ANESTHESIOLOGY

Validation of a Deep Learning–based Automatic Detection Algorithm for Measurement of Endotracheal Tube–to–Carina Distance on Chest Radiographs

Min-Hsin Huang, M.D., M.S., Chi-Yeh Chen, Ph.D., Ming-Huwi Horng, Ph.D., Chung-I Li, Ph.D., I-Lin Hsu, M.D., Ph.D., Che-Min Su, M.D., Yung-Nien Sun, Ph.D., Chao-Han Lai, M.D., Ph.D.

ANESTHESIOLOGY 2022; 137:704–15

Scan for CME exam

EDITOR'S PERSPECTIVE

What We Already Know about This Topic

- Deep learning image classification techniques are changing the interpretation process in a range of radiology settings
- It is unclear whether automated detection of a misplaced endotracheal tube can perform similarly to critical care clinicians

What This Article Tells Us That Is New

- A deep learning–based algorithm developed using portable chest radiographs from 1,842 adult intubated patients can identify the endotracheal tube tip, carina, and endotracheal tube tip–to–carina distance with a measurement error of 2.6 mm, 3.6 mm, and 4.0 mm, respectively
- The algorithm performed as well as, if not better than, 11 critical care clinicians in identifying these portable chest radiograph landmarks

ABSTRACT

Background: Improper endotracheal tube (ETT) positioning is frequently observed and potentially hazardous in the intensive care unit. The authors developed a deep learning–based automatic detection algorithm detecting the ETT tip and carina on portable supine chest radiographs to measure the ETT–carina distance. This study investigated the hypothesis that the algorithm might be more accurate than frontline critical care clinicians in ETT tip detection, carina detection, and ETT–carina distance measurement.

Methods: A deep learning–based automatic detection algorithm was developed using 1,842 portable supine chest radiographs of 1,842 adult intubated patients, where two board-certified intensivists worked together to annotate the distal ETT end and tracheal bifurcation. The performance of the deep learning–based algorithm was assessed in 4-fold cross-validation (1,842 radiographs), external validation (216 radiographs), and an observer performance test (462 radiographs) involving 11 critical care clinicians. The performance metrics included the errors from the ground truth in ETT tip detection, carina detection, and ETT–carina distance measurement.

Results: During 4-fold cross-validation and external validation, the median errors (interquartile range) of the algorithm in ETT–carina distance measurement were 3.9 (1.8 to 7.1) mm and 4.2 (1.7 to 7.8) mm, respectively. During the observer performance test, the median errors (interquartile range) of the algorithm were 2.6 (1.6 to 4.8) mm, 3.6 (2.1 to 5.9) mm, and 4.0 (1.7 to 7.2) mm in ETT tip detection, carina detection, and ETT–carina distance measurement, significantly superior to that of 6, 10, and 7 clinicians (all $P < 0.05$), respectively. The algorithm outperformed 7, 3, and 0, 9, 6, and 4, and 5, 5, and 3 clinicians (all $P < 0.005$) regarding the proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error in ETT tip detection, carina detection, and ETT–carina distance measurement, respectively. No clinician was significantly more accurate than the algorithm in any comparison.

Conclusions: A deep learning–based algorithm can match or even outperform frontline critical care clinicians in ETT tip detection, carina detection, and ETT–carina distance measurement.

(ANESTHESIOLOGY 2022; 137:704–15)

Among adverse events associated with endotracheal intubation, improper endotracheal tube (ETT) positioning is frequently observed and potentially hazardous if not promptly recognized and managed.[1–3] An ETT placed

This article has been selected for the Anesthesiology CME Program (www.asahq.org/JCME2022DEC). Learning objectives and disclosure and ordering information can be found in the CME section at the front of this issue. This article is featured in "This Month in Anesthesiology," page A1. This article is accompanied by an editorial on p. 664. Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org). This article has an audio podcast. This article has a visual abstract available in the online version. Y.-N.S. and C.-H.L. contributed equally to this work.

Submitted for publication November 3, 2021. Accepted for publication September 6, 2022. Published online first on September 21, 2022.

Min-Hsin Huang, M.D., M.S.: Department of Surgery, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.

Chi-Yeh Chen, Ph.D.: Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan; MOST AI Biomedical Research Center, Tainan, Taiwan.

Ming-Huwi Horng, Ph.D.: Department of Computer Science and Information Engineering, National Pingtung University, Pingtung, Taiwan.

Chung-I Li, Ph.D.: Department of Statistics, College of Management, National Cheng Kung University, Tainan, Taiwan.

I-Lin Hsu, M.D., Ph.D.: Department of Surgery, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.

Che-Min Su, M.D.: Department of Surgery, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.

Yung-Nien Sun, Ph.D.: Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan; MOST AI Biomedical Research Center, Tainan, Taiwan.

Chao-Han Lai, M.D., Ph.D.: Department of Surgery, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan; Department of Biochemistry and Molecular Biology, College of Medicine, National Cheