Appendix300: A multi-institutional laparoscopic appendectomy video dataset for computational modeling tasks

Fiona R. Kolbinger[1,2,3], Max Kirchner[4], Kevin Pfeiffer[2], Sebastian Bodenstedt[4,5], Alexander C. Jenke[4], Julia Barthel[1], Matthias Carstens[1], Karolin Dehlke[1,11], Sophia Dietz[1,11], Sotirios Emmanouilidis[6], Guido Fitze[7], Martin Freitag[8], Fabian Holderried[6], Thorsten Jacobi[9], Weam Kanjo[8], Linda Leitermann[10], Sören Torge Mees[10], Steffen Pistorius[1,11], Conrad Prudlo[7], Astrid Seiberth[10], Jurek Schultz[7], Karolin Thiel[6], Daniel Ziehn[9], Stefanie Speidel[4,5], Jürgen Weitz[1,5], Jakob Nikolas Kather[2,12,13], Marius Distler[1], Oliver Lester Saldanha[2,13]

Affiliations

[1] Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[2] Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[3] Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA
[4] 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), Germany
[5] Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden University of Technology, Dresden, Germany
[6] Department of General, Visceral and Thoracic Surgery, St. Elisabethen-Klinikum Ravensburg, Ravensburg, Germany
[7] Department of Pediatric Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany
[8] Krankenhaus St. Joseph-Stift Dresden GmbH, Dresden, Germany
[9] Diakonissenkrankenhaus Dresden, Dresden, Germany
[10] Department of General and Visceral Surgery, Dresden-Friedrichstadt General Hospital, Dresden, Germany
[11] Asklepios-ASB Krankenhaus Radeberg, Radeberg, Germany
[12] Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, 01307 Dresden, Germany
[13] Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Corresponding author(s): Fiona R. Kolbinger (fiona.kolbinger@uniklinikum-dresden.de), Oliver Lester Saldanha (oliverlester.saldanha@tu-dresden.de)

Abstract

The limited availability of diverse and representative training data poses a critical barrier to the development of clinically relevant computational tools for intraoperative surgical decision support. As surgical interventions are not routinely recorded and data annotation often requires domain knowledge, existing open-access surgical video datasets with high-quality annotations are available only for a restricted number of surgical procedures, such as cholecystectomy, and typically originate from a single institution. Appendix300 is a comprehensive dataset of video footage from 330 surgical procedures, including 312 full-length recordings of laparoscopic appendectomies in pediatric and adult patients treated at five German centers, along with 18 control recordings of non-inflamed appendices collected during non-appendectomy laparoscopic surgeries. Clinical metadata for each recording includes patient demographics, medical history, clinical symptoms, preoperative laboratory parameters, and postoperative histopathological findings. Additionally, we provide annotations of the grade of appendicitis based on the intraoperative presentation. This dataset enables new and clinically relevant validation tasks for computer vision in laparoscopic surgery, thereby enhancing the breadth and translational relevance of AI-based surgical video analysis.

Background & Summary

Computational analysis of medical imaging and clinical data has the potential to improve diagnostic accuracy and treatment stratification across medical fields. Increasing evidence from prospective clinical trials supports the clinical applicability and benefit of Artificial Intelligence (AI)-based tools.[1] For example, computational tools have demonstrated expert-level performance for polyp detection during colonoscopy[2], the detection of pulmonary nodules in chest X-rays[3], and mammogram interpretation for breast cancer screening[4].

Computational model performance and generalizability depend on the availability of large and diverse sets of clinical training and test data. Some clinical fields produce imaging as a means of documentation that becomes a component of patient records and is repetitively reviewed after acquisition for diagnostic follow-up or therapeutic decision-making (i.e., electrocardiograms in cardiology, X-rays in radiology, pictures of polyps in colonoscopy). In contrast, surgical video data is more complex to acquire and store in a standardized way due to its large file size. Reviewing and annotating surgical video data are time-consuming and require surgical domain knowledge.[5] While surgical video recordings represent an objective documentation of a surgical procedure and capture information that could be used for educational and prognostic purposes[6–8], surgeries are not routinely recorded, and retrospectively written operation notes remain the sole standard at most clinical institutions[9]. The largest available open-access datasets of intraoperative imaging comprise data from about 100 patients and cover a limited variety of procedures, including cholecystectomy (i.e., Cholec80[10], CholecT50[11]) and cataract surgeries (i.e., Cataract-101[12], CATARACTS[13]). Most datasets exclusively comprise spatial or categorical annotations of the surgical scene (i.e., presence of tools and anatomical structures in a video frame) rather than patient-level or clinical data, allowing for a limited number of surgical AI validation tasks that insufficiently represent the breadth of clinical reasoning tasks in surgery.

