NEJM
AI

NEJM AI 2024; 1 (2)
DOI: 10.1056/AIoa2300088

ORIGINAL ARTICLE

Artificial Intelligence Identifies Factors Associated with Blood Loss and Surgical Experience in Cholecystectomy

Josiah G. Aklilu, B.S.,[1] Min Woo Sun, B.S.,[1] Shelly Goel, B.S.,[2] Sebastiano Bartoletti, M.D.,[3] Anita Rau, Ph.D.,[1] Griffin Olsen, M.D., M.S.,[4] Kay S. Hung, M.D.,[3] Sophie L. Mintz, C.R.N.P.,[3] Vicki Luong, P.A.-C.,[3] Arnold Milstein, M.D., M.P.H.,[5] Mark J. Ott, M.D.,[6] Robert Tibshirani, Ph.D., F.R.S., F.R.S.C.,[1,7] Jeffrey K. Jopling, M.D.,[8] Eric C. Sorenson, M.D.,[6] Dan E. Azagury, M.D., F.A.C.S.,[3] and Serena Yeung-Levy, Ph.D.[1,2,5,9]

Received: August 18, 2023; Revised: October 16, 2023; Accepted: November 14, 2023; Published: January 25, 2024

Abstract

BACKGROUND Laparoscopic surgery videos offer valuable insights into the intraoperative skills of surgeons. Traditionally, skill assessment has focused on trainees, but analyzing the operative techniques of established surgeons can reveal behaviors that are associated with surgical expertise. Computer vision (CV), a domain of artificial intelligence (AI), facilitates scalable, video-based assessment, enabling the discovery of novel associations between surgical skill and clinical outcomes. For this study, we developed an AI-powered CV model capable of autonomously recognizing fine-grained surgical actions in laparoscopic videos and uncovering associations between these actions and operative blood loss and surgical experience.

METHODS We utilized a dataset of laparoscopic surgical videos from 243 patients who underwent cholecystectomy. We used a subset of these videos to train an AI-powered CV model to recognize 150 fine-grained surgical action triplets (SATs) comprising unique combinations of three components: surgical instruments (16 total), motions (13), and anatomical structures (19). We then used the trained AI model to recognize these SATs in all 243 case videos. We considered estimated blood loss, as reported postoperatively by the performing surgeon, and refined this measure using retrospective video review by experienced surgeons, yielding operative blood loss. We also considered surgeon experience, defined as the number of postresidency years of the operating surgeon. We used a logistic regression model to infer blood loss and surgical experience on the basis of AI-identified surgical actions in the laparoscopic videos. We subsequently analyzed the relationships among surgical actions, operative blood loss, and surgical experience.

Drs. Azagury and Yeung-Levy contributed equally to this article.

The author affiliations are listed at the end of the article.

Dr. Yeung-Levy can be contacted at syyeung@stanford.edu or at Stanford University, 350 Jane Stanford Way, Stanford, CA 94305.

RESULTS The operating surgeons in the video dataset had 8 to 31 years of surgical experience. Estimated operative blood loss among patients ranged from 0 to 175 ml. Our model predicted binary blood loss (low vs. moderate) with an area under the receiver operator characteristic (AUROC) of 0.81 and binary surgical experience (low vs. high) with an AUROC of 0.78. Higher blood loss was significantly associated with increased duration of use of a laparoscopic suction irrigator to dissect the cystic pedicle (P=0.04) and with use of the irrigator to aspirate blood (P=0.03) or irrigate the cystic pedicle (P=0.04). High surgical experience was moderately associated with longer duration of dissection of connective tissue with L-hook electrocautery (P=0.07) and with total duration of the case (P=0.07). High surgical experience was strongly associated with elective cases (P<0.001).

CONCLUSION This study demonstrates the capability of AI CV models to analyze intricate surgical activity in large volumes of video data. By training the CV model on a set of laparoscopic cholecystectomy videos and then deploying it to recognize surgical actions in a larger cohort, we obtained novel and scalable insights without labor-intensive manual review. We specifically demonstrate the capability of AI-powered CV models to correlate surgical experience and technique with intraoperative outcomes (blood loss). (Funded by the Stanford Clinical Excellence Research Center and others.)

Introduction

The impact of operative skill on postoperative outcomes has been widely recognized, with population-based studies establishing associations between surgeons and outcomes such as mortality and readmission.[1-3] However, understanding the precise influence of specific procedural maneuvers on outcomes remains elusive. In 2013, Birkmeyer et al.[4] made significant strides toward that end by establishing a correlation between surgical skill and clinical outcomes in a pivotal study on complication rates following bariatric surgery. That study relied on a manual review of video footage to obtain peer ratings of intraoperative technical skill; the labor-intensive annotation process limited data acquisition to a set of only 20 cases in a year. Despite revealing the connection between lower aggregate peer ratings of surgical skill and higher complication rates and mortality, further exploration has been hindered by the time-consuming nature of manual video annotation.

In the decade since the Birkmeyer et al.[4] study, advancements in artificial intelligence (AI) have revolutionized data analysis capabilities. We hypothesize that AI can unravel more intricate associations between surgeon techniques and patient outcomes compared with prior manual review. The growing prevalence of laparoscopic procedures,[5-10] with their easily recorded and stored video feeds, presents an ideal setting for training deep neural network–based computer vision (CV) models. These models can autonomously recognize and document surgical events and actions from videos to enable the correlation of such events with clinical outcomes across multiple surgeons and institutions on an unprecedented scale.

