

Identification of Appendicitis Using Ultrasound with the Aid of Machine Learning

Kentaro Hayashi, MD,^{1,2,i} Tetsuya Ishimaru, MD, PhD,^{1,ii} Jaesung Lee, BSEE,³
Shun Hirai, MPhys,³ Tomoaki Ooke, BPEd,³ Takahiro Hosokawa, MD,⁴ Kanako Omata, MD,¹
Youhei Sanmoto, MD,¹ Tomo Kakiyama, MD,¹ and Hiroshi Kawashima, MD¹

Abstract

Background: Diagnosing pediatric appendicitis by ultrasonography (US) is difficult because US requires significant training and skill. We evaluated whether artificial intelligence (AI) can augment US.

Materials and Methods: Among 70 abdominal ultrasound videos containing 85–347 images each, 50 were used to train the AI neural network. Each video was categorized based on the detection percentage and percent accuracy: most (>50%), partial (10–50%), and none (<10%). Test 1 involved verification of appendix detection by AI using the remaining 20 videos. Test 2 involved the evaluation of the effect of AI utilization on pediatricians.

Results: From 50 videos, 6914 images were used to train the AI network. In test 1, 3 pediatric surgeons judged 10 (50.0%), 4 (20.0%), and 6 (30.0%) videos as “most,” “partial,” and “none,” respectively, regarding the detection percentage; 7 (35.0%), 7 (35.0%), and 6 (30.0%) videos were judged, respectively, concerning the percent accuracy. Five (83.3%) of six test videos with a scan area depth of 8 cm were judged as “none” for both detection and accuracy. In test 2, six videos were also judged as “none” for both categories, showing a negative effect on the participants (5 pediatric residents and 5 pediatric intensive-emergency fellows), but the other categories showed little negative effect.

Conclusions: Appendicitis in a shallow US scan area can be easily identified with AI support. Even with the detection of a partial appendicitis shadow, AI is still helpful. However, if AI does not detect appendicitis at all, examiners may be negatively affected.

Keywords: artificial intelligence, machine learning, appendicitis, ultrasonography, diagnostic imaging

Introduction

APPENDICITIS IS OFTEN diagnosed in children by ultrasonography (US).¹ US performed by experts shows high visualization, accuracy, sensitivity, and specificity.² However, performing US is not easy for all pediatricians or emergency physicians. The use of US to diagnose appendicitis requires significant personnel training and skills,³ leading to excessive utilization of computed tomography (CT) or magnetic resonance imaging (MRI), especially in non-pediatric facilities where patients suspected of having appendicitis are often assessed by pediatricians or emergency

physicians without the aid of US experts.⁴ To compensate for the lack of experienced US experts, we attempted to create a new system to support appendicitis diagnosis by US with artificial intelligence (AI).

For further improvement using AI, we conducted this preliminary study in two phases: in test 1, we aimed to describe the current performance quality of the AI system with US in diagnosing appendicitis and assessed the factors influencing the AI performance. In test 2, we evaluated whether potential users would appreciate AI assistance and how the examiners' clinical judgments changed.

¹Department of Pediatric Surgery, Saitama Children's Medical Center, Saitama, Japan.

²Department of Pediatric Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan.

³Morpho, Inc., Tokyo, Japan.

⁴Department of Radiology, Saitama Children's Medical Center, Saitama, Japan.

Presentation: A summary of this study was presented at the IPEG conference 2021.

ⁱORCID ID (<https://orcid.org/0000-0001-8360-7630>).

ⁱⁱORCID ID (<https://orcid.org/0000-0002-2908-7416>).

Materials and Methods

Ethics

A hybrid study design was used. First, we constructed an AI network and performed test 1 with retrospective data, then we performed test 2 as a prospective study to evaluate the effects of AI support on pediatric residents and fellows.

The study protocol complied with the Declaration of Helsinki and was approved by the Research Ethics Committee of Saitama Children's Medical Center (2018-03-13 and 2020-05-016). The requirement for written informed consent from the patients was waived because of the retrospective nature of the study, and written consent from the participant pediatricians in test 2 was obtained.