To increase the diversity of annotated surgical video data and validation tasks for surgical data science[14] approaches, we curated the Appendix300 dataset, comprising 330 video recordings of laparoscopic appendectomies and healthy control surgeries (i.e., appendix recordings from non-appendectomy laparoscopic surgeries) performed at five centers, as well as matched clinical metadata and annotations of the intraoperative grade of appendicitis (**Figure 1**).

Appendix300 currently represents the largest publicly available collection of surgical video data with patient metadata and the first curated dataset of laparoscopic appendectomies. By combining patient-level clinical data and pathology-related annotations with laparoscopic video data, this dataset connects a range of structured data types along the diagnostic and therapeutic pathway, facilitating pathophysiology-related investigations. Furthermore, the multicentric nature of the dataset allows for the evaluation of decentralized learning approaches. In the context of acute appendicitis, potential future applications include the development of intraoperative assistance systems for quality control and a harmonization of the intraoperative inflammation grade and the histopathological phenotype, which are often discordant[15]. Relation with preoperative imaging[16] and postoperative outcome data (i.e., surgical complications) are possible future extensions that are beyond the scope of this clinical use case but will need to be generated for other clinical use cases.

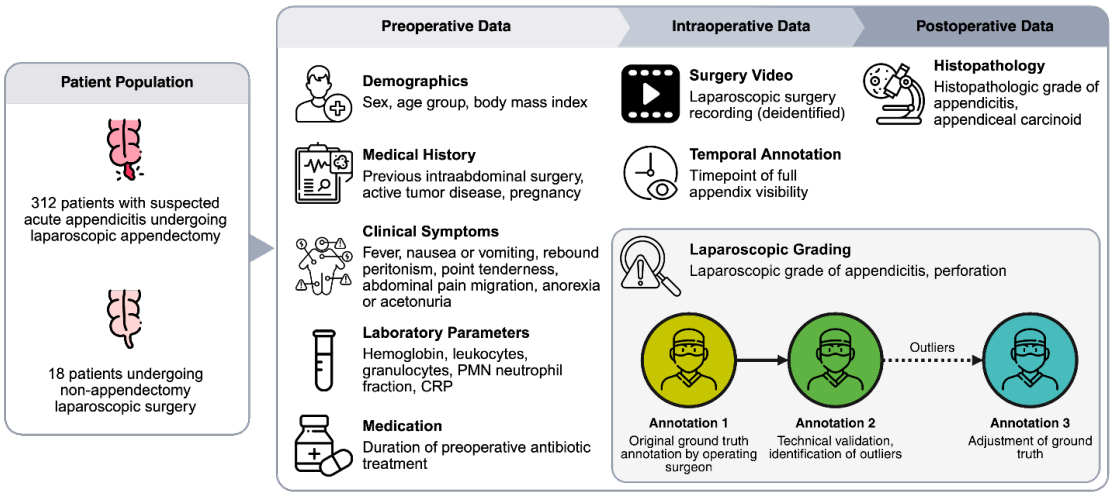

Figure 1. Dataset Overview. The Appendix300 dataset covers preoperative, intraoperative, and postoperative data of 312 patients with suspected appendicitis undergoing laparoscopic appendectomy and 18 patients undergoing non-appendectomy laparoscopic surgery at five surgical centers. Preoperative data comprise demographics, select medical history details, clinical symptoms of acute appendicitis, laboratory parameters, and recent antibiotic medication history. The surgery video recording was temporally annotated with regard to the timepoint of complete appendix visibility before appendix dissection, the laparoscopic grade of appendicitis, and perforation. Two surgeons independently annotated the laparoscopic grade and major disagreements were resolved through a third independent annotation. Postoperative metadata include the histopathological grade of appendicitis as well as the presence or absence of an appendiceal carcinoid. Abbreviations: CRP (C-reactive protein), PMN (polymorphonuclear).

Methods

This dataset comprises laparoscopic video recordings of 312 laparoscopic appendectomies and 18 non-appendectomy laparoscopic surgeries, matched clinical metadata along the preoperative and postoperative treatment pathway, and annotations of the intraoperative grade of appendicitis (**Figure 1**, **Figure 2**).

Video Recording

Between June 2023 and September 2024, video data from 312 laparoscopic appendectomies and 18 non-appendectomy laparoscopic surgeries were gathered at five German surgical centers. The participating centers comprise four public hospitals (Asklepios-ASB Krankenhaus Radeberg, Diakonissenkrankenhaus Dresden, Krankenhaus St. Joseph Stift Dresden, St. Elisabethen-Krankenhaus Ravensburg) and one academic surgical department (Department of Pediatric Surgery, University Hospital Carl Gustav Carus Dresden). All included patients had a clinical indication for the surgical procedure.

