Contents lists available at ScienceDirect

European Journal of Surgical Oncology

journal homepage: www.ejso.com

Surgical workflow analysis for Surgomics and context-aware assistance in robot-assisted minimally invasive esophagectomy (RAMIE): a retrospective, single-arm, multicenter annotation and machine learning study

Johanna M. Brandenburg [a,b,c,d], André Schulze [b,c,d], Alexander C. Jenke [c,e], Nithya Bhasker [c,e], Noelle Bleser [b,c], Denise Junger [f], Antonia Stern [g], Dominik Rivoir [c,d,e], Hamid Naderi [h], Fleur Fritz-Kebede [h], Oliver Burgert [f], Lena Maier-Hein [i], Lars Mündermann [g], Sebastian Bodenstedt [c,d,e], Stefanie Speidel [c,d,e], Vladimir J. Lozanovski [j], Peter P. Grimminger [j], Adrian Billeter [k], Rosa Klotz [a,l], Jürgen Weitz [b,c,d], Marius Distler [b,c,d], Beat P. Müller-Stich [k], Martin Wagner [b,c,d,*]

[a] *Department of General, Visceral, and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany*
[b] *Department of Visceral, Thoracic, and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany*
[c] *National Center for Tumor Diseases (NCT/UCC), Dresden, Germany*
[d] *Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, Dresden, Germany*
[e] *Department of Translational Surgical Oncology, National Center for Tumor Diseases (NCT), NCT/UCC Dresden, a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany*
[f] *Reutlingen University, School of Informatics, Research Group Computer Assisted Medicine (CaMed), Reutlingen, Germany*
[g] *Advanced Technology, Karl Storz SE & Co KG, Tuttlingen, Germany*
[h] *Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany*
[i] *Department of Intelligent Medical Systems (IMSY), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[j] *Department of General-, Visceral- and Transplant Surgery, University Medical Center of the Johannes Gutenberg University, Mainz, Germany*
[k] *University Digestive Healthcare Center Basel, Basel, Switzerland*
[l] *The Study Center of the German Surgical Society (SDGC), Heidelberg University Hospital, Heidelberg, Germany*

ARTICLE INFO

Keywords:
Artificial intelligence
Robot-assisted surgery
Precision medicine
Machine learning
Surgical data science
Surgomics

ABSTRACT

Introduction: Robot-assisted minimally invasive esophagectomy (RAMIE) is a complex procedure that may benefit from workflow analysis for context-aware assistance and surgical data science. This study aimed to model the RAMIE workflow, validate the applicability of the obtained workflow model in the operating room (OR) and retrospectively assess its generalizability across three academic centers using video data and automated workflow analysis with machine learning (ML).

Methods: A RAMIE workflow model was developed based on currently available literature, participatory OR observation, and expert interviews. This model was formalized to be included into a checklist tool to document the workflow live in the OR. To investigate generalizability of the workflow model, the surgical phases of 36 RAMIE videos from three different academic hospitals were retrospectively annotated. Based on this data set, a ML model was trained and tested within a six-fold cross validation.

Results: Ten surgical phases with 60 underlying steps were identified for RAMIE. The applicability of the workflow model was validated with live documentation in the OR. Multicenter video annotations revealed significant inter-institutional differences in the duration of all ten RAMIE phases. The ML model for automatic phase recognition showed an accuracy of 0.872 ± 0.091 and an f1-score of 0.872 ± 0.082 over all videos. The center with the best performing videos achieved a mean accuracy of 0.919 ± 0.036.

* Corresponding author. Department of Visceral, Thoracic, and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307, Dresden, Germany.

E-mail address: martin.wagner@ukdd.de (M. Wagner).

https://doi.org/10.1016/j.ejso.2025.111174
Received 10 August 2025; Received in revised form 4 November 2025; Accepted 6 November 2025
Available online 7 November 2025

Conclusion: The RAMIE workflow was successfully modeled and validated in a retrospective multicenter setting. Despite high variability in phase duration between surgical centers, ML-based phase recognition achieved highly promising results.

1. Introduction

The success of a complex surgical procedure depends on technical abilities such as surgical skill [1], but also on precise coordination and communication among surgical team members. Therefore, these interactions, as well as the operating room (OR) itself as a high-stake environment, require thorough investigation [2,3]. Surgical workflow analysis can enhance efficiency and safety in the OR and offers the potential to provide context-aware real-time information during the procedure. This is particularly beneficial in cases of frequent changes in team constellation or complex patient cases, such as those encountered in academic hospitals. The concept of Surgomics [3] aims to quantify and analyze intraoperative processes by automatically deriving surgomic features as procedure characteristics, such as the amount of blood in the surgical field, the pattern of instrument usage, or the duration of certain surgical phases, with the help of machine learning (ML) [4]. Thus, to automate the aspired workflow analysis, ML is applied by means of surgical data science methods [5]. These methods can process and analyze data from a variety of sources in the OR, such as videos of minimally invasive procedures or medical devices, and then enable an automatic identification and tracking of different surgical phases and steps. Several approaches that use ML for automatic phase recognition have already been employed [6–8]. Monitoring the progress of a procedure and providing real-time, context-aware information [9] can result in workflow optimization, reduced delay, automated performance assessment, and perhaps even improved patient outcomes [10]. A prerequisite for ML-based analysis is to establish a standardized workflow model of surgical phases and steps to facilitate consistent data annotation, particularly considering the variability across surgeons, techniques, and surgical centers. ML algorithms must therefore be trained and tested using data from multiple surgical centers [11]. Robot-assisted minimally invasive esophagectomy (RAMIE) is a particularly complex surgical procedure with a potentially high risk of postoperative complications, such as anastomotic leak and conduit necrosis, as well as a considerable risk of death after surgical treatment [12]. First approaches regarding standardization of the RAMIE procedure from a surgical perspective have already been taken [13]. The procedure thus offers the potential for workflow-based assistance. Furthermore, models of the RAMIE workflow [14,15] as well as an automatic phase recognition for the thoracic part [16] or the abdominal part [17] of RAMIE have been introduced. However, no comprehensive ML-based analysis of the whole procedure has been performed, nor has it been compared between different surgical centers.

In the present multicenter comparative study, we aimed to leverage ML for a surgical procedure of high complexity and variability, which is a challenge from both the surgical and the data science perspective. We employed the following approaches:

1. We modeled the RAMIE workflow considering existing literature [13,14]. We then validated the applicability of this surgical workflow model live in the OR.
2. In a next step, we validated the generalizability of the workflow model phases by applying them to videos from three different clinical centers with potentially different workflows.
3. We trained a multi-stage ML model and validated it on the annotated videos to investigate whether an ML model can automatically recognize the defined phases in a multicentric data set.

2. Methods

In this manuscript, the term '(RAMIE) workflow model' refers to the formal explicit specification of the RAMIE procedure, while the term 'ML model' denotes the machine learning algorithm trained for automatic phase recognition. Our methodological approach followed the applicable parts of the DECIDE-AI reporting guideline for early-stage clinical evaluation of artificial intelligence (AI)-driven decision support systems [18], with particular emphasis on comprehensive data description, precise ML model training implementation, validation, and consideration of clinical implications and limitations. An overview of the study is depicted in the graphical abstract.

