A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video

Krystel Nyangoh Timoh, Arnaud Huaulmé, Kevin Cleary, Myra A. Zaheer, Vincent Lavoué, Dan Donoho, Pierre Jannin

HHS Public Access

Author manuscript

Surg Endosc. Author manuscript; available in PMC 2023 June 21.

Published in final edited form as:

Surg Endosc. 2023 June ; 37(6): 4298–4314. doi:10.1007/s00464-023-10041-w.

A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video

Krystel Nyangoh Timoh[1,2,3,7], **Arnaud Huaulme**[2], **Kevin Cleary**[4], **Myra A. Zaheer**[5], **Vincent Lavoué**[1], **Dan Donoho**[6], **Pierre Jannin**[2]

[1]Department of Gynecology and Obstetrics and Human Reproduction, CHU Rennes, Rennes, France

[2]INSERM, LTSI - UMR 1099, University Rennes 1, Rennes, France

[3]Laboratoire d'Anatomie et d'Organogenèse, Faculté de Médecine, Centre Hospitalier Universitaire de Rennes, 2 Avenue du Professeur Léon Bernard, 35043 Rennes Cedex, France

[4]Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC 20010, USA

[5]George Washington University School of Medicine and Health Sciences, Washington, DC, USA

[6]Division of Neurosurgery, Center for Neuroscience, Children's National Hospital, Washington, DC 20010, USA

[7]Department of Obstetrics and Gynecology, Rennes Hospital, Rennes, France

Abstract

Background—Annotated data are foundational to applications of supervised machine learning. However, there seems to be a lack of common language used in the field of surgical data science.

The aim of this study is to review the process of annotation and semantics used in the creation of SPM for minimally invasive surgery videos.

Methods—For this systematic review, we reviewed articles indexed in the MEDLINE database from January 2000 until March 2022. We selected articles using surgical video annotations to describe a surgical process model in the field of minimally invasive surgery. We excluded studies focusing on instrument detection or recognition of anatomical areas only. The risk of bias was evaluated with the Newcastle Ottawa Quality assessment tool. Data from the studies were visually presented in table using the SPIDER tool.

Results—Of the 2806 articles identified, 34 were selected for review. Twenty-two were in the field of digestive surgery, six in ophthalmologic surgery only, one in neurosurgery, three in gynecologic surgery, and two in mixed fields. Thirty-one studies (88.2%) were dedicated to phase, step, or action recognition and mainly relied on a very simple formalization (29, 85.2%). Clinical

✉Krystel Nyangoh Timoh, krystel.NYANGOH.TIMOH@chu-rennes.fr.

Disclosure Dr. Krystel NYANGOH TIMOH, Dr. Arnaud HUAULMÉ, Dr. Kevin CLEARY, Mrs. Myrah A ZAHEER, Dr. Dan DONOHO, and Dr. Pierre JANNIN have no conflict of interest or financial ties to disclose. Pr. Vincent LAVOUÉ has a contract with Intuitiv® for proctoring.

information in the datasets was lacking for studies using available public datasets. The process of annotation for surgical process model was lacking and poorly described, and description of the surgical procedures was highly variable between studies.

Conclusion—Surgical video annotation lacks a rigorous and reproducible framework. This leads to difficulties in sharing videos between institutions and hospitals because of the different languages used. There is a need to develop and use common ontology to improve libraries of annotated surgical videos.

Keywords

Surgical data science; Ontology; Surgical process model; Annotation; Surgical video; Minimally invasive surgery

With the rise of minimally invasive surgery—including endoscopic, laparoscopic, and robotically assisted procedures—the new domain of *surgical data science* is emerging to improve the consistency and, hopefully, quality of care of the patient [1, 2]. Surgical data science consists of the scientific characterization of digital surgical information to improve patient outcomes. Among other areas of interest, two important goals of this field are to analyze *surgical workflows* and develop *context-aware systems,* both of which are significant features of the *operating room of the future* [3]. In the case of minimally invasive surgeries using endoscopy, both goals require surgical videos to be manually labeled in a spatial and/or temporal way following a *surgical process model (SPM)* [4], which is the cornerstone of surgical data science. Surgical process modeling consists of an analytical reduction of the surgical procedure in a formal or semi-formal representation defining phases, steps, and actions. It has already been developed for open surgery [5, 6]. *Computer vision using machine learning* has recently been used successfully for phase/step recognition or, in some cases, to estimate how long the surgical procedure will last [7, 8]. The vast majority of these approaches requires labeled surgical videos.

However, the availability of a large volume of labeled data represents a major bottleneck for machine learning applied to surgical video analysis [8], and partly explains why the application of machine learning in surgery is limited compared to medical imaging. It is therefore important to share data between different institutions to increase the data pool and accelerate research. An important limitation to this sharing is the lack of a standard vocabulary for video annotation.

Although the Society of Gastrointestinal and Endoscopic Surgeons (SAGES) has recently provided guidelines for video annotation, a review that reproduces prior processes in surgical video annotation with robust formalization is required to ensure that surgical data are machine-readable *and clinically meaningful* [9]. Recent literature reviews of surgical data science focus mostly on artificial intelligence and machine learning techniques [8]. To date, no reviews have summarized the process of choice and language used for surgical video annotation.

Therefore, the aim of this study is to review the process of annotation and semantics used in the creation of SPM for minimally invasive surgery videos.

To present a review of the literature using annotation for creating SPM for minimally invasive surgery.

Materials and methods

Search strategy

We conducted a systematic review according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis guidelines (PRISMA) 2020 [10].

Two investigators (KNT and MZ) performed an English literature search on Medline (Pubmed) from January 2000 to March 2022.

We selected articles in English including the use of labeled surgical videos to contextualize the SPM for minimally invasive surgeries.

The following keywords were used: minimally invasive surgery OR surgery AND machine learning OR deep learning OR computer vision AND surgical workflow OR surgical process model OR video annotation.

After the exclusion of duplicate articles, all the articles were screened by the two investigators. Titles and abstracts were initially assessed for eligibility before conducting a second selection based on the full text to exclude inappropriate articles. Any discrepancies were resolved by consensus.

Inclusion and exclusion criteria

Studies were included if they used surgical video labeling in minimally invasive surgery.

All fields of surgical specialty were considered. Studies focusing on instrument detection or recognition of anatomical areas only were excluded. We excluded commentaries, editorials, expert consensus, reviews, abstracts, and pure bioengineering research.

Definitions

- An *SPM* is defined as “a simplified pattern of a surgical process that reflects a predefined subset of interest of the surgical process in a formal or semi-formal representation” [4].
- The *granularity level* for the temporal description of the procedure is defined as the level of abstraction at which the surgical procedure is described. The highest level is *the procedure* itself. The procedure is composed of a list of phases. They must occur sequentially in the following order: Access, Execution of Surgical Objectives, Closure. Each phase is composed of several steps. As described by Lalys et al. [4] and the recommendations of SAGES [9], steps represent a sequence of activities used to achieve a clinically meaningful surgical objective. They are procedure specific. An *activity* is defined as a physical task including an action verb, instrument, and anatomy with an origin and a destination. It has starting and ending times, as well as the body part(s) performing the action. Each activity is composed of a list of motions.

- A *granularity level* for the spatial description of the anatomy is defined with a basic hierarchical organization of anatomic spatial features from high level to low level: (1) Anatomic region (e.g., upper or lower abdomen, pelvis, retroperitoneum), (2) General anatomy (e.g., veins, arteries, muscle), and (3) Specific anatomy (e.g., liver, gallbladder, stomach, cystic artery, common bile duct) [9].
- A surgical procedure is described through the annotation of videos. This description is available through a representation at a certain level of formalization describing the level at which the description is represented: heavyweight ontology is a representation with the highest formalization level based on a hierarchy of concepts and relations, lightweight ontologies are represented by Unified Modeling Language (UML) class diagrams and/or eXtensible Markup Language (XML) Schema. At the lower level, hierarchical decomposition, sequential or non-sequential lists are also used, suggesting a list of words to represent one or many levels of the surgery's granularity [4, 9].