Appendectomy recordings and clinical metadata were consecutively gathered in clinical routine. Patients with an indication of laparoscopic appendectomy for suspected acute appendicitis and documented histopathologic inflammation grade were included in the appendectomy dataset. Exclusion criteria comprised conversion to open surgery, incomplete surgery recordings, and cases with corrupted files. Surgeries were performed, recorded, and saved using locally available laparoscopic hardware (**Table 1**).

Table 1: Recording infrastructure and technical details of the dataset for all contributing centers.

	Center 1	**Center 2**	**Center 3**	**Center 4**	**Center 5**
Clinical Institution	public	academic	public	public	public
Patient Demographics	adult	pediatric	adult	pediatric and adult	adult
Number of Recordings	51	49	42	71	118
Laparoscopic System	Arthrex Synergy UHD4, Arthrex SynergyID	STORZ IMAGE1 S	STORZ IMAGE1 S	STORZ IMAGE1 S, B. Braun AESCULAP EinsteinVision	Richard Wolf ENDOCAM 4K
Recorded Frame Rate [frames per second]	30	25	25	50	50
Resolution [pixels]	1920 x 1080	1920 × 1080 or 3840 × 2160	1920 × 1080 or 3840 x 2160	1920 x 1080	1920 x 1080
File Type	mp4	mp4	mp4	mp4	mp4
File Output [single file or multiple files]	single file	single or multiple files	single file	single file	single or multiple files

Acquisition of Clinical Labels and Annotations

A graphical user interface (**Supplementary Material 1**) was implemented to save raw surgery recordings, corresponding clinical labels, and original annotations of the laparoscopic grade of appendicitis on local hardware in anonymized form. This process was identical for laparoscopic appendectomy recordings and appendix recordings from non-appendectomy laparoscopic surgeries.

The graphical user interface was implemented in Python using tkinter and integrates a custom video playback component based on PyAV and Pillow for frame-accurate seeking, stereo video handling, and consistent frame rate rendering. It supports hospital-specific configurations (e.g., multi-video workflows, stereo cropping) via a centralized config structure and allows for timestamp annotation and structured metadata input through predefined form elements. All annotations and associated videos were saved to a defined output structure, with the metadata stored in a standardized CSV file, enabling reproducible and site-adaptive data extraction.

Clinical labels were derived from routine documentation based on a clinical labeling protocol (**Supplementary Material 2**). Following this protocol, the following parameters were preoperatively gathered: Sex, age, body-mass index (BMI), clinical symptoms based on the Alvarado score[17] (migration of abdominal pain to the right lower quadrant, anorexia or acetone in the urine, nausea/vomiting, point tenderness in the right lower quadrant, rebound peritonism in the right iliac fossa, body temperature), preoperative laboratory parameters (hemoglobin, leukocytes, granulocytes, proportion of polymorphonuclear cells, C-reactive

protein), select medical history details (history of intraabdominal surgery, active tumor disease, pregnancy), and the duration of preoperative antibiotic treatment (**Figure 1**).

Using the graphical user interface's video review function (**Supplementary Material 1**), a timestamp, at which the appendix is fully visible prior to invasive preparation, was documented by the contributing surgery residents and attendings (FRK, JB, KD, SD, FH, WK, LL, JS, AS, DZ). Based on this timestamp, a 100-second video snippet was extracted from the full video recording, which covers 50 seconds before and after the timepoint of full appendix visibility (**Figure 2 a**).

The intraoperative grade of appendicitis and the presence of perforation were classified by the surgical teams carrying out the procedure based on an annotation protocol (**Supplementary Material 3**, **Figure 2 b**), which follows the intraoperative classification proposed by Gomes *et al.*[18] (annotation 1). Annotation 1 represents the original ground truth annotation.