Construction of the AI network

We retrospectively gathered 70 ultrasound videos recorded at the time of appendicitis diagnosis at Saitama Children's Medical Center. All videos were recorded between December 2015 and October 2019. In accordance with standard procedures at our children's hospital, trained radiologists performed all sonographies to diagnose appendicitis—which guaranteed the quality of the videos. Each video consisted of a continuous series of 85–347 (mean: 139, median: 104) images. All the videos used showed clinically or pathologically diagnosed appendicitis and enlarged appendixes. Videos with poor quality that made retrospective diagnosis of appendicitis difficult for radiologists or pediatric surgeons were not included in these 70 videos, and they contained no marks or comments by the radiologists. We used LOGIQ E9 (GE Healthcare, Chicago, IL) until July 2019 for 58 videos and LOGIQ E10 (GE Healthcare) in August 2019 for the remaining 12 videos. All patients were examined without any specific preparations. Of the 70 abdominal ultrasound videos, 50 were chosen for AI neural network training. The appendixes in each image were annotated by 5 pediatric surgeons (K.H., T.I., K.O., Y.S., and T.K.).

We developed an AI with Morpho, Inc. (Tokyo, Japan). Medical image segmentation involves separating objects of interest in medical images. Our target was to detect the appendix from an ultrasonogram. Many convolutional neural network architectures have been proposed for medical image

segmentation. We used U-Net⁵-based architecture, which is widely used in medical image segmentation. Because very little training data were available—which can cause overfitting—we adapted data augmentation (i.e., cropping, flipping, and adding noise) to increase the data. In the training phase, we normalized all the images to stabilize the training model. In general, the appendix is quite small, and a cost function that can handle such a problem should be used. When we trained the model, we used dice loss⁶ as a cost function.

Finally, we constructed the AI system. When the AI detected the appendix, it highlighted the structure in yellow (Fig. 1).

Test 1: Evaluation of the AI performance quality. In test 1, we aimed to describe the performance quality of the AI system with US in diagnosing appendicitis and assessed the factors influencing its performance using the remaining 20 videos (test videos). These were evaluated by 3 other pediatric surgeons. In using the 20 test videos, the evaluated surgeons did not perform sonography by themselves, and the test videos did not contain any marks or comments by the radiologists or pediatric surgeons.

First, we played the plain test videos without any information from the AI system, and then played the test videos with AI suggestions. After comparing the plain videos and AI videos, the 3 evaluated surgeons estimated the detection percentage and percent accuracy based on their own judgment (Fig. 2). Detection percentage was defined as the area of true appendixes successfully detected by the AI system divided by the total area of appendixes visualized through a continuous series of images on a video; percent accuracy was defined as the area of true appendixes successfully detected by the AI system divided by the total area of appendixes detected by the AI. Each answer was categorized into “most” (>50%), “partial” (10%–50%), and “none” (<10%), respectively (Fig. 3A–C). Three surgeons were evaluated independently at different times. The categorized answers that matched 2 or all the evaluated surgeons' answers were adopted as the results. If the answers differed among the 3 surgeons, they re-watched the test videos and discussed the answers. This test was observed by the test proctor (K.H.) to check whether appendicitis was correctly detected.

After the results were achieved, we evaluated and described which factors—such as patients' age, sex, diagnosis, or details of US examination—heavily affected AI performance.

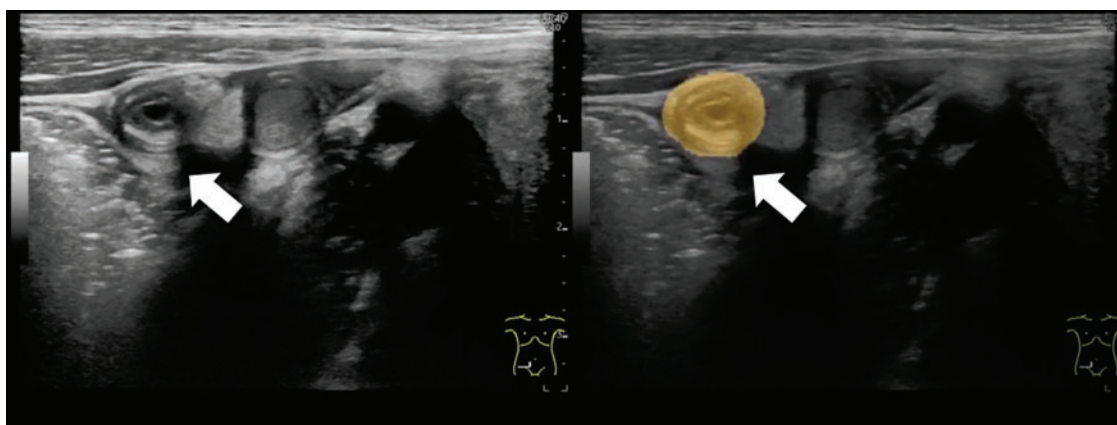


FIG. 1. Appendicitis (indicated by white arrows) detected using AI (right picture). AI, artificial intelligence. Color images are available online.