2.1. RAMIE workflow model

Based on the previous work of Egberts et al. [13] and Harris et al. [14] as well as on participatory observation in the OR of Heidelberg (HD), a comprehensive workflow model of the RAMIE procedure was developed (Appendix A.1) according to the "Hierarchy of events for segmentation of events in surgical video" [19]. The phases and steps of the workflow model were iteratively formalized and expanded through semi-structured interviews with an experienced esophageal surgeon with several years of expertise in RAMIE (B.P.M.-S.) and a board-certified surgeon undergoing substantial training in RAMIE (A. B.). The procedure consists of two parts, the thoracic and the abdominal part, which in between require closure of the abdominal trocar wounds, undocking of the robot, repositioning of the patient, and robotic redocking. For workflow analysis, the procedure was then partitioned into ten phases, five in the abdominal and five in the thoracic part (Appendix A.1). In the next step, specialized declarative knowledge about the procedure, such as numbering of respective lymph nodes stations [20], components of the intervention [21], and mandatory vs. optional steps were added. Context-specific procedural knowledge including images for potential assistance for specific steps such as port placement or gastric tube construction was identified (Appendix A.1). To validate the applicability of our workflow model in the OR of HD, a previously developed checklist tool [22] was expanded to enable tailored support [23]. The surgical workflow model was translated from a spreadsheet into a formal notation in Business Model and Notation (BPMN) serving as the configuration file for the checklist tool and allowing for decision branches [24]. The additional procedural knowledge was accessible via icons displayed next to the corresponding checkbox, which revealed the information when clicked. The checklist tool was used by surgeons and medical students in the OR to validate the workflow model and assess practicability in a clinical setting (Fig. 1).

2.2. Multicentric workflow model validation

For the multicentric workflow model validation, videos of RAMIE procedures of 36 patients were included in the final data set, 12 each from three academic surgical centers in Germany including HD, Dresden (DD), and Mainz (MZ). Ethics approval was granted by Heidelberg University under number S-248/2021, by Technical University Dresden under number BO-EK-177032021, and by University Medicine Mainz under number 2022–16815. For retrospectively collected data written informed consent was not necessary, because no change in treatment occurred and the anonymity of the patient was ensured. For prospectively collected data, written informed consent was provided by all patients. The patients underwent RAMIE as the primary treatment for esophageal cancer.

To validate the workflow model on multicentric data, retrospective manual reference annotation of surgical video phases was performed (Fig. 2). An annotation protocol (Table 1) was designed in collaboration with a board-certified surgeon (M.W.), based on the previously described RAMIE workflow model (Appendix A.1), and included a detailed description of the start point of each phase. These start points were defined according to clearly distinguishable surgical aspects of the workflow and at the same time considering the aim of providing ML algorithms with recognizable structures or actions in the video. The start point of the following phase automatically marks the end point of the current phase, thus avoiding undefined sections. Retrospective annotation of all 36 videos was performed manually by two independent medical students with specific surgical knowledge about the RAMIE procedure (J.M.B. and N.B.), taking all available frames of the video into account. Disagreement in the frame-based annotation was solved in discussion of the two annotators. The patient repositioning between the abdominal and thoracic part was also annotated, however excluded from training and analysis due to center-specific variations: DD and MZ left out the repositioning in most of the procedure videos, whereas HD kept that part.

2.3. Statistical analysis

Distribution of the phase durations was calculated using the median and interquartile range (IQR) of the reference annotations. Some phases were not captured on video due to truncation or technical limitations. Therefore, only phases with a recorded duration were included in this analysis. The calculation in minutes was performed after extracting one frame per second (fps, i.e., individual still images from the video stream), accounting for the varying fps values of each video. Phase differences between the three centers were assessed with the Kruskal-Wallis test on the reference annotations [25]. P-values were adjusted through post-hoc Dunn-Bonferroni pairwise comparisons [26,27]. Differences with a p-value below 0.05 were considered significant. Inter-rater agreement was assessed by using the Fleiss' Kappa score on all available annotations [28].

2.4. Workflow analysis with machine learning

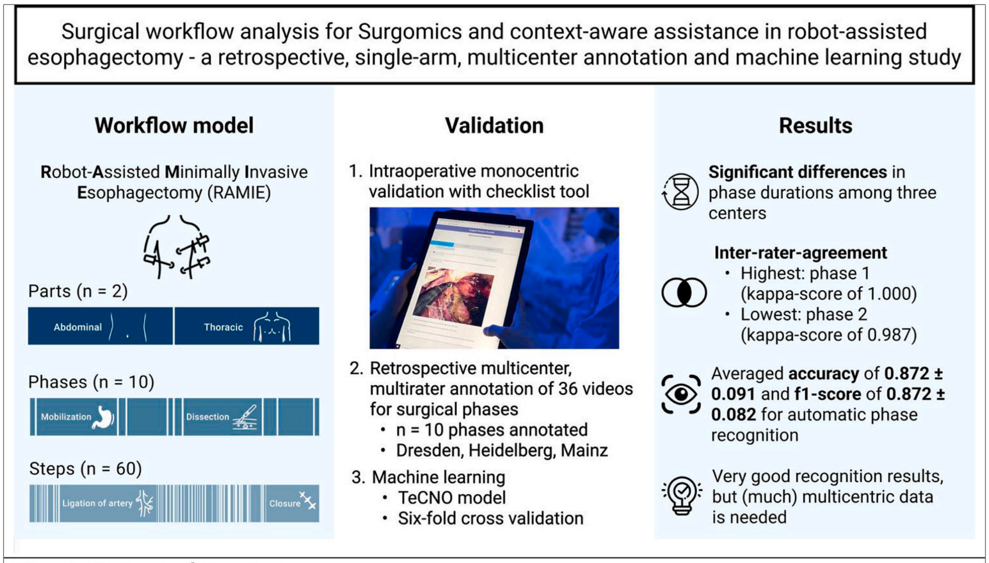
To automate surgical workflow analysis based on our workflow model, we trained and tested a ML model. The Temporal Convolutional Network for the Operating room (TeCNO) [29] is a multi-stage learning approach, which was used for training. TeCNO was trained with a ResNet50 and two stages, using the similar hyperparameter settings as in the publication by Rivoir et al. [30], which were determined prior to training and remained unchanged throughout. The implementation was done using PyTorch v2.1.0 and training was done using an NVIDIA RTX A5000. The 36 original videos (twelve from each center) were down-sampled to one fps and fed into the ML model together with the reference annotations of surgical phases. We employed six-fold cross-validation on our dataset to assess our ML model's performance. During each iteration of the cross-validation process, one fold (consisting of six videos) was used as the test set, while the remaining five folds (consisting of 30 videos) were used for training the ML model. This process was repeated six times so that each fold acted as the test set once. The respective test set of each fold consisted of two videos from each center (six videos in total) to ensure balanced representation. Cross-validation allowed us to obtain a more accurate estimate of the performance by minimizing the impact of potential outliers in the multiple test sets created within the cross-validation process. To maximize the use of available data for the cross-validation, there was no additional external test set used.

The performance of the ML model was evaluated by comparing the predictions to the manual reference annotations. The evaluation metrics accuracy, precision, recall, and f1-score were computed over the collection of all test-time predictions, providing a comprehensive assessment of our ML model's performance. The standard deviation for each metric was then calculated across the individual test videos, not across the folds, providing insight into performance variability at the video level. Additionally, 95 % confidence intervals for performance metrics were estimated using hierarchical bootstrapping [31]. Finally, the overlap score was calculated providing insight into how accurately the ML model can predict the exact timing and duration of different phases. The calculation, as described in the work of Lea et al., was performed accordingly [32]. The code is publicly available at https://gitlab.com/nct_tso_public/ramie-workflow.