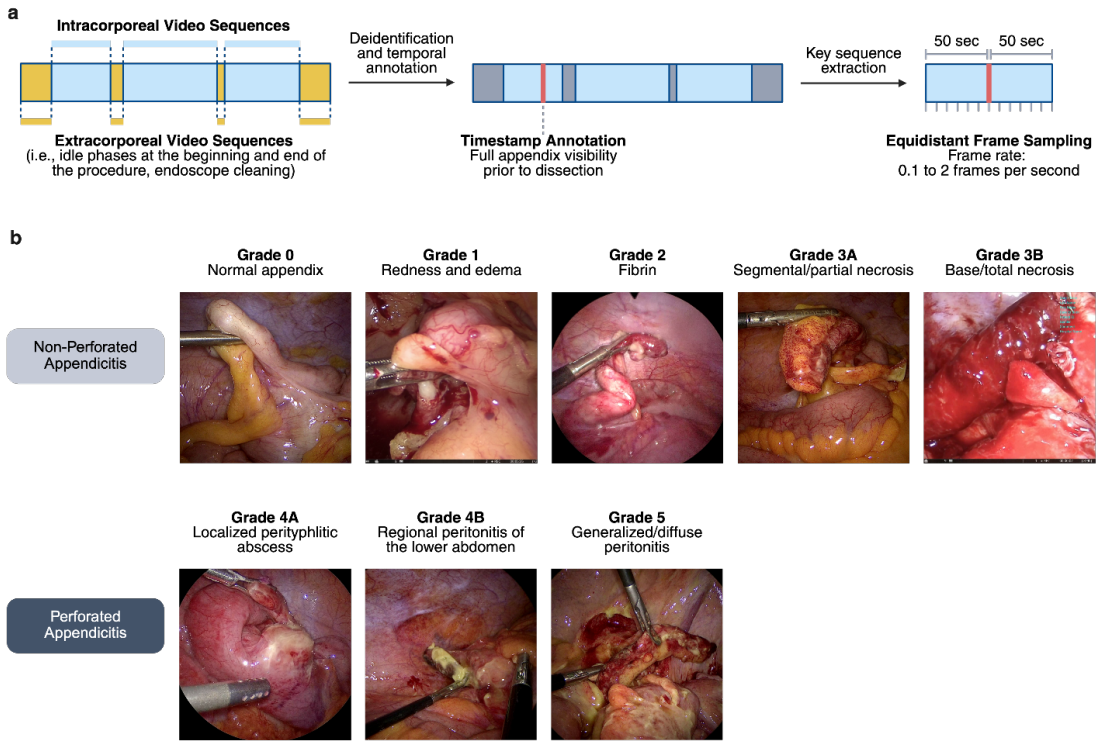

Figure 2. Data preprocessing and annotation of the intraoperative grade of appendicitis. (a) Schematic representation of the data processing steps. Following the deidentification of extracorporeal laparoscopic video sequences, each appendectomy recording was labeled with a temporal label (timestamp) marking the beginning of appendiceal dissection. Up to 200 equidistant frames were sampled from a 100-second key video sequence spanning 50 seconds before and after the timestamp. **(b)** Intraoperative grades of appendicitis were defined based on Gomes *et al.* (2012)[18]. Based on an annotation protocol (**Supplementary Material 3**), each video was classified by the surgical team. For technical validation, a second reviewer performed a second annotation of the intraoperative grade of appendicitis. In case of major disagreement, a third independent annotation was acquired to adjust the ground truth where necessary.

For technical validation and to identify outliers and possible annotation errors, a second reviewer (general surgery resident with 4 years of experience in laparoscopic surgery) independently annotated the intraoperative grade of appendicitis for all recordings (annotation 2). Outliers were defined as a disagreement of at least three appendicitis grades between annotation 1 and annotation 2 (e.g., grade 2 and grade 5; grade 3A and grade 4B). We report the interrater agreement between annotation 1 and annotation 2 quantitatively

using weighted Cohen's kappa[19] (**Table 2**, **Figure 3 a**), and qualitatively through a review of divergent ratings (**Figure 3 b**).

To adjust the ground truth where necessary, outliers were subjected to a third independent annotation by a general surgery resident with 7 years of experience in laparoscopic surgery (annotation 3). In case of a major disagreement between annotation 1 and annotation 3 (i.e., a disagreement of 3 or more appendicitis grades), the ground truth was adjusted to annotation 3. If annotation 3 was in agreement with annotation 1 (i.e., disagreement of up to 2 appendicitis grades), the ground truth was not adjusted (**Figure 3 b**).

Table 2 summarizes patient characteristics and appendicitis grade annotations for all surgeries.