Data extraction (Fig. 1)

The principal data items extracted and analyzed from the articles are as follows:

– Application

- *Surgical specialties* In this review, we individualized surgical specialties according to the organ system targeted: digestive surgery, ophthalmologic surgery, neurosurgery, gynecologic surgery, and ear-nose-throat surgery.
- *Clinical applications* phase, step, or action recognition, surgery time prediction, surgical quality, context-aware systems, robotic assistance, and automatic generation report.
- *Quality criteria of the dataset* To assess the quality of the dataset and avoid the high risk of bias, as described by Anteby et al. [8], we reported the following items: description of the population, clinical information, ethics committee approval, clinical selection criteria of patients eligible for surgery, selection criteria of videos, inclusion timeline, and consecutive cases.
- *Modeling* Modeling describes and explains the *work domain* which is identified by the granularity level at which the procedure is studied, the operator involved, and the formalization[4].
- The granularity levels (defined above).
  - We measured the semantic strength of each phase, i.e., the median number of words used to define the phase.
- Annotation creation was the methodology employed to provide SPM: generation based on local expert consensus; or based on international consensus, literature, or upper ontology.
- The formalization level (defined above)

- *Acquisition* The collection of data on which the models are built and the first step toward creating an SPM. We extracted information about:

- The videos: number
- The annotation software: name and availability
- The surgeons: number and level of expertise
- The annotators: number, specific training, quality of annotators and study of inter-/intra-annotator variability.

Quality assessment

For assessing the risk of bias to comply with Prisma criteria, the Newcastle–Ottawa quality assessment scale was used. Even though not all domains of Newcastle–Ottawa could be applied, Results are presented in Appendix.

Results

Study selection and characteristics of the studies

The initial search yielded a total of 2806 articles of which 374 were screened and 75 were eligible. Finally, 34 were selected for the review including a total of 2588 annotated videos (Fig. 2). All the studies were published between 2011 and 2022 (Table 1).

Figure 1 represents the framework of analysis of the articles.

Application

Surgical field—Of the selected studies, 22 were in the field of digestive surgery, six in ophthalmologic surgery only, one in neurosurgery, three in gynecologic surgery, and two in mixed fields (1 of ophthalmologic and digestive surgeries, and 1 of digestive and gynecologic surgeries) (Fig. 3). Only two of the 34 studies studied robotic-assisted surgery.

Clinical application and results—Most of the studies (31, 88.2%) were dedicated to phase, step, or action recognition. Three (8.8%) studies focused on surgical quality, and two (6.1%) on procedure duration. Only one (2.9%) study provided a clinical correlation with surgical procedure annotation.

Thirty-three (97.1%) studies used machine learning techniques.

Quality criteria of the dataset

The population involved in the dataset was fully described in two (5.9%) studies only. In the studies of Kitugachi et al. and Khan et al. [11, 12], information such as colorectal tumor score and histologic nature of the pituitary gland tumors were provided. Clinical information was lacking for studies using available public datasets. Clinical information was confronted to annotation analysis results in Cheng et al. only (2.9%) [13]. They correlated the severity of cholecystitis with the duration of the surgical procedure [13]. However, no other clinical information was available in their study. Selection clinical criteria of patients for surgery

were detailed for one (2.9%) study [11] only. Selection criteria of videos for inclusion in the dataset were detailed in four (11.8%) studies [11, 13–15].

Only thirteen (41.2%) studies mentioned ethics committee approval [11–18].

Only Huaulmé et al. indicated that cases were consecutively included in the dataset during the inclusion time [18], decreasing the bias associated with non-consecutive case inclusion.

The inclusion timeline was often extensive (median duration of 35 (5–125) months) implying a possible shift in surgical guidelines during the study period.

Eleven (32.4%) studies [12, 17, 19–27] used one or several available public datasets (Table 2) and the remaining 23 (67.7%) used private datasets.

Modeling

Formalization—Most of the studies were based on a very simple formalization: a 2D graph with a sequential list in 24 (70.6%) studies and a non-sequential list in five (14.7%). A more complex formalization was used in two studies with a hierarchical decomposition in three (8.8%) and a diagram in one.

Work domain

Surgical procedures—The studied *surgical procedures* were cholecystectomy (17, 50%), cataract surgery (7, 20.6%), bypass (2, 5.9%), sleeve gastrectomy (3, 8.8%), hysterectomy (2, 5.8%), pituitary gland removal (1, 2.9%), rectopexy (1, 2.9%), and sacrocolpopexy (1, 2.9%). At this highest level, only 53% (18) of the studies clearly described the specific surgical procedure.

Annotation creation was described in 23 (68%) studies and was based on one surgeon (9, 26.5%), literature exclusively (4, 11.8%), local consensus by several surgeons exclusively (4, 11.8%), both local consensus and literature (4, 11.8%), cognitive task analysis with engineer and surgeon experts (1, 2.9%), or upper ontology (1, 2.9%).

The *lower granularity level* included phases (29, 87.9%), steps (5, 14.7%), actions (7, 20.6%), and instruments (23, 68%).

Phases—Two (2/29, 6.9%) studies used the term phases in coherence with the previous consensual definition [13, 28]. We observed that the terms step and phase were employed indiscriminately in many articles.

The median semantic strength (i.e., number of words used to describe a phase) was 2 but was highly variable (0–33). Four (13.8%) studies provided additional information like the start and ending of a phase [12, 13, 29, 30].

Eleven (38%) studies provided pictures to illustrate phases.

For the same surgical procedure, we observed high heterogeneity in the number of phases described from one study to another (Fig. 4): from 6 to 14 for cholecystectomies; from 3 to

12 for cataract procedures; 7 and 10 for the two hysterectomy studies; and 7 and 8 for the sleeve gastrostomy procedures.

Three studies focusing on the whole procedure did not detail the number of phases [15, 18, 26].

In the study by Guedon et al., excessive bleeding was described as a possible additional phase that could occur at any time [31].

Nine studies annotated idle times, [15, 17, 23, 24, 30, 32–35] providing additional data on phase annotation.

Steps—Five (14.3%) studies described the *steps* [11, 14, 23, 24, 28], and used the term in coherence with the previous consensual definitions [1, 4]. Semantic strength varied considerably with a median of 3.5 (2–36). None of the studies provided additional information (like the start and end of a step), or pictures to illustrate the steps.

Activities: actions, instruments, anatomy—*Activities* were characterized in nine (26.5%) of studies [4, 12, 16, 29, 30, 36–38], with one (11.1%) providing pictures to illustrate the activities [39].

The study by Derathé et al. [16], analyzing laparoscopic sleeve gastrectomy, focused on a single phase: the exposure of the surgical scene with a view to assessing its quality. Therefore, they annotated activities during this phase: actions, instruments, and anatomy.

Instruments were characterized in 23 (68%) studies. In 19 (82.6%) studies, specific instruments were described with a median of 12 (2–21) instruments specified per study. Four (17.4%) studies [12, 19, 20, 23] provided pictures to illustrate the instruments.

Anatomy was described in eight (23%) studies [16, 17, 36–40], and specific anatomy described in three [38–40]. Anatomical characteristics (normal or pathologic) were never reported.

Other useful information—Some authors, such as Derathé et al. [16], focused on the *surgical quality* within a phase with a quality-oriented annotation.

Mascagani et al. focused on the identification of the critical view of safety [38, 41].

Other studies added *events* and classified them as “normal or abnormal” such as Hashimoto et al. [14] during sleeve gastrectomy, and Huaulmé et al. [18] during rectopexy.

Finally, Malpani et al. [30] reported additional data provided by da Vinci Surgical System as tool identity, tool changes, endoscopic movements, repositioning of the manipulator, and a head-in indicator identifying whether the surgeon was working at the console.

Acquisition

Surgeons—The number of surgeons performing or involved in the surgeries was listed in 24 (70.6%) studies. The median number of surgeons involved was 6 (1–28). Twenty

(58.8%) studies reported the expertise level of surgeons although the definition of expertise was poorly detailed and heterogeneous: a trainee in one study, an expert in 10, and mixed trainee and surgeon in nine. Two studies reported that the surgeries could be performed by two operators spontaneously: an expert and a fellow [30, 32]. Twenty (58.8%) of the studies mentioned that the surgery took place in an affiliated institution, and five studies involved multiple institutions.