Although recent works have shown the promise of AI in recognizing surgical actions in videos, they have yet to explore the relationship between such actions and surgical outcomes. To our knowledge, this study introduces the first AI-powered investigation of a large, multi-institutional dataset of surgical videos, focusing on correlating specific surgical actions in laparoscopic cholecystectomy with intraoperative surgical outcomes. We used our AI model to temporally segment laparoscopic videos into identified surgical actions. We then conducted a statistical analysis to reveal which summary features, derived from these surgical actions, best explained reported clinical outcomes and attributes. We focused specifically on estimated blood loss and surgeon experience. Our analysis identified key markers of effective surgery at a level of granularity previously unattainable because of the necessity for laborious, case-by-case, manual review.

Methods

In this retrospective study, we examined patients who underwent laparoscopic cholecystectomy within the Intermountain Health (IH) system between July 2021 and November 2022. Our primary objective was to identify specific surgical maneuvers associated with positive indicators of surgical performance and high surgical skill. Operative blood loss was used as an indicator of good surgical performance (adjusted for severity of cholecystitis), and surgical experience in years was used as a proxy for surgical skill. We gathered laparoscopic video footage and relevant clinical data for each patient. The dataset comprised more than 90 hours of laparoscopic video from 243 patients, 24 surgeons, and 7 hospitals (Table S1 in the Supplementary Appendix). Utilizing a CV AI model, we conducted an in-depth analysis of surgical activity in the videos. Using a statistical model, we then associated the findings with operative blood loss and surgical experience. The study received approval from the Institutional Review Boards of Stanford University and IH.

PARTICIPATING SURGEONS AND ANNOTATORS

To facilitate the training of an AI model to autonomously recognize detailed surgical activity in laparoscopic videos, 12 annotators labeled occurrences of surgical actions in 114 of the 243 patient cases. The annotators — surgical residents, nurse practitioners, and physician assistants from Stanford Health Care (SHC) and IH — participated in two training sessions to become familiar with the cholecystectomy ontology[11-15] utilized in this study (Fig. S1). One attending surgeon from SHC, one from IH, and one from Johns Hopkins University, together with an IH surgical resident, arrived at a consensus on the components of laparoscopic cholecystectomy to be labeled, building on prior work[11,12,16] in AI-based surgical activity recognition. These prior works have illustrated the segmentation of laparoscopic surgery into sequences of shorter surgical action components.[15]

This study defines surgical action triplets (SATs) as unique combinations of three components, consistent with prior work[11,12,17]: the instrument in use, the instrument's motion, and the manipulated anatomy. We identified 16 instruments, 13 motions, and 19 anatomical structures, leading to 150 clinically pertinent combinations (Table S3), filtered to prevent invalid combinations such as "Suction Irrigation, Clip, Gallbladder." Our team of expert clinicians and highly trained research personnel provided manual labels of fine-grained SATs on 114 of the 243 laparoscopic videos available, yielding 4057 minutes of labeled video, encompassing 8470 labeled actions. Using the expert annotations, we trained a state-of-the-art AI-powered CV model to identify actions from video. Details of this model are provided in Figure 1 and Section 6 of the Supplementary Appendix (see Figs. S3-S7). We subsequently used the model to detect the prevalent SAT

