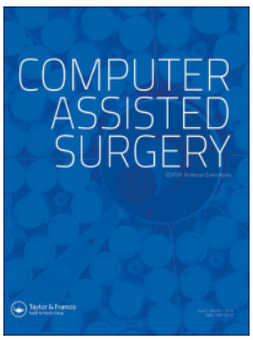

Computer Assisted Surgery

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/icsu21

Challenges in surgical video annotation

Thomas M. Ward, Danyal M. Fer, Yutong Ban, Guy Rosman, Ozanan R. Meireles & Daniel A. Hashimoto

To cite this article: Thomas M. Ward, Danyal M. Fer, Yutong Ban, Guy Rosman, Ozanan R. Meireles & Daniel A. Hashimoto (2021) Challenges in surgical video annotation, Computer Assisted Surgery, 26:1, 58-68, DOI: 10.1080/24699322.2021.1937320

To link to this article: https://doi.org/10.1080/24699322.2021.1937320

Published online: 14 Jun 2021.

COMPUTER ASSISTED SURGERY
2021, VOL. 26, NO. 1, 58–68
https://doi.org/10.1080/24699322.2021.1937320

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Challenges in surgical video annotation

Thomas M. Ward[a], Danyal M. Fer[b], Yutong Ban[a,c], Guy Rosman[a,c], Ozanan R. Meireles[a] and Daniel A. Hashimoto[a]

[a]Surgical AI & Innovation Laboratory, Department of Surgery, Massachusetts General Hospital, Boston, MA, USA; [b]Department of Surgery, University of California San Francisco East Bay, Hayward, CA, USA; [c]Distributed Robotics Laboratory, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

Annotation of surgical video is important for establishing ground truth in surgical data science endeavors that involve computer vision. With the growth of the field over the last decade, several challenges have been identified in annotating spatial, temporal, and clinical elements of surgical video as well as challenges in selecting annotators. In reviewing current challenges, we provide suggestions on opportunities for improvement and possible next steps to enable translation of surgical data science efforts in surgical video analysis to clinical research and practice.

KEYWORDS

Surgical video; Annotation; Image classification; Object detection; Semantic segmentation; temporal annotation; inter-rater reliability

Introduction

Annotation of visual data is important as it provides ground truth labels of real-world objects, scenes, and events that can then be utilized to train computer vision algorithms. While everyday tasks such as face recognition and some types of object recognition in classification can be performed with a high degree of accuracy, there remains quite a gap when comparing state-of-the-art computer vision performance to that of humans. This gap is particularly pronounced in surgery.

Indeed, several elements about surgery present challenges when considering methods for annotation of surgical data, particularly surgical video. Understand that operative events can be defined temporally, over the amount of time in which they occur; visually, as discrete spatial elements; or a combination of both. Algorithms should be designed to impart clinically significant outputs, which, therefore, requires clinically significant temporal and spatial annotations [1]. However, what constitutes a clinically significant event that requires annotation? In the case of bleeding, is bleeding an event that happens in discrete time periods that can be easily annotated? Which episodes of bleeding are worth annotation as clinically notable, and which can be considered 'normal' surgical oozing that does not require annotation? Such questions highlight the importance of considering and addressing challenging annotation questions upfront in a given project.

This paper reviews some of the challenges that are currently being tackled in annotation of surgical video and offers a clinical perspective on considerations to be made by researchers when defining annotation their schemas. We review challenges that arise in the selection of annotators, in spatial annotation, phase annotation (annotation of operative steps), annotation of clinically meaningful events, and annotations of surgical performance (Table 1).

Challenges with annotators

In considering the domain experience of annotators, there is a consideration of not just an annotator's experience in annotating video but also their experience in surgery. When selecting annotators, balancing expertise in these two domains is a challenge that will require further research to determine which attributes of annotators result in the generation of clinically relevant and consistent annotations. Clinical experience can be an initial discriminator when classifying different annotators resulting in 'Clinical Expert', 'Clinical Trainee', 'Layperson', or 'Crowd' (Table 2). While only a handful of papers in computer vision have been published to date regarding the differences in annotations

CONTACT Daniel A. Hashimoto dahashimoto@mgh.harvard.edu Surgical AI & Innovation Laboratory, Department of Surgery, Massachusetts General Hospital, Boston, MA, USA

Table 1. Overview of some of the challenges, key points, and recommendations to overcome those challenges in surgical video annotation.

Challenge	Key challenge points	Overcoming the challenge
Varying clinical expertise of annotators and cost of clinical experts	Different clinical phenomena require different levels of clinical expertise to identify	Determine clinical complexity of the phenomena to be annotated and determine appropriate relevant clinical expertise 'Two-pass' annotation with clinical experts verifying or augmenting annotations of trainees or laypersons
Poor inter-annotator reproducibility (i.e. high variability amongst annotators)	Multiple annotators may be imprecisely annotating clinical phenomena, resulting in significantly variability in annotations Multiple annotators may not conceptualize clinical phenomena in a similar manner	Precise definitions of clinical phenomena *via* annotation guides Careful selection of appropriate metrics of reliability to assess quality of annotations Qualitative review of variable annotations may identify which situations lead to variability due factors such as clinical difficulty
Incomplete representation of spatial clinical phenomena	Annotations of spatial data may be biased toward clear situations that are not clinically reproducible	Establishment of clear *a priori* clinical phenomenon of interest for spatial annotation
Selection of identification vs. segmentation methods	Spatial annotations with bounding boxes may incompletely capture the clinical phenomenon of interest Spatial annotations with semantic annotation may be too time-consuming for the clinical phenomenon of interest	Clearly establish clinical phenomenon of interest on which to base annotation strategy
Selection of causal vs. acausal methods of workflow analysis	Performance in workflow analysis can be higher with acausal approaches but may limit clinical utility	Define the intended clinical use case for the proposed work – decision-making and other uses that require online analysis require causal approaches while posthoc uses may benefit from acausal approaches
Annotation of clinically meaningful events	Defining 'clinically meaningful' events can be difficult as different clinicians have different conceptions of what is meaningful	Engage in *a priori* discussion with clinical experts to define clinically meaningful events in the context of the research project
Assessment of surgical performance has few objective measures	Existing performance assessments such as OSATS are subjective and require rater training for good inter-rater reliability	Performance should ultimately be assessed by patient outcomes
Expanding types of annotations can make existing annotations incomplete or obsolete	An existing set of annotations is unlikely to be exhaustive and may require updating	Version control enables layering of annotations to improve, update, or augment existing labels

Table 2. Annotator classes and relevant clinical experience.

