JAMA Surgery | Original Investigation

Automatic Surgical Skill Assessment System Based on Concordance of Standardized Surgical Field Development Using Artificial Intelligence

Takahiro Igaki, MD; Daichi Kitaguchi, MD; Hiroki Matsuzaki, MS; Kei Nakajima, MD; Shigehiro Kojima, MD, PhD; Hiro Hasegawa, MD, PhD; Nobuyoshi Takeshita, MD, PhD; Yusuke Kinugasa, MD, PhD; Masaaki Ito, MD, PhD

Invited Commentary

Supplemental content

IMPORTANCE Automatic surgical skill assessment with artificial intelligence (AI) is more objective than manual video review–based skill assessment and can reduce human burden. Standardization of surgical field development is an important aspect of this skill assessment.

OBJECTIVE To develop a deep learning model that can recognize the standardized surgical fields in laparoscopic sigmoid colon resection and to evaluate the feasibility of automatic surgical skill assessment based on the concordance of the standardized surgical field development using the proposed deep learning model.

DESIGN, SETTING, AND PARTICIPANTS This retrospective diagnostic study used intraoperative videos of laparoscopic colorectal surgery submitted to the Japan Society for Endoscopic Surgery between August 2016 and November 2017. Data were analyzed from April 2020 to September 2022.

INTERVENTIONS Videos of surgery performed by expert surgeons with Endoscopic Surgical Skill Qualification System (ESSQS) scores higher than 75 were used to construct a deep learning model able to recognize a standardized surgical field and output its similarity to standardized surgical field development as an AI confidence score (AICS). Other videos were extracted as the validation set.

MAIN OUTCOMES AND MEASURES Videos with scores less than or greater than 2 SDs from the mean were defined as the low- and high-score groups, respectively. The correlation between AICS and ESSQS score and the screening performance using AICS for low- and high-score groups were analyzed.

RESULTS The sample included 650 intraoperative videos, 60 of which were used for model construction and 60 for validation. The Spearman rank correlation coefficient between the AICS and ESSQS score was 0.81. The receiver operating characteristic (ROC) curves for the screening of the low- and high-score groups were plotted, and the areas under the ROC curve for the low- and high-score group screening were 0.93 and 0.94, respectively.

CONCLUSIONS AND RELEVANCE The AICS from the developed model strongly correlated with the ESSQS score, demonstrating the model's feasibility for use as a method of automatic surgical skill assessment. The findings also suggest the feasibility of the proposed model for creating an automated screening system for surgical skills and its potential application to other types of endoscopic procedures.

JAMA Surg. doi:10.1001/jamasurg.2023.1131
Published online June 7, 2023.

Author Affiliations: Surgical Device Innovation Office, National Cancer Center Hospital East, Kashiwanoha, Kashiwa, Chiba, Japan (Igaki, Kitaguchi, Matsuzaki, Nakajima, Kojima, Hasegawa, Takeshita, Ito); Department of Colorectal Surgery, National Cancer Center Hospital East, Kashiwanoha, Kashiwa, Chiba, Japan (Igaki, Kitaguchi, Nakajima, Kojima, Hasegawa, Takeshita, Ito); Department of Gastrointestinal Surgery, Tokyo Medical and Dental University Graduate School of Medicine, Yushima, Bunkyo-Ku, Tokyo, Japan (Igaki, Kinugasa).

Corresponding Author: Masaaki Ito, MD, PhD, Surgical Device Innovation Office, National Cancer Center Hospital East, 6-5-1, Kashiwanoha, Kashiwa, Chiba 277-8577, Japan (maito@east.ncc.go.jp).

Surgical skill is one of the most important factors directly involved in patient outcomes.[1,2] Several surgical skill assessment studies have been conducted,[3-13] among which the classic method is the system most widely used. In the classic method, an expert surgeon evaluates a surgical trainee's operation based on a surgical skill assessment tool, such as the Objective Structured Assessment of Technical Skill or the Global Operative Assessment of Laparoscopic Skills.[3,4] The validity of these tools as an objective assessment of surgical performance has been evaluated in a variety of environments.[14,15] However, these assessments require the time and resources of expert surgeons or trained raters and rely on the judgments of individuals, in which subjectivity is inevitable. Therefore, automating the surgical skill assessment with artificial intelligence (AI), which is more objective and reduces human burden, has been proposed to address these problems.

In recent years, AI, especially deep learning, has greatly contributed to the medical field, automating human cognitive functions, such as problem-solving and decision-making.[16,17] Artificial intelligence can learn the features of objects in images and videos to make predictions for new data to identify, detect, and classify. There have been many reports regarding the use of deep learning not only in diagnostic areas, such as mammography for breast cancer and polyp detection for endoscopy,[18,19] but also in surgical areas, such as prostate, uterine, and surgical tool detection in surgery.[20,21] From this point of view, the introduction of AI is expected to further promote skill assessment.

Standardization of the surgical field development is one of the most important surgical skills, and it has been the target of recent skill assessment tools regarding the accuracy of surgical field exposure and dissection.[22-24] In the critical view of safety (CVS) in cholecystectomy,[25] a consensus on surgical field development has been achieved to ensure the safety of surgical procedures.[26] A CVS in laparoscopic right hemicolectomy was also proposed to secure a standardized procedure that minimizes technical hazards and facilitates teaching.[27] The laparoscopic sigmoid colon resection (Lap-S) procedure was divided into step-by-step phases based on a clear definition proposed in previous studies by some of us.[28,29] There is no clear critical view report in Lap-S; however, based on the previously defined phase classification,[28,29] each phase has a specific surgical field.

Therefore, we conducted this study to develop a deep learning model that can recognize standardized surgical fields for each phase in Lap-S. We also evaluated the feasibility of an automatic surgical skill assessment based on concordance of the standardized surgical field development using the proposed deep learning model.

Key Points

Question Can a deep learning model be used to recognize the standardized surgical fields in laparoscopic sigmoid colon resection and perform automatic surgical skill assessment?

Findings In this diagnostic study reviewing 120 intraoperative videos of laparoscopic colorectal surgery in Japan, the artificial intelligence confidence score from the developed model strongly correlated with the Endoscopic Surgical Skill Qualification System score.

Meaning These results suggest the model's feasibility for use as a method of automatic surgical skill assessment and the feasibility of an automated screening system for surgical skills.

Methods

Data Set

The protocol for this diagnostic study was reviewed and approved by the ethics committee of National Cancer Center Hospital East. Informed consent was obtained from all participants in the form of an opt-out option following the Good Clinical Practice Guidelines of the Ministry of Health and Welfare of Japan. The study conformed to the provisions of the Declaration of Helsinki[30] in 1964 (as revised in Brazil in 2013). We followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

In this study, 650 intraoperative videos of Lap-S that were submitted to the Japan Society for Endoscopic Surgery (JSES) for the Endoscopic Surgical Skill Qualification System (ESSQS) between August 2016 and November 2017 were collected and used as training data. The JSES introduced the ESSQS in 2004, and it has since contributed greatly to maintaining and improving the skills of surgeons.[31-33] To receive certification as a technically qualified surgeon based on the ESSQS, candidates submit nonedited videos for examination, which are assessed by 2 judges in a double-blind manner, using strict criteria for the qualification of candidates (eTables 1 and 2 in Supplement 1). The surgeons' technical skills are scored from 0 to 100, with higher scores indicating greater surgical skill. Only 20% to 30% of examinees are considered suitable for certification, and fewer than 10% of surgeons in Japan are ESSQS certified.[34] Our data set also included the scores assigned to each video by the ESSQS.

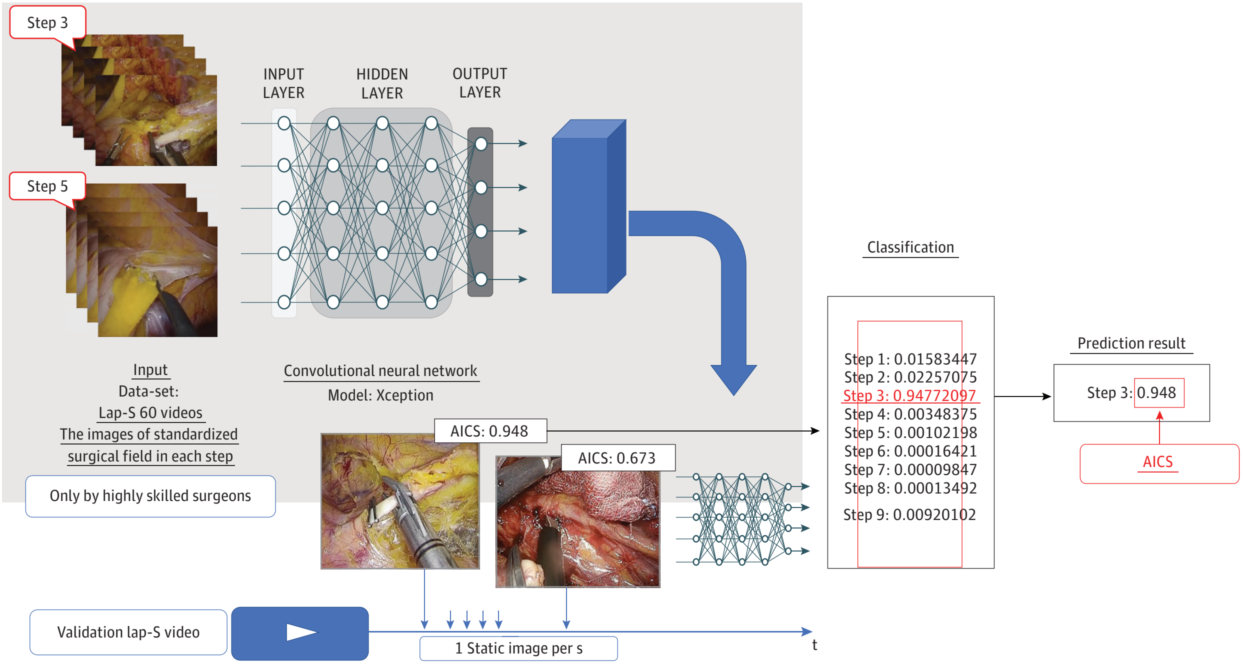
Recognition Model of the Standardized Surgical Field

We extracted 60 videos of Lap-S with ESSQS scores higher than 75 points to construct a model to recognize standardized surgical field development. We incorporated the ESSQS score as an inclusion criterion because the ESSQS assesses the quality of surgical fields; therefore, it can be assumed that a surgeon with a high ESSQS score had used a standardized surgical field. The model was constructed by learning the surgical field development for each phase with deep learning. Each surgical phase was defined in a previous article[28] (eTable 3 in Supplement 1). In addition, the arbitrariness associated with manual view selection was also eliminated by dividing all scenes in each phase and inputting all of them, rather than arbitrarily selecting certain scenes in each phase to use for training.

AI Confidence Score

When performing classification tasks, as in this study, softmax functions are used in the output layer. To determine to

Figure 1. Learning Process of the Deep Learning–Based Standardized Surgical Field Recognition Model and the Artificial Intelligence Confidence Score (AICS) Calculation Process



which class a given datum belongs, its probability is outputted along with the class result. All the output results are in the range of 0 to 1, and the sum of those output results is 1. The proposed deep learning model can recognize a standardized surgical field as a static image, analyze its similarity to a surgical field development, and output a percentage reflecting that similarity (ie, a number between 0 and 1), which we termed the *AI confidence score* (AICS). In other words, images of surgical field development closer to that of the highly skilled surgeons in the training data set would have a higher AICS; however, when the surgical field was dissimilar to the training data set and unfamiliar with the model, the model would not recognize the surgical field with a high AICS. For example, when the model recognizes the development of the surgical field in phase X, if the development is clearly unique to phase X, the model will output phase X without hesitation and with 100% confidence. However, if the development of phase X is atypical, the model will probably be unsure whether it is phase X, Y, or Z. For instance, if the output is phase X with 40% confidence, phase Y with 30% confidence, and phase Z with 30% confidence, the final output result will be phase X, which is correct, albeit the confidence level is only 40%. The process of this model is illustrated in **Figure 1**.

Model Optimization

Convolutional neural network (CNN)-based deep learning, which is a type of neural network that automates the extraction of visual features to optimize the performance of a given task, such as object identification based on an image, was used in the current study. The task of this deep learning model was to perform classification for surgical steps and hyperparameters, which were optimized to maximize the classification accuracy applied to the model. To evaluate the performance of the model, hold-out validation was adopted, and the overall accuracy was used as the metric. The CNN features were extracted using the Xception implemented in TensorFlow.[35] The calculations were performed in Python, version 3.6 (Python Software Foundation).

Validation

From the data set of 650 cases, another 60 videos were extracted prior to model construction for validation. To allow for variability in results, the data set was divided into 3 groups by ESSQS scores, and the cases were selected from each group in a balanced and random manner. Based on ESSQS scores (range, 0-100 points), the 650 intraoperative videos were divided into the following 3 score groups: scores less than 2 SDs from the mean (low-score group; 15 cases), scores within the range spanning the mean plus or minus the 2 SDs (middle-score group; 30 cases), and scores greater than 2 SDs from the mean (high-score group; 15 cases). Each validation video was divided into frame units as static images of 1 frame per second, and those images were applied to the model. The AICS was calculated for each static image, and the mean AICS was calculated by dividing the total AICS of all static images by the number of all static images.

Mean AICS Compared With ESSQS for Each Case

The receiver operating characteristic (ROC) curve was measured to determine the discriminating power of the proposed model. The sensitivity, specificity, and area under the ROC (AUROC) were calculated to screen the low- and high-score cases.

Table. ESSQS Score Distribution for the Entire Cohort

Group	Cases, No.	ESSQS score, mean (SD)
All	650	66.2 (8.6)
Model construction	60	79.3 (3.5)
Validation	60	62.4 (13.8)
High score	15	79.3 (4.0)
Middle score	30	63.9 (4.4)
Low score	15	42.7 (5.1)

Abbreviation: ESSQS, Endoscopic Surgical Skill Qualification System.

Statistical Analysis

Data were analyzed from April 2020 to September 2022. The Spearman rank correlation coefficient was adopted to compare the mean AICS and each ESSQS score for validation videos. All *P* values were 2-sided, and $P < .05$ was considered statistically significant. All statistical analyses were performed using EZR (Saitama Medical Center, Jichi Medical University),[36] which is a graphical user interface for R, version 2.13.0 (R Project for Statistical Computing). More precisely, it is a modified version of R commander designed to add statistical functions frequently used in biostatistics.