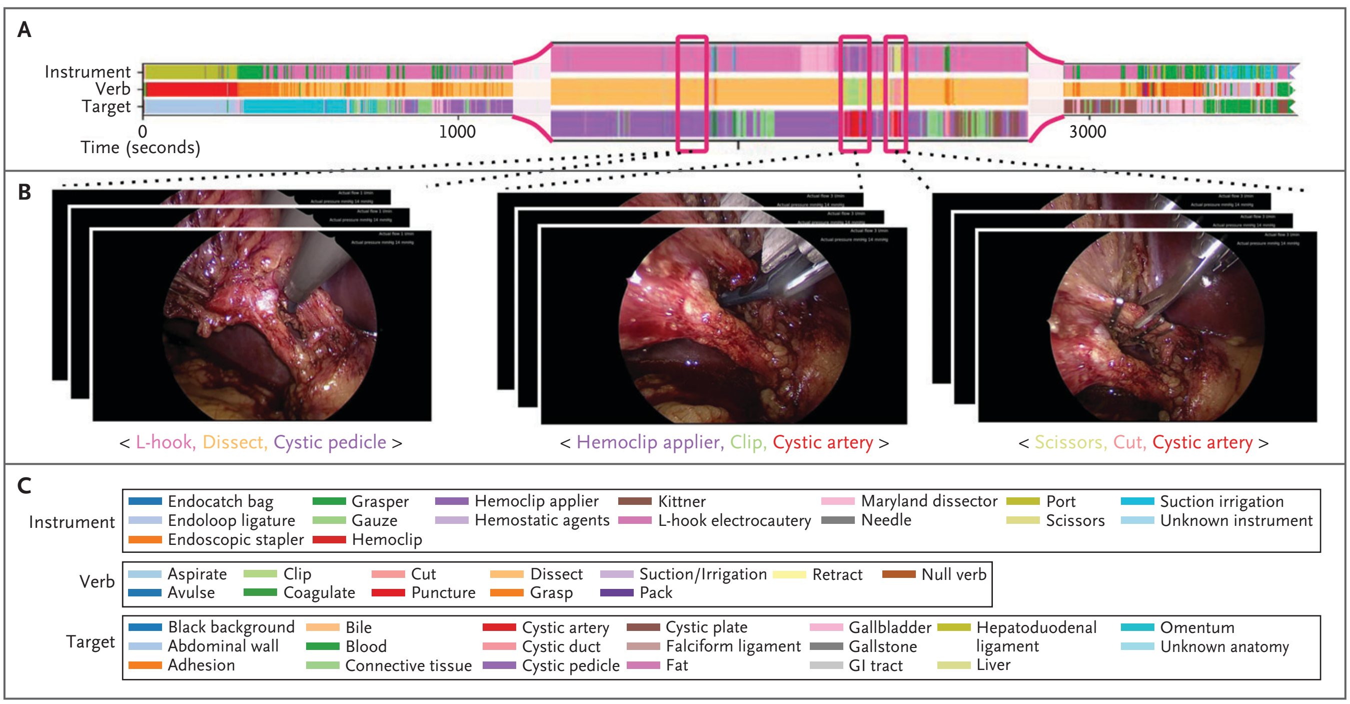

Figure 1. Automated Computer Vision Analysis of Laparoscopic Cholecystectomy Videos.

Panel A shows an example timeline of surgical action triplets (SATs; i.e., instrument, verb, target interactions) detected by our model for one laparoscopic cholecystectomy case (a portion of the timeline is enlarged for visibility). Panel B shows laparoscopic video frames at several time points during the case along with the corresponding detected SATs. The legend in Panel C indicates the set of SAT components detected by our model. GI denotes gastrointestinal.

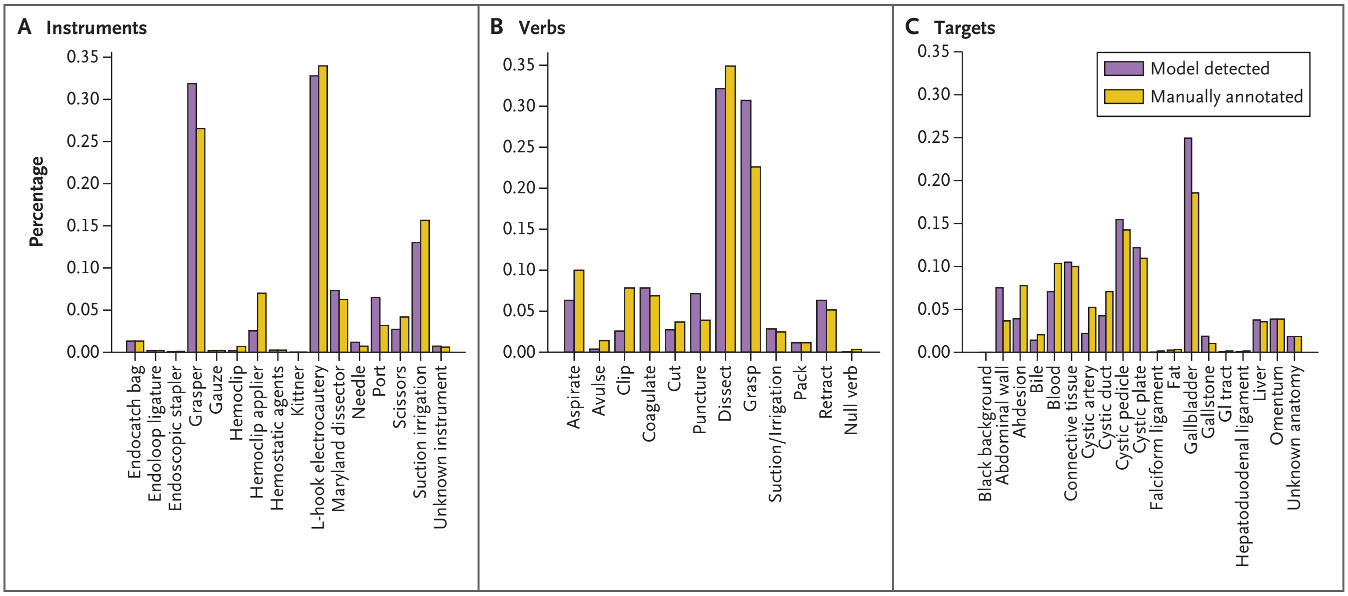

Figure 2. Distribution of Surgical Action Triplets (SATs) Detected by Our Model on All Cases versus the Distribution of Human Annotations on the Labeled Subset.

This figure shows a comparison of our model's detected SAT labels on the entire cohort with the ground truth annotations provided by human annotators on a subset of our cohort. Panel A shows the percentage of SAT labels with each laparoscopic instrument. Panel B shows the percentage of SAT labels involving a particular motion. Panel C shows the percentage of SAT labels with particular anatomy as the target of the action. We demonstrated that our artificial intelligence model is able to capture a similar distribution of key surgical elements as a human annotator. GI denotes gastrointestinal.

at every moment in the original 114 cases and the remaining 129 videos, generating an additional 5792 minutes of SAT annotations without requiring any human labor (Fig. 2).

We calculated intermediate, human-interpretable derivatives from these actions (e.g., average duration of an activity) and used them to fit a statistical model for estimating operative blood loss and surgical expertise. Operative blood loss was estimated by the surgeon in the postoperative report and refined through a standardized review of the operative video by an independent surgeon. We also recognized that disease severity acts as a confounder in appropriately assessing surgical skill. Therefore, we adjusted for severity using the Parkland Grading Scale[18] (PGS). PGS ratings (see Table S2 and Fig. S2), labeled by two attending surgeons and two surgical residents for all 243 cases, were used as the primary control for case complexity.

