Artificial Intelligence for Surgeons

Subjects: Surgery

Contributor: Pietro Mascagni, Giuseppe Quero, Fiona R. Kolbinger, Claudio Fiorillo, Davide De Sio, Fabio Longo, Carlo Alberto Schena, Vito Laterza, Fausto Rosa, Roberta Menghi, Valerio Papa, Vincenzo Tondolo, Caterina Cina, Marius Distler, Juergen Weitz, Stefanie Speidel, Nicolas Padoy, Sergio Alfieri

Computer vision (CV) is a field of artificial intelligence (AI) that deals with the automatic analysis of videos and images. Recent advances in AI and CV methods coupled with the growing availability of surgical videos of minimally invasive procedures have led to the development of AI-based algorithms to improve surgical care.

Keywords: artificial intelligence ; surgery ; surgical data science

---

AI concepts and terms

AI is an umbrella term referring to the study of machines that emulate traits generally associated with human intelligence, such as perceiving the environment, deriving logical conclusions from these perceptions, and performing complex actions. AI applications in medicine are steadily increasing, and have already demonstrated clinical impact in various fields including dermatology, pathology, and endoscopy.

Medical decisions are usually not binary, but highly complex and adaptable with regard to timing (i.e., oncological treatment course, timing of diagnostic procedures), invasiveness (i.e., extent of surgery), and depend on available human and technological resources. In most cases, such choices are made not only on the basis of logical rules and guidelines, but also integrate professional experience. Given the plethora of variation possibilities, it would be extremely complex, if not impossible, to explicitly program machines to perform complex medical tasks, such as understanding free text in electronic health records to stratify patients or interpreting radiological images to make diagnoses. However, the cornerstone of AI is the ability of machines to learn with experience. In machine learning (ML), “experience” corresponds to data. In fact, ML algorithms are designed to iterate over large-scale datasets, identify patterns, and optimize their parameters to better solve a specific problem. While the term strong or general AI relates to the aspiration to create human-like intellectual competences and abstract thinking patterns, currently available AI applications—not only in the field of medicine—are limited to very specific (and in many cases simplified) problems, generally referred to as *weak* or *narrow* AI. In the last two decades, deep learning (DL), a subset of ML, has shown unprecedented performances in the analysis of complex, unstructured data such as free text and images. DL uses multilayer artificial neural networks (ANNs), collections of artificial neurons or perceptrons inspired by biological neural networks, to derive conclusions based on patterns in the input data. In medicine and surgery, a large amount of data is visual, in the form of images (e.g., radiological, histopathological) or videos (e.g., endoscopic and surgical videos). In addition, videos natively guide minimally invasive surgical procedures and could be analyzed for intraoperative assistance and postoperative evaluations. This brief introduction will hence focus on CV, the subfield of AI focusing on machine understanding of visual data.

key steps and considerations for surgical AI research

Based on the schematic introduction of key AI-related concepts and terms, the following section will provide a brief overview of a typical surgical AI pipeline in the field of CV (**Figure 1**). While automated surgical video analysis will be used as an example in the following section, similar approaches can be applied to other types of medical imaging and, in modified structure, to medical data in general.

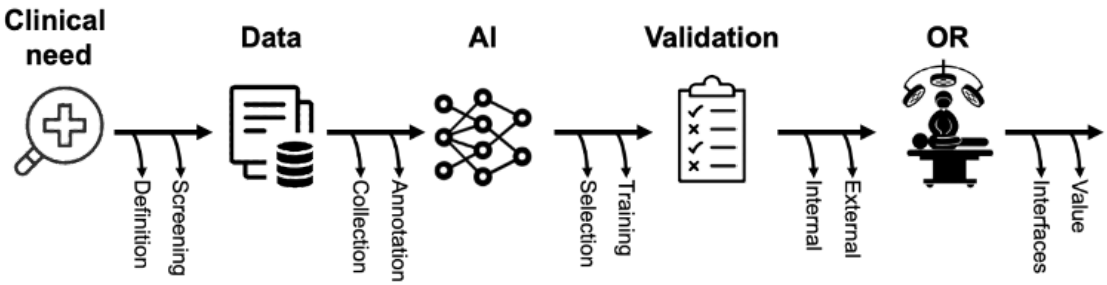

Figure 1. Schematic representation of the phases of surgical AI research.