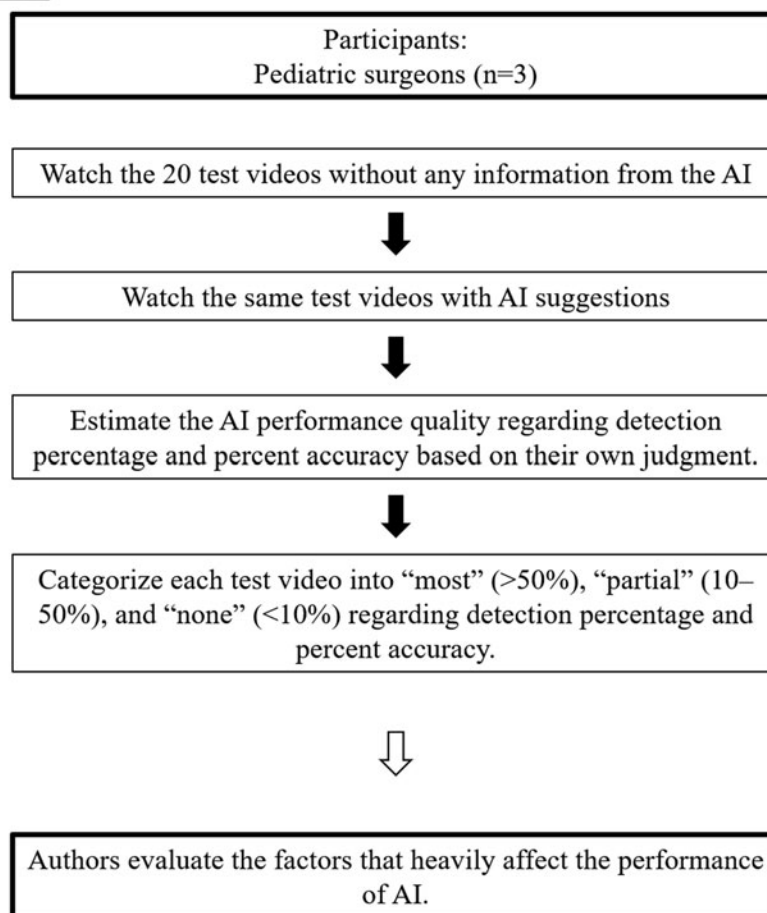
Test 1

FIG. 2. A flow chart of test 1. The evaluated surgeons did not perform sonography by themselves, and the test videos did not contain any marks or comments by the radiologists or pediatric surgeons.

Test 2: The effects of AI support on the pediatric residents and fellows. In test 2, we evaluated whether potential users would appreciate AI assistance and how the examiners' clinical judgments changed. The recruited participants included 5 pediatric residents and 5 pediatric intensive-

emergency fellows who regularly cared for emergency patients at our institution. We showed the same 20 test videos used in test 1 to the participants and tested them independently at different times. The following three questions were posed: Q1 “Which structure do you think is the appendix?,”

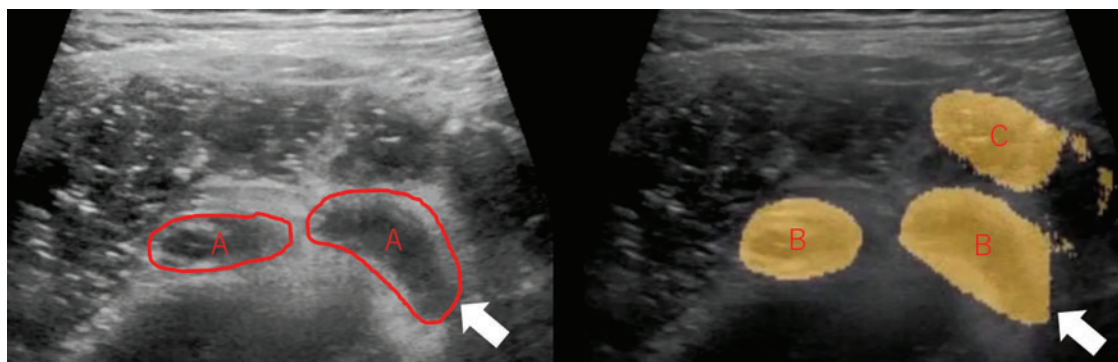


FIG. 3. The appendix as identified by the surgeon (A, left). The appendix as identified by AI (right). (B) True positive. (C) False positive. Discrepancy between the surgeon-identified true appendix and the AI-detected true appendix is indicated by the white arrows. Detection percentage = B/A ; percent accuracy = $B/(B + C)$. In this figure, detection percentage and percent accuracy are $\sim 100\%$ and 60% , respectively. However, each percentage was estimated using not only one image but using a continuous series of images on a video. Each of the 3 evaluated surgeons estimated these values. AI, artificial intelligence. Color images are available online.