Separately, as an additional clinical covariate, a surgical resident labeled binary critical view of safety (CVS) achievement in 69 cases, which given annotation resources, was sufficient for training an effective model to recognize CVS achievement in all cases. Through this fine-grained association of surgical actions and clinical covariates with relevant outcomes, we were able to shed light on the relationship between positive patient outcomes and surgeon technique, paving the way for automated and standardized surgical skill assessment.

OUTCOMES

The primary outcomes in this study were operative blood loss and surgical experience. We categorized operative blood loss into two groups: minimal blood loss (≤10 ml) and significant blood loss (>10 ml). This threshold was chosen to capture the difference in the nature of bleeding in laparoscopic cholecystectomy. For example, 100 ml of blood loss may be reflective of an adverse intraoperative event, such as involuntary injury of a major blood vessel, even if the absolute magnitude of this blood loss does not lead to any significant postoperative consequence. In addition, it is standard practice for 10 ml of blood to be marked in the operative report if there is minimal bleeding. Thus, after an extensive review of operative blood loss in our cohort, we chose 10 ml or less of blood loss as an indicator that no significant bleeding was present in the case, whereas larger values indicated technical error and/or case severity.

Similarly, we divided surgical experience into two categories: low experience and high experience, with the threshold set at 15 years of experience since commencing general surgery residency. All supervising attending surgeons in our cohort perform at least 50 laparoscopic cholecystectomy procedures each year. The group of surgeons in the hospitals included in our study cohort perform over 3000 laparoscopic cholecystectomy procedures each year. A secondary outcome was the achievement of the CVS, which was used as a covariate in a statistical model of operative blood loss.

STATISTICAL ANALYSIS

Surgical actions offer a detailed perspective of surgeries yet are not directly interpretable. Consequently, we processed the detected SATs into interpretable measures. These serve as a proxy for components of widely adopted surgical skill rating systems, such as Global Operative Assessment of Laparoscopic Skills (GOALS)[19] and Objective Structured Assessment of Technical Skill (OSATS).[20] For example, the GOALS system includes a scale that measures efficiency, with low ratings indicating uncertain, inefficient efforts characterized by numerous tentative movements and constant focus shifts. To encapsulate this surgical behavior, we computed the frequency of specific surgical actions, which, unlike total duration, captures focus shifts. The OSATS scale assesses operative flow, which we capture through the average idle time between detected actions.

Measures considered in this study include total video duration, total durations of each unique SAT, the time until the first clipping and cutting actions, the total time between the first and last port insertion, the average idle time between labeled actions, the number of times an L-hook electrocautery or grasper is used within 10 seconds of bleeding, the number of times a clip is applied immediately before and after an instance of bleeding, the number of L-hook or grasper dissections of the gallbladder or cystic pedicle immediately before gallstone or bile spillage, and the frequency of actions 1 minute before an adverse bleeding or bile spillage event. All durations and time periods were measured in seconds (Table S4). Additional clinical attributes were included as covariates: elective status (urgent/nonurgent), preoperative diagnosis, assisting resident level, PGS severity, and achievement of CVS. We controlled for disease severity in our analysis using expert-labeled PGS, which grades acute cholecystitis severity on the basis of intraoperative images assessing inflammation and anatomy (Supplementary Appendix, Section 4). We also trained our algorithm to detect CVS achievement as an additional clinical covariate. Following prior work on automated CVS detection,[21,22] we constructed a CV model capable of capturing CVS achievement (Supplementary Appendix, Section 4).

We conducted a comprehensive statistical analysis of the SATs detected by our AI model. We fit four multivariate logistic regression models with an L1 or Lasso penalty[23] on features derived from SATs and relevant clinical features to classify the minimal versus significant operative blood loss in all cases, mild PGS 1 to 2 cases only, moderate to severe PGS 3 to 5 cases only, and low versus high surgical experience in all cases. We used 10-fold crossvalidation to find the optimal shrinkage factor lambda for the Lasso penalty. The receiver operator characteristic curves were constructed using predictions from the 10-fold crossvalidation procedure. Only predictions when the samples were held out in the test folds were used. Lasso regularization aids in preventing overfitting and reduces the number of variables from the large number of SATs and other derived measures to a meaningful subset. To account for feature selection, we used selective inference to correct the P value and confidence interval estimation for the model coefficients.[24] Note that odds ratios depend on the chosen unit of measurement (1 second) and scale with the duration of an observation. This is of particular relevance when considering that laparoscopic cholecystectomy operative times vary from tens of minutes to multiple hours.