3. Results

3.1. RAMIE workflow model

A total of 10 surgical phases (Table 1) with 60 underlying steps were identified (Appendix A.1). Of the 60 steps, 57 were mandatory and three



Surgical workflow analysis for Surgomics and context-aware assistance in robot-assisted esophagectomy - a retrospective, single-arm, multicenter annotation and machine learning study

Workflow model

Robot-Assisted Minimally Invasive Esophagectomy (RAMIE)

Parts (n = 2)

Phases (n = 10)

Steps (n = 60)

Validation

1. Intraoperative monocentric validation with checklist tool
2. Retrospective multicenter, multirater annotation of 36 videos for surgical phases
   - n = 10 phases annotated
   - Dresden, Heidelberg, Mainz
3. Machine learning
   - TeCNO model
   - Six-fold cross validation

Results

Significant differences in phase durations among three centers

Inter-rater-agreement
- Highest: phase 1 (kappa-score of 1.000)
- Lowest: phase 2 (kappa-score of 0.987)

Averaged **accuracy** of **0.872 ± 0.091** and **f1-score** of **0.872 ± 0.082** for automatic phase recognition

Very good recognition results, but (much) multicentric data is needed

Visual abstract of the study

optional (detachment of abdominal adhesions, cholecystectomy, drainage insertion in phase 5). Furthermore, of the 60 steps eight steps were identified as not visible in the surgical video and four steps as only sometimes visible in the surgical video. For 18 steps, context-specific procedural knowledge such as instructions on how to perform the step, numbering of respective lymph node stations [20], warnings, images, or information on trocar positioning was provided (Appendix A.1).

3.2. *Intraoperative live validation*

Using the modified checklist tool, ten surgeries were documented live in the OR (Fig. 1). For the last two surgeries the abdominal part only was documented, because the thoracic part was performed with an open approach. Based on the RAMIE BPMN workflow model, the checklist tool displays the surgical steps as checklist items. When documenting live, it was not always possible to finish the checklist completely due to OR member misunderstandings, a restricted view of the surgical field, or the need for the checklist-operator to assist the surgeon for a certain period. The surgical phases of five procedures were fully documented using the checklist tool, and the median surgery duration for these cases was calculated to be 438.32 min (IQR 382.8–582.95). An overview with the results of the phase documentation with the checklist can be found in Table 2.

3.3. *Multicenter comparative workflow validation*

3.3.1. *Phase length and frequency*

The ten surgical phases of all 36 videos were annotated by two independent medical students. Disagreement was solved in discussion. Overall, the 36 videos showed a median procedure duration of 284.58 min (IQR 208.05–361.32), with HD having the longest median duration (374.45 min, IQR 345.85–412.84) and MZ having the shortest (195.41 min, IQR 163.68–206.97). The longest phase within all centers was phase 7 “Dissection of esophagus with regional lymphadenectomy” in the thoracic part with a median phase duration of 72.14 min (IQR 61.33–88.5). The shortest phase within all centers was phase 5 “Completion of Abdominal Part” with 1.97 min (IQR 1.17–3.81). A detailed overview of the phase length of all centers is shown in Table 2.

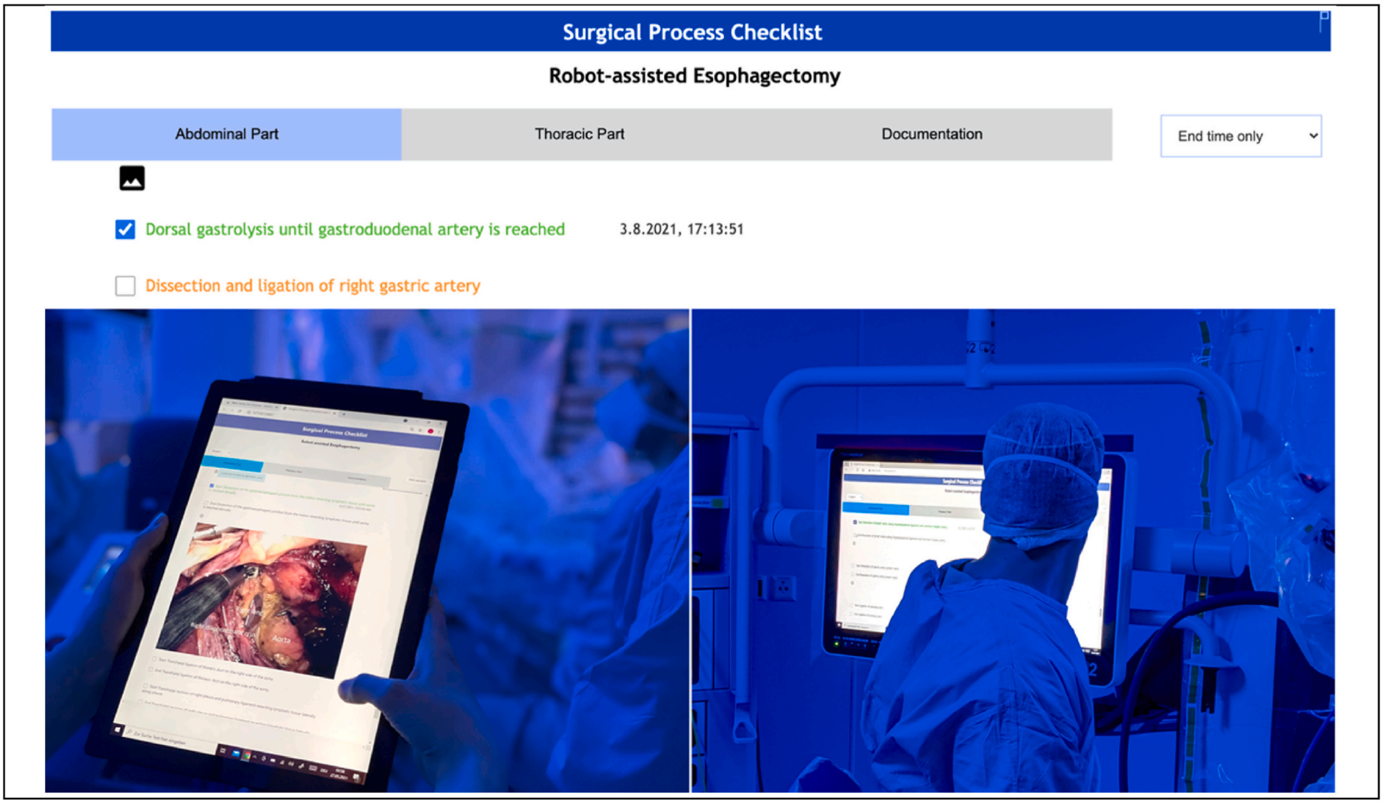
There were significant differences between the lengths of all phases across the three centers. Overall, MZ videos were significantly shorter compared to HD ($p < 0.001$) and DD ($p = 0.004$) videos (Table 2). The surgical phases with the most interruptions by another phase were phase 2 “Gastric and distal esophagus mobilization” (up to 4 occurrences) and phase 4 “Completion of abdominal lymphadenectomy” (up to 3 occurrences) (Fig. 2). This switching between phases was especially observed in videos from DD.

3.3.2. *Inter-rater-agreement*

Prior to extracting one fps for phase duration calculations and ML training, every video frame was annotated by two independent annotators. The inter-rater-agreement was evaluated using the Fleiss’ Kappa score. The phases with the lowest annotation agreement were phase 2 “Gastric and distal esophageal mobilization” (Kappa-score of 0.987) and phase 4 “Completion of abdominal lymphadenectomy” (Kappa-score of 0.985). The phase with the highest agreement was phase 1 with a Kappa-score of 1.000. The center with the lowest annotation agreement in the videos was DD with a Kappa score of 0.984. HD had the highest agreement with a Kappa-score of 1.000.

3.4. *Workflow analysis with machine learning*

Although high variability in phase duration between the surgical centers was observed, suggesting the need for more multicentric data, the performance of the TeCNO ML model showed an accuracy of 0.872



Surgical Process Checklist

Robot-assisted Esophagectomy

☑ Dorsal gastrolysis until gastroduodenal artery is reached 3.8.2021, 17:13:51

☐ Dissection and ligation of right gastric artery

Fig. 1. Live validation of the robot-assisted minimally invasive esophagectomy workflow model in the operating room with the modified checklist tool. An extract of the checklist tool itself is displayed in the upper part of the figure. The use of the tool live in the operating room by an assistant outside the sterile field (left) or the bedside assisting surgeon (right) is shown in the lower part.