Once a clinical need has been clearly defined, an appropriate, large-scale, and representative dataset needs to be generated. To verify data appropriateness, it is good practice to see if subject-matter experts (i.e., surgeons) routinely acquire such data and can consistently solve the identified problem using this type of data. For instance, if we want to train a machine to automatically assess the critical view of safety in videos of laparoscopic cholecystectomy, it is important to verify surgeons' inter-rater agreement in assessing such view and, eventually, devise strategies to formalize and improve such assessments. The inter-rater agreement of experts can also be used to roughly estimate the amount of data necessary to train and test an AI model, as lower inter-rater agreements are generally found in more complex problems that require larger datasets to solve. Finally, since AI performance is heavily dependent on the quality of data used during training, it should be verified that the dataset accurately represents the setting of foreseen clinical deployment. Using the same example of laparoscopic cholecystectomy, acute and chronic cholecystitis cases should be included in the dataset if we want the AI to work in both scenarios.

A further, essential step in generating a dataset for AI is annotation. The term annotation describes the process of labeling data with the information the AI should learn to predict. The type of information to annotate depends on the problem the algorithm is intended to solve. For instance, temporal annotations (e.g., timestamps) are needed to train an AI model to classify surgical steps while spatial annotations (e.g., bounding boxes or segmentations) are required to train an AI model to detect anatomical structures within an image. Regardless of the use case, high-quality annotations are essential for training AI using supervised learning approaches, currently the most common type of learning, as contrasting annotations will significantly hamper training of an AI algorithm. In the context of evaluating the accuracy of an AI algorithm for image recognition, it is important to consider that annotations also serve as "*ground truth*" for comparison. In fact, predictions of the previously trained AI are compared to experts' annotations to compute performance metrics. The greater the overlap between the annotations and the predictions, the better the algorithm is. Consequently, the reliability of annotations defines the validity of AI assessments. The development and improvement of methods to assess the quality of annotations are subject to ongoing scientific discussion. Generally, reporting annotation protocols, details on annotators' expertise, as well as integrating a thorough annotation review process involving multiple annotators and expert reviewers while reporting inter-rater agreements allow to scrutinize annotations.

The annotated dataset should then be split into a training set, used to develop the AI algorithm through multiple iterations, and a test set, used to evaluate the AI performance on unseen data. Split ratio can vary, but it is important to prevent data leaks between training and testing subsets. Of primary importance, test data should not be exposed during training. In addition, testing data should remain as independent as possible from the training dataset. Specifically, this means that not only all image data from one surgical video should be assigned to either the training or the test dataset, but also that serial examinations from one patient (i.e., multiple colonoscopy videos over time) should be treated as a coherent sequence that should not be separated between the training and test dataset.

At this stage, the dataset and task of interest will be explored to select the best AI architecture or algorithm to then refine, train, and test. In most cases, healthcare professionals and computer scientists collaborate in this process. Interdisciplinary education is, therefore, critical to enable all partners to understand both the clinical and the algorithmic perspectives, to critically appraise related literature, and to overall facilitate a constructive interdisciplinary collaboration. Specifically, involved healthcare professionals should understand and participate in the selection of metrics used to evaluate AI performance. The most commonly used metrics to evaluate how well an AI solves a given task describe the overlap between the true outcome or the annotated "*ground truth*" and the AI prediction. An important challenge in metric selection is the fact that these overlap metrics are merely surrogate parameters for the clinical benefit. This underlines the need for continuous clinical feedback during the entire process of conceptualization and evaluation of AI applications. Since events to be predicted are often rare (i.e., surgery complications), datasets are commonly unbalanced towards positive or negative cases and require balanced metrics for reliable AI performance assessment. In addition, different clinical applications should optimize different metrics. For instance, screening applications where the cost of a false negative is high, as in computer-aided detection of polyps during screening colonoscopy, should value sensitivity over

specificity. In turn, when assessing safety measures such as the critical view in laparoscopic cholecystectomy, the cost of a false positive is high, which is why specificity should be favored over sensitivity. Similar to reporting of annotations, the selected metrics should be transparently reported including specifications about the computing process and underlying assumptions about measured (surrogate) parameters. This is particularly important, as purely technical metrics often fail to predict actual clinical value and ongoing research is looking at developing evaluation methods and metrics specifically for surgical AI applications.

Regardless of how well surgical AIs have been developed and tested, external validation and translational studies are essential to evaluate the clinical potential. Since AI performance is notably dependent on training data, testing on multicentric data reflecting different acquisition modalities, patient populations, and hospital settings is necessary to evaluate how well AI systems generalize outside of the development setting. However, very few external validations studies have been performed to date since most open-access datasets only contain data from single centers. In such scenarios, multi-institutional collaboration is one of the most influential prerequisites for the development of clinically relevant AI applications.

To conclude, well designed implementation studies looking at how to integrate such technology in complex clinical and surgical workflows and assessing how these changes impact patient care are crucial to measure actual value for patients and healthcare systems. Translational studies exploring the clinical value of surgical AI still remain to be published, but currently available guidelines can help designing protocols, early assessments, and reporting of AI-based interventions.