Videos—The median number of videos used per study was 45 (7–461).

Videos came from robotic surgery in two studies [28, 35]. One study mixed videos from robotic and laparoscopic surgeries [35].

The annotation software was reported in 12 (34.3%) studies (Table 3).

Annotators—Information about annotators was available in 23 (65.7%) studies. The median number of annotators was 2 (1–4). The annotators had been specifically trained in three studies [14, 15, 31]. Thirteen studies indicated the expertise level of annotators: surgeons in nine studies [11–14, 17, 23, 33, 35] [40], non-clinical researchers in two [15, 18], mixed physicians and trained annotators in one [42], and mixed scientists and expert surgeons in one [16]. Only one study studied inter-annotator variability [13] (Table 4). None of the studies described the learning curve of the annotators.

Annotation was corrected by an expert surgeon annotator in four studies [14, 15, 18, 23]

In one study the annotator was changed for the same dataset [15].

Quality assessment and risk of bias

Manuscripts were ascribed a high risk of bias because of failure to report ethics committee approval and to describe the study population. The Newcastle–Ottawa quality assessment scale was used (Appendix).

Evolution of quality of annotation over the years

The quality of annotation seems to have slightly improved over the 14-year time span.

Nine studies used literature and cross-referenced previous studies in the process of creating an annotation. These studies were all recent (from 2014 to now).

Before 2020 (i.e., between 2008 and 2020), among 19 articles, only 6 (31.6%) studies had ethics committee approval. Only one (5%) study had clinical data within the dataset. Information about annotators was reported in 9 (47.4%) studies (surgeons in 3 cases, physicians and trained annotators in 1 case, and scientists and surgeons in 1 case). The median semantic strength of phase was 4 (2.2–6.6), and of steps was 4.5 (3–6.5).

After 2020 (i.e., between 2020 and 2022), among 15 articles, 9 (60%) studies had ethics committee approval. Three (20%) studies had clinical data within the dataset. Information about annotators was reported in 7 (47%) studies (surgeons in 6 cases, data engineers in 1

case, and scientists in 1 case). The median semantic strength of phase was 6.2 (3.2–10), and of steps was 2.8 (1–3).

Discussion

The present review highlights that the process of surgical video annotation for minimally invasive surgeries is highly variable between studies. Surgical procedure description through the SPM lacks robust and consistent formalization illustrating relationships between concepts. The methodology employed to choose semantics and vocabulary is rarely standardized and not reproducible, leading to heterogeneity in the generation of SPM among studies. These results may explain the current lack of success stories in the field of surgical data science.

Video labeling is a matter of high interest in the surgical data science area. The semantics used are the fundamental basis for generating SPM [43], i.e., detailed descriptions of surgical procedures [4]. SPM is an original approach to establish a solid basis for analysis of various aspects of surgical procedures [44]. Its usage could improve surgical workflow management in the modern operating room and help optimize and improve the procedures [3]. Additionally, semantic coherence facilitates data sharing and collaboration between institutions which would result in gathering enough surgical cases to ensure the representativeness of pathologies and procedures. Data are the foundation of machine learning, and the lack of annotated data is a limiting factor for improving deep learning performance [8]. Machine learning mainly uses supervised learning; thus, raw data are of little utility without annotation [45]. Finally, a robust representation of SPM or formalization is needed to represent surgical knowledge, and make it explicit, and consequently shareable. According to the recent SAGES recommendations, formalization must be universal, scalable, machine-readable, clinically applicable, and flexible [14]. Ontologies are key to creating standardized SPMs [37, 43]. Therefore, choosing appropriate vocabulary to annotate surgical videos is crucial to raising artificial intelligence surgical data science.

In this review, we found that laparoscopic cholecystectomy is the most studied surgical procedure as in the review by Garrow et al. [45]. This is probably because there are many publicly available datasets which focus on this simple, and often minimally invasive, procedure.

In the current review, one-third of the studies did not describe the annotation creation employed to generate SPM. When described, only five studies were based on both local consensus and literature or on an upper ontology. Huaulmé et al. used a cognitive task analysis between engineer and surgeon experts [18]. The Delphi methodology for SPM generation may be a good option as many expert surgeons from different institutions are involved in the process, making the results more broadly accepted [46].

We also observed a high variability and heterogeneity in SPMs across the studies. There was considerable variability between the number of phases or steps for the same surgical procedures. For example, depending on the study, the cataract procedure consisted of 4 to

14 phases. Also, the terms “steps” and “phases” were often used incorrectly, making study comparison difficult, although clear definitions of phases and steps exist [4, 14].

We also investigated the quality of the definitions of the elements of SPM by measuring the mean number of words used (called semantic strength) and found it to be highly variable for phases and steps. We noted that four studies added information at the start and end of a phase, increasing the accuracy of definitions. Also, additional pictures to illustrate phases, steps, or instruments were provided in some studies.

Most of the studies in the present review used very simple formalization such as a sequential or non-sequential list. This lightweight formalization does allow visualization of relationships between elements compared to heavyweight formalization. While international healthcare terminology standards for biomedical data science are well established (such as the Foundational Model of Anatomy (FMA) [47], Gene Ontology (GO) [48], and others), ontologies to describe activities and other aspects of interventional care processes are rare [1, 2]. However, In the surgical field, recent insights have provided clues to create a robust formalization. OntoSPM [43] and LapOntoSPM [37] are the first specific ontologies focusing on the modeling of the entities of surgical process models. OntoSPM [43] is now organized as a collaborative action associating a dozen European research institutions, gathering the basic vocabulary to describe surgical actions, instruments, actors, and their roles. This project is promising as initiatives like the OBO Foundry [49] (a project that focuses on biology and biomedicine) provided evidence that building and sharing interoperable ontologies stimulate data sharing within a domain [1]. Widespread broad ontology for surgical application is thus fundamental, built upon close collaboration between surgeons and engineers [50]. It improves the clinical relevance of the terms used as well as promoting the use of vocabulary familiar to surgeons. The formalization of the SPM has already been used to describe open surgical procedures. This allows interesting analyses such as distinguishing expert and junior performance in discectomy, for example [6].

However, there are some disadvantages of ontology including a lack of flexibility, a considerable initial effort [44], and its so-called complexity. These factors explain why ontology is not widely used in machine learning despite its interest. Representing relationships between elements can be time-consuming and challenging in spite of software like Protégé, a free open-source ontology editor [51].

All the studies included in this review used endoscopic or microscopic video for data acquisition. These techniques give a good view of the surgical site, suitable for low-level data acquisition, but no access to operative room ergonomics or insight into team interactions [44]

Beyond the question of ontology, video annotation is associated with several challenges. Practical considerations need to be taken into account when sharing surgical videos between countries and between hospitals. The use of a dataset of quality is fundamental. Before gathering videos within a dataset, ethics committee approval is required, and patient consent must be obtained. In this review, several publications failed to provide information about ethics committee approval, the surgeons involved, number of institutions, and clinical data.

These results are consistent with those of Anteby et al. [8]. Furthermore, public datasets provide less clinical information. A pseudo-anonymization should be performed to be able to use a video retrospectively. When integrating a shared database, the question of copyright of the video can also be an issue [31]. Furthermore, a major bottleneck for data annotation is the lack of access to expert knowledge. However, dedicated software can help structure expert knowledge through SPM [52]. It is therefore crucial to use specifically trained annotators. In this review, information about annotators was often lacking, and especially inter-annotator reliability. Moreover, labeling videos is timeconsuming, as demonstrated by Huaulmé et al. [36], and resource-intensive. One solution may be crowdsourcing, where the annotation task is outsourced to an anonymous untrained crowd [2]. Ultimately, data should be collected as a matter of best practice in a consistent, longitudinal manner using tools that are smoothly integrated into the clinical workflow. Workers in the field need to identify allies and clear short-term “win scenarios” that will build interest and trust in the area so that hospitals, insurers, and practitioners all see the value of creating these resources, which will ultimately advance the profession [2]. Surgeons must be reserved for high-performance annotation such as identifying anatomical features and assessing the quality of a dissection [50].

In this review, only two studies reported cases using robotic surgery, and one also focused on data about optics. Robotic assistance surgery offers the possibility to extract additional data from robot kinematics and event data. Hung’s team has introduced automated performance metrics (APM) with machine learning to assess a surgeon’s performance, recognize the surgical activity, and even anticipate surgical outcomes [53–55]. Additionally, in a recent review, our team demonstrated that APMs could be considered objective tools for technical skills assessment, even though associations between APMs and clinical outcomes remain to be confirmed by further studies, particularly outside the field of urology [56].

Limitations of the current review include the fact that we focused exclusively on minimally invasive surgery based on video. Studies focusing on element recognition, such as anatomy or instruments, were excluded in order to focus on the whole surgical procedure.

We put forward several propositions to support the generation of consistent and shareable annotation of surgical video labeling. First, each project should be supported by a dedicated team with a real partnership between surgeons and engineers with a protocol and process of annotation decided ahead of time. Second, the objective of the project should be clearly determined with a specific question and a specific outcome. This would help define the granularity level of the SPM required in both temporal and spatial dimensions during the annotation. Third, the creation of a high-quality database is crucial including both ethics committee approval and patient consent, and the collection of anonymized patient clinical data. Fourth, a clear methodology for annotation with SPM should be established based on worldwide expert surgeon consensus and literature with robust formalization applied to enhance the relationships between the different elements. Finally, trained annotators must be carefully chosen for a specific task according to their abilities to minimize inter-annotator variability. An annotation review by experts can be added when needed.

We conclude that most of the studies in this review failed to adhere to a rigorous, reproducible, and understandable framework for surgical video annotation. This results in the use of different languages and hinders the sharing of videos between institutions and hospitals, resulting in difficulties for widespread dissemination of surgical data science. There is an urgent need to follow rigorous and formal methodologies in surgical video annotation, including common ontology.

Acknowledgements

The authors want to acknowledge Felicity Neilson, native English speaker specialized in scientific writing for English editing.

Appendix: Risk of bias and the Newcastle–Ottawa quality assessment scale

Study	Selection Ascertainment of exposure Secure record (e.g., surgical records)	Outcome Assessment of outcome Independent blind assessment, record linkage
Blum et al. [32]	*	*
Bodenstedt et al. [19]	*	*
Bodenstedt et al. [20]	*	*
Cheng et al. [13]	*	*
Derathé et al. [16]	*	*
Dergachyova et al. [21]	*	*
Guedon et al. [31]	*	*
	*	*
Hashimoto et al. [14]	*	*
Huaulmé et al. [18]	*	*
Jalal et al. [56]	*	*
Jin et al. [22]	*	*
Katic et al. [37]	*	*
Khan et al. [11]	*	*
Kitugachi et al. [12]	*	*
Kitugachi et al. [29]	*	*
Pangal et al. [55]	*	*
Lalys et al. [39]	*	*
Lalys et al. [40]	*	*
Lecuyer et al. [23]	*	*
	*	*
Malpani et al. [30]	*	*
Mascagani et al. [38]	*	*
Mascagani et al. [41]	*	*
Meuwssen et al. [28]	*	*
Nespolo et al. [17]	*	*
Guerin et al. [56]	*	*
Quellec et al. [33]	*	*

Study	Selection Ascertainment of exposure	Outcome Assessment of outcome
	Secure record (e.g., surgical records)	Independent blind assessment, record linkage
Ramesh et al. [24]	*	*
Shi et al. [25]	*	*
	*	*
Twinanda et al. [26]	*	*
	*	*
Twinanda et al. [27]	*	*
	*	*
Yeh et al. [15]	*	*
Yu et al. [42]	*	*
Zhang et al. [35]	*	*
Zhang et al. [34]	*	*

References

1. Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A, Fallert J, Feussner H, Giannarou S, Mascagni P, Nakawala H, Park A, Pugh C, Stoyanov D, Vedula SS, Cleary K, Fichtinger G, Forestier G, Gibaud B, Grantcharov T, Hashizume M, Heckmann-Nötzel D, Kenngott HG, Kikinis R, Mündermann L, Navab N, Onogur S, Roß T, Sznitman R, Taylor RH, Tizabi MD, Wagner M, Hager GD, Neumuth T, Padoy N, Collins J, Gockel I, Goedeke J, Hashimoto DA, Joyeux L, Lam K, Leff DR, Madani A, Marcus HJ, Meireles O, Seitel A, Teber D, Ückert F, Müller-Stich BP, Jannin P, Speidel S (2022) Surgical data science - from concepts toward clinical translation. Med Image Anal 76:102306 [PubMed: 34879287]
2. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S, Hashizume M, Katic D, Kenngott H, Kranzfelder M, Malpani A, März K, Neumuth T, Padoy N, Pugh C, Schoch N, Stoyanov D, Taylor R, Wagner M, Hager GD, Jannin P (2017) Surgical data science for next-generation interventions. Nat Biomed Eng 1:691–696 [PubMed: 31015666]
3. Cleary K, Kinsella A (2005) OR 2020: the operating room of the future. J Laparoendosc Adv Surg Tech A 15(495):497–573
4. Lalys F, Jannin P (2014) Surgical process modelling: a review. Int J Comput Assist Radiol Surg 9:495–511 [PubMed: 24014322]
5. Jannin P, Raimbault M, Morandi X, Riffaud L, Gibaud B (2003) Model of surgical procedures for multimodal image-guided neurosurgery. Comput Aided Surg 8:98–106 [PubMed: 15015723]
6. Riffaud L, Neumuth T, Morandi X, Trantakis C, Meixensberger J, Burgert O, Trelhu B, Jannin P (2010) Recording of surgical processes: a study comparing senior and junior neurosurgeons during lumbar disc herniation surgery. Neurosurgery 67:325–332 [PubMed: 21099555]
7. Moglia A, Georgiou K, Georgiou E, Satava RM, Cuschieri A (2021) A systematic review on artificial intelligence in robot-assisted surgery. Int J Surg 95:106151 [PubMed: 34695601]
8. Anteby R, Horesh N, Soffer S, Zager Y, Barash Y, Amiel I, Rosin D, Gutman M, Klang E (2021) Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surg Endosc 35:1521–1533 [PubMed: 33398560]
9. Meireles OR, Rosman G, Altieri MS, Carin L, Hager G, Madani A, Padoy N, Pugh CM, Sylla P, Ward TM, Hashimoto DA (2021) SAGES consensus recommendations on an annotation framework for surgical video. Surg Endosc 35:4918–4929 [PubMed: 34231065]
10. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC,

Welch VA, Whiting P, McKenzie JE (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372:n160 [PubMed: 33781993]