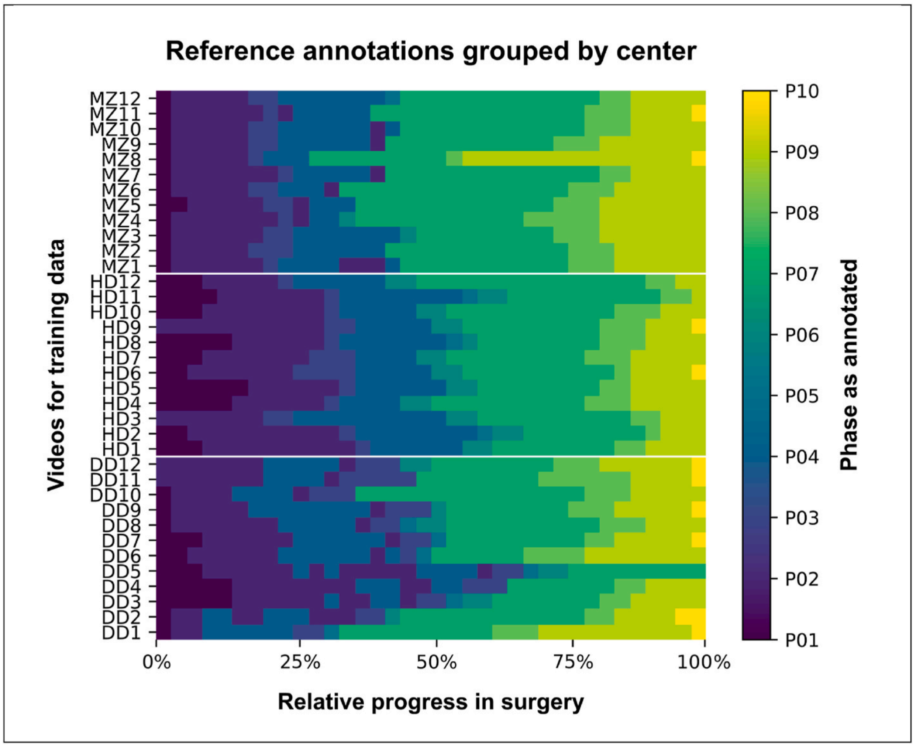


Fig. 2. Overview of reference annotations grouped by center. On the left y-axis train sets videos (n = 36) and on the x-axis relative progress of surgery [%] with a stream of phases are depicted with a specific color for each phase (P01-P10). DD = Dresden, HD = Heidelberg, MZ = Mainz.

± 0.091 [0.840, 0.900], an f1-score of 0.872 ± 0.082 [0.844, 0.899], and a balanced accuracy (each phase equally weighted, irrespective of duration) of 0.771 ± 0.110, averaged across the 36 videos from all three centers. The surgical videos of MZ showed the best results with a mean accuracy of 0.919 ± 0.036 [0.899, 0.937] and a mean f1-score of 0.919 ± 0.035 [0.900, 0.936]. The mean overlap score for all videos was 0.560 ± 0.113 with MZ videos achieving the best results (0.602 ± 0.071). Further performance metrics like precision and recall as well as differences in performance for each center are provided in Appendix A.2. The phase with the highest mean accuracy was phase 7 “Dissection of esophagus with regional lymphadenectomy” with 0.958 ± 0.070. The phase with the lowest mean accuracy was phase 5 “Completion of abdominal part” with 0.394 ± 0.295. A distribution of the accuracy per phase and center is shown in Appendix A.3. Normalized confusion matrices for each center, as well as for all videos combined, show the true positive rate of classified frames for each phase along the diagonal (Fig. 3). To visualize the predictive accuracy of the TeCNO ML model, the stream of phases for the three best performing and the three worst performing videos (one of each center) are shown in Fig. 4.

4. Discussion

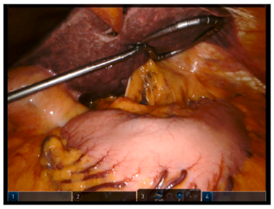
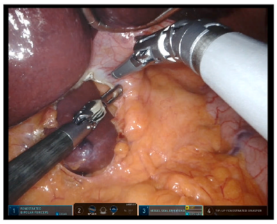
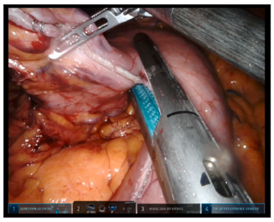
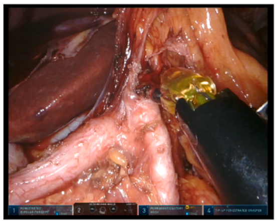
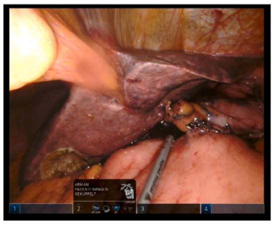
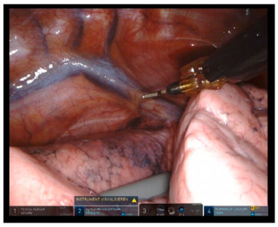
4.1. Towards context-aware assistance

Our workflow model, based on prior RAMIE/esophagectomy workflow models [13,14], includes ten surgical phases and 60 steps. The workflow model was integrated into a checklist tool for intraoperative live documentation, demonstrating its feasibility even under challenging OR conditions. The implementation of live workflow documentation via a structured checklist represents a promising strategy to bridge the gap between procedural safety tools and routine surgical practice, as checklists have been shown to reduce adverse events in surgery and significantly improve the quality of care [24,33,34]. The developed checklist tool thus transforms the workflow model from an analytical concept into an actively used intraoperative tool, addressing a major limiting factor in current practice [24,34]. Our developed checklist tool captures steps not visible in the surgical video, offering a more comprehensive view of procedure duration, including pre- and post-laparoscopy phases.

Additional context-specific information was provided for corresponding checklist items, including detailed instructions, warnings, illustrative images, and numbering of respective lymph node stations [20]. In this way, the checklist extends its role beyond task documentation to become a comprehensive reference for intraoperative guidance and educational purposes. Looking ahead, live workflow documentation with additional procedural knowledge also establishes the foundation for intelligent digital assistance. We have already developed an ML model for automatic surgical phase recognition, and the next step will be the automatic identification of individual surgical steps with direct linkage to the checklist. This would enable context-aware support in the OR, where relevant procedural knowledge is displayed automatically and potentially tailored to individual team members [35]. In future work we aim to supplement as well as validate the context-specific procedural knowledge as a basis for an AI-driven intraoperative assistance system. To complement the existing insights into user-specific information needs, conducting semi-structured interviews with additional OR team members, such as anesthesiologists and scrub nurses, is further necessary to capture perspectives shaped by their individual workflows, goals, and tasks [36,37]. Moreover, to support future automatic recognition, additional data sources such as microphones or external cameras could be utilized [2].

