *Issues Sci. Technol.* **38**, 70–74 (2022).

149. Song, Y. et al. 3D-printed epifluidic electronic skin for machine learning-powered multimodal health surveillance. *Sci. Adv.* **9**, eadi6492 (2023).

150. Kim, S., Brooks, A. K. & Groban, L. Preoperative assessment of the older surgical patient: honing in on geriatric syndromes. *Clin. Inter. Aging* **10**, 13–27 (2015).

151. Kim, K. M., Yefimova, M., Lin, F. V., Jopling, J. K. & Hansen, E. N. A home-recovery surgical care model using AI-driven measures of activities of daily living. *NEJM Catal. Innov. Care Deliv.* (2022).

152. Modelling and AI using sensor data to personalise rehabilitation following joint replacement (MAPREHAB). *ClinicalTrials.gov* https://clinicaltrials.gov/study/NCT04289025 (2024).

153. Chen, E., Prakash, S., Janapa Reddi, V., Kim, D. & Rajpurkar, P. A framework for integrating artificial intelligence for clinical care with continuous therapeutic monitoring. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-023-01115-0 (2023).

154. McLean, K. A. et al. Readiness for implementation of novel digital health interventions for postoperative monitoring: a systematic review and clinical innovation network analysis. *Lancet Digit. Health* **5**, e295–e315 (2023).

155. Bilbro, N. A. et al. The IDEAL reporting guidelines: a Delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. *Ann. Surg.* **273**, 82–85 (2021).

156. Skivington, K. et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *Br. Med. J.* **374**, n2061 (2021).

157. Horton, R. Surgical research or comic opera: questions, but few answers. *Lancet* **347**, 984–985 (1996).

158. Bagenal, J. et al. Surgical research-comic opera no more. *Lancet* **402**, 86–88 (2023).

159. Saeidi, H. et al. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Sci. Robot* **7**, eabj2908 (2022).

160. Olavarria, O. A. et al. Robotic versus laparoscopic ventral hernia repair: multicenter, blinded randomized controlled trial. *Br. Med. J.* **370**, m2457 (2020).

161. Kawka, M., Fong, Y. & Gall, T. M. H. Laparoscopic versus robotic abdominal and pelvic surgery: a systematic review of randomised controlled trials. *Surg. Endosc.* **37**, 6672–6681 (2023).

162. Maier-Hein, L. et al. Surgical data science—from concepts toward clinical translation. *Med. Image Anal.* **76**, 102306 (2022).

163. Janssen, B. V., Kazemier, G. & Besselink, M. G. The use of ChatGPT and other large language models in surgical science. *BJS Open* **7**, zrad032 (2023).

164. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).

165. Sandal, L. F. et al. Effectiveness of app-delivered, tailored self-management support for adults with lower back pain-related disability: a SELFBACK randomized clinical trial. *JAMA Intern. Med.* **181**, 1288–1296 (2021).

166. Strömblad, C. T. et al. Effect of a predictive model on planned surgical duration accuracy, patient wait time, and use of presurgical resources: a randomized clinical trial. *JAMA Surg.* **156**, 315–321 (2021).

167. Auloge, P. et al. Augmented reality and artificial intelligence-based navigation during percutaneous vertebroplasty: a pilot randomised clinical trial. *Eur. Spine J.* **29**, 1580–1589 (2020).

168. Tsoumpa, M. et al. The use of the Hypotension Prediction Index integrated in an algorithm of goal directed hemodynamic treatment during moderate and high-risk surgery. *J. Clin. Med.* **10**, 5884 (2021).

169. Garrow, C. R. et al. Machine learning for surgical phase recognition. *Ann. Surg.* **273**, 684–693 (2021).

170. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at https://arxiv.org/abs/2204.06125 (2022).

171. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (NIPS, 2017).

Author contributions

C.V., E.M.H., G.O.G. and E.J.T. conceived of, planned and drafted the article.

Competing interests

The authors declare no competing interests.

Additional information