Results

ESSQS Score

The mean (SD) ESSQS score of all 650 intraoperative videos was 66.2 (8.6) points, and those of the 60 model-construction and 60 validation videos were 79.3 (3.5) points and 62.4 (13.8) points, respectively. The mean (SD) ESSQS scores for the high-, middle-, and low-score groups were 79.3 (4.0), 63.9 (4.4), and 42.7 (5.1) points, respectively. The ESSQS score distribution of the 60 intraoperative videos is shown in the **Table** and the eFigure in Supplement 1. The validation cohort was similar to all cohort populations.

Recognition Model of Standardized Surgical Field Development

In this study, a deep learning–based classification task for each step in Lap-S was performed to develop a CNN model that recognizes each step of standardized surgical field development. Thus, we used 50 of the 60 videos in the model construction data set for the training algorithm, and the remaining 10 videos were used as the test data to evaluate the classification accuracy of the developed model. The videos used in training were not included in the test data. In the surgical step classification task, the overall accuracy of the model was 78.2%. The results are shown in **Figure 2** as a confusion matrix.

AICS

The results of the AICS based on the recognition model of the standardized surgical field development for each video are shown in **Figure 3**. The Spearman rank correlation coefficient between the AICS and ESSQS score was 0.81 ($P < .001$).

Automatic Screening for ESSQS Score Groups

The ROC curves for the screening of the low- and the high-score groups are shown in **Figure 4**. The specificity and sensitivity for the screening of the low-score group were 93.3% and 82.2%, respectively, when the threshold value was 0.88, and the intraoperative video was judged as belonging to the low-score group. The AUROC for the screening of the low-score group was 0.93. The specificity and sensitivity for the screening of the high-score group were 93.3% and 86.7%, respectively, when the threshold value was 0.91, and the intraoperative video was judged as belonging to the high-score group. The AUROC for the screening of the high-score group was 0.94.

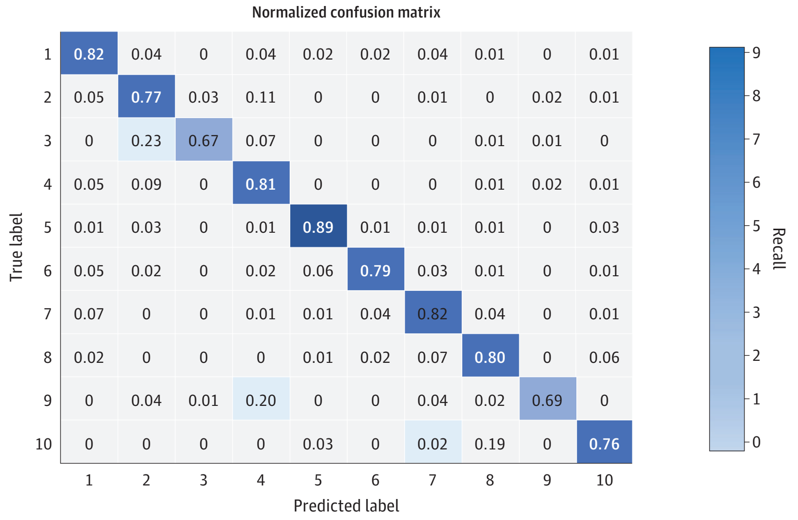
Discussion

In this study, we developed a deep learning model that could recognize standardized surgical field development and showed that the AICS output from the proposed model was strongly correlated with ESSQS scores. Furthermore, the model could automatically screen for the low-score group with 93.3% specificity and 82.2% sensitivity and for the high-score group with 93.3% specificity and 86.7% sensitivity. These results suggest that an automatic surgical skill assessment with surgical field recognition is feasible, and the proposed model has potential for use as 1 of the skill assessment items in an automatic skill assessment system.

Skill assessments based on human video review are time consuming and labor intensive. For example, concerning the JSES, approximately 1000 surgeon candidates submit for ESSQS each year, and the full surgical videos of all of them must be graded by expert surgeons. If at least the high- and low-score groups could be distinguished and filtered automatically with AI, the burden on the human raters would be substantially reduced. The screening accuracy of this study was relatively high, although still insufficient, and the results are promising for the feasibility of automated screening. Although AICS alone will not be sufficient for complete screening, other complementary approaches for screening could be incorporated to create an automatic screening system with high accuracy and robustness.

Previous attempts at automatic surgical skill assessment were mainly performed by tracking hand or surgical tool motion[37-43]; however, motion tracking mainly evaluates instrument handling or dexterity alone. Although instrument handling and dexterity are important elements for automatic surgical skill assessment,[3,4] surgical skill assessment needs to be evaluated from various perspectives. Recent surgical skill assessment tools have focused on the evaluation of the quality of standardized surgical field development, such as exposure to surgical fields with appropriate traction and cooperation with assistants, quality of dissection layer, and demonstration of landmarks.[22-24] The most famous standardized surgical field development is the CVS in laparoscopic cholecystectomy (LC). Strasberg et al[25] introduced the CVS in 1995 to promote the recognition of gallbladder elements that reduce the risk of bile duct injury and to avoid mistakes due to anatomical alterations and altered visual perception. In addition, laparoscopic right hemi-