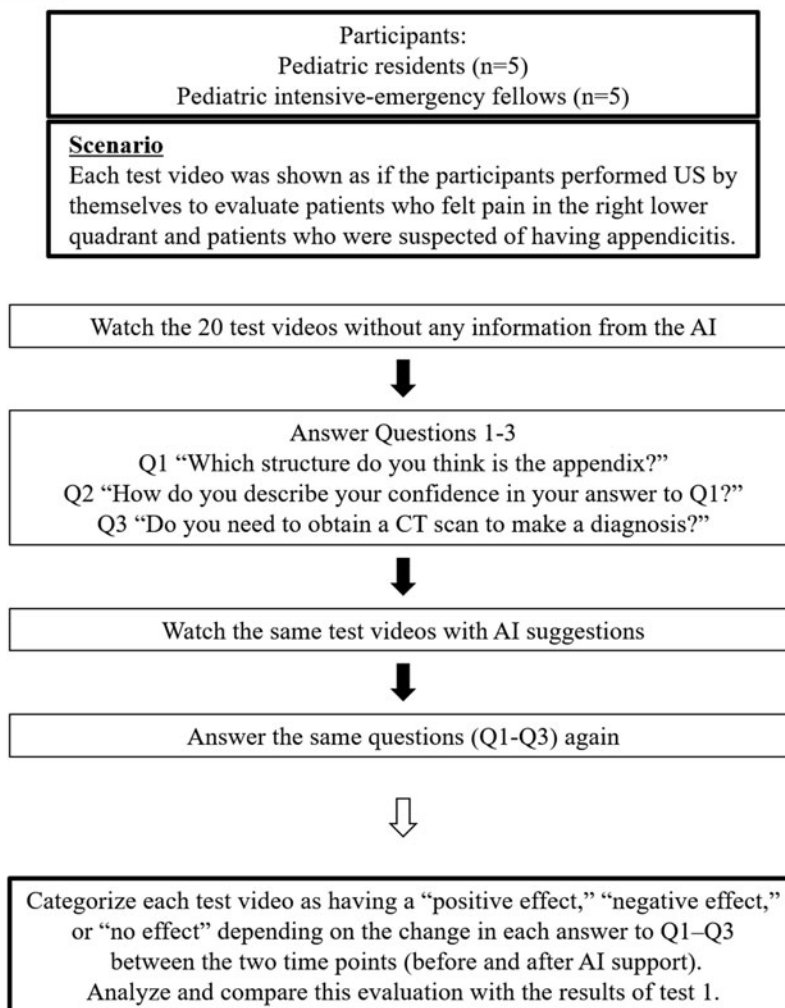
Test 2

FIG. 4. A flow chart of test 2. The participants did not perform sonography by themselves, and the test videos did not contain any mark or comments by the radiologists, or pediatric surgeons. In Q1 ("Which structure do you think is the appendix?"), the answers from the participants were judged as "correct" or "wrong" by the test proctor (K.H.): "Correct": the participant correctly identified the appendix. "Wrong": the participant pointed out structures other than the appendix or judged that the appendix could not be visualized. "Unknown": the participant could not find which structure was the appendix or whether the appendix was visualized. In Q2 ("How do you describe your confidence in your answer to Q1?"), the participants described their confidence level on a scale from 0 (no confidence at all) to 100 (highly confident in the diagnosis). In Q3 ("Do you need to obtain a CT scan to make a diagnosis?"), the participant answered "yes" if the participant required a CT scan, or "no" if the participant did not require a CT scan. CT, computed tomography.

Q2 "How do you describe your confidence in your answer to Q1?" and Q3 "Do you need to obtain a CT scan to make a diagnosis?" First, the plain test videos were presented and questions 1 through 3 (Q1-Q3) were answered, and then the AI test videos were presented and Q1-Q3 were answered again. The change in each answer (Q1-Q3) between the two time points (before and after AI support) was evaluated. This evaluation was then analyzed and compared with the results of test 1 (Fig. 4).

To prevent random guesses from the participants, they were blinded as to whether the appendix was visualized or whether the patients were diagnosed with appendicitis. Similar to test 1, the participants did not perform sonography by themselves, and the test videos did not contain any marks or comments by the radiologists or pediatric surgeons.