Clinical annotator class	Experience
Clinical Expert	Completed residency/fellowship
Clinical Trainee	In residency training
Layperson	No clinical background/training
Crowd	Crowd platform based (e.g. Mechanical Turk) groups of annotators, typically without clinical background/training

between clinical experts (or clinical trainees) and crowd annotators, studies in the field of surgical education have investigated differences between such annotators in identifying surgical anatomy or rating surgical performance [2]. Prior work has demonstrated the feasibility of utilizing crowd annotators to annotate elements of video ranging from anatomic structures to performance. For simple tasks with well-defined criteria, such as identifying surgical instruments, layperson and crowd annotators can annotate at a level similar to that of surgeons [3–7]. We caution that in many of these studies, videos were pre-edited by clinical experts to only show brief video clips of the procedure's critical portions. This pre-editing makes annotations by lay annotators easier to perform as it constrains the data to be more clinically relevant and serves to filter some of the noise in the data.

Furthermore, crowd annotators may exploit class imbalances in the data to maximize their percentage of correct annotations, preferentially selecting labels that are more likely to be prevalent in the data. For example, crowd annotators may liberally annotate the presence of a 'grasper' in a video because graspers are such common instruments. In an additional illustration of this point, Deal et al. demonstrated that while crowd annotators and clinical expert annotators had a good degree of correlation in their assessment of the quality of the critical view of safety (CVS) attained in laparoscopic cholecystectomy, crowd annotators were less likely to recognize high CVS scores and more likely to give an 'average' score (3 or 4 out of 6) than clinical expert annotators. Furthermore, crowd annotators were less likely to be able to identify poor quality CVS when compared to clinical expert annotators and favored again giving an average score [4,8]. The 'average' score is less likely to be perceived as incorrect or otherwise flagged in statistical analysis

for outliers in annotations. Thus, while simple tasks may be handled reasonably by crowd annotators, for more complex tasks such as identifying anatomy and the quality of a dissection, crowd annotators' results can vary from those of clinical expert and trainee annotators.

Careful selection of annotators is, therefore, necessary. Experienced surgeons are costly – both from a financial perspective if reimbursing them for their time spent on annotations and from an opportunity cost perspective, where time spent annotating video is time away from treating patients. Thus, while it may be sufficient to have the crowd annotate low level items of interest such as tools, it is likely necessary to have clinical experts or trainees, to annotate more complex phenomena. Clinical expert and trainee annotators have surgical experience and possess a broader understanding of surgical principles that allows them to interpret segments of video that may not cleanly fit annotation definitions. Additionally, more experienced clinical annotators can provide insight into new labels that may need to be created. For example, an annotator with limited surgical experience may either incorrectly label (e.g. believing a bladder neck reconstruction is part of an urethrovesical anastomosis) or be unable to label a segment. A clinically experienced annotator would instead see the related set of actions as being distinct and requiring a different, novel label.

One possible solution to the cost is a 'two-pass' method to video annotation [2]. On the first pass, a layperson annotates the video to the fullest of best of their ability and highlights areas in which they have uncertainty. This first annotation set could be generated from a crowd annotator, or even an automated machine learning model trained on a small amount of data. As proof of concept, some automated models can identify surgical phases with datasets of under 100 videos; however, these works do not all explicitly define the qualifications of the annotators they utilized [9] .On the second annotation 'pass,' the clinical expert annotator could rapidly annotate the video by only reviewing areas of uncertainty, the boundaries of start and end of phases, and areas requiring expert knowledge (e.g. the steps of an intracorporeal anastomosis). To ensure quality of the 'first-pass,' clinical expert annotators could be used to review and verify (i.e. audit) a subset of the first-pass annotations. Additional research is needed to determine how much of the data would need to be audited to ensure data quality. Such audits also raise the issue of how best to assess the agreement or inter-rater reliability between annotators.

While early work in the field utilized a single clinical expert annotator to ensure consistency of annotations across all data [10], subsequent research has since incorporated multiple clinical expert annotators – both to spread the burden of annotation and to allow for measurement of the potential reliability of annotations [11]. Assessing differences between annotators allows one to determine whether the definitions of the phenomena of interest were appropriately applied or understood. Depending on the type of annotation under investigation, different metrics can be calculated to assess inter-annotator reliability. At perhaps the most basic level, a simple percent agreement can be calculated. However, this does not account for potential agreement that can occur by chance alone. Therefore, various statistical measures of agreement can be considered, such as Cohen or Fleiss's κ, Krippendorf's α, or intraclass correlation coefficient [11,12]. An in-depth discussion of the appropriateness of individual metrics for a given situation is outside the scope of this article; however, it is important to consider aspects such as the number of annotators and the prevalence of a given annotation [13,14]. Reassuringly, in a preliminary study, pooling annotations from multiple clinical expert annotators did not result in a decrease of the trained model's performance [15]. To help reduce variation across annotators, it is critical to precisely define the phenomenon of interest that is to be annotated. The critical challenge in developing annotation guidelines is that they require the annotator to know, from the video alone, the surgeon's intent. Therefore, surrogates of surgical intent from video cues alone must be identified for accurate video annotation. These surrogates, often referred to as anchors or definitions, are used by annotators to determine how to classify a procedure's video segments (e.g. from time t_1 to t_2 the surgeon completes a gastrojejunal anastomosis) or spatial elements (e.g. the pixel at x_1,y_1,z_1 denotes part of structure A). Defining these anchors often incurs a trade-off between inter-annotator reproducibility (and, therefore, an algorithm's performance) and capturing clinically meaningful phenomenon. Consider labeling the process of creating a gastrojejunal anastomosis. This step 'starts' when the surgeon decides to begin the anastomosis, for which there are no visible video cues. An annotator would have to guess or otherwise infer when the surgeon makes this decision, creating significant variability. To create a reproducible annotation, the step's start could be when an instrument that creates the enterotomies first touches the tissue. However, while