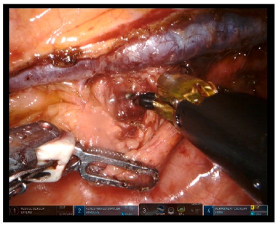
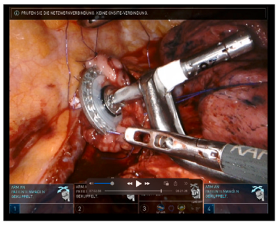
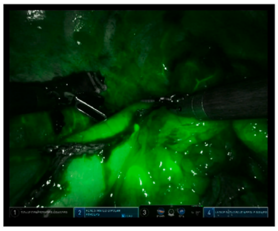
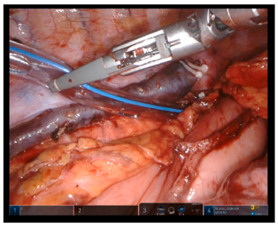
The considerable variability in surgical workflows poses a major challenge for intraoperative live documentation. It remains to be determined how well the proposed RAMIE workflow model from the HD OR can be transferred to other centers and whether adaptations will be necessary. Accordingly, future work should focus on validating the checklist both at the level of individual steps and across different surgical centers. Moreover, research should investigate potential correlations between workflow patterns, team experience, and intra- and postoperative outcomes. Such analyses would provide valuable evidence

Table 1
Annotation protocol
The ten surgical phases of RAMIE are presented with a definition of the start point of the respective phase as well as an example frame.

No.	Phase name	Definition "Start of phase"	Example frame
Abdominal part			
1	Preparation for abdominal part	Starts when the camera is initially inserted into the trocar for the abdominal part (or, if unavailable, at the first video signal in the abdomen) and includes the initial adhesiolysis. If a cholecystectomy is performed at the beginning of the surgery, this step is also considered part of the "Preparation for abdominal part" phase. When performed later during surgery, the cholecystectomy may also be part of phase 2, 3, or 4.	
2	Gastric and distal esophagus mobilization	Starts with the first incision in the lesser omentum. If the initial incision is not made in the lesser omentum, the first incision in tissue followed by gastric and esophageal mobilization is defined as the start of phase 2. Should phases 3/4 commence before the completion of phase 2, phase 2 continues with the first incision made in the mobilization area after phases 3/4.	
3	Gastric tube construction	Starts when the stapler for the gastric tube construction is visible for the first time. Can include sutures for the gastric tube and ICG perfusion check as well as sutures of the greater omentum or hiatus.	
4	Completion of abdominal lymphadenectomy	Begins with the first incision in the tissue along the major branches of the celiac trunk. This phase may be interrupted to complete phase 2 or perform phase 3. It then continues with the first cut made in the lymphatic tissue along the major branches of the celiac trunk or the profound tissue to reach the aorta after completing phase 2. Phase 4 further includes coronary vein ligation and ligation of the left gastric artery.	
5	Completion of abdominal part	Begins upon completion of phase 4 (typically after lymphadenectomy along the pericardium) or phase 3, usually with an overview of the abdomen. This phase may include irrigation and haemostasis, insertion of drainage, final inspection of the gastric tube, and robot undocking.	
Thoracic part			
6	Preparation for thoracic part	Starts when the camera is initially inserted into the trocar for the thoracic part (or, if unavailable, at the first video signal in the thoracic cavity).	

(continued on next page)

Table 1 (*continued*)

No.	Phase name	Definition "Start of phase"	Example frame
7	Dissection of esophagus with regional lymphadenectomy	Starts with the first incision in the tissue around the azygos vein or azygos vein arcus. If the initial incision is not made in this area, the phase begins with the first cut made for thoracic tissue mobilization.	
8	Insertion of anvil in the esophageal stump and stomach pull-up	Starts when the first robotic instrument is undocked, followed by the remaining instruments (second undocking during surgery). If the camera is outside the patient prior to undocking, the phase commences with the last frame in the trocar.	
9	Circular stapled anastomosis	Starts after the final pull on the stomach in the video, moving the stomach outside the body for stapler insertion.	
10	Completion of thoracic part	Starts after the last suture over the anastomosis is set tight or the last suture is cut off. If no oversewing is performed, the phase begins with the first frame after the ICG perfusion check.	

of the clinical impact of workflow quality and further emphasize the potential benefits of real-time checklist support.

Despite these challenges, integrating automatic AI-based phase and step recognition with an intraoperative checklist tool could ultimately combine the strengths of structured checklists and AI-driven workflow analysis, automating checklist completion and removing the need for manual documentation. Beyond enhancing safety and efficiency, this approach offers opportunities to compare surgical techniques and analyze procedural approaches.

4.2. Towards Surgomics

Automated phase recognition represents an established field in surgical data science [38]. Prior work has demonstrated that ML models can drive surgical workflows towards greater standardization, efficiency, and objectiveness [10]. Such progress is particularly relevant for minimally invasive esophagectomy, where high variability in technique and execution has been documented [39]. However, few studies currently exist on AI-based automatic phase recognition for robotic surgery [16]. Examples of phase recognition approaches in minimally invasive gastrointestinal procedures include laparoscopic cholecystectomy [40], sleeve gastrectomy [6], peroral endoscopic myotomy [7], Roux-en-Y gastric bypass surgery [11], the thoracic part of RAMIE [16], and the abdominal part of RAMIE [17], most of which were conducted on monocentric datasets.

In this context, the present study developed an automatic ML-based phase recognition model for RAMIE based on 36 videos from three centers. The model achieved very promising results with a mean accuracy of 0.872 ± 0.091 [0.840, 0.900] over all videos. Given the high variability in surgical workflows, the use of multi-center datasets is crucial, improving model generalizability across diverse patient populations and procedural variations [11].

Beyond standardization, automatic phase recognition provides the foundation for AI-based surgical quality assessment at the level of specific phases or critical steps [14]. Furthermore, when combined with instrument tracking, workflow recognition may enable objective assessment of surgical proficiency and learning curves [16].

Overall, the developed ML-based phase recognition model integrates seamlessly into the Surgomics concept [3], including features like “Duration of surgery/surgical phases” and “Order of performed phases/steps.” Future work should integrate AI-based workflow analysis at the surgical-step level and incorporate additional surgomic features, such as instrument use and field characteristics (e.g., bloodiness and smokiness) [4], potentially enhancing insights into surgical techniques and complications. Additionally, exploring the feasibility of inferring surgical phase from specific instruments or key anatomical structures [41] and assessing its alignment with automatic phase recognition, could potentially enhance the automated phase recognition. We observed significant differences in phase duration among centers, likely due to factors such as center-specific surgical standards, surgeon experience and skill level, or patient anatomy. These differences should be further investigated with respect to their potential impact on

Table 2
Differences in reference phase annotations
Results of the surgical phase durations of the three centers DD (Dresden), HD (Heidelberg), and MZ (Mainz) based on the reference annotations of all 36 videos and of the modified checklist tool.
Phase differences between the centers were assessed with the Kruskal-Wallis test [25]. P-values were adjusted through post-hoc Dunn-Bonferroni pairwise comparisons [26,27]. Significant differences (p-value <0.05) are bold.