209 **Table 2: Summary of patient characteristics and surgery-related data.**

	Center 1	Center 2	Center 3	Center 4	Center 5	All
Sex						
Female	30	25	21	33	64	173
Male	21	24	21	38	53	157
Body Mass Index						
Underweight (BMI < 18.5)	0	15	0	7	0	22
Normal weight (BMI 18.5 – 24.9)	0	22	0	30	47	99
Pre-obesity (BMI 25 – 29.9)	0	1	0	16	38	55
Obesity class I (BMI 30 – 34.9)	0	3	0	9	14	26
Obesity class II (BMI 35 – 39.9)	0	2	0	3	5	10
Obesity class III (BMI > 40)	0	1	0	0	0	1
N/A	51	5	42	6	13	117
Age group						
Under 5 years	0	4	0	1	0	5
5 to 9 years	0	14	0	3	0	17
10 to 14 years	0	21	0	9	0	30
15 to 19 years	3	10	2	2	6	23
20 to 24 years	2	0	9	8	21	40
25 to 29 years	2	0	3	4	12	21
30 to 34 years	4	0	1	3	10	18
35 to 39 years	7	0	35	2	11	25
40 to 44 years	10	0	4	8	9	31
45 to 49 years	2	0	1	5	9	17
50 to 54 years	3	0	5	5	10	23
55 to 59 years	5	0	4	6	7	22
60 to 64 years	6	0	2	0	7	15
65 to 69 years	4	0	0	4	1	9
70 to 74 years	0	0	1	3	3	7
75 to 79 years	3	0	3	3	4	13
80 years and over	0	0	2	5	7	14
History of intraabdominal surgery						
None	45	44	34	57	99	279
Minor	6	0	5	8	15	34
Major	0	0	3	5	2	10
N/A	0	5	0	1	1	7
Surgery type						
Laparoscopic appendectomy	38	44	42	71	117	312
Non-appendectomy laparoscopic surgery	13	5	0	0	0	18
Laparoscopic grading of appendicitis (final ground truth)						
0 (normal looking)	3	11	1	1	1	17
1 (redness and edema)	12	6	9	5	22	54
2 (fibrin)	8	21	4	11	19	63
3A (segmental/partial necrosis)	7	4	11	19	24	65
3B (base/total necrosis)	11	0	1	7	25	44
4A (perityphlitic abscess)	5	4	0	10	11	30
4B (regional peritonitis of the lower abdomen)	3	3	15	11	12	44
5 (generalized/diffuse peritonitis)	2	0	1	7	3	13
Histopathologic grade of appendicitis						
No histopathologic signs of appendicitis	2	5	1	0	0	8
Mild	3	9	10	6	18	46
Intermediate	29	29	13	54	67	192
Severe	17	5	18	11	32	83
N/A	0	1	0	0	0	1
Interrater Agreement [weighted Cohen's κ, Annotation 1 vs. Annotation 2]	0.527	0.578	0.636	0.509	0.646	0.615
Video duration [min, mean ± SD]	5.43 ± 4.09	29.52 ± 19.81	33.81 ± 12.42	33.94 ± 17.52	30.12 ± 14.40	27.51 ±17.67

Ethics Statement

This dataset was collected in accordance with the Declaration of Helsinki and its later amendments. All data were compiled in an anonymized fashion using the described user interface. The responsible Institutional Review Boards reviewed and approved this study on August 4th, 2022 (ethics committee at the Technical University Dresden, approval number BO-EK-332072022), September 13th, 2023 (ethics committee of the Sächsische Landesärztekammer, approval number EK-BR-75/23-1), and December 23rd, 2023 (ethics committee of the Landesärztekammer Baden-Württemberg, approval number B-F-2023-023). The trial, in the context of which this dataset was acquired, was prospectively registered at the German Clinical Trials Register (Deutsches Register Klinischer Studien, DRKS) on December 9th, 2022 (trial registration ID: DRKS00030874). Patients were informed about the anonymized acquisition, analysis, and publication of data from their inpatient treatment. Following local legislation, no written informed consent was required for anonymized acquisition, analysis, and publication of clinical data.

Data Records

The Appendix300 dataset comprises 100-second video snippets from 330 surgery video recordings, along with patient metadata and annotations. These video snippets cover 50 seconds before and after the annotated timepoint of full appendix visibility. Due to the large file size of the deidentified full-length surgery recordings, they are made available to interested researchers upon reasonable request. To access the full-length surgery recordings, data requestors must sign a data use agreement.

Data will be made publicly available following the peer review process for non-commercial use under the Creative Commons Attribution CC-BY. If readers wish to use or reference this dataset, they should cite this paper.

Description of the data folder structure

The dataset is structured into two main directories: “Video_Snippets” and “Images”. Each of these directories is organized by contributing center, with data grouped into subfolders for each patient, such as “Center1_001”. The “Video_Snippets” directory contains patient subfolders with the respective 100-second video snippet covering 50 seconds before and after the annotated timepoint of full appendix visibility, following the same folder structure. In the “Images” directory, each patient subfolder includes a frames.zip archive with 200 frames extracted from the corresponding video snippet at 2 frames per second. At the top level of the dataset, an “overall_merged.csv” file is available to provide a consolidated view of all centers and patients, including metadata and annotations.

Technical Validation

For technical validation of the annotations of the intraoperative grade of appendicitis, an independent secondary annotation was acquired. The overall interrater agreement (weighted Cohen’s kappa) was 0.614, with variations across the contributing centers (**Table 2**, **Figure 3 a**). Of 330 cases, 30 (9.1%) were classified as outliers or potential annotation errors based on a major disagreement between annotation 1 and annotation 2. Upon a third independent review of these 30 outlier cases (annotation 3), ground truths were adjusted to annotation 3 for 14 cases (4.2% of all cases) and original grade labels (annotation 1) were retained for 16 cases (**Figure 3 b**). These findings provide a baseline for interrater variability in laparoscopic grading of appendicitis.

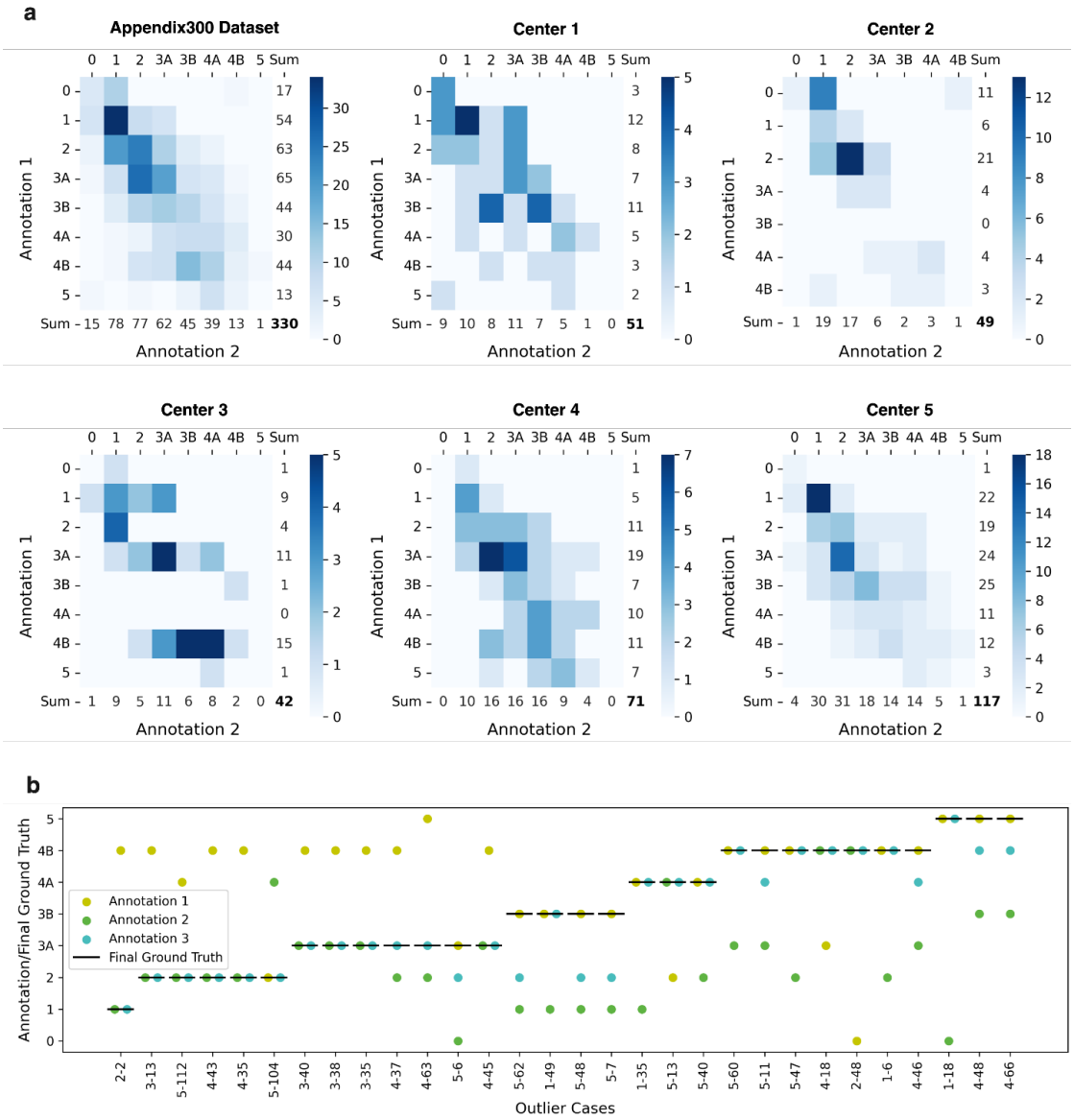

Figure 3. Technical validation of the laparoscopic grade of appendicitis annotations. Confusion matrices **(a)** indicate the agreement between the original annotation by the operating surgeon (annotation 1) and the independent second annotation (annotation 2) for all patients. **(b)** Of all 330 cases, 30 outlier cases with a disagreement of more than two appendicitis grades between annotation 1 and annotation 2 were subjected to a third independent annotation, and ground truth labels were adjusted as necessary. Case indices (x-axis ticks) comprise the center number and the case number, separated by a dash.