reproducible, this fails to capture intent and results in a definition most surgeons would say is too narrow and loses important information (e.g. orienting and selecting the ideal loop of bowel). Finally, the wide variety of surgical techniques can result in visually distinctive segments between different surgeons and institutions that may need to be classified separately.

As detailed in the subsequent challenges on spatial and temporal annotation, the balance between having flexible definitions that preserve clinical relevance and precise definitions that improve inter-rater reliability can be optimized by considering the specific phenomena of interest. Some variability in annotating clinical phenomena may be unavoidable as such variability may reflect inherent differences in the conceptualization of such phenomena by surgeons. For example, surgeons may differ in their interpretation of the correct surgical plane (i.e. the potential space between two structures through which a dissection can be performed) or in the amount of bleeding that qualifies as clinically significant. These underlying differences could provide clues on the difficulty of a surgical situation (e.g. significant adhesions or inflammation) and may require additional annotation from human experts. While high variability between annotators in such edge cases might threaten a project seeking to utilize automated methods, it can also serve as a useful metric to more closely study a clinical phenomenon through other methods that may be more appropriate (e.g. qualitative methods).

Challenges in spatial annotation

Spatial annotation refers to the annotation of the spatial information (e.g. position, region of interest) of specific elements such as anatomy, tools, or visually salient events (e.g. blood) without necessarily including a consideration of the temporal manner in which such elements may arise. At first glance, annotation of structures along a spatial coordinate system would seem to be straightforward; however, several considerations arise when evaluating annotations created with minimal guidance.

As with any spatial annotation task, the phenomenon of interest for a research task should be well-defined *a priori*. The importance of determining first the *phenomenon* of interest as opposed to the phase or workflow of interest is exemplified by the task of identifying the critical view of safety (CVS) in laparoscopic cholecystectomy. The CVS is defined by the Society of American Gastrointestinal and Endoscopic Surgeons as a method to identify the cystic duct and artery during laparoscopic cholecystectomy. More specifically, the view that must be obtained is defined by achieving the following three criteria: (1) the hepatocystic triangle is cleared of fat and fibrous tissue, (2) the lower one third of the gallbladder is separated from the liver to expose the cystic plate, and (3) two and only two structures should be seen entering the gallbladder. For researchers interested in annotating the CVS, they must consider how these criteria will be applied based on the phenomenon or question of interest.

For classification tasks, it may be sufficient to simply collect a dataset containing images of CVS. One must then consider what quality of CVS has been attained as not all critical views are created equal and a grading system has been proposed to identify different qualities of CVS (Figure 1). Examples of high-quality CVS may rarely be found in existing datasets.

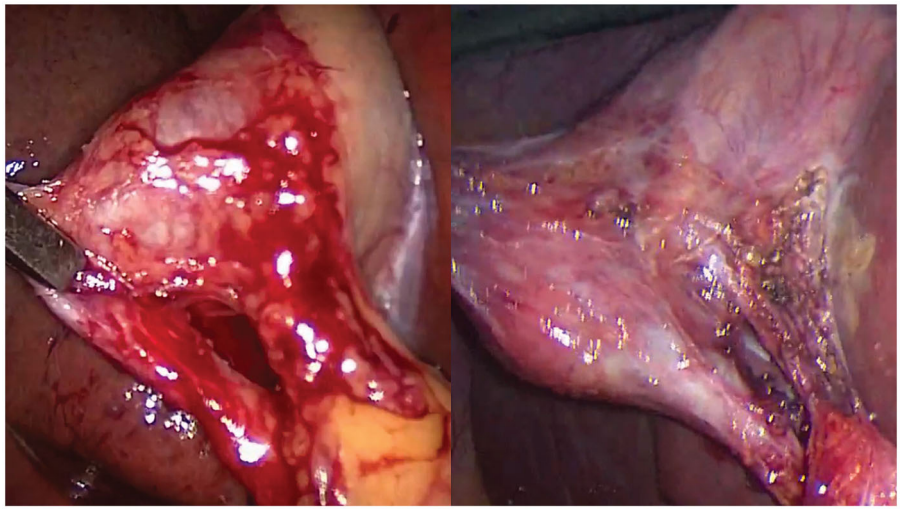

Figure 1. Comparison of two different views of the hepatocystic triangle illustrating different levels of dissection that can be performed in attempting to obtain a critical view of safety.