11. Khan DZ, Luengo I, Barbarisi S, Addis C, Culshaw L, Dorward NL, Haikka P, Jain A, Kerr K, Koh CH, Layard Horsfall H, Muirhead W, Palmisciano P, Vasey B, Stoyanov D, Marcus HJ (2021) Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0). J Neurosurg. 10.1016/j.bas.2021.100580
12. Kitaguchi D, Takeshita N, Matsuzaki H, Oda T, Watanabe M, Mori K, Kobayashi E, Ito M (2020) Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research. Int J Surg 79:88–94 [PubMed: 32413503]
13. Cheng K, You J, Wu S, Chen Z, Zhou Z, Guan J, Peng B, Wang X (2022) Artificial intelligence-based automated laparoscopic cholecystectomy surgical phase recognition and analysis. Surg Endosc 36:3160–3168 [PubMed: 34231066]
14. Hashimoto DA, Rosman G, Witkowski ER, Stafford C, NavaretteWelton AJ, Rattner DW, Lillemoe KD, Rus DL, Meireles OR (2019) Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. Ann Surg 270:414–421 [PubMed: 31274652]
15. Yeh HH, Jain AM, Fox O, Wang SY (2021) PhacoTrainer: a multicenter study of deep learning for activity recognition in cataract surgical videos. Transl Vis Sci Technol 10:23
16. Derathé A, Reche F, Moreau-Gaudry A, Jannin P, Gibaud B, Voros S (2020) Predicting the quality of surgical exposure using spatial and procedural features from laparoscopic videos. Int J Comput Assist Radiol Surg 15:59–67 [PubMed: 31673963]
17. Garcia Nespolo R, Yi D, Cole E, Valikodath N, Luciano C, Leiderman YI (2022) Evaluation of artificial intelligence-based intraoperative guidance tools for phacoemulsification cataract surgery. JAMA Ophthalmol 140:170–177 [PubMed: 35024773]
18. Huaulmé A, Jannin P, Reche F, Faucheron JL, Moreau-Gaudry A, Voros S (2020) Offline identification of surgical deviations in laparoscopic rectopexy. Artif Intell Med 104:101837 [PubMed: 32499005]
19. Bodenstedt S, Rivoir D, Jenke A, Wagner M, Breucha M, Müller-Stich B, Mees ST, Weitz J, Speidel S (2019) Active learning using deep Bayesian networks for surgical workflow analysis. Int J Comput Assist Radiol Surg 14:1079–1087 [PubMed: 30968355]
20. Bodenstedt S, Wagner M, Mündermann L, Kenngott H, Müller-Stich B, Breucha M, Mees ST, Weitz J, Speidel S (2019) Prediction of laparoscopic procedure duration using unlabeled, multimodal sensor data. Int J Comput Assist Radiol Surg 14:1089–1095 [PubMed: 30968352]
21. Dergachyova O, Bouget D, Huaulmé A, Morandi X, Jannin P (2016) Automatic data-driven real-time segmentation and recognition of surgical workflow. Int J Comput Assist Radiol Surg 11:1081–1089 [PubMed: 26995598]
22. Jin Y, Li H, Dou Q, Chen H, Qin J, Fu CW, Heng PA (2020) Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med Image Anal 59:101572 [PubMed: 31639622]
23. Lecuyer G, Ragot M, Martin N, Launay L, Jannin P (2020) Assisted phase and step annotation for surgical videos. Int J Comput Assist Radiol Surg 15:673–680 [PubMed: 32040704]
24. Ramesh S, Dall'Alba D, Gonzalez C, Yu T, Mascagni P, Mutter D, Marescaux J, Fiorini P, Padoy N (2021) Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. Int J Comput Assist Radiol Surg 16:1111–1119 [PubMed: 34013464]
25. Shi X, Jin Y, Dou Q, Heng PA (2021) Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition. Med Image Anal 73:102158 [PubMed: 34325149]
26. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36:86–97 [PubMed: 27455522]
27. Twinanda AP, Yengera G, Mutter D, Marescaux J, Padoy N (2019) RSDNet: learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE Trans Med Imaging 38:1069–1078 [PubMed: 30371356]

28. Meeuwsen FC, van Luyn F, Blikkendaal MD, Jansen FW, van den Dobbelsteen JJ (2019) Surgical phase modelling in minimal invasive surgery. Surg Endosc 33:1426–1432 [PubMed: 30187202]
29. Kitaguchi D, Takeshita N, Matsuzaki H, Takano H, Owada Y, Enomoto T, Oda T, Miura H, Yamanashi T, Watanabe M, Sato D, Sugomori Y, Hara S, Ito M (2020) Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. Surg Endosc 34:4924–4931 [PubMed: 31797047]
30. Malpani A, Lea C, Chen CC, Hager GD (2016) System events: readily accessible features for surgical phase detection. Int J Comput Assist Radiol Surg 11:1201–1209 [PubMed: 27177760]
31. Guédon ACP, Meij SEP, Osman K, Kloosterman HA, van Stralen KJ, Grimbergen MCM, Eijsbouts QAJ, van den Dobbelsteen JJ, Twinanda AP (2021) Deep learning for surgical phase recognition using endoscopic videos. Surg Endosc 35:6150–6157 [PubMed: 33237461]
32. Blum T, Padoy N, Feußner H, Navab N (2008) Workflow mining for visualization and analysis of surgeries. Int J Comput Assist Radiol Surg 3:379–386
33. Quellec G, Lamard M, Cochener B, Cazuguel G (2014) Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Trans Med Imaging 33:2352–2360 [PubMed: 25055383]
34. Zhang B, Ghanem A, Simes A, Choi H, Yoo A (2021) Surgical workflow recognition with 3DCNN for sleeve gastrectomy. Int J Comput Assist Radiol Surg 16:2029–2036 [PubMed: 34415503]
35. Zhang Y, Bano S, Page AS, Deprest J, Stoyanov D, Vasconcelos F (2022) Large-scale surgical workflow segmentation for laparoscopic sacrocolpopexy. Int J Comput Assist Radiol Surg 17:467–477 [PubMed: 35050468]
36. Huaulmé A, Despinoy F, Perez SAH, Harada K, Mitsuishi M, Jannin P (2019) Automatic annotation of surgical activities using virtual reality environments. Int J Comput Assist Radiol Surg 14:1663–1671 [PubMed: 31177422]
37. Katić D, Schuck J, Wekerle AL, Kenngott H, Müller-Stich BP, Dillmann R, Speidel S (2016) Bridging the gap between formal and experience-based knowledge for context-aware laparoscopy. Int J Comput Assist Radiol Surg 11:881–888 [PubMed: 27025604]
38. Mascagni P, Alapatt D, Urade T, Vardazaryan A, Mutter D, Marescaux J, Costamagna G, Dallemagne B, Padoy N (2021) A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. Ann Surg 274:e93–e95 [PubMed: 33417329]
39. Lalys F, Bouget D, Riffaud L, Jannin P (2013) Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures. Int J Comput Assist Radiol Surg 8:39–49 [PubMed: 22528057]
40. Lalys F, Riffaud L, Bouget D, Jannin P (2012) A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Trans Biomed Eng 59:966–976 [PubMed: 22203700]
41. Mascagni P, Alapatt D, Laracca GG, Guerriero L, Spota A, Fiorillo C, Vardazaryan A, Quero G, Alfieri S, Baldari L, Cassinotti E, Boni L, Cuccurullo D, Costamagna G, Dallemagne B, Padoy N (2022) Multicentric validation of EndoDigest: a computer vision platform for video documentation of the critical view of safety in laparoscopic cholecystectomy. Surg Endosc. 10.1007/s00464-022-09112-1
42. Yu F, Silva Croso G, Kim TS, Song Z, Parker F, Hager GD, Reiter A, Vedula SS, Ali H, Sikder S (2019) Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. JAMA Netw Open 2:e191860 [PubMed: 30951163]
43. Gibaud B, Forestier G, Feldmann C, Ferrigno G, Gonçalves P, Haidegger T, Julliard C, Katić D, Kenngott H, Maier-Hein L, März K, de Momi E, Nagy D, Nakawala H, Neumann J, Neumuth T, Rojas Balderrama J, Speidel S, Wagner M, Jannin P (2018) Toward a standard ontology of surgical process models. Int J Comput Assist Radiol Surg 13:1397–1408 [PubMed: 30006820]
44. Gholinejad M, Loeve AJ, Dankelman J (2019) Surgical process modelling strategies: which method to choose for determining workflow? Minim Invasive Ther Allied Technol 28:91–104 [PubMed: 30915885]

45. Garrow CR, Kowalewski KF, Li L, Wagner M, Schmidt MW, Engelhardt S, Hashimoto DA, Kenngott HG, Bodenstedt S, Speidel S, Müller-Stich BP, Nickel F (2021) Machine learning for surgical phase recognition: a systematic review. Ann Surg 273:684–693 [PubMed: 33201088]

46. Marcus HJ, Khan DZ, Borg A, Buchfelder M, Cetas JS, Collins JW, Dorward NL, Fleseriu M, Gurnell M, Javadpour M, Jones PS, Koh CH, Layard Horsfall H, Mamelak AN, Mortini P, Muirhead W, Oyesiku NM, Schwartz TH, Sinha S, Stoyanov D, Syro LV, Tsermoulas G, Williams A, Winder MJ, Zada G, Laws ER (2021) Pituitary society expert Delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. Pituitary 24:839–853 [PubMed: 34231079]