Usage Notes

The Appendix300 dataset can be used for various purposes in the field of machine learning, either on its own or in combination with other, already existing datasets. Used on its own, it facilitates the training of computational models identifying the intraoperative grade of appendicitis as a classification task and differentiation of perforated and non-perforated appendicitis (binary classification task), which are new clinical use cases for algorithm validation in surgical data science.

The multi-institutional nature of the Appendix300 dataset also facilitates the evaluation of decentralized and federated machine learning models for the above-mentioned validation tasks. We report the results of a benchmarking study evaluating Swarm Learning for

decentralized, privacy-preserving collaboration in surgical video analysis in a separate publication.

The Appendix300 dataset has three key limitations: First, some clinical data are incompletely available due to the manual entry of anonymized metadata at contributing centers. For example, no information about the BMI could be obtained from center 1 and center 3 (**Table 2**). This limitation represents a common scenario in multi-institutional data collection efforts and could be overcome through direct access to electronic health records, which was not feasible in this work due to legal and administrative restrictions at the contributing centers. Second, the dataset does not include any preoperative imaging data or imaging-derived variables that have been identified as predictors of a complicated course, e.g., the presence of free intraabdominal air or an intra-abdominal abscess.[20] Similarly, no data on postoperative complications are available in this dataset. Third, annotations of the laparoscopic grade of appendicitis followed the Gomes classification, which, in itself, has limitations.[18] For example, this classification requires the presence of macroscopic necroses for the assignment of grade 3, and the presence of encapsulated abscesses or leakage of pus into the abdominal cavity for the assignment of grade 4 (**Figure 2 b**, **Supplementary Material 3**). While it is applicable to most laparoscopic presentations of appendicitis, some intermediate cases may be particularly ambiguous to classify. For example, an appendix presenting with fibrin coverage, mild regional peritonitis, and small amounts of opaque (i.e., likely purulent) ascites, yet without macroscopic necroses or encapsulated abscesses, may fall into either grade 2 (fibrin) or grade 4B (regional peritonitis). This ambiguity and potential non-linearity of the clinical progression of appendicitis and its laparoscopic presentation have been acknowledged by previous research.[21,22]

Despite these limitations, the Appendix300 dataset represents a considerable step towards clinically relevant applications for computational surgical video analysis, as it connects demographic, symptom-related, laparoscopic, and histopathological data for a large, multi-institutional patient cohort, and introduces a novel use case that will help diversify the clinical applications of research efforts in surgical data science.

Code Availability

No custom code was used in the generation or processing of this dataset.

Acknowledgements

The authors gratefully acknowledge administrative support from Sandra Korn, Ulrike Neckmann, Christian Praetorius, and Anika Stützer (Clinical Trial Center, Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus Dresden, Dresden, Germany) and technical support from all participating institutions in locally setting up data collection infrastructure.

This work was supported by the German Cancer Research Center (CoBot 2.0) and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of Germany's Excellence Strategy (EXC 2050/1, Project ID 390696704) within the Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of the Dresden University of Technology. FRK receives support from the Joachim Herz Foundation (Add-On Fellowship for Interdisciplinary Life Science), the Central Indiana Corporate Partnership AnalytiXIN Initiative, the Evan and Sue Ann Werling Pancreatic Cancer Research Fund, and the Indiana Clinical and Translational Sciences Institute (EPAR4157) funded, in part, by Grant Number UM1TR004402

from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. MK and ACJ are supported by the European Union through NEARDATA under grant agreement ID 101092644. JNK is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the National Institutes of Health, the NIHR, or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor other granting authorities can be held responsible for them.

Author contributions

Conceptualization: FRK, JNK, MD, OLS; Methodology: FRK, SB, JS, SS, JNK, OLS; Software: MK, KP, SB, ACJ, OLS; Validation: FRK, MK, KP, SB, ACJ, MC, OLS; Formal analysis: FRK, MK, KP, OLS; Investigation: FRK, KP, SB, JB, MC, KD, SD, SE, FH, WK, LL, CP, AS, JS, DZ, MD, OLS; Resources: GF, MF, TJ, STM, SP, JS, KT, SS, JW, JNK, MD; Data Curation: FRK, MK, KP, SB, JB, MC, KD, SD, SE, FH, WK, LL, CP, AS, JS, DZ, MD, OLS; Writing - Original Draft: FRK; Writing - Review & Editing: MK, KP, SB, ACJ, JB, MC, KD, SD, SE, GF, MF, FH, TJ, WK, LL, STM, SP, CP, AS, JS, KT, DZ, SS, JW, JNK, MD; Visualization: FRK, MK; Supervision: FRK, SS, JW, JNK, MD, OLS; Project administration: FRK, OLS; Funding acquisition: FRK, SB, SS, JW, JNK, MD.