The change in each answer (Q1-Q3) between the two time points (before and after AI support) was categorized as having a positive effect, negative effect, or no effect according to the algorithm given in Table 1.

Data collection

We collected the following data on the clinical characteristics of the patients from whom we obtained training and test US videos: age, sex, and details of US videos such as the scan area depth.

Statistical analysis

Categorical variables are presented as percentages (%), and continuous variables are presented as medians (ranges). Because there were few previous reports or insufficient

TABLE 1. COMPREHENSIVE ANALYSIS OF THE EFFECTS OF ARTIFICIAL INTELLIGENCE

<i>Q1 Point out appendicitis correctly</i>		<i>Q2 Confidence in the answer to Q1</i>	<i>Q3 Need for CT scan?</i>		<i>Effect of AI</i>
<i>Before referring to AI</i>	<i>After referring to AI</i>	<i>Change in confidence level >10% before and after referring to AI</i>	<i>Before referring to AI</i>	<i>After referring to AI</i>	
Correct	Correct	Increase	Yes	Yes	Positive effect
			Yes	No	Positive effect ^a
			No	Yes	Positive effect
		Decrease	No	No	Positive effect
			Yes	Yes	Negative effect ^a
			Yes	No	Negative effect
		Unchanged (change ≤10%)	No	Yes	Negative effect
			No	No	Negative effect
			Yes	Yes	No effect
			Yes	No	Positive effect
			No	Yes	Negative effect
			No	No	No effect
Unknown	Unknown	Increase	Yes	Yes	Negative effect
	Wrong				Negative effect
	Correct				Positive effect
	Unknown				No effect
	Wrong				Negative effect
	Correct				Positive effect
Wrong	Wrong	Increase	Yes	Yes	Positive effect
			Yes	No	Negative effect
			No	Yes	Negative effect ^a
		Decrease	No	No	Negative effect
			Yes	Yes	Positive effect
			Yes	No	Positive effect ^a
		Unchanged (change ≤10%)	No	Yes	Positive effect
			No	No	Positive effect
			Yes	Yes	No effect
			Yes	No	Negative effect
			No	Yes	Positive effect
			No	No	No effect

Example 1: The participant initially pointed out appendicitis correctly, then changed to a wrong answer after referring to AI → Negative effect.

Example 2: The participant pointed out appendicitis correctly but confidence level reduced after referring to AI → Negative effect.

Example 3: The participant pointed out appendicitis correctly and had unchanged confidence level (change within 10%), but cancelled CT → Positive effect.

Example 4: The participant pointed out appendicitis wrongly, but confidence level decreased after referring to AI → Positive effect.

^aThese are conflicts (e.g., the confidence decreased, but the participant cancelled CT, or vice versa), but no one chose these options.

AI, artificial intelligence; CT, computed tomography.

hypotheses to test, we only observed and reported our experience; we did not test a specific hypothesis or compare results with previous studies. Statistical analyses were performed using the R package version 3.3.1 (The R Foundation for Statistical Computing, Vienna, Austria).

Results

Construction of the AI network

From the 50 videos, 6914 images were used to train the neural network, and 3454 images contained annotated appendixes. The characteristics of the patients from whom the training videos were obtained were as follows: males, 37 (74.0%); age (year, median [range]), 10 [5–14]; and scan area depth (cm, median [range]), 5 [3–10].

Test 1. All evaluated surgeons could correctly identify appendicitis in all the videos. All videos achieved at least 2 of

the 3 evaluated surgeons' consensus. Regarding the detection percentage among the 20 test videos, 10 (50.0%), 4 (20.0%), and 6 (30.0%) videos were judged as "most," "partial," and "none," respectively. Concerning the percent accuracy, 7 (35.0%), 7 (35.0%), and 6 (30.0%) videos were judged as "most," "partial," and "none," respectively. Regarding the scan area depth, 6 (30%) videos had scan areas that were 8 cm deep; 2 (10%), 7 cm deep; 7 (35%), 6 cm deep; 2 (10%), 5 cm deep; 2 (10%), 4 cm deep; and 1 (5%), 3.5 cm deep.