Phase no.	Phase name	Center/checklist (and no. of annotated phases with checklist)	Median [min] (Interquartile range [min])	Centers for statistical testing		P-value
				Center 1	Center 2	
1	Preparation for abdominal part	DD	10.92 (2.84–21.76)	DD	HD	0.179
		HD	27.92 (18.82–39.27)	**HD**	**MZ**	**<0.001**
		MZ	2.88 (1.43–4.79)	MZ	DD	0.125
		All centers	5.92 (2.76–23.74)			
		Checklist (n = 10)	50.88 (37.36–66.59)			
2	Gastric and distal esophagus mobilization	DD	66.23 (51.5–74.21)	DD	HD	0.998
		HD	72.92 (60.9–97.15)	**HD**	**MZ**	**<0.001**
		MZ	33.07 (28.15–36.08)	**MZ**	**DD**	**0.004**
		All centers	59.68 (34.72–73.18)			
		Checklist (n = 9)	65.22 (54.49–89.97)			
3	Gastric tube construction	DD	21.01 (18.41–24.54)	DD	HD	0.336
		HD	14.98 (11.21–17.03)	**HD**	**MZ**	**0.003**
		MZ	8.13 (6.3–8.68)	**MZ**	**DD**	**<0.001**
		All centers	13.55 (8.72–19.74)			
		Checklist (n = 10)	12.13 (9.2–16.16)			
4	Completion of abdominal lymphadenectomy	DD	40.33 (32.48–45.9)	DD	HD	0.189
		HD	63.05 (52.53–67.9)	**HD**	**MZ**	**<0.001**
		MZ	24.75 (21.03–33.02)	MZ	DD	0.060
		All centers	39.28 (27.76–56.78)			
		Checklist (n = 10)	57.94 (28.88–76.46)			
5	Completion of abdominal part	DD	2.4 (1.61–3.88)	DD	HD	0.941
		HD	3.88 (2.12–5.59)	**HD**	**MZ**	**0.002**
		MZ	1.05 (0.8–1.81)	**MZ**	**DD**	**0.049**
		All centers	1.97 (1.17–3.81)			
		Checklist (n = 10)	15.79 (10.97–23.55)			
6	Preparation for thoracic part	DD	3.92 (2.29–13.01)	DD	HD	0.061
		HD	17.87 (15.58–18.71)	**HD**	**MZ**	**<0.001**
		MZ	1.11 (0.64–1.57)	MZ	DD	0.051
		All centers	4.51 (1.55–16.36)			
		Checklist (n = 8)	42.58 (34.17–52.27)			
7	Dissection of esophagus with regional lymphadenectomy	DD	67.08 (60.58–75.19)	**DD**	**HD**	**0.028**
		HD	96.03 (81.4–111.7)	**HD**	**MZ**	**<0.001**
		MZ	61.87 (51.85–73.08)	MZ	DD	0.834
		All centers	72.14 (61.33–88.5)			
		Checklist (n = 8)	79.89 (73.0–122.25)			
8	Insertion of anvil in the esophageal stump and stomach pull-up	DD	21.82 (18.98–27.36)	DD	HD	1.0
		HD	20.81 (17.55–27.53)	**HD**	**MZ**	**0.023**
		MZ	14.34 (11.97–16.13)	**MZ**	**DD**	**0.003**
		All centers	18.61 (14.26–22.18)			
		Checklist (n = 7)	26.40 (22.72–32.55)			
9	Circular stapled anastomosis	DD	40.68 (35.02–52.22)	DD	HD	0.383
		HD	36.22 (29.02–44.18)	HD	MZ	0.696
		MZ	29.71 (27.41–32.09)	**MZ**	**DD**	**0.021**
		All centers	35.32 (29.31–44.4)			
		Checklist (n = 5)	57.08 (32.49–85.15)			
10	Completion of thoracic part	DD	8.25 (6.04–9.93)	DD	HD	1.0
		HD	6.8 (5.09–7.31)	HD	MZ	0.063
		MZ	1.59 (0.83–4.1)	**MZ**	**DD**	**0.004**
		All centers	5.88 (3.15–8.25)			
		Checklist (n = 5)	26.72 (19.11–38)			
Total	Complete surgery	DD	284.58 (247.8–361.32)	DD	HD	0.472
		HD	374.45 (345.85–412.84)	**HD**	**MZ**	**<0.001**
		MZ	195.41 (163.68–206.97)	**MZ**	**DD**	**0.004**
		All centers	284.58 (208.05–361.32)			
		Checklist (n = 5)	438.32 (382.8–582.95)			

complications and clinical outcomes. Notably, the SEVERE score has been shown to correlate with overall procedure duration [42]. To enable a comprehensive comparison and in-depth analysis of these variances, more multicenter data with an analysis of step order and duration is needed and further surgomic features must be systematically developed. Regarding phase transitions, DD videos had the most frequent switches between phases in the reference annotations, while HD videos showed none, suggesting higher standardization. However, the workflow model phases were developed in HD and did not undergo multicentric consensus before data-based validation, potentially biasing it towards HD-specific practices. Phases 2 and 4 had the most interruptions, indicating they are the least standardized, while the thoracic part showed no interruptions, suggesting higher standardization. Future work should examine specific steps within these phases to identify where interruptions occur.

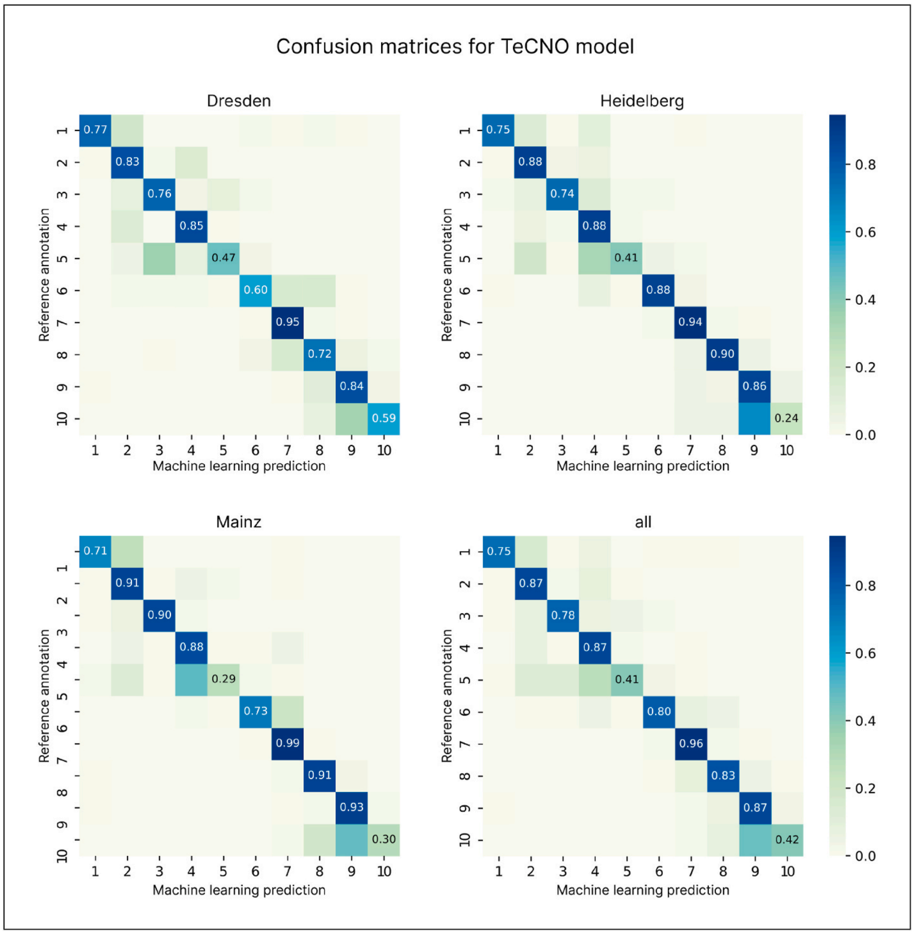


Fig. 3. Normalized confusion matrices of each center and of all videos together for the TeCNO machine learning model phase predictions during testing of each fold.

4.3. Limitations

Despite significant differences between centers, automatic phase recognition using ML methods was feasible and showed promising results. A standardized surgical workflow can greatly aid automatic phase recognition, and in turn, phase recognition can contribute to procedure standardization, ensuring consistency across surgical teams. However, this study focused on German centers, and exploring workflow variances across Europe or globally would be most valuable. With respect to annotation quality, inter-rater agreement between the two annotators was evaluated using Fleiss' Kappa, an established and widely used measure. The exceptionally high Kappa values indicate that the annotation protocol provided a clear and consistent framework, leading to robust agreement across the long duration of phases. Disagreements were mainly concentrated at phase transitions, which are less strongly reflected in the overall Kappa score. The ML model's variable performance between centers and phases highlights the need for more training data, especially cases with deviations and adverse events, to improve generalization. Although cross-validation allowed us to minimize the impact of potential outliers by leveraging multiple validation splits, it does not replace the need for an independent test set to evaluate true ML model performance in the future. In this work, we consciously opted against holding out a separate test set to maximize the data available for training. In terms of technical improvements, evaluating additional network architectures, such as end-to-end ML models [30], may further optimize performance. Furthermore, the ML model exhibited flickering in its predictions, which indicates rapid and frequent phase switching. Reducing this flickering by applying a smoothing technique to the predictions could significantly improve metrics such as the overlap score, which could be explored in future steps. Building on the current ML-based phase recognition, it would also be highly beneficial to extend annotation and automatic recognition to the 60 surgical steps within the RAMIE workflow model to enable more accurate workflow analysis and more specific tailored support. Future work will therefore focus on annotating the visible surgical steps in the videos and retraining the ML model with these fine-grained annotations. However, several challenges remain for effective AI training. Step annotation requires considerable effort and depends on expert knowledge to establish consistent annotation guidelines, which may best be achieved through a Delphi consensus process.