Figure 2. Confusion Matrix of Results of Surgical Field Development Recognition



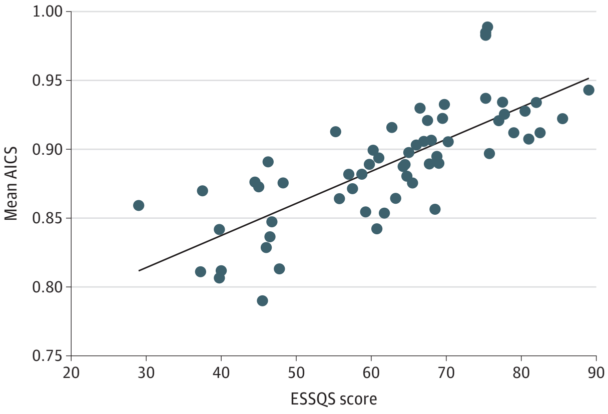
	1	2	3	4	5	6	7	8	9	10
1	0.82	0.04	0	0.04	0.02	0.02	0.04	0.01	0	0.01
2	0.05	0.77	0.03	0.11	0	0	0.01	0	0.02	0.01
3	0	0.23	0.67	0.07	0	0	0	0.01	0.01	0
4	0.05	0.09	0	0.81	0	0	0	0.01	0.02	0.01
5	0.01	0.03	0	0.01	0.89	0.01	0.01	0.01	0	0.03
6	0.05	0.02	0	0.02	0.06	0.79	0.03	0.01	0	0.01
7	0.07	0	0	0.01	0.01	0.04	0.82	0.04	0	0.01
8	0.02	0	0	0	0.01	0.02	0.07	0.80	0	0.06
9	0	0.04	0.01	0.20	0	0	0.04	0.02	0.69	0
10	0	0	0	0	0.03	0	0.02	0.19	0	0.76

The overall accuracy of the model was 78.2%.

colectomy, the procedure of complete mesocolic excision, is technically demanding and carries the risk of serious complications. The critical-view concept was thus recommended to minimize technical hazards and facilitate teaching.[27] Therefore, both dexterity for instruments and standardized surgical field development are important items in the surgical skill assessment. A comprehensive surgical skill assessment system should be constructed by combining these and other surgical skill assessment factors, such as efficiency and tissue handling. We believe the model proposed in this study can play a part in this comprehensive system.

This study focused on the AICS. In our standardized surgical field development recognition model, this score is just a number that represents the similarity between the still and training images. However, by limiting the video in the training data set to only highly skilled surgeons, we developed a surgical field recognition model specific to skillful and standardized surgical field development and thus created a model with the potential to evaluate surgical skills. Mascagni et al[44] presented automatic assessment of the CVS in LC using deep learning. Static images from LC videos were annotated with CVS criteria and hepatocystic anatomy segmentation. A deep neural network comprising a segmentation model to highlight hepatocystic anatomy and a classification model to predict CVS criteria achievement was developed. This system evaluates whether the CVS criteria are met as achieved (score of 1) or not achieved (score of 0). However, many other procedures cannot be evaluated on a 0 or 1 basis because the criteria for surgical field development are not well defined, unlike CVS in LC. Therefore, we believe that our proposed similarity-based concept is applicable to more types of surgical procedures. In addition, although AICS can be automatically obtained at the softmax layer in common deep neural network architectures, to our knowledge, no previous studies have used this value for surgical skill assessment.

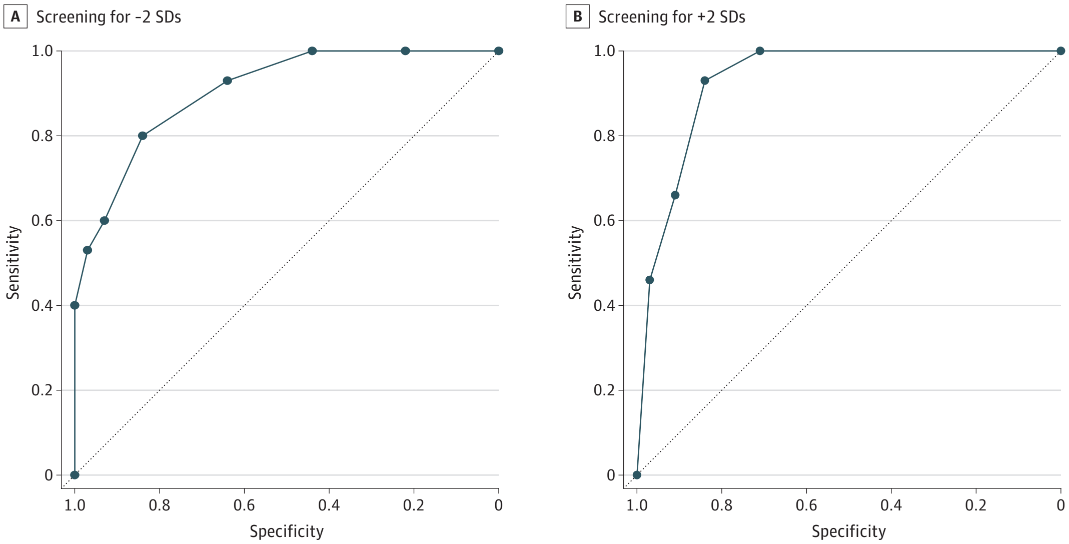
Figure 3. Correlation Between the Artificial Intelligence Confidence Score (AICS) and Endoscopic Surgical Skill Qualification System (ESSQS) Score



The Spearman rank correlation coefficient was 0.81 ($P < .001$).