According to the evaluated surgeons' assessments, the test videos were categorized into the following groups based on the detection percentage and percent accuracy: group 1, "most" (detected) and "most" (accurate); group 2, "most" and "partial"; group 3, "partial" and "partial," and group 4, "none" and "none," respectively. Each group comprised 7 (35%), 3 (15%), 4 (20%), and 6 (30%) patients, respectively. Each factor seemed to be related to AI performance (Table 2). In the group 1 videos, the median age of patients was 9 years

TABLE 2. RESULTS OF TEST 1: ARTIFICIAL INTELLIGENCE PERFORMANCE AND RELATED FACTORS

Group	AI performance		No. of patients (%)	Sex	Age, years (median [range])	Scan area depth, cm (median [range])
	Detection percentage	Percent accuracy		Male (%)		
1	Most	Most	7 (35)	4 (57.1)	9 [6–10]	5 [3.5–8]
2	Most	Partial	3 (15)	1 (33.3)	10 [8–14]	6 [4–6]
3	Partial	Partial	4 (20)	3 (75.0)	12 [6–16]	6 [6–7]
4	None	None	6 (30)	3 (50.0)	12 [7–14]	8 [6–8]

According to the evaluated surgeons' assessments, the test videos were categorized into the following groups based on the detection percentage and percent accuracy: group 1, most and most; group 2, most and partial; group 3, partial and partial, and group 4, none and none, respectively.

The scan area depth is deeper in group 4 than in other groups.

AI, artificial intelligence.

and the median scan area depth was 5 cm; in the group 4 videos, the median age was 12 years and the median scan area depth was 8 cm. Moreover, 5 (83.3%) of the six test videos with a scan area depth of 8 cm, were in group 4.

Test 2. Regarding Q1, 10 (50%) [5 (25%)–15 (75%)] of the 20 test videos were correctly assessed before AI support, and 12 (60%) [7 (35%)–16 (80%)] were after AI support. Regarding Q3, 2 of the 5 pediatric intensive-emergency fellows answered that they perform CT every time they attend to a patient with right lower quadrant pain (regardless of US results) in case they missed an appendicitis diagnosis.

In groups 1–3, AI was found to have a negative effect for only 0%–20% of the pediatricians. However, for group 4, a negative effect was found for 20%–80% of the pediatricians (Table 3).

Discussion

Many studies about AI have been conducted in the medical field, and some studies tried to diagnose appendicitis through AI.^{7–9} In many of them there were attempts to diagnose appendicitis using clinical symptoms or predictive scores without imaging studies,⁹ whereas some of them used CT.⁸ Imaging is a central tool in appendicitis diagnosis that is used to reduce negative appendectomies,¹⁰ and we cannot ignore the unfavorable malignant side effects of radiation.¹¹ We could not find AI research using MRI; however, MRI is expensive, time consuming, and requires sedation in children.

Therefore, we still need to encourage appendicitis diagnosis through US. Few studies have attempted to diagnose appendicitis through US,⁷ and whether AI can help diagnose appendicitis by US and the clinical effect of AI on the examiners remain unknown.

Summary of the results

In this study, we summarized our experience with a newly constructed AI system to support appendicitis diagnosis, the relationship between the clinical factors and AI performance, and the clinical effects of AI on the examiners. According to test 1, AI may easily assess appendicitis with an US with a shallow scan area. Regarding test 2, even if only a part of the appendix is detected, AI may still be helpful to examiners. However, if almost none of the appendix is detected, AI may have a negative effect on cognitive level (i.e., perception or decision-making) of the examiners.

Performance of AI and related factors affecting its performance

We assessed the factors that are related to AI performance, and the scan area depth seemed to affect AI performance more than sex or the severity of appendicitis. Whether shallow US scans are more accurate is unknown, but it may be because noise increases more as the scan area becomes larger and deeper. To improve AI performance, reducing unnecessary structures from the input image may be efficient in decreasing the noise. In a previous study on CT, the authors

TABLE 3 RESULTS OF TEST 2: RELATIONSHIP BETWEEN ARTIFICIAL INTELLIGENCE PERFORMANCE AND EFFECTS ON THE PEDIATRICIANS

Group	AI performance			Effects of AI on the pediatricians		
	Detection percentage	Percent accuracy	No. of patients (%)	Negative effect, % (median [range])	No effect, % (median [range])	Positive effect, % (median [range])
1	Most	Most	7 (35)	0 [0–20]	10 [0–70]	90 [20–90]
2	Most	Partial	3 (15)	0 [0–0]	40 [10–60]	60 [40–90]
3	Partial	Partial	4 (20)	10 [0–20]	35 [30–50]	50 [40–70]
4	None	None	6 (30)	40 [20–80]	55 [10–80]	5 [0–10]

According to test 1, the test videos are categorized into the following groups based on the detection percentage and percent accuracy: group 1, most and most; group 2, most and partial; group 3, partial and partial, and group 4, none and none, respectively.

Group 4 (AI performance are none and none) shows the highest proportion of negative effects on the pediatricians, but the others are associated with almost no negative effects.