5. Conclusion

In this study, we introduced a comprehensive workflow model of the RAMIE procedure and trained and tested an ML model for automatic phase recognition with video data of three academic surgical centers. The workflow model was validated with manual live documentation using a checklist tool intraoperatively. Multicentric retrospective validation of the workflow model phases, based on video annotations from three surgical centers, showed statistically significant differences in phase duration across the centers. In the next step we plan to integrate and analyze further multicentric data. Based on these extended data sets, automatic phase recognition with ML will likely be able to pave the way for Surgomics-based complication prediction and context-aware process assistance.

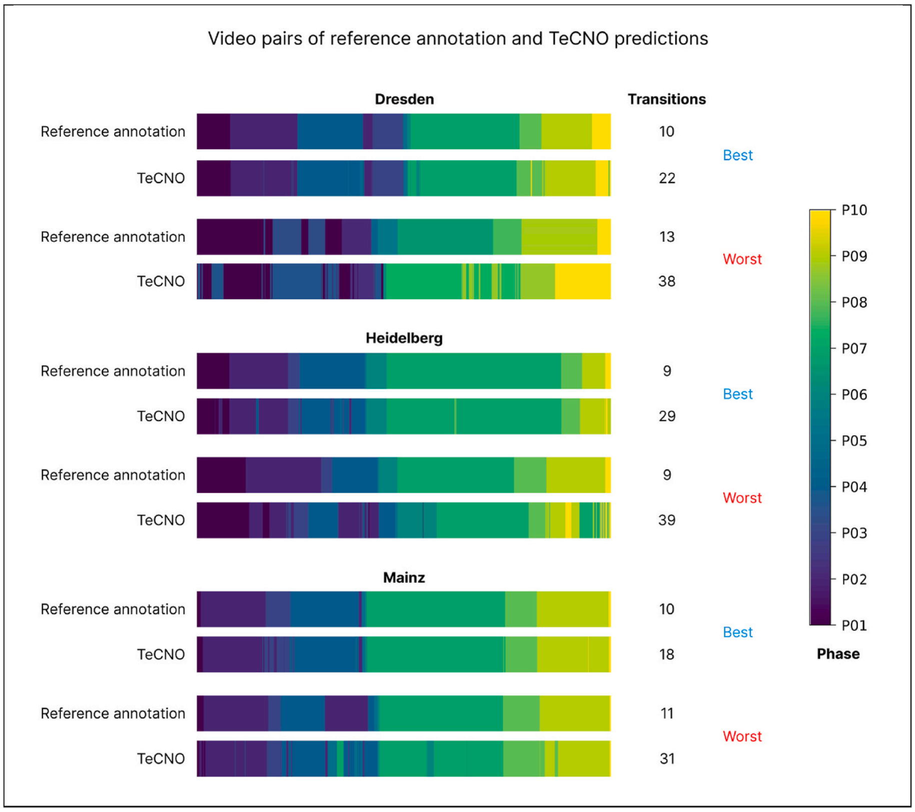


Fig. 4. Examples of relative chronological visualizations of the ten surgical phases of the robot-assisted minimally invasive esophagectomy from six videos (the best and the worst performance of the machine learning of each center) for reference annotations (upper row) and TeCNO prediction (lower row). The videos were divided into 1000 parts for visualization purposes. For each part, the phase that occurred most frequently in this section was selected. Furthermore, the number of phase transitions of the reference annotations as well as of the predictions was added. For the predictions a minimal duration threshold was set at 15 s.

Disclosures

A. Stern and L. Mündermann are employees of KARL STORZ SE & Co. KG. M. Wagner, R. Klotz, S. Bodenstedt, B. P. Müller-Stich, M. Distler, and S. Speidel are project leaders of the Surgomics project, funded by the German Federal Ministry of Health (grant number 2520DAT82) with medical device manufacturer KARL STORZ SE & Co. KG being a project partner. M. Distler, L. Maier-Hein, S. Speidel, S. Bodenstedt, R. Klotz, and M. Wagner are project partners in the Surgical AI Hub Germany project funded by the German Federal Ministry of Education and Research (grant number 02K223A110).

M. Wagner received a speaker fee from KARL STORZKARL STORZ SE & Co. KG. O. Burgert and D. Junger are funded by the Ministry of Science, Research and Arts Baden-Württemberg and the European Fund for Regional Development (EFRE) (grant number FEIH_KMU_1099897). A. C. Jenke was funded by the European Union through NEARDATA (grant agreement ID 101092644). J. M. Brandenburg, A. Schulze, A. C. Jenke, N. Bhasker, D. Rivoir, H. Naderi, N. Bleser, F. Fritz-Kebede, V. J. Lozanovski, P. P. Grimminger, A. Billeter, and J. Weitz have no conflicts of interest or financial ties to disclose.

Credit author statement

M. Wagner, B. P. Müller-Stich, M. Distler, J. Weitz, R. Klotz, A. Billeter, P. P. Grimminger, S. Speidel, L. Maier-Hein, L. Mündermann, O. Burgert, and F. Fritz-Kebede conceptualized the study. M. Wagner and J. M. Brandenburg designed the study. J. M. Brandenburg, A. Schulze, N. Bleser, D. Junger, A. Billeter, V. J. Lozanovski, R. Klotz, B. P. Müller-Stich, J. Weitz, P. P. Grimminger, and M. Distler contributed to data acquisition. Quality control of data and algorithms was performed by J. M. Brandenburg, A. Schulze, A. C. Jenke, D. Rivoir, and H. Naderi. Data analysis and interpretation were conducted by J. M. Brandenburg, A. Schulze, A. C. Jenke, N. Bhasker, S. Speidel, L. Maier-Hein, O. Burgert, and F. Fritz-Kebede. N. Bhasker, A. Stern, and J. M. Brandenburg performed the statistical analysis. J. M. Brandenburg, A. Schulze, and A. C. Jenke prepared the manuscript. J. M. Brandenburg, A. Schulze, M. Wagner, and A. C. Jenke were responsible for manuscript editing. All authors, including J. M. Brandenburg, A. Schulze, A. C. Jenke, N. Bhasker, N. Bleser, D. Junger, A. Stern, D. Rivoir, H. Naderi, F. Fritz-Kebede, O. Burgert, L. Maier-Hein, L. Mündermann, S. Bodenstedt, S. Speidel, V. J. Lozanovski, P. P. Grimminger, A. Billeter, R. Klotz, J. Weitz, M. Distler, B. P. Müller-Stich, and M. Wagner, reviewed the manuscript and approved the final version of the paper.

Funding information

This work has been funded by the German Federal Ministry of Health (BMG) within the “Surgomics” project (grant number BMG 2520DAT82), by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany’s Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence “Centre for Tactile Internet with Human-in-the-Loop” (CeTI) of Technische Universität Dresden and by the National Center for Tumor Diseases Heidelberg within the cancer therapy program “Data Science driven Surgical Oncology”.