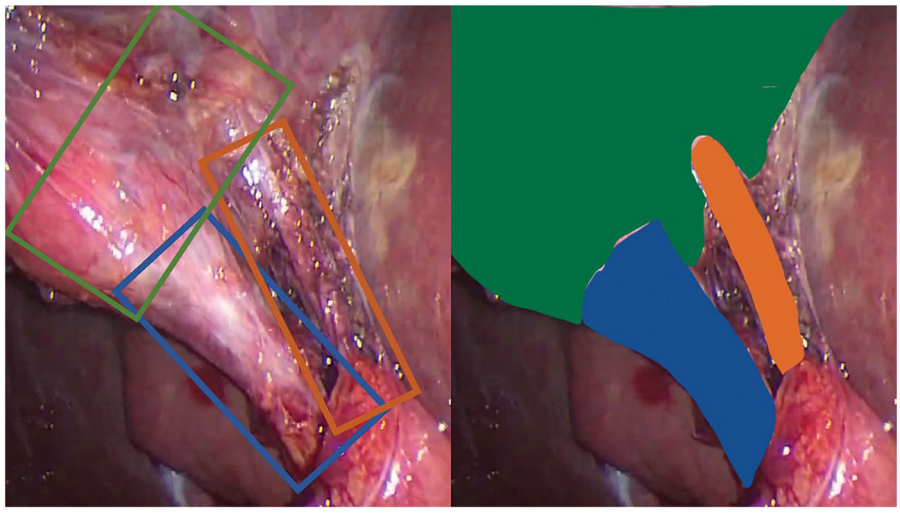

Figure 2. Use of bounding boxes (left) may result in overlapping identification of structures versus semantic segmentation (right). Consideration should be given to which approach more faithfully reflects the clinical phenomenon of interest. Green denotes gallbladder, blue denotes cystic duct, orange denotes cystic artery.

Furthermore, many surgeons rarely strive to pursue the highest quality CVS, instead choosing to obtain a CVS sufficient for their level of comfort in identifying the key structures. Thus, in this example, one should determine a priori whether the goal is to classify high quality CVS only or a range of CVS quality.

To further extend the example of identifying CVS, consideration should be made for the granularity of spatial annotation necessary. Bounding boxes may be sufficient to identify large aspects of anatomy such as whole organs (e.g. the gallbladder) or tools (e.g. Maryland grasper), but in the case of CVS, semantic segmentation may be more appropriate where the use of bounding boxes may lead to incorrect or overlapping annotation of structures (Figure 2). Further, for granular annotations of anatomy, some structures may have clearly demarcated 'starts' and 'ends' (i.e. the edge of the liver) whereas other structures may be less discrete. For example, when annotating the cystic artery, the connective tissue surrounding the artery may make labeling the structure difficult as the border between the artery and the gallbladder may be 'fuzzy'. Approaches borrowed from surgical education, such as visual concordance testing, may help to better delineate these 'fuzzy' borders to arrive at a consensus annotation [16] .Clinical expert annotators may be able to better evaluate this border but there will likely be bias in how an image is labeled, particularly in datasets where videos are labeled by a small group of annotators.

Spatial annotation in video can be tedious as objects may have to be tracked over large periods of time with the average video consisting of 25–30 frames per second (fps). Certainly, there may be no need to sample frames in real-time, and some phenomena can be sampled at only 1–2 fps. This again calls for consideration of the specific clinical phenomena for which the annotations are being generated. Software tools that assist with automated tracking of objects can be used but may also require auditing and correction.

One should also consider how to annotate some of the more abstract spatial characteristics that are perceived by surgeons, including the concepts of surgical planes – the potential, avascular interface/space that exists between structures or different types of tissues, retraction, and exposure. While these are largely spatial concepts, each of these can change in slight but important ways with time. As such, the challenges in annotating these characteristics will be described separately below.

Challenges in temporal annotation

An area of particular interest within the surgical community is understanding surgical workflow. Initial forays into workflow analysis (also known as surgical process modeling) were established by the work of Pierre Jannin and his group and has subsequently been extended with the goal of pursuing a common ontology [17]. As with spatial annotation, temporal annotation brings with it many challenges to carefully consider prior to implementing and sinking time into annotating a large number of videos.

The importance of determining first the *phenomenon* of interest as opposed to the phase or workflow of interest is again highlighted and drives how temporal annotations may be defined. One needs to

consider if the consumer of the annotations is acausal or causal. An *acausal* consumer has access to the entire video and can use past *and* future video frames to help identify the phenomenon of interest in the current video frame. *Causal* consumers, unlike an acausal one, only know the past and current frames of the video, so they cannot use future video events to help with decision making. An example of an acausal consumer is an algorithm that automatically labels the steps of a video found in a surgical video library. The algorithm can use information from the entire video to precisely label the start and end times of phases. For example, it often is hard to identify when a phase is finished since there are instrument exchanges in the field and no clear visual cues to a phase transition. Knowing exactly when the next phase starts in the future (since it has access to the future video frames) allows the acausal labeling model to accurately determine the phase's end. A causal consumer, on the other hand, might be an algorithm used in the operating room in real-time to help surgeons with their decision processes. Just like the surgeon, this causal AI model will not know the future events, and therefore need to 'think like a surgeon' using only information from the past video frames.

The intended use of the annotations, be it for a causal or acausal consumer, will heavily impact the definition of annotations and the process for generating them. For example, if one defines 'Dissection of Calot's Triangle' as the appearance of a dissecting tool on screen during active dissection, during an operation the surgeon may change from dissection to using the tool to remove an adhesion. To label this phase transition, the annotator, knowing only the video frame, will then need to rewind, end their annotation of 'Dissection of Calot's triangle,' and then relabel the next portion as 'Remove adhesion.' It is important to realize that by rewinding and modifying their previous annotation, they are now performing an acausal annotation. If this annotation is used to train a causal real-time algorithm, sub-optimal identification around the boundary of steps can occur, since the causal algorithm will not be able to account for future events. If the algorithm is tweaked and made acausal, better performance may be seen [18]. Both causal and acausal annotations are acceptable depending on the ultimate goal of their application. This phenomenon leads to a rule of thumb to achieve maximal algorithm performance: acausal annotations should only be used for acausal applications, while causal annotations can be used for both causal and acausal algorithms.