Competing interests

FRK declares advisory roles for Radical Healthcare, USA; and the Surgical Data Science Collective (SDSC), USA. JNK declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Scailyte, Switzerland; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI, Synagen, and Ignition Labs, Germany, has received a research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Eisai, Janssen, MSD, BMS, Roche, Pfizer and Fresenius. MD declares advisory roles for AESCULAP AG, Germany; and Intuitive Surgical, USA, and has received honoraria from Medtronic. All other authors declare no competing interests.

References

1. Han, R. *et al.* Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health* **6**, e367–e373 (2024).
2. Repici, A. *et al.* Artificial intelligence and colonoscopy experience: lessons from two randomised trials. *Gut* **71**, 757–765 (2022).
3. Nam, J. G. *et al.* AI Improves Nodule Detection on Chest Radiographs in a Health Screening Population: A Randomized Controlled Trial. *Radiology* **307**, e221894 (2023).
4. Lång, K. *et al.* Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).
5. Nyangoh Timoh, K. *et al.* A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video. *Surg. Endosc.* **37**, 4298–4314 (2023).
6. Jin, A. *et al.* Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 691–699 (IEEE, 2018).
7. Brown, J. A. *et al.* Video review reveals technical factors predictive of biliary stricture and cholangitis after robotic pancreaticoduodenectomy. *HPB* **23**, 144–153 (2021).
8. Varban, O. A. *et al.* Evaluating the Effect of Surgical Skill on Outcomes for Laparoscopic Sleeve Gastrectomy: A Video-based Study. *Ann. Surg.* **273**, 766–771 (2021).
9. Yiu, A., Lam, K., Simister, C., Clarke, J. & Kinross, J. Adoption of routine surgical video recording: a nationwide freedom of information act request across England and Wales. *EClinicalMedicine* **70**, 102545 (2024).
10. Twinanda, A. P. *et al.* EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans. Med. Imaging* **36**, 86–97 (2016).
11. Nwoye, C. I. *et al.* Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Med. Image Anal.* **78**, 102433 (2022).
12. Schoeffmann, K. *et al.* Cataract-101: video dataset of 101 cataract surgeries. in *Proceedings of the 9th ACM Multimedia Systems Conference* 421–425 (Association for Computing Machinery, New York, NY, USA, 2018).
13. Al Hajj, H. *et al.* CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. *Med. Image Anal.* **52**, 24–41 (2019).
14. Maier-Hein, L. *et al.* Surgical data science – from concepts toward clinical translation. *Med. Image Anal.* **76**, 102306–102306 (2022).
15. Bolmers, M. D. M. *et al.* Discrepancies between Intraoperative and Histological Evaluation of the Appendix in Acute Appendicitis. *J. Gastrointest. Surg.* **24**, 2088–2095 (2020).
16. Rajpurkar, P. *et al.* AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining. *Sci. Rep.* **10**, 3958 (2020).
17. Alvarado, A. A practical score for the early diagnosis of acute appendicitis. *Ann. Emerg. Med.* **15**, 557–564 (1986).
18. Gomes, C. A., Nunes, T. A., Fonseca Chebli, J. M., Junior, C. S. & Gomes, C. C. Laparoscopy grading system of acute appendicitis: new insight for future trials. *Surg. Laparosc. Endosc. Percutan. Tech.* **22**, 463–466 (2012).
19. Yilmaz, A. E. & Demirhan, H. Weighted kappa measures for ordinal multi-class classification performance. *Appl. Soft Comput.* **134**, 110020 (2023).
20. Scheijmans, J. C. G. *et al.* Development and validation of the Scoring System of Appendicitis Severity 2.0. *JAMA Surg.* **159**, 642–649 (2024).
21. Gomes, C. A. *et al.* Acute appendicitis: proposal of a new comprehensive grading system based on clinical, imaging and laparoscopic findings. *World J. Emerg. Surg.* **10**, 60 (2015).
22. Temple, C., Huchcroft, S. & Temple, W. The natural history of appendicitis in adults A prospective study. *Ann. Surg.* **221**, 278–281 (1995).