In this study, we only used the ESSQS score for validation and did not evaluate videos based on the Objective Structured Assessment of Technical Skill or the Global Operative Assessment of Laparoscopic Skills. The ESSQS score is scored by 2 or 3 skilled surgeons trained to eliminate subjectivity as much as possible. There are many reports that this score is valid and reflects patient outcomes[28,29,31,32]; therefore, we believe that it is the most suitable score for validating the performance of the developed model in this study. Furthermore, the JSES videos were collected from various institutions, which allowed us to build an unbiased model for a variety of surgical procedures, instead of procedures specific to a certain hospital, and may make the model more generally useful across institutions and procedures.

Figure 4. Receiver Operating Characteristic Curves for Low- and High-Score Groups



A, The screening threshold for the group with scores less than the 2 SDs (low-score group) was 0.880 (95% CI, 0.859-0.993); sensitivity, 0.933; specificity, 0.822; and area under the receiver operating characteristic curve, 0.93. B, The screening threshold for the group with scores greater than the 2 SDs (high-score group) was 0.907 (95% CI, 0.883-0.996); specificity, 0.933; sensitivity, 0.867; and area under the receiver operating characteristic curve, 0.94.

Limitations

This study has several limitations. First, it was exploratory and not prospectively validated. In addition, due to the small sample size, comprising only 60 cases each for the training and validation sets, the results of this study lack robustness. Furthermore, mist and dirt reduce recognition accuracy. There are many measures to improve recognition accuracy in such difficult situations. One way to solve this problem is to add such images to the training data set. Another method is to perform preprocessing called data augmentation, which applies image processing, such as blurring and rotation, to the training data set to improve the model's recognition performance for such images.

Second, this study does not directly lead to improved patient outcomes. This is only 1 measure of surgical skill assessment. It is possible to have a perfectly standardized surgical field development and yet inflict organ damage due to a surgeon's extremely poor dissection skills. Similarly, a nonstandardized surgical field development can be free of complications. This study is still in the proof-of-concept phase, and we have not yet evaluated the relationship with clinical outcomes, so this is an issue for the future. However, as the safety of laparoscopic cholecystectomy has increased substantially since surgeons became aware of the CVS, we believe that proper surgical field development is important not only as a skill assessment item but also in terms of improving safety. Third, this system is intended to be used as a scoring-based evaluation rather than as feedback. We did not evaluate whether a particular scene was accomplished, as in the CVS for cholecystectomy; rather, we evaluated whether the surgery progressed throughout the entire procedure with an appropriate view. Thus, the system does not indicate what kind of surgical field development is standardized, and it also cannot give the surgeon feedback on how to improve the surgical field development. Finally, surgical field development is just 1 of the evaluation items, and it is not possible to completely evaluate surgical skills with this item alone. In the future, it will be necessary to develop a comprehensive system by verbalizing and automating the evaluation of skills from various perspectives, such as dissection ability, autonomy, dexterity, and efficiency, and this proposal is only 1 of the factors.

Conclusions

In this study, the AICS output from the developed model was strongly correlated with the ESSQS score, and the results showed the feasibility of this model for automatic surgical skill assessment. The proposed model also suggested the feasibility of an automated screening system for surgical skills. We believe that the methods of this study can be applied to other types of endoscopic procedures and can therefore be expanded to multiple areas in the future.

ARTICLE INFORMATION

Accepted for Publication: January 28, 2023.

Published Online: June 7, 2023. doi:10.1001/jamasurg.2023.1131

Author Contributions: Dr Ito had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Igaki, Kitaguchi, Kojima, Hasegawa, Takeshita, Kinugasa, Ito.
Acquisition, analysis, or interpretation of data: Igaki, Kitaguchi, Matsuzaki, Nakajima, Kojima, Takeshita.
Drafting of the manuscript: Igaki, Kitaguchi, Matsuzaki, Nakajima.
Critical revision of the manuscript for important intellectual content: Igaki, Kitaguchi, Kojima,

Hasegawa, Takeshita, Kinugasa, Ito.
Statistical analysis: Igaki.
Obtained funding: Takeshita.
Administrative, technical, or material support: Kitaguchi, Kojima, Ito.
Supervision: Kitaguchi, Kojima, Hasegawa, Takeshita, Kinugasa, Ito.