Temporal annotations, like all annotations, are difficult to define in a manner that leads to consistent annotations from multiple annotators. Extremely precise start and stop times (e.g. only when the instrument is touching the tissue), can make causal annotations more reproducible between annotators. However, these styles of annotations are not only tedious to perform but may also be of limited clinical utility. Another possible solution is to consider whether some overlap in phases is acceptable or whether some phases can be combined. Consider the case of isolating the cystic duct and cystic artery in laparoscopic cholecystectomy. Dissecting the fatty, fibrous tissue between the duct and artery may help to isolate both structures. In this case, phases could be combined into 'isolation of cystic duct and artery' or could be further divided into 'isolation of cystic duct,' 'isolation of cystic artery,' and 'isolation of cystic duct and artery.' Datasets such as Cholec80 take the approach of having more general phases to work around such difficulties [10]; however, this may limit their application to more precise clinical challenges such as decision support. Once again, clearly defining the phenomenon of interest is important to determine the level of annotation that is required.

Anchoring phases around the presence of instruments can provide concrete cues to annotators about the start/end of a phase. However, the presence of a surgical instrument alone does not define an operative phase in the mind of a surgeon. Rather, the tool is selected to achieve the goals of the phase. Thus, there may be situations such as in cholecystectomy when a scissor is introduced not to cut the cystic duct or artery but to open more of the peritoneum overlying the gallbladder. Consider phases to be less about which instruments are in the video and more about how instruments interact with tissue in the operative field to yield a given phenomenon (e.g. dissection, exposure, resection, etc.) [19]. Such consideration should allow for more clinically applicable annotations.

Additionally, temporal structure in an operation can be considered hierarchical. That is, a phase may consist of different steps which are performed by engaging in various actions. Prior work has described atomic surgical gestures, also known as surgemes, in terms of kinematics of robotic procedures [20]. Such gestures can fit into a temporal hierarchy of workflow as gestures can combine to yield an action, which is performed as part of an operative step. There is ongoing work at the Society of American Gastrointestinal and Endoscopic Surgeons to create clinically grounded

definitions for a hierarchical structure of temporal events in an operative video.

Challenges in annotating clinically meaningful events and characteristics

Annotation of clinically meaningful events is one of the foremost challenges in surgical video annotation given that there is limited consensus on what constitutes clinically meaningful. Consider the case of bleeding as an example. Bleeding occurs when blood moves from the lumen of a blood vessel into the surgical field; however, clinical context is extremely important in judging if this movement of blood is potentially deleterious to the patient or an expected ooze of little consequence to clinical outcome, especially if bleeding is to be used as an event to help guide or assess surgeons. For example, while performing an anastomosis, there may be bleeding from the edge of an enterotomy: this can be an expected or even positive sign, as it indicates the tissue has adequate perfusion. However, if bleeding occurs a few millimeters away, say from tearing of the bowel or the mesentery by a grasper, this could be considered an adverse event. Other questions include: how much blood loss is potentially harmful to the patient? What rate of blood loss can be temporarily ignored or otherwise expected to be self-limited? This necessitates understanding of context as, for example, the amount of expected bleeding in an appendectomy is significantly different than a liver resection. Additionally, even in the context of a single procedure type, patient factors such as inflammation can significantly affect the judgment of which episodes of bleeding are considered expected versus unexpected.

Jung et al. demonstrated that in procedures where there is limited bleeding to be expected, consensus on classifying bleeding events and their severity can be achieved with highly trained raters [21]. However, this does not address the issue of scalability in annotating events that require a significant amount of clinical knowledge from the annotator or the consideration that this could be an extremely tedious annotation to perform. Leveraging active learning and semi-supervised approaches can reduce the burden of labeling events, anatomy, tools, and phases [22–24]; however, training a model to recognize clinically relevant events can be problematic. How does a model discriminate between pooled blood and a slow ooze? Does every bleeding event necessitate review by a trained evaluator? Such considerations must be clarified ahead of annotation to ensure inter-rater reliability and consistent detection of ground truth phenomena.

The annotation of intraoperative adverse events (iAEs) other than bleeding presents similar challenges. Most of the work around identifying iAEs has been completed using data from large claims databases or prospective clinical registries [25,26]. Intraoperative adverse events that may be detected in operations include clear examples such as enterotomy, inadvertent thermal injury, or other unintentional damage to an organ (e.g. ureter, spleen, liver, bile duct, etc.) and more nuanced examples such as serosal injury to bowel (e.g. from excessive traction) during adhesiolysis or inadvertent spillage of bile from the gallbladder during cholecystectomy. Many surgeons may not consider spillage of bile from the gallbladder itself to be an adverse event while others cite a possible increased risk of postoperative fluid collection as reason to consider spillage to be adverse [27]. Similarly, the perception of operative planes, adequacy of retraction and exposure, and the characterization of tissues can vary across surgeons [28]. The identifiability of clinically meaningful phenomena also presents a tremendous challenge for their annotation and successful deployment in an AI model. Phenomena are identifiable, if – based on the data, labels, and AI model – they can reproducibly be identified. The patient's favorite color, for example, is a non-identifiable phenomenon in the context of surgical video analysis. Many surgical events and areas of interest, unfortunately, are either non-identifiable or poorly identifiable due to their limited visual recognizability. For example, it is often difficult to recognize, even for advanced trainees, the subtle difference between being in the correct versus incorrect surgical plane. We see this daily in the operating room manifested as the 'sixth sense' of the expert surgeon. Thus, we reiterate the importance of defining the clinical phenomena of interest that are visually identifiable in advance and in a manner that is clear to annotators. With these types of annotations, having additional annotator training beyond just providing annotation guides and definitions can allow for iterative improvement in annotation quality across a group of annotators [2].

Challenges in annotating surgical performance

Annotation of surgical video for surgical performance has been a longstanding practice in surgical education. The goal of these annotations, historically, has not been to train an algorithm to perform automated assessment but rather, to provide structured, formative

feedback to surgeons. Methods of assessment can be divided into global and procedure/task-specific assessments, with several different assessment tools that have been validated for the purposes of distinguishing experienced and inexperienced surgeons. Kinematic approaches such as assessment of motion have also been used to differentiate experienced from novice surgeons and to track the learning curve [29,30].

The Objective Structured Assessment of Technical Skills (OSATS) is perhaps the most reported assessment tool used in surgery. It provides a global rating scale for assessment of surgical skill, including measures such as 'respect for tissue,' 'instrument handling', and 'flow of operation.' Each of these elements is rated on a 5-point Likert scale with anchors at 1 (poor performance), 3 (acceptable performance), and 5 (superior performance). As demonstrated in Table 3, while anchors are provided to help annotators better understand the performance sought, scoring is ultimately subjective and can be influenced by the surgical experience and expectations of the annotators. Similar issues exist for more domain specific assessment tools in laparoscopic and robotic surgery. Thus, while research is underway to develop summative assessment tools for specific operations [31,32], assessment of performance that is based on rating scales will continue to have elements of subjectivity. As long as it is appropriately considered in analysis, such subjectivity could actually enrich assessment if it allows for feedback to be provided to the surgeon.

Full, unedited video serves as the raw data source for assessing surgical performance and allows for assessment of the entire procedure, including both technical and non-technical aspects of a surgeon's performance. However, annotation and analysis of entire cases can be time-consuming, with operative times ranging from 20 min for short procedures to several hours for complex cases. The use of short segments of cases, edited together as a synopsis, has been explored in the surgical education literature, but reports have noted that assessments of edited videos have poor inter-rater reliability and low discriminative ability in distinguishing trained versus untrained participants [33]. Thus, while short segments may be sufficient when assessing specific tasks such as intracorporeal suturing, full videos are likely necessary to appropriately annotate performance on a procedure as a whole.

Ultimately, assessment of performance will likely be tied to clinical outcomes. The correlation between ratings of surgical performance and clinical outcomes has been well documented [34,35], raising the possibility of eventually annotating performance based on expected clinical outcome for patients rather than the subjective rating of human (or machine) graders. However, limitations in reporting of outcomes relative to surgical performance have thus far affected its application. The heterogeneity in assessing rate of learning of skills across surgeons and specialties has made it difficult to specifically identify learning curves for many procedures [36], and lack of clarity around where a surgeon may sit on the learning curve may affect expected outcomes. Variation in performance by a single, experienced surgeon across cases may lead to differences in outcome as can differences in the complexity of a patient's presentation [37]. Finally, it is difficult to attribute causality in outcome of a patient to the intraoperative phase of care and surgeon's performance alone. Factors such as the patient's comorbidities, postoperative care, and effect of other providers, play a role in the clinical outcome of the patient. Therefore, approaches that isolate a surgeon's contributions only - whether through rating scales, kinematics, or computer vision - may only provide a partial contribution to a patient's expected clinical outcome. Given these challenges, annotation of surgical performance relative to clinical outcomes alone remains an elusive ideal that requires significant further investigation.

Challenges in annotation tools

Armed with a video dataset and well-planned annotation schema, the surgical annotator must then put

Table 3. An excerpt of part of the Objective Structured Assessment of Technical Skills from Martin et al. [28].

Time and motion				
1 Many unnecessary moves	2	3 Efficient time/motion but some unnecessary moves	4	5 Clear economy of movement and maximum efficiency
Flow of operation				
1 Frequently stopped operating and seemed unsure of next move	2	3 Demonstrated some forward planning with reasonable progression of procedure	4	5 Obviously planned course of operation with effortless flow from one move to the next

theory into practice and annotate videos. Many annotation tools and software exist, each trying to take human labels and translate them into machine-understandable inputs from which an AI model can learn. Some tools focus on temporal annotations alone, like Anvil, while others, like Visual Object Tagging Tool (VoTT) and *interactive* Video Annotation Tool (*i*VAT), can only annotate spatial features. More recently developed tools incorporate both labeling abilities, such as the publicly available VCG Image Annotator (VIA) [38]. Rudimentary annotation can even be performed with spreadsheets, simply by entering data and timestamps into data cells.

A user-friendly and efficient annotation tool can make-or-break the annotation process. Always evaluate a software prior to committing to its use for a project. Features to consider include a program's interface, process for loading videos, annotation export formats, and ability to integrate AI models to facilitate annotation. Its interface must be user friendly to the annotator (be it layperson, clinical trainee or clinical expert,) and run across different operating systems. In order to annotate, it must have access to videos and be able to play a wide range of video formats. Some software can even load videos from a centralized video repository rather than requiring annotators to have video copies present on their computers. This centralized storage keeps videos in one secure location, which minimizes chain-of-custody issues with regard to privacy laws. The software must also be able to export the data in a format that the AI model can preferably directly load, and if not, at least a format with bindings in common programming languages so it can be easily modified for model input. Additionally, enabling version control of annotations and the data dictionary is important as both will need continuous updating. Often, determinations are made to change the way a dataset is being annotated in order to improve algorithm development. For example, annotations may be too generic for current machine learning technology to learn from them; or annotations may not be purposely suited to the problem being solved. Ensuring efficient and accurate updates to labels will enable more accurate data and reduce the need to re-annotate entire datasets. Lastly, if the annotation task is to be performed on a large-scale, AI models, as mentioned previously, can 'pre-annotate' the dataset. Some software facilitates this pre-annotation task, allowing it to happen directly in the model, which allows for substantially faster creation of annotated datasets.