47. Rosse C, Mejino JL Jr (2003) A reference ontology for biomedical informatics: the foundational model of anatomy. J Biomed Inform 36:478–500 [PubMed: 14759820]

48. Lomax J, McCray AT (2004) Mapping the gene ontology into the unified medical language system. Comp Funct Genomics 5:354–361 [PubMed: 18629164]

49. Grenon P, Smith B, Goldberg L (2004) Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 102:20–38 [PubMed: 15853262]

50. Moglia A, Georgiou K, Morelli L, Toutouzas K, Satava RM, Cuschieri A (2022) Breaking down the silos of artificial intelligence in surgery: glossary of terms. Surg Endosc 36:7986–7997 51. [PubMed: 35729406]

51. Protégé. https://protege.stanford.edu/

52. Huaulmé A, Dardenne G, Labbe B, Gelin M, Chesneau C, Diverrez JM, Riffaud L, Jannin P (2022) Surgical declarative knowledge learning: concept and acceptability study. Comput Assist Surg (Abingdon) 27:74–83 [PubMed: 35727207]

53. Hung AJ, Ma R, Cen S, Nguyen JH, Lei X, Wagner C (2021) Surgeon automated performance metrics as predictors of early urinary continence recovery after robotic radical prostatectomy-a prospective bi-institutional study. Eur Urol Open Sci 27:65–72 [PubMed: 33959725]

54. Ma R, Lee RS, Nguyen JH, Cowan A, Haque TF, You J, Robert SI, Cen S, Jarc A, Gill IS, Hung AJ (2022) Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills-an unblinded pilot randomized controlled trial. J Urol. 10.1097/JU.0000000000002691

55. Pangal DJ, Kugener G, Cardinal T, Lechtholz-Zey E, Collet C, Lasky S, Sundaram S, Zhu Y, Roshannai A, Chan J, Sinha A, Hung AJ, Anandkumar A, Zada G, Donoho DA (2021) Use of surgical video-based automated performance metrics to predict blood loss and success of simulated vascular injury control in neurosurgery: a pilot study. J Neurosurg 137(3):840–849. 10.3171/2021.10.JNS211064

56. Guerin S, Huaulmé A, Lavoue V, Jannin P, Timoh KN (2022) Review of automated performance metrics to assess surgical technical skills in robot-assisted laparoscopy. Surg Endosc 36:853–870 [PubMed: 34750700]