Conflict of Interest Disclosures: Mr Matsuzaki reported receiving personal fees from Jmees Inc outside the submitted work. Dr Takeshita reported receiving grants from the Japan Agency for Medical Research and Development during the conduct of the study. Dr Kinugasa reported receiving personal fees and honoraria for lectures from Medtronic and Johnson & Johnson. Dr Ito reported receiving personal fees from Asahi Intecc Co, Ltd, Olympus Corporation, Johnson & Johnson KK Medical Company, Covidien Japan Inc, Kaken Pharmaceutical Co, Ltd, Applied Medical Japan, Taiho Pharmaceutical Co, Ltd, MC Medical Inc, Bayer Yakuhin, Ltd, Nippon Covidien, Ltd, Fugifilm Corporation, and Nippon Zoki Pharmaceuticals Co, Ltd; receiving grants from SB-Kawasumi Laboratories, Inc, Hirotsu Bio Science Inc, Sumitomo Chemical Co, Ltd, A-Traction Inc, Fujita Medical Instruments Co, Ltd, Sumitomo Bakelite Co, Ltd, Surg Storage, Guardant Health Inc, Astellas Pharma Inc, Olympus Corporation, Kawasumi Laboratories, Inc, Johnson & Johnson KK Medical Company, Indivumed GmbH, Fugifilm Corporation, EBM Corporation, and Nippon Covidien, Ltd; and having patents with Mitsubishi Precision Co, Ltd, BMG Inc, Muranaka Medical Instruments Co Ltd, EBM Corporation, Niti-On Co, Ltd, Sumitomo Bakelite Co, Ltd, Fujita Medical Instruments Co, Ltd, Takano Company Ltd, and Olympus Corporation. No other disclosures were reported.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Birkmeyer JD, Finks JF, O'Reilly A, et al; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(15):1434-1442. doi:10.1056/NEJMsa1300625

2. Stulberg JJ, Huang R, Kreutzer L, et al. Association between surgeon technical skills and patient outcomes. *JAMA Surg*. 2020;155(10):960-968. doi:10.1001/jamasurg.2020.3007

3. Martin JA, Regehr G, Reznick R, et al. Objective Structured Assessment of Technical Skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278.

4. Wilson MS, Middlebrook A, Sutton C, Stone R, McCloy RF. MIST VR: a virtual reality trainer for laparoscopic surgery assesses performance. *Ann R Coll Surg Engl*. 1997;79(6):403-404.

5. Francis NK, Hanna GB, Cuschieri A. Reliability of the Advanced Dundee Endoscopic Psychomotor Tester for bimanual tasks. *Arch Surg*. 2001;136(1):40-43. doi:10.1001/archsurg.136.1.40

6. Smith SG, Torkington J, Brown TJ, Taffinder NJ, Darzi A. Motion analysis. *Surg Endosc*. 2002;16(4):640-645. doi:10.1007/s004640080081

7. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am J Surg*. 2002;184(1):70-73. doi:10.1016/S0002-9610(02)00891-7

8. Feldman LS, Sherman V, Fried GM. Using simulators to assess laparoscopic competence: ready for widespread use? *Surgery*. 2004;135(1):28-42. doi:10.1016/S0039-6060(03)00155-7

9. Dosis A, Aggarwal R, Bello F, et al. Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Arch Surg*. 2005;140(3):293-299. doi:10.1001/archsurg.140.3.293

10. Egi H, Okajima M, Yoshimitsu M, et al. Objective assessment of endoscopic surgical skills by analyzing direction-dependent dexterity using the Hiroshima University Endoscopic Surgical Assessment Device (HUESAD). *Surg Today*. 2008;38(8):705-710. doi:10.1007/s00595-007-3696-0

11. Xeroulis G, Dubrowski A, Leslie K. Simulation in laparoscopic surgery: a concurrent validity study for FLS. *Surg Endosc*. 2009;23(1):161-165. doi:10.1007/s00464-008-0120-9

12. Tokunaga M, Egi H, Hattori M, et al. Approaching time is important for assessment of endoscopic surgical skills. *Minim Invasive Ther Allied Technol*. 2012;21(3):142-149. doi:10.3109/13645706.2011.596547

13. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113. doi:10.1016/j.amjsurg.2005.04.004

14. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg*. 2008;247(2):372-379. doi:10.1097/SLA.0b013e318160b371

15. van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg*. 2010;97(7):972-987. doi:10.1002/bjs.7115

16. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical decision-making. *JAMA Surg*. 2020;155(2):148-158. doi:10.1001/jamasurg.2019.4917

17. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism*. 2017;69S:S36-S40. doi:10.1016/j.metabol.2017.01.011

18. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6

19. Berzin TM, Topol EJ. Adding artificial intelligence to gastrointestinal endoscopy. *Lancet*. 2020;395(10223):485. doi:10.1016/S0140-6736(20)30294-4

20. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Computer-assisted real-time automatic prostate segmentation during TaTME: a single-center feasibility study. *Surg Endosc*. 2021;35(6):2493-2499. doi:10.1007/s00464-020-07659-5

21. Madad Zadeh S, Francois T, Calvet L, et al. SurgAI: deep learning for computerized laparoscopic image understanding in gynaecology. *Surg Endosc*. 2020;34(12):5377-5383. doi:10.1007/s00464-019-07330-8