AI, artificial intelligence.

manually limited the input area to three-dimensional isotropic cubes ($4 \times 4 \times 4 \text{ cm}^3$).⁸ Another study on US labeled the fascia area to remove noise.⁷ To limit the area, studies need to be conducted regarding whether learning other structures such as the vessels, muscles, intestines, and the peritoneal membrane improves the performance of AI.

Young age may also contribute to the shallowness of the scan area. In pediatric patients, the appendix is often found in shallower areas than in adult patients; the body size or thinness of the subcutaneous fat of pediatric patients may allow us to use a high-frequency probe to visualize appendicitis clearly. Therefore, AI seems to be more useful in pediatric patients than in adult patients.

Moreover, other secondary signs of appendicitis may also be useful in identifying the appendix. Upon reviewing the structures misidentified by AI as appendicitis, low echoic areas (i.e., the iliac vessels, parts of muscles, the ileum) were included. If we can add any information from Doppler US to AI, it may help to distinguish vasculature from appendicitis. Therefore, further improvements in AI are warranted.

Relationship between AI performance and effects on the pediatricians

Although all the evaluated surgeons in test 1 could correctly identify appendicitis, each participant in test 2 could only correctly point out 10 (50%) with the same test videos. These pediatricians usually see emergency cases, and they represent potential users of the AI system for supporting appendicitis diagnosis. This discordance precisely reflects our motivation as to why we need to construct the AI system for diagnosing appendicitis to support pediatricians who perform US without experts.

Ideally, AI should detect appendicitis 100% correctly, but this may be very hard to achieve, and it may take a long time. However, the hurdle may be lower than we first thought because AI seems to be helpful and has few negative effects on examiners (at least when appendicitis is partially identified by AI), which may encourage the promoters or researchers of AI. One of the problems with utilizing AI in clinical situations is the absence of personal accountability in cases when AI leads to a misdiagnosis.¹² In light of this view, the negative effect of AI utilization on pediatricians may be critical. However, according to test 2, AI affected the examiners negatively only when nearly none of the shadows of the appendix were detected, which is an important problem that needs to be solved to bring AI into practical use.

First, we created this AI to support the detection of the shadows of the appendix because we hypothesized that many pediatricians would have difficulties in identifying the appendix, and that they would often become disoriented and miss the appendix; sometimes they did not even notice that that appendicitis was visualized by US. Therefore, AI should be highly sensitive. However, many participants in this study stated that they were relieved if what they pointed out coincided with what AI detected, which means the examiners used AI as the answer. In this situation, AI requires high specificity. The discordance with what AI required may be owing to the test situation (already recorded videos were shown). Therefore, we need to study whether AI helps examiners to detect appendicitis in a more realistic situation to confirm whether the sensitivity or specificity of AI needs improvement.

We started to construct AI to decrease the frequency of unnecessary CT scans, but some examiners, especially more experienced pediatricians, seem to have made their own decision-making rules to obtain CT scans for all patients suspected of having appendicitis, regardless of the information suggested by AI. This means that AI may not quickly decrease unnecessary CT scans unless AI becomes widespread and achieves high reliability. For AI to be widely accepted, a strategy to improve its performance and evaluate its reliability is needed.

Limitations

This study has several limitations. First, in test 1, AI performance was analyzed qualitatively, not quantitatively, depending on the surgeon's interpretation. Although we analyzed the area quantitatively by measuring the area, there was a large discrepancy between the surgeon's impression and the results of the calculation depending on the measured area. This may be owing to the noise (false positives) of the AI suggestion. Through videos consisting of many images, AI tries to detect appendicitis in each image and causes many small noises. However, these small noises did not seem to be a concern for surgeons because they could imagine the shape of the appendix through the videos. To evaluate the effect more clinically, we adopted the results of the surgeon's interpretation. We believe that this qualitative assessment is a more practical and clinically oriented assessment. Second, we did not train AI with a normal appendix, only that exhibiting appendicitis. Therefore, we could not determine the extent to which AI can detect a normal appendix, which may be a future direction for research. Moreover, we chose suitable videos for training and testing from previous data stored in our institution, but there may be selection bias. In addition, AI should be improved and evaluated to achieve generalizability to emergency department settings. Third, because this study was conducted as a preliminary study for further improvement of this AI system and the sample size was small, we only described our results and did not test any hypotheses or perform any statistical analyses. However, we believe this study raised some meaningful points and suggestions for those trying to construct AI systems for abdominal sonography.

Conclusions

According to our study's findings, US with a shallow scan area may easily assess appendicitis with AI. Even if only a partial shadow of the appendix is detected, AI may still be helpful to the examiners. However, if almost none of the shadows of the appendix are detected, AI may have a negative effect on the examiners. Further improvement in AI is warranted.