Unfortunately, the ultimate annotation software has yet to be created. No publicly available software can annotate images and videos with spatial and temporal annotations, by multiple annotators for the same video, from a centralized video repository, with easy annotation export, and annotation assistance by AI models. Until such a tool is created, current users must either use nonpublic industry tools (if they have access to one) or the limited publicly available ones, considering the tradeoffs listed above.

Next steps forward

Given many of the challenges we have reviewed above, there is clearly a need to establish consensus around the development and use of surgical annotation. Efforts have been underway to bring together the surgical data science community with the goal of moving forward from concepts in data science to actions required for the translation to clinical investigation and ultimate application to patient care [39,40].

We highlighted the importance of clearly defining the clinical phenomena of interest in executing annotations of surgical video. Clearly outlining and defining the concepts that exist within and across operations in a manner that is accessible to all researchers is a key component of enabling multi-institutional, multidisciplinary research that can be compared, contrasted, or combined. OntoSPM is an ambitious project that aims to outline a core ontology for surgical process models to enable large scale research efforts across groups [17]. In 2020, the Society of American Gastrointestinal and Endoscopic Surgeons convened a consensus group of surgeons and engineers to draft recommendations on the annotation of minimally invasive surgical videos, including both spatial and temporal annotations, the results of which are pending publication.

It is important to consider, however, that some annotator variability may not be the fault of poorly defined phenomena. Rather, such variability may reflect the 'fuzzy' nature of a phenomenon itself [41]. Even experienced surgeons may differ in their conceptualization of some phenomena, such as safe and unsafe zones of dissection or identification of specific anatomic structures [16,42,43]. Thus, combining annotations to serve as a fuzzier ground truth or to establish thresholds of agreement as ground truth may serve to either enhance modeling of clinical phenomena that are, by nature, fuzzy or provide a more realistic benchmark for model performance (i.e. to compare

agreement of a model to multiple annotators vs. a single annotator) [11,16,44].

One must also consider the downstream biasing effects of data and annotations. All models, ML or not, are inherently biased: they are simplified, compressed, representations of reality learned from limited information. Even unsupervised learning models that do not use annotations have bias, as they learn from an inevitably unrepresentative subset of surgical videos. Tremendous thought must be put into building diverse, representative datasets, not just those from 'perfect' cases. Similar care must be given to defining widely applicable annotation labels. We do caution that, even with the best of efforts, these models will be biased. Studies into the effects of bias is an active and critical research area that will ensure the fair and effective deployment of AI into the operating room.

Finally, some elements of annotation of surgical video remain to be clearly defined (e.g. clinically meaningful events ranging from bleeding to bowel injury to retraction and exposure). While these types of events can be defined internally within a given study, scaling research efforts to enable translation to clinical practice will require at least some consensus on how such events should be annotated. Partnership between surgical data scientists, practicing surgeons, and health services and surgical education researchers could yield fruitful discussion and consensus on how to handle these types of events to enable consistent annotation across fields.

Conclusions

The rigorous application of surgical video annotation will be important to further advance the field of surgical data science, particularly as it relates to research on development of computer vision applications. In designing research to develop and validate such applications, researchers should consider carefully the specific phenomena of interest to determine whether ground truth annotations appropriately represent those phenomena or whether they represent alternative phenomena outside the scope of interest. Additional work will be required to build consensus across disciplines on annotation of clinically meaningful events and surgical performance, as these concepts across disciplines ranging from surgical data science to surgical education and health services research. Consensus efforts across disciplines offer an opportunity to impact a wider scope of work beyond automated surgical video analysis.

Disclosure statement

TMW, YB, GR, ORM, and DAH receive research support from Olympus Corporation. TM, GR, ORM, DAH have received research support from CRICO Risk Management Foundation. DMF is a consultant for Johnson & Johnson. GR receives research support from Toyota Research Institute. ORM is a consultant for Medtronic and Olympus Corporation. DAH is a consultant for Johnson & Johnson and Verily Life Sciences. DAH has received research support from the Intuitive Foundation and the Society of American Gastrointestinal and Endoscopic Surgeons.

References

[1] Hashimoto DA, Rosman G, Rus D, et al. Artificial Intelligence in surgery: promises and perils. Ann Surg. 2018;268(1):70–76.

[2] Deal SB, Stefanidis D, Brunt LM, et al. Development of a multimedia tutorial to educate how to assess the critical view of safety in laparoscopic cholecystectomy using expert review and crowd-sourcing. Am J Surg. 2017;213:988–990.

[3] Maier-Hein L, Mersmann S, Kondermann D, et al. Can masses of non-experts train highly accurate image classifiers? In: Golland P, Hata N, Barillot C, Hornegger J, Howe R, editors. Medical image computing and computer-assisted intervention – MICCAI 2014. Cham: Springer International Publishing; 2014. p. 438–445.

[4] Deal SB, Stefanidis D, Telem D, et al. Evaluation of crowd-sourced assessment of the critical view of safety in laparoscopic cholecystectomy. Surg Endosc. 2017;31(12):5094–5100.

[5] Malpani A, Vedula SS, Chen CCG, et al. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. Int J Comput Assist Radiol Surg. 2015;10(9):1435–1447.

[6] Holst D, Kowalewski TM, White LW, et al. Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. J Endourol. 2015;29(10):1183–1188.

[7] Chen C, White L, Kowalewski T, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res. 2014;187(1):65–71.