Acknowledgements

Open Access funding was enabled and organized by Projekt DEAL. This work has been funded by the German Federal Ministry of Health (BMG) within the “Surgomics” project (Grant Number BMG 2520DAT82); by the German Research Foundation (DFG, Deutsche

Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden; by the National Center for Tumor Diseases Heidelberg within the cancer therapy program "Data Science driven Surgical Oncology"; by the German Federal Ministry of Education and Research within the "Surgical AI Hub Germany" project (grant number BMBF 02K223A110) and within the program "Souverän. Digital. Vernetzt.", joint project 6G-life (project identification number: 16KISK001K); and by the Ministry of Science, Research and Arts Baden-Württemberg as well as the European Fund for Regional Development (EFRE) (grant number FEIH_KMU_1099897).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejso.2025.111174.

References

[1] Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434–42.

[2] Jung JJ, Jüni P, Lebovic G, et al. First-year analysis of the operating room Black box study. Ann Surg 2020;271:122–7.

[3] Wagner M, Brandenburg JM, Bodenstedt S, et al. Surgomics: personalized prediction of morbidity, mortality and long-term outcome in surgery using machine learning on multimodal data. Surg Endosc 2022;36:8568–91.

[4] Brandenburg JM, Jenke AC, Stern A, et al. Active learning for extracting surgomic features in robot-assisted minimally invasive esophagectomy: a prospective annotation study. Surg Endosc 2023;37:8577–93.

[5] Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. Nat Biomed Eng 2017;1:691–6.

[6] Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. Ann Surg 2019;270:414–21.

[7] Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic myotomy. Surg Endosc 2021;35:4008–15.

[8] Wagner M, Müller-Stich B-P, Kisilenko A, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. Med Image Anal 2023;86:102770.

[9] Rivoir D, Bodenstedt S, Funke I, et al. Rethinking anticipation tasks: uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. arXiv, https://doi.org/10.48550/arXiv.2007.00548; 2007.

[10] Garrow CR, Kowalewski K-F, Li L, et al. Machine learning for surgical phase recognition: a systematic review. Ann Surg 2021;273:684–93.

[11] Lavanchy JL, Ramesh S, Dall'Alba D, et al. Challenges in multi-centric generalization: phase and step recognition in Roux-en-Y gastric bypass surgery. Int J Comput Assist Radiol Surg 2024;19:2249–57.

[12] Low DE, Alderson D, Cecconello I, et al. International consensus on standardization of data collection for complications associated with esophagectomy: esophagectomy complications consensus group (ECCG). Ann Surg 2015;262:286–94.

[13] Egberts J-H, Biebl M, Perez DR, et al. Robot-assisted oesophagectomy: recommendations towards a standardised ivor lewis procedure. J Gastrointest Surg 2019;23:1485–92.

[14] Harris A, Butterworth J, Boshier PR, et al. Development of a reliable surgical quality assurance system for 2-stage esophagectomy in randomized controlled trials. Ann Surg 2022;275:121–30.

[15] Kingma BF, Read M, van Hillegersberg R, et al. A standardized approach for the thoracic dissection in robotic-assisted minimally invasive esophagectomy (RAMIE). Dis Esophagus 2020;33:doaa066.

[16] Takeuchi M, Kawakubo H, Saito K, et al. Automated surgical phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence. Ann Surg Oncol 2022;29:6847–55.

[17] Eckhoff JA, Ban Y, Rosman G, et al. TEsoNet: knowledge transfer in surgical phase recognition from laparoscopic sleeve gastrectomy to the laparoscopic part of ivor lewis esophagectomy. Surg Endosc 2023;37:4040–53.

[18] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: decide-ai. Nat Med 2022;28:924–33.

[19] Meireles OR, Rosman G, Altieri MS, et al. SAGES consensus recommendations on an annotation framework for surgical video. Surg Endosc 2021;35:4918–29.

[20] Hagens ERC, van Berge Henegouwen MI, van Sandick JW, et al. Distribution of lymph node metastases in esophageal carcinoma [TIGER study]: study protocol of a multinational observational study. BMC Cancer 2019;19:662.

[21] Blencowe NS, Mills N, Cook JA, et al. Standardizing and monitoring the delivery of surgical interventions in randomized clinical trials. Br J Surg 2016;103:1377–84.

[22] Just E, Schaumann K, Junger D, et al. Towards automated surgical documentation using automatically generated checklists from BPMN models. Curr Dir Biomed Eng 2021;7:135–9.

[23] Junger D, Just E, Brandenburg JM, et al. Toward an interoperable, intraoperative situation recognition system via process modeling, execution, and control using the standards BPMN and CMMN. Int J Comput Assist Radiol Surg 2024;19:69–82.

[24] Ryniak C, Burgert O. Automatic generation of checklists from business process model and notation (BPMN) models for surgical assist systems. Curr Dir Biomed Eng 2020;6(1):20200005.

[25] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.

[26] Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Seeber; 1936.

[27] Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961;56:52–64.

[28] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.

[29] Czempiel T, Paschali M, Keicher M, et al. TeCNO: surgicalphase recognition with multi-stage temporal convolutional networks. Lect Notes Comput Sci LNCS 2020; 12263:343–52.

[30] Rivoir D, Funke I, Speidel S. On the pitfalls of batch normalization for end-to-end video learning: a study on surgical workflow analysis. Med Image Anal 2024;94:103126.

[31] Saravanan V, Berman GJ, Sober SJ. Application of the hierarchical bootstrap to multi-level data in neuroscience. Neuron Behav Data Anal Theory 2020;3(5). https://nbdt.scholasticahq.com/article/13927.

[32] Lea C, Vidal R, Hager GD. Learning convolutional action primitives for fine-grained action recognition. In: 2016 IEEE international conference on robotics and automation (ICRA); 2016. p. 1642–9.

[33] Hales B, Terblanche M, Fowler R, Sibbald W. Development of medical checklists for improved quality of patient care. Int J Qual Health Care 2008;20(1):22–30.

[34] de Vries EN, Hollmann MW, Smorenburg SM, Gouma DJ, Boermeester MA. Development and validation of the SURgical PAtient safety system (SURPASS) checklist. Qual Saf Health Care 2009;18(2):121–6.

[35] Wong HW, Forrest D, Healey A, Shirafkan H, Hanna GB, Vincent CA, et al. Information needs in operating room teams: what is right, what is wrong, and what is needed? Surg Endosc 2011;25(6):1913–20.

[36] Joeres F, Schindele D, Luz M, Blaschke S, Russwinkel N, Schostak M, et al. How well do software assistants for minimally invasive partial nephrectomy meet surgeon information needs? A cognitive task analysis and literature review study. PLoS One 2019;14(7):e0219920.

[37] Jalote-Parmar A, Badke-Schaub P. Workflow integration matrix: a framework to support the development of surgical information systems. Des Stud 2008;29(4):338–68.

[38] Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A, et al. Surgical data science - from concepts toward clinical translation. Med Image Anal 2022;76:102306.

[39] Hanna GB, Arya S, Markar SR. Variation in the standard of minimally invasive esophagectomy for cancer–systematic review. Semin Thorac Cardiovasc Surg 2012; 24(3):176–87.

[40] Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO, Navab N. Statistical modeling and recognition of surgical workflow. Med Image Anal 2012;16(3):632–41.

[41] den Boer RB, Jaspers TJM, de Jongh C, et al. Deep learning-based recognition of key anatomical structures during robot-assisted minimally invasive esophagectomy. Surg Endosc 2023;37:5164–75.

[42] Jung JJ, Jüni P, Gee DW, Zak Y, Cheverie J, Yoo JS, et al. Development and evaluation of a novel instrument to measure severity of intraoperative events using video data. Ann Surg 2020;272(2):220–6.