Results

Our study dataset comprised 243 unique patient cases of laparoscopic cholecystectomy from IH. Among the 243 patients, 72 had a clinician-reported blood loss of 10 ml or less, and 171 had blood loss greater than 10 ml. Surgeons had 15 years of experience or less in 128 cases and more than 15 years in the remaining 115 cases. Annotators provided manual labels of fine-grained SATs on 114 of the 243 laparoscopic videos. This subset was used to train our AI model, which we then used to detect SATs in all 243 cases. A surgical resident labeled binary CVS achievement in 69 cases; these labels were used to train a separate model for recognizing CVS achievement for all cases. PGS

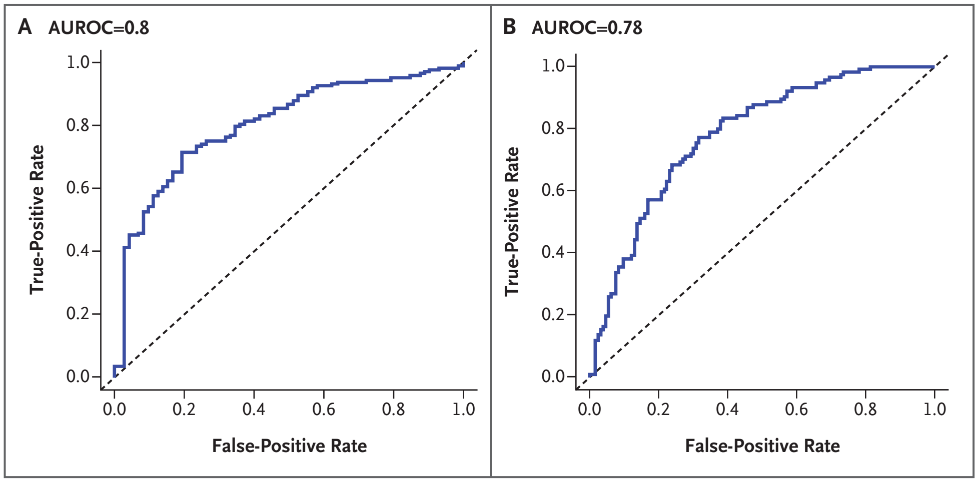

Figure 3. Receiver Operating Characteristic (ROC) Curves for the Operative Blood Loss Model and the Surgeon Experience Model.

Panel A shows the ROC curve for our statistical model fitted on clinical covariates, such as disease severity (as measured by the Parkland Grading Scale) and artificial intelligence–derived video-based features for classifying operative blood loss. Panel B shows the ROC curve for our model of surgeon experience. Twenty covariates were included in the final fitted model for the full cohort, which showed a strong association between higher blood loss and longer durations of the laparoscopic suction irrigator aspirating blood (P=0.02) or suction-irrigating blood (P=0.04) (Table 1). AUROC denotes area under the receiver operating characteristic.

disease severity ratings, labeled by two attending surgeons and two surgical residents, were used for all 243 cases.

OPERATIVE BLOOD LOSS

We used our CV model to detect SATs in all 243 patient videos. We then extracted features from these AI-detected surgical behaviors for use in a statistical model of outcomes. A multivariate logistic regression model for estimating blood loss using these features achieved an area under the receiver operator characteristic (AUROC) of 0.81 on the full cohort, and regression models for the PGS 1 to 2 and PGS 3 to 5 cohorts both achieved an AUROC of 0.73 (Fig. 3).

In cases with a gallbladder of PGS 1 or 2, higher operative blood loss was strongly associated with longer durations of the suction irrigator aspirating blood (P=0.01) and coagulation of blood through L-hook electrocautery (P=0.02). Additionally, in cases with PGS 3 disease severity or higher, higher intraoperative blood loss was found to be significantly associated with longer usage of the suction irrigator for dissection of the cystic pedicle (P=0.04). Moreover, in moderate to severe cases, higher blood loss was associated with longer usage of the irrigator to irrigate the cystic pedicle (P=0.04) and with longer durations of the irrigator aspirating blood (P=0.03).

SURGICAL EXPERIENCE

For surgical experience, we fit another logistic regression model using the AI-derived features, which achieved an AUROC of 0.78. Higher surgical experience was strongly associated with the elective status of the cholecystectomy (P<0.001), whereas lower experience was moderately associated with longer duration of dissection of connective tissue through L-hook electrocautery (P=0.07) and total duration of the case (P=0.07).

Discussion

In this study, we analyzed 243 laparoscopic cholecystectomy cases from multiple institutions and developed a CV model to detect surgical activity. We then extracted features from the detected actions and used a statistical model to examine the association between surgical activity and outcomes. Our cohort of laparoscopic cases with extensively labeled surgical actions represents the most extensive dataset of unique patient cases examined with AI. Whereas previous studies have explored the development of AI models for recognizing surgical actions,[11,12,25] most have been restricted to datasets derived from the Cholec80 laparoscopic surgery dataset,[15] which consists of only 80 unique videos and lacks postoperative data.

Derivative datasets such as CholecT40 and CholecT45,[17,25,26] which have been used in work on recognizing SATs, are limited to less than 50 unique patients. None of these prior studies have leveraged the automation capabilities of CV to analyze a large number of surgical videos and investigate relationships between surgical activity and clinical outcomes.

We demonstrate that our framework is capable of systematically uncovering associations between surgical activity and intraoperative adverse outcomes and can generalize automated analysis capabilities across laparoscopic videos from multiple institutions in a large health care network. A unique feature of this study is the metadata that accompany the laparoscopic video data. Surgical experience and the skills of practicing surgeons are difficult to assess from most existing datasets.[17,25] Many, if not all, video datasets may include portions of the operation performed by trainees. Our unique dataset, in collaboration with multiple institutions, allowed us to identify a subset of cases without trainee involvement.