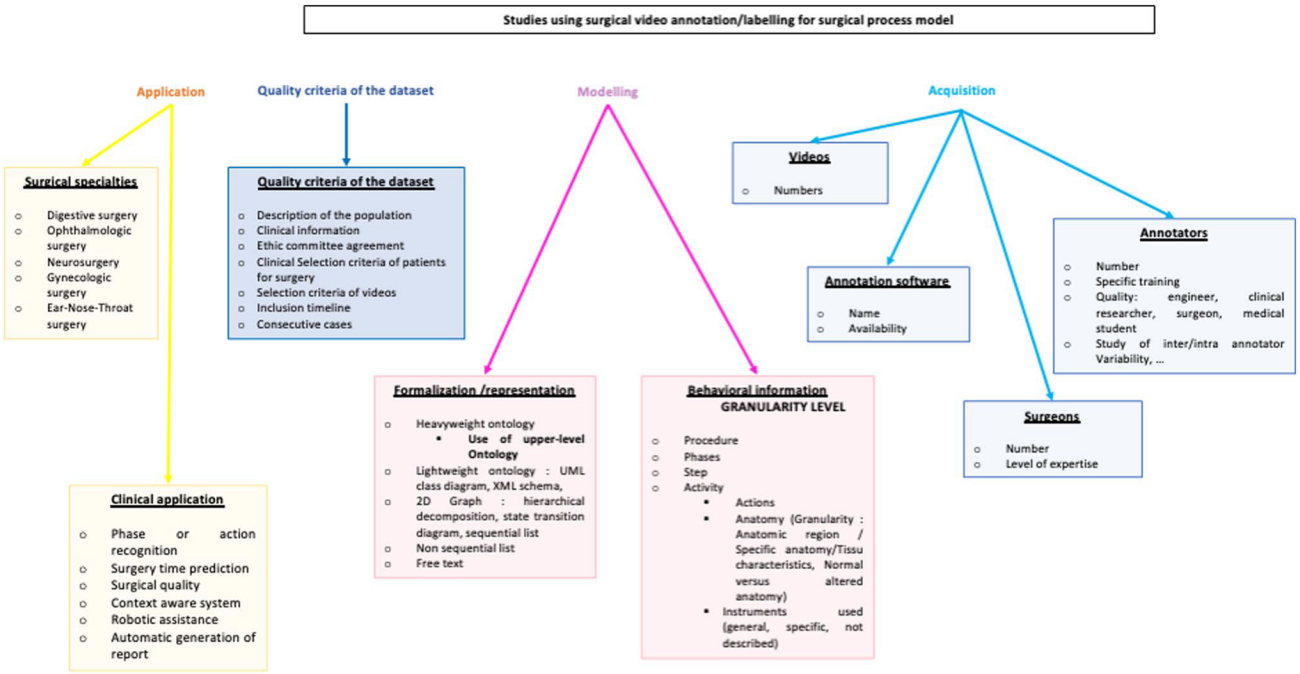

Fig. 1.
Framework of analysis of articles

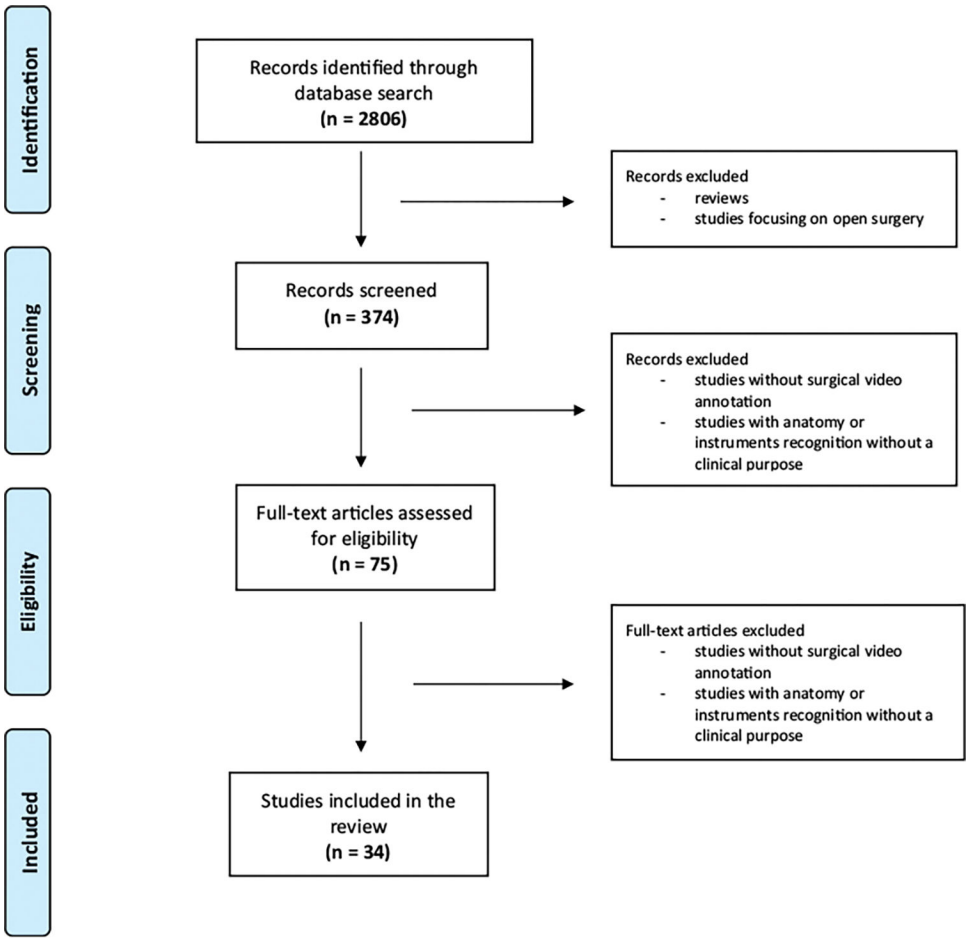

Fig. 2.
Flowchart

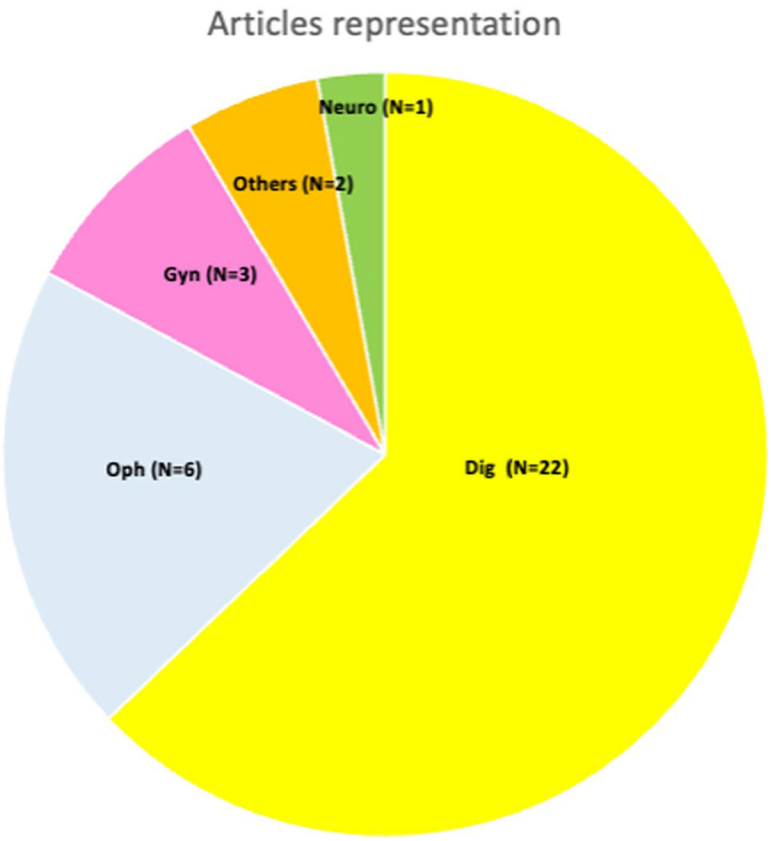

Fig. 3.
Distribution of the articles

Fig. 4.

Heterogeneity of phases for cataract surgical procedure

Table 1

Articles included in the review

Study	Sample	Phenomena of interest	Design				Evaluation	Research
			Applications	Platform	Algorithms	Input features		
Blum et al. [32]	10 surgical videos 2 surgeons and assistants	Laparoscopic cholecystectomies	Automatic generation and visualization of the workflow of surgery	Laparoscopy	HMM model merging	Video frames (instruments signals)	Generation of a graphic unit user Average duration. Probability of reaching node, probability of an instrument being used	Qualitative Quantitative
Bodenstedt et al. [19]	80 surgical videos (CHOLEC80)	Laparoscopic cholecystectomies	Surgical workflow analysis	Laparoscopy	Deep Bayesian networks	Video frames	Variance, variation ratio, entropy, mutual information, the weighted LI score, accuracy	Quantitative
Bodenstedt et al. [20]	80 surgical videos (CHOLEC80)	Laparoscopic cholecystectomies	To directly predict and refine the duration of laparoscopic interventions	Laparoscopy	recurrent CNNs	surgical device. Endoscopic image data. Both	Mean absolute error. Mean relative error	Quantitative
Cheng et al. [13]	163 surgical videos 4 medical centers 7 different surgeons with different levels	Laparoscopic cholecystectomies	Automatic surgical phases recognition	Laparoscopy	CNN LTSM	Video frames	precision, recall, LI score, and overall accuracy	Quantitative
Derathé et al. [16]	29 surgical videos Two confirmed surgeons	Laparoscopic Sleeve Gastrectomy	Automatic surgical phase recognition (to predict the exposure of the surgical scene)	Laparoscopy	Algorithm pipeline	Video frames	Accuracy, sensitivity, specificity	Quantitative
Dergachyova et al. [21]	7 surgical videos	Laparoscopic Cholecystectomies	Automatic phase recognition	Laparoscopy	Adaboost classifier Hidden semi-Markov Model	Video frames Instrument usage signals	Accuracy, Precision, recall	Quantitative
Guedon et al. [31]	33 surgical videos 5 surgeons 35 surgical videos 3 surgeons	Laparoscopic Cholecystectomies Laparoscopic hysterectomies	Automatic phase recognition	Laparoscopy	InceptionV3 network with ResNet50	Video frames	Accuracy, precision, recall	Quantitative
Hashimoto et al. [14]	88 surgical videos	Laparoscopic sleeve gastrectomies	Automatic step recognition	Laparoscopy	Sleevenet (combined ResNet and LTSM)	Video frames	Accuracy	Quantitative
Huaulmé et al. [18]	11 Surgical videos 1 surgeon expert	Laparoscopic rectopexies	Automatic detection of surgical process	Laparoscopy	Multi-dimensional non-linear temporal scaling with a hidden semi-Markov model	Video frames	Accuracy, recall, precision	Quantitative

Study	Sample	Phenomena of interest	Design				Evaluation	Research
			Applications	Platform	Algorithms	Input features		
Jalal et al. [56]	80 Surgical videos	80 Surgical videos	Automatic phase recognition	Laparoscopy	CNN and NARX	Video frames	Accuracy and precision	Quantitative
Jin et al. [22]	80 Surgical videos	80 Surgical videos	Automatic phase recognition and surgical tool detection	Laparoscopy	MTRCNet-CL	Video frames	Precision, recall, accuracy, mean average precision	Quantitative
Katie et al. [37]	16 surgical videos	16 multiple surgical procedures(11 pancreatic resections and 5 adrenalectomies) Several surgeons	Automatic phase recognition	Laparoscopy	Random forest Cultural optimization	Video frames	recognition rate, variance of the recognition rate the average time to recognize a phase	Qualitative Quantitative
Khan etal. [11]	50 surgical videos 1 attending surgeon and 1 subspecialty fellow	Endoscopic pituitary surgery	Automatic phase recognition	Endoscopy	CNN and Recurrent CNN	Video frames	Precision, recall, accuracy, LI score	Quantitative
Kitugachi et al. [12]	300 surgical videos 19 high-volume endoscopic centers Numerous different surgeons	Laparoscopic Sigmoidectomies	Automatic phase recognition	Laparoscopy	CNN	Video frames	Overall accuracies. Intersection over union	Quantitative
Kitugachi et al. [29]	71 surgical videos 19 surgeons	Laparoscopic Sigmoidectomies	Automatic phase recognition	Laparoscopy	Inception-ResNet-v2 LightGBM	Video frames	Precision, recall, LI score, and overall accuracy	Quantitative
Pangal et al. [55]	10 surgical videos	Laparoscopic Cholecystectomies	Evaluation of the reliability of real-time workflow recognition in	Laparoscopy	Intelligent Workflow analysis and Prediction software	Continuous sensor-based data acquisition	Correlation coefficient, percentage of over or under estimation	Quantitative
Lalys et al. [39]	20 surgical videos 2 experts surgeons	Cataract	automatic detection of low-level surgical tasks, that is, the sequence of activities in a surgical procedure	Microscopy	SVM DTW	Video frames	percentage of frames correctly recognized in one video within a frame-by-frame analysis, global accuracy, and recognition rates per activity pair	Quantitative
Lalys et al. [40]	20 surgical videos 3 experts surgeons	Cataract	automatic recognition of high-level surgical tasks	Microscopy	DTW HMM	Videos frames	frequency recognition rate, accuracy	Quantitative
Lecuyer et al. [23]	80 surgical videos 50 surgical videos	Laparoscopic Cholecystectomy Cataract	Automatic phase recognition	Laparoscopy	VG19, Inception V3, Resnet50	Video frames	Accuracy, Duration of annotation using assistance system compared to duration without	Quantitative
Malpani et al. [30]	24 surgical videos 6 faculty residents with more than 20 residents	Robotic hysterectomies	Automatic phase recognition	Laparoscopy robot-assisted	Semi-Markov conditional random field	Video frames Tool motion data System (console) events	Precision, recall, accuracy, the normalized Leven-shtein distance	Quantitative