22. Miskovic D, Ni M, Wyles SM, et al; National Training Programme in Laparoscopic Colorectal Surgery in England. Is competency assessment at the specialist level achievable? a study for the national training programme in laparoscopic colorectal surgery in England. *Ann Surg*. 2013;257(3):476-482. doi:10.1097/SLA.0b013e318275b72a

23. Champagne BJ, Steele SR, Hendren SK, et al. The American Society of Colon and Rectal Surgeons assessment tool for performance of laparoscopic colectomy. *Dis Colon Rectum*. 2017;60(7):738-744. doi:10.1097/DCR.0000000000000817

24. Curtis NJ, Foster JD, Miskovic D, et al. Association of surgical skill assessment with clinical outcomes in cancer surgery. *JAMA Surg*. 2020;155(7):590-598. doi:10.1001/jamasurg.2020.1004

25. Strasberg SM, Hertl M, Soper NJ. An analysis of the problem of biliary injury during laparoscopic cholecystectomy. *J Am Coll Surg*. 1995;180(1):101-125.

26. Strasberg SM, Brunt LM. Rationale and use of the critical view of safety in laparoscopic cholecystectomy. *J Am Coll Surg*. 2010;211(1):132-138. doi:10.1016/j.jamcollsurg.2010.02.053

27. Strey CW, Wullstein C, Adamina M, et al. Laparoscopic right hemicolectomy with CME: standardization using the "critical view" concept. *Surg Endosc*. 2018;32(12):5021-5030. doi:10.1007/s00464-018-6267-0

28. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc*. 2020;34(11):4924-4931. doi:10.1007/s00464-019-07281-0

29. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research. *Int J Surg*. 2020;79:88-94. doi:10.1016/j.ijsu.2020.05.015

30. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191-2194. doi:10.1001/jama.2013.281053

31. Akagi T, Endo H, Inomata M, et al. Clinical impact of Endoscopic Surgical Skill Qualification System (ESSQS) by Japan Society for Endoscopic Surgery (JSES) for laparoscopic distal gastrectomy and low anterior resection based on the National Clinical Database (NCD) registry. *Ann Gastroenterol Surg*. 2020;4(6):721-734. doi:10.1002/ags3.12384

32. Shibasaki S, Suda K, Nakauchi M, et al. Impact of the Endoscopic Surgical Skill Qualification System on the safety of laparoscopic gastrectomy for gastric cancer. *Surg Endosc*. 2021;35(11):6089-6100. doi:10.1007/s00464-020-08102-5

33. Aoyama S, Inoue Y, Ohki T, Itabashi M, Yamamoto M. Usefulness of the endoscopic surgical skill qualification system in laparoscopic colorectal surgery: short-term outcomes: a single-center and retrospective analysis. *BMC Surg*. 2019;19(1):90. doi:10.1186/s12893-019-0528-2

34. Ichikawa N, Homma S, Funakoshi T, et al. Impact of technically qualified surgeons on laparoscopic colorectal resection outcomes: results of a propensity score-matching analysis. *BJS Open*. 2020;4(3):486-498. doi:10.1002/bjs5.50263

35. Chollet F. Xception: deep learning with depthwise separable convolutions. *ArXiv*. Preprint posted online April 4, 2017. doi:10.1109/CVPR.2017.195

36. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplant*. 2013;48(3):452-458. doi:10.1038/bmt.2012.244

37. Zia A, Sharma Y, Bettadapura V, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int J Comput Assist Radiol Surg*. 2016;11(9):1623-1636. doi:10.1007/s11548-016-1468-2

38. Azari DP, Frasier LL, Quamme SRP, et al. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann Surg*. 2019;269(3):574-581. doi:10.1097/SLA.0000000000002478

39. Kanumuri VV, Ameen B, Tarabichi O, Kozin ED, Lee DJ. Semiautomated motion tracking for objective skills assessment in otologic surgery: a pilot study. *OTO Open*. 2019;3(1):X19830635. doi:10.1177/2473974X19830635

40. Al Hajj H, Lamard M, Charriere K, Cochener B, Quellec G. Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network. *Annu Int Conf IEEE Eng Med Biol Soc*. 2017;2017:2002-2005. doi:10.1109/EMBC.2017.8037244

41. Zia A, Essa I. Automated surgical skill assessment in RMIS training. *Int J Comput Assist Radiol Surg*. 2018;13(5):731-739. doi:10.1007/s11548-018-1735-5

42. Lee D, Yu HW, Kwon H, Kong HJ, Lee KE, Kim HC. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *J Clin Med*. 2020;9(6):1964. doi:10.3390/jcm9061964

43. Levin M, McKechnie T, Khalid S, Grantcharov TP, Goldenberg M. Automated methods of technical skill assessment in surgery: a systematic review. *J Surg Educ*. 2019;76(6):1629-1639. doi:10.1016/j.jsurg.2019.06.011

44. 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(5):955-961. doi:10.1097/SLA.0000000000004351