A decade since the seminal work of Birkmeyer et al.[4] demonstrated manual video review to assess surgical skill, our work revisits surgical skill assessment through the lens of modern AI. Our findings demonstrate the value of using AI to recognize fine-grained surgical behaviors in large numbers of laparoscopic videos to predict surgical outcomes and assess surgical skill. In our cohort, our AI model indicates that, in cases classified as moderate to severe on the basis of disease severity, longer usage durations of the suction irrigator for cystic pedicle dissection are significantly associated with higher intraoperative blood loss. This implies a relationship between irrigator usage and case difficulty because these cases tend to be more inflamed and prone to bleeding, prompting the use of the irrigator as a less traumatic blunt dissector. In milder cases (PGS 1 to 2), our model indicates that aspiration or coagulation of blood is strongly correlated with high blood loss, a confirmatory finding illustrating the potential for CV to effectively and autonomously identify expected surgical maneuvers in response to adverse intraoperative events, such as bleeding. Additionally, we find that AI can detect clear associations between intraoperative outcomes and identifiers of disease severity. Specifically, more inflamed gallbladders, as measured by PGS, substantially increase the risk of high blood loss, whereas CVS achievement decreases the odds of blood loss (Table 1). Although these findings did not reach statistical significance, we believe that this is in part because of our use of selective inference, which makes the threshold for significance more stringent, and that those findings will be confirmed with larger datasets in the future. These findings have clinical importance because the CVS method was designed to minimize the risk of common bile duct injury, a life-altering complication, during laparoscopic cholecystectomy. No prior findings, to our knowledge, showed an increase in blood loss with poor CVS achievement, but our findings point to the fact that the more thorough dissection required to achieve CVS is also protective against increased blood loss.

With respect to surgical experience, a statistical model can reliably utilize known correlates of surgical experience with case information, such as elective status (e.g., nonelective cases are less routinely performed by experienced practitioners, who tend to perform more elective surgeries). Additionally, the model can also identify associative features detected by AI that provide insight into relationships between surgical maneuvers and experience. For example, our model finds that longer time spent using the L-hook electrocautery for dissection of the gallbladder off the liver bed is moderately associated with less surgical experience. This is consistent with notions of surgical proficiency because the time taken to address gallbladder removal can be indicative of a novice or expert surgeon. These types of insights could significantly impact surgical practice because AI-enabled, large-scale analysis could reveal hidden relationships between surgical actions and proficiency (Table 2).

These preliminary findings have direct implications for initiatives like the American College of Surgeons' National Surgical Quality Improvement Program (NSQIP).[27] Launched in 1994, NSQIP aims to systematically gather preoperative, intraoperative, and postoperative data for patients undergoing surgery at participating institutions. This multi-institutional benchmarking facilitates efforts to decrease postoperative mortality and morbidity.[27] As storage and analysis of laparoscopic video data become increasingly available at these institutions, incorporating objective intraoperative features derived from AI could serve as a natural extension to electronic health record–based predictions of complications.[28] Current NSQIP data from individual patient records lack the detailed intraoperative analysis that AI can potentially provide on a large scale. By integrating AI-derived, objective intraoperative features, existing NSQIP data could be enriched with granular intraoperative

Table 1. Selected Features Associated with Operative Blood Loss for All Cases, Mild Severity Cases, and Moderate- to High-Severity Cases*

Feature	Odds Ratio	P Value	CI (Low)	CI (High)
Full Cohort				
PGS	1.2087	0.16	0.6703	2.0374
Endoloop Ligature Puncture Abdominal Wall	1.0366	0.31	0.7580	1.3573
Port Retract Abdominal Wall	1.0121	0.42	0.7588	1.3612
Suction Irrigation Suction/Irrigation Blood	1.0103	0.04	0.9949	1.0746
Maryland Dissector Grasp adhesion	1.0088	0.88	0.0000	1.8994
Grasper Grasp Cystic Pedicle	1.0073	0.50	0.8272	1.0769
L-hook Electrocautery Coagulate Gallbladder	1.0053	0.73	0.3127	1.4376
Maryland Dissector Grasp Gallbladder	1.0052	0.60	0.4734	1.5232
Grasper Retract Liver	1.0048	0.16	0.9817	1.0394
Suction Irrigation Aspirate Blood	1.0041	0.02	1.0009	1.0267
Hemoclip Applier Clip Cystic Artery	1.0038	0.55	0.8960	1.0450
L-hook Electrocautery Coagulate Connective Tissue	1.0026	0.54	0.8922	1.0844
Suction Irrigation Dissect Cystic Pedicle	1.0025	0.31	0.9815	1.0261
L-hook Electrocautery Coagulate Blood	1.0018	0.83	0.7374	1.0464
Maryland Dissector Grasp Cystic Duct	1.0012	0.79	0.6549	1.0759
Total Duration	1.0004	0.49	0.9990	1.0013
L-hook Electrocautery Coagulate Cystic Plate	0.9997	0.43	0.9979	1.0016
L-hook Electrocautery Dissect Adhesion	0.9990	0.73	0.9845	1.0835
Grasper Grasp Fat	0.9757	0.71	0.6765	4.4333
CVS	0.6720	0.17	0.0266	2.4193
Mild Severity Cases (PGS 1–2)				
L-hook Electrocautery Coagulate Blood	1.0446	0.02	1.0100	2.7803
Maryland Dissector Grasp Gallbladder	1.0404	0.14	0.5832	44.1470
Maryland Dissector Dissect Gallbladder	1.0391	0.81	0.0000	860.0842
Port Retract Abdominal Wall	1.0296	0.14	0.9050	1.5246
L-hook Electrocautery Coagulate Cystic Pedicle	1.0259	0.74	0.1936	1.3443
Grasper Grasp Liver	1.0206	0.51	0.5520	1.4687
Grasper Grasp Cystic Pedicle	1.0108	0.70	0.4584	1.1919
Suction Irrigation Aspirate Blood	1.0082	0.01	1.0032	1.0333
L-hook Electrocautery Coagulate Liver	1.0080	0.07	0.6890	1.1401
Maryland Dissector Dissect Adhesion	1.0078	0.90	0.0000	2.8643
Suction Irrigation Dissect Cystic Pedicle	1.0032	0.28	0.9661	1.0717
Grasper Retract Liver	1.0026	0.67	0.8375	1.0786
Grasper Grasp Gallbladder	1.0006	0.39	0.9762	1.0149
Endocatch Bag Pack Gallbladder	0.9985	0.70	0.9562	1.1610
Suction Irrigation Dissect Connective Tissue	0.9434	0.64	0.6972	4.4759
Grasper Dissect Connective Tissue	0.5004	0.61	0.0197	260,314.0929
Moderate- to High-Severity Cases (PGS 3–5)				
Suction Irrigation Suction/Irrigation Cystic Pedicle	1.2100	0.04	0.9477	2.0717
Scissors Cut Gallbladder	1.0163	0.72	0.1554	1.3154
Suction Irrigation Dissect Cystic Pedicle	1.0037	0.04	0.9990	1.0151
Suction Irrigation Aspirate Blood	1.0016	0.03	0.9999	1.0071
Grasper Retract Liver	1.0006	0.72	0.9512	1.0080
L-hook Electrocautery Dissect Connective Tissue	0.9995	0.34	0.9848	1.0104
Grasper Grasp Connective Tissue	0.9946	0.67	0.8488	1.8145