[8] Pugh CM, Hashimoto DA, Korndorffer JRJ. The what? How? And Who? Of video based assessment. Am J Surg. 2021;221(1):13–18.

[9] Twinanda AP. Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos. [updated online 2017 Jan 27; cited 2020 Sept 10]. https://www.theses.fr/2017STRAD005.

[10] Twinanda AP, Shehata S, Mutter D, et al. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging. 2016;36(1):86–97.

[11] Hashimoto DA, Rosman R, Witkowski E, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. Ann Surg. 2019;270(3):414–421.

[12] Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic

myotomy. Surg Endosc. 2020. DOI:10.1007/s00464-020-07833-9
[13] Gwet KL. Testing the difference of correlated agreement coefficients for statistical significance. Educ Psychol Meas. 2016;76(4):609–637.
[14] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–549.
[15] Ward TM, Hashimoto D, Ban Y, et al. Training with pooled annotations from multiple surgeons has no effect on a deep learning artificial intelligence model's performance. J Am Coll Surg. 2020;231(4):e203.
[16] Madani A, Namazi B, Altieri MS, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Ann Surg. 2020. DOI:10.1097/SLA.0000000000004594
[17] Gibaud B, Forestier G, Feldmann C, et al. Toward a standard ontology of surgical process models. Int J Comput Assist Radiol Surg. 2018;13(9):1397–1408.
[18] Ban Y, Rosman G, Ward T, et al. Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows. ArXiv200900681 Cs. [updated online 2020 Dec 3; 2021 cited 28]. http://arxiv.org/abs/2009.00681.
[19] Nwoye CI, Gonzalez C, Yu T, et al. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: Medical image computing and computer-assisted intervention – MICCAI. Heidelberg (Germany): Springer International Publishing; 2020.
[20] Lin HC, Shafran I, Yuh D, et al. Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput Aided Surg. 2006;11(5):220–230.
[21] Jung JJ, Jüni P, Gee DW, et al. Development and evaluation of a novel instrument to measure severity of intraoperative events using video data. Ann Surg. 2020;272(2):220–226.
[22] Vezhnevets A, Buhmann JM, Ferrari V. Active learning for semantic segmentation with expected change. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012. p. 3162–3169.
[23] Kim T, Lee K, Ham S, et al. Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT. Sci Rep. 2020;10(1):366.
[24] Rocha C. d C, Padoy N, Rosa B. Self-supervised surgical tool segmentation using kinematic information. In: 2019 International Conference on Robotics and Automation (ICRA). 2019. p. 8720–8726.
[25] Kaafarani HM, Mavros MN, Hwabejire J, et al. Derivation and validation of a novel severity classification for intraoperative adverse events. J Am Coll Surg. 2014;218(6):1120–1128.
[26] Bohnen JD, Mavros MN, Ramly EP, et al. Intraoperative adverse events in abdominal surgery: what happens in the operating room does not stay in the operating room. Ann Surg. 2017;265(6):1119–1125.
[27] Rice DC, Memon MA, Jamison RL, et al. Long-term consequences of intraoperative spillage of bile and gallstones during laparoscopic cholecystectomy. J Gastrointest Surg. 1997;1(1):85–91.
[28] Martin J, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84(2):273–278.
[29] Hashimoto DA, Sirimanna P, Gomez ED, et al. Deliberate practice enhances quality of laparoscopic surgical performance in a randomized controlled trial: from arrested development to expert performance. Surg Endosc. 2015;29(11):3154–3162.
[30] Levin M, McKechnie T, Khalid S, et al. Automated methods of technical skill assessment in surgery: a systematic review. J Surg Educ. 2019;76(6):1629–1639.
[31] Feldman LS, Pryor AD, Gardner AK, et al. SAGES Video-Based Assessment (VBA) program: a vision for life-long learning for surgeons. Surg Endosc. 2020;34:3285–3288.
[32] Ritter EM, Gardner AK, Dunkin BJ, et al. Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative performance assessment. Surg Endosc. 2020;34(7):3176–3178.
[33] Scott DJ, Rege RV, Bergen PC, et al. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech. 2000;10(4):183–190.
[34] Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369(15):1434–1442.
[35] Curtis NJ, Foster JD, Miskovic D, et al. Association of surgical skill assessment with clinical outcomes in cancer surgery. JAMA Surg. 2020;155(7):590–598.
[36] Soomro N, Hashimoto DA, Porteous A, et al. Systematic review of learning curves in robot-assisted surgery. BJS Open. 2020;4(1):27–44.
[37] Curtis NJ, Stevenson ARL, Francis NK. Assessment of surgical skill and performance variability-reply. JAMA Surg. 2020;155(12):1175.
[38] Bianco S, Ciocca G, Napoletano P, et al. An interactive tool for manual, semi-automatic and automatic video annotation. Comput Vis Image Underst. 2015;131:88–99.
[39] Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. Nat Biomed Eng. 2017;1(9):691–696.
[40] Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science–from concepts to clinical translation. ArXiv Prepr ArXiv201102284. Published online 2020.
[41] Arbelaez P, Maire M, Fowlkes C, et al. Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell. 2010;33(5):898–916.
[42] Madani A, Watanabe Y, Bilgic E, et al. Measuring intra-operative decision-making during laparoscopic cholecystectomy: validity evidence for a novel interactive Web-based assessment tool. Surg Endosc. 2017;31(3):1203–1212.
[43] Madani A, Grover K, Watanabe Y. Measuring and teaching intraoperative decision-making using the visual concordance test: deliberate practice of advanced cognitive skills. JAMA Surg. 2020;155(1):78–79.
[44] Krause J, Perer A, Bertini E. A user study on the effect of aggregating explanations for interpreting machine learning models. In: ACM Workshop on Interactive Data Exploration and Analytics. New York (NY): ACM; 2018.