Study	Sample	Phenomena of interest	Design				Evaluation	Research
			Applications	Platform	Algorithms	Input features recorded by the Da Vinci		
Mascagani et al. [38]	155 surgical videos 12 surgeons	Laparoscopic Cholecystectomies	Automatic location of critical events and provide short video clips documenting the critical view of safety	Laparoscopy	Endodigest	Video frames	Mean Absolute Error, the percentage of video clips in which the CVS was assessable by surgeons (relevance)	Quantitative
Mascagani et al. [41]	144 surgical videos 4 italian centers	Laparoscopic Cholecystectomies	Automatic phase recognition (external validation)	Laparoscopy	EndoDigest	Video frames	mean error, percentage of the automatically extracted short video clips documented CVS effectively	Quantitative
Meuwssen et al. [28]	40 surgical videos	Laparoscopic Hysterectomies	Automatic phase recognition	Laparoscopy	Random forest	Video frames	Accuracy, mean absolute error, the end time prediction	Quantitative
Nespolo et al. [17]	10 surgical videos Attending physicians Surgical trainee	Cataract	Automatic phase recognition	Microscopy	Faster R-CNN	Video frames	Area under the receiver operating characteristic curve, mean proceeding speed	Quantitative
Guerin et al. [56]	16 surgical videos 4 surgeons of various skills	Laparoscopic Cholecystectomies	Automatic phase recognition	Laparoscopy	DTW Hidden Markov	Videos from endoscopic views and two external views	accuracy, average recall, and average precision	Quantitative
Quellec et al. [33]	186 surgical videos 10 different surgeons	Cataract	Automatic phase recognition	Microscopy	Conditional random fields Unary potential	Video frames	Accuracy, area under the receiver operating characteristic (ROC) curve	Quantitative
Ramesh et al. [24]	40 surgical videos 7 experts surgeons	Bypass 40	Automatic phase and step recognition	Laparoscopy	MTMS-TCN CNN	Video frames	accuracy, precision, and recall	Quantitative
Shi et al. [25]	80 surgical videos 41 surgical videos	Laparoscopic cholecystectomies	Automatic phase recognition	Laparoscopy	Two-stage Semi-Supervised Learning	Video frames	accuracy, precision and recall, jaccard	Quantitative
Twinanda et al. [26]	80 surgical videos 7 surgical videos	Laparoscopic cholecystectomies	Automatic phase recognition and tool detection	Laparoscopy	Endonet	Video frames	Average precision, average recall, accuracy.	Quantitative
Twinanda et al. [27]	120 surgical videos 170 surgical videos	Laparoscopic cholecystectomies Laparoscopic bypass	Automatic estimation of the remaining surgery duration	Laparoscopy	RSDNet	Video frames	Mean absolute error, overestimation, underestimation	Quantitative
Yeh et al. [15]	298 surgical videos 12 resident surgeons	Cataract	Automatic phase/ step recognition	Microscopy	VGG16 VGG16 followed by CNN and RCNN	Video frames	Accuracy, Micro-averaged area under receiver operating characteristic curves	Quantitative

Study	Sample	Phenomena of interest	Design				Evaluation	Research
			Applications	Platform	Algorithms	Input features		
Yu et al. [42]	100 surgical videos 1 Faculty and 1 trainee surgeons	Cataract	Automatic phase recognition	Microscopy	1) SVM 2) a recurrent neural network (RNN) input 3) a CNN 4) a CNN-RNN input with a time series of images; and (5) a CNN-RNN input with time series of images and instrument labels	Video frames	Accuracy, area under the receiver operating characteristic curve, sensitivity, specificity, and precision	Quantitative
Zhang et al. [35]	14 surgical videos A group of surgeons CholecSO	Laparoscopic sacrocol-popexies	Automatic phase recognition	Laparoscopy	seq2seq LTSM and transformers	Video frames	accuracy, FI score. Ward metric	Quantitative
Zhang et al. [34]	461 surgical videos	Robotic laparoscopic sleeve gastrectomies	Automatic phase recognition	Robotic and laparoscopy	Inflated 3D ConvNet (I3D)	Video frames	accuracy, precision, and recall weighted jaccard score	Quantitative

Table 2

Available public datasets and articles related to this current review

Datasets	Field	Interventions	Number of Videos	Studies	Application
CHOLEC80http://camma.u-strasbg.fr/datasets	Digestive surgery	Laparoscopic cholecystectomy	80	Shi et al. [25]	Surgical workflow recognition
				Jin et al. [22]	Tool presence detection and phase recognition
				Twinanda et al. [26]	Tool presence detection and phase recognition
				Bodenstedt et al. [19]	Surgical workflow recognition
				Bodenstedt et al. [20]	Procedure duration prediction
				Jalal et al. [56]	Surgical workflow recognition
				Lecuyer et al. [23]	Surgical workflow recognition
CHOLEC120 = CHOLEC80 + 40 cholecystectomies https://github.com/CAMMA-public/Surgical-Phase-Recognition	Digestive surgery	Laparoscopic cholecystectomy	120	Twinanda et al. [27]	Prediction of remaining surgery duration
CATARACT 101 http://ftp.itec.aau.at/datasets/ovid/cat-101/	Ophthalmology surgery	Cataract	101	Nespolo et al. [17]	Surgical workflow recognition
CATARACT https://arxiv.org/abs/1906.11586	Ophthalmology surgery	Cataract	50	Lecuyer et al. [23]	Surgical workflow recognition

Table 3

Annotation software used in the articles

Annotation Software	Academic or corporate	Studies
Nous (COSMONiO) ®	Private	Guédon, et al. [31]
Annotate ®	Private	Derathé et al. [16] Huaulmé et al. [18] Lecuyer et al. [23]
Anvil research tool annotation software ®	Public	Cheng et al. [13] Hashimoto et al. [14]
Swansuite ®	Private	Katic et al. [37]
Touchsurgery ®	Private	Khan et al. [11]
Via Software ®	Private	Yeh et al. [15]

Table 4

Experience of annotators

Study	Level of experience of annotators
Blum et al. [32]	Not mentioned
Bodenstedt et al. [19]	Not mentioned
Bodenstedt et al. [20]	Not mentioned
Cheng et al. [13]	surgeons
Derathé et al. [16]	Surgeons and scientists
Dergachyova et al. [21]	Not mentioned
Guedon et al. [31]	Instructed students and author
	Instructed students and author
Hashimoto et al. [14]	surgeons
Huaulmé et al. [18]	Scientist
Jalal et al. [56]	Not mentioned
Jin et al. [22]	Not mentioned
Katic et al. [37]	Not mentioned
Khan et al. [11]	surgeons
Kitugachi et al. [12]	surgeons
Kitugachi et al. [29]	Not mentioned
Pangal et al. [55]	Not mentioned
Lalys et al. [39]	Not mentioned
Lalys et al. [40]	Surgeon
Lecuyer et al. [23]	Surgeon
	Surgeon 1
Malpani et al. [30]	Not mentioned
Mascagani et al. [38]	Not mentioned
Mascagani et al. [41]	Not mentioned
Meuwssen et al. [28]	Not mentioned
Nespolo et al. [17]	surgeons
Guerin et al. [56]	Not mentioned
Quellec et al. [33]	surgeons
Ramesh et al. [24]	Not mentioned
Shi et al. [25]	Not mentioned
	Not mentioned
Twinanda et al. [26]	Not mentioned
	Not mentioned
Twinanda et al. [27]	Not mentioned
	Not mentioned
Yeh et al. [15]	Non-clinical researcher
Yu et al. [42]	Physicians and trained annotators
Zhang et al. [35]	surgeons
Zhang et al. [34]	Not mentioned