* The odds ratios for the features selected (i.e., nonzero coefficients) were determined by the Lasso logistic regression model for measuring operative blood loss in the full cohort, the PGS 1–2 group, and the PGS 3–5 group. CI denotes confidence interval; CVS, critical view of safety; and PGS, Parkland Grading Scale.

Table 2. Selected Features Associated with Surgeon Experience*				
Feature	Odds Ratio	P Value	CI (Low)	CI (High)
Elective	2.3582	<0.001	1.8984	9.1678
Total Duration	0.9999	0.07	0.9995	1.0001
L-hook Electrocautery Dissect Cystic Pedicle	0.9997	0.56	0.9967	1.0087
L-hook Electrocautery Coagulate Connective Tissue	0.9994	0.84	0.9821	1.2866
Suction Irrigation Suction/Irrigation Blood	0.9989	0.41	0.9809	1.0256
L-hook Electrocautery Dissect Connective Tissue	0.9988	0.07	0.9915	1.0011
Assist	0.9211	0.63	0.1435	78.5546

* The odds ratios for the features selected were determined by the Lasso logistic regression model for measuring surgeon experience in the full cohort. CI denotes confidence interval.

insights, bolstering both surgeons' and hospitals' capacity to understand and prevent avoidable complications.

Despite these promising results, our study has limitations. The primary limitation is the exclusive focus on cholecystectomy; the approach should be validated for other laparoscopic procedures to demonstrate its clinical utility. Ultimately, the goal is to identify intraoperative factors that lead to clinically relevant postoperative consequences, and this is challenging for cholecystectomies given the very low complication rates of this procedure. We also used surgical experience as a surrogate for surgical proficiency, which may represent a bias, because more clinical experience does not necessarily equate to higher laparoscopic proficiency. Additionally, although laparoscopic video is abundant, patient privacy concerns and data collection challenges could hinder the development of effective deep learning models for surgical activity recognition.

In conclusion, our study highlights the potential of AI in facilitating detailed and large-scale analyses of laparoscopic surgical videos and in uncovering intricate relationships between surgical activity and operative outcomes. An AI model trained to recognize fine-grained surgical actions in laparoscopic cholecystectomy produced interpretable features associated with adverse intraoperative events, like bleeding. This study represents the first investigation into the link between AI-derived, visual surgical activity features and operative outcomes in laparoscopic surgery. By examining the activity of surgical practitioners at scale, we can gain insights into key markers of successful surgery in one of the most common procedures. This study not only confirms the potential for AI models to scale and reveal new insights on large surgical cohorts, it also may inspire new research into the use of surgical videos to enhance surgical training and reduce complications.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Partially supported by the Stanford Clinical Excellence Research Center, Stanford Medicine, and Intermountain Healthcare through a collaboration grant and the National Science Foundation Future of Work Investment (award number 2026498). Dr. Yeung-Levy is a Chan Zuckerberg Biohub–San Francisco Investigator. Dr. Rau acknowledges support from the Isackson Family Foundation and the Stanford Head and Neck Surgery Research Fund. Mr. Aklilu acknowledges support from the Stanford Graduate Fellowship. Mr. Sun acknowledges support from the National Library of Medicine (grant number 5T15LM007033-40). Dr. Tibshirani acknowledges support from the National Institutes of Health (grant number 5R01 EB001988-16) and the National Science Foundation (grant number 19 DMS1208164).