Authors' Contributions

K.H., T.I., and T.H.: Substantial contributions to the conception, design of the work, the acquisition, analysis, and interpretation of data for the work. K.H.: Drafting the work and revising it critically for important intellectual content. K.H., T.I., L.J., S.H., T.O., T.H., K.O., Y.S., T.K., and H.K.: Final approval of the version to be published. K.H.: Agreement to be accountable for all aspects of the work in ensuring

that questions related to the accuracy and integrity of any part of the work are appropriately investigated and resolved. T.I., L.J., S.H., T.O., T.H., K.O., Y.S., T.K., and H.K.: Revising the article critically for important intellectual content. T.I., L.J., S.H., T.H.: Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy, and integrity of any part of the work are appropriately investigated and resolved. L.J., S.H., T.O.: Substantial contributions to the analysis and the interpretation of data for the work. T.O., K.O., Y.S., T.K., and H.K.: Agreement to be accountable for all integrity of any part of the work are appropriately investigated and resolved. K.O., Y.S., T.K., H.K.: Substantial contributions to the acquisition of data for the work.

Acknowledgments

The authors thank the other pediatric surgery team members, radiology members of our hospital, and the members of Morpho, Inc., The authors also thank the Medical-Industrial HUB Organization of Tokyo for introducing members of Saitama Children's Medical Center to the members of Morpho, Inc. The authors thank Editage (www.editage.com) for editing and reviewing this article for English language.

Disclosure Statement

L.J., S.H., and T.O., are employees of Morpho, Inc., (Tokyo, Japan), a company that manufactures AI software systems. K.H., T.I., T.H., K.O., Y.S., T.K., and H.K. have no conflicts of interest or financial ties to disclose.

Funding Information

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Expert Panel on Pediatric Imaging, Koberlein GC, Trout AT, et al. ACR appropriateness Criteria® suspected appendicitis-child. *J Am Coll Radiol* 2019;16:S252–S263.
2. Cundy TP, Gent R, Frauenfelder C, Lukic L, Linke RJ, Goh DW. Benchmarking the value of ultrasound for acute appendicitis in children. *J Pediatr Surg* 2016;51:1939–1943.
3. Trout AT, Sanchez R, Ladino-Torres MF, Pai DR, Strouse PJ. A critical evaluation of US for the diagnosis of pediatric acute appendicitis in a real-life setting: How can we improve the diagnostic value of sonography? *Pediatr Radiol* 2012;42:813–823.
4. Jones RE, Gee KM, Preston SC, Babb JL, Beres AL. Diagnostic utilization and accuracy of pediatric appendicitis imaging at adult and pediatric Centers. *J Surg Res* 2019; 240:97–103.
5. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-assisted Intervention, 2015. <https://arxiv.org/abs/1505.04597> (accessed October 25, 2021)
6. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, et al. (eds). *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. London, UK: Springer, 2017, pp. 240–248.
7. Kim KB, Park HJ, Song DH. Semi-dynamic control of FCM initialization for automatic extraction of inflamed appendix from ultrasonography. *Curr Med Imaging Rev* 2019;15:810–816.
8. Park JJ, Kim KA, Nam Y, Choi MH, Choi SY, Rhie J. Convolutional-neural-network-based diagnosis of appendicitis via CT scans in patients with acute abdominal pain presenting in the emergency department. *Sci Rep* 2020;10: 9556.
9. Yoldas O, Tez M, Karaca T. Artificial neural networks in the diagnosis of acute appendicitis. *Am J Emerg Med* 2012; 30:1245–1247.
10. Hendriks IG, Langen RM, Janssen L, Verrijth-Wilms IM, Wouda S, Janzing HM. Does the use of diagnostic imaging reduce the rate of negative appendectomy? *Acta Chir Belg* 2015;115:393–396.
11. Lee KH, Lee S, Park JH, et al. Risk of hematologic malignant neoplasms from abdominopelvic computed tomographic radiation in patients who underwent appendectomy. *JAMA Surg* 2021;156:343–351.
12. Liew C. The future of radiology augmented with artificial intelligence: A strategy for success. *Eur J Radiol* 2018;102: 152–156.

Address correspondence to:

Kentaro Hayashi, MD
 Department of Pediatric Surgery
 Saitama Children's Medical Center
 1-2, Shintoshin, Chuo-ku
 Saitama City 330-8777
 Saitama Prefecture
 Japan

E-mail: 884kentaro@gmail.com