Author Affiliations

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA

[2] Department of Computer Science, Stanford University, Stanford, CA

[3] Department of Surgery, Stanford University School of Medicine, Stanford, CA

[4] Intermountain Healthcare Delivery Institute, Intermountain Healthcare, Murray, UT

[5] Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA

[6] Department of Surgery, Intermountain Medical Center, Murray, UT

[7] Department of Statistics, Stanford University, Stanford, CA

[8] Department of Surgery, Johns Hopkins Medicine, Baltimore

[9] Department of Electrical Engineering, Stanford University, Stanford, CA

References

1. Wallis CJD, Jerath A, Coburn N, et al. Association of surgeon-patient sex concordance with postoperative outcomes. JAMA Surg 2022;157:146-156. DOI: 10.1001/jamasurg.2021.6339.

2. Wallis CJ, Ravi B, Coburn N, Nam RK, Detsky AS, Satkunasivam R. Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. BMJ 2017;359:j4366. DOI: 10.1136/bmj.j4366.

3. Satkunasivam R, Klaassen Z, Ravi B, et al. Relation between surgeon age and postoperative outcomes: a population-based cohort study. CMAJ 2020;192:E385-E392. DOI: 10.1503/cmaj.190820.
4. Birkmeyer JD, Finks JF, O'Reilly A, et al.. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434-1442. DOI: 10.1056/NEJMsa1300625.
5. Fuchs KH. Minimally invasive surgery. Endoscopy 2002;34:154-159. DOI: 10.1055/s-2002-19857.
6. Nguyen NT, Nguyen B, Shih A, Smith B, Hohmann S. Use of laparoscopy in general surgical operations at academic centers. Surg Obes Relat Dis 2013;9:15-20.10.1016/j.soard.2012.07.002.
7. Xiong W, Xu Y, Chen T, et al. Laparoscopic *vs.* open surgery for gastrointestinal stromal tumors of esophagogastric junction: a multicenter, retrospective cohort analysis with propensity score weighting. Chin J Cancer Res 2021;33:42-52. DOI: 10.21147/j.issn.1000-9604.2021.01.05.
8. Xiangfei M, Yinzhe X, Yingwei P, Shichun L, Weidong D. Open versus laparoscopic hepatic resection for hepatocellular carcinoma: a systematic review and meta-analysis. Surg Endosc 2019;33:2396-2418. DOI: 10.1007/s00464-019-06781-3.
9. Agha R, Muir G. Does laparoscopic surgery spell the end of the open surgeon? J R Soc Med 2003;96:544-546. DOI: 10.1177/014107680309601107.
10. Rassweiler J, Frede T, Guillonneau B. Advanced laparoscopy. Eur Urol 2002;42:I-XII. DOI: 10.1016/S0302-2838(02)00366-4.
11. Nwoye CI, Yu T, Gonzalez C, et al. Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med Image Anal 2022;78:102433. DOI: 10.1016/j.media.2022.102433.
12. Sharma S, Nwoye CI, Mutter D, Padoy N. Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. Int J Comput Assist Radiol Surg 2023;18:1053-1059. DOI: 10.1007/s11548-023-02914-1.
13. Kiyasseh D, Ma R, Haque TF, et al. A vision transformer for decoding surgeon activity from surgical videos. Nat Biomed Eng 2023;7:780-796. DOI: 10.1038/s41551-023-01010-8.
14. Garrow CR, Kowalewski K-F, Li L, et al. Machine learning for surgical phase recognition: a systematic review. Ann Surg 2021;273:684-693. DOI: 10.1097/SLA.0000000000004425.
15. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 2017;36:86-97. DOI: 10.1109/TMI.2016.2593957.
16. Jin A, Yeung S, Jopling J, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. 2018 IEEE Winter Conference on Applications of Computer Vision 2018:691-699. DOI: 10.1109/WACV.2018.00081.
17. Nwoye CI, Alapatt D, Yu T, et al. CholecTriplet2021: a benchmark challenge for surgical action triplet recognition. Med Image Anal 2023;86:102803. DOI: 10.1016/j.media.2023.102803.
18. Madni TD, Leshikar DE, Minshall CT, et al. The Parkland grading scale for cholecystitis. Am J Surg 2018;215:625-630. DOI: 10.1016/j.amjsurg.2017.05.017.
19. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 2005;190:107-113. DOI: 10.1016/j.amjsurg.2005.04.004.
20. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997;84:273-278. DOI: 10.1046/j.1365-2168.1997.02502.x.
21. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Ann Surg 2022;275:955-961. DOI: 10.1097/sla.0000000000004351.
22. Mascagni P, Alapatt D, Laracca GG, et al. Multicentric validation of EndoDigest: a computer vision platform for video documentation of the critical view of safety in laparoscopic cholecystectomy. Surg Endosc 2022;36:8379-8386. DOI: 10.1007/s00464-022-09112-1.
23. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996;58:267-288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.
24. Taylor J, Tibshirani RJ. Statistical learning and selective inference. Proc Natl Acad Sci USA 2015;112:7629-7634. DOI: 10.1073/pnas.1507583112.
25. Nwoye CI, Yu T, Sharma S, et al. CholecTriplet2022: Show me a tool and tell me the triplet — an endoscopic vision challenge for surgical action triplet detection. Med Image Anal 2023;89:102888. DOI: 10.1016/j.media.2023.102888.
26. Nwoye CI, Gonzalez C, Yu T, et al. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. Medical Image Computing and Computer Assisted Intervention — MICCAI 2020 2020:364-374. DOI: 10.1007/978-3-030-59716-0_35.
27. ACS. History. 2023 (https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/about-acs-nsqip/history).
28. Ren Y, Loftus TJ, Datta S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA Netw Open 2022;5:e2211973. DOI: 10.1001/jamanetworkopen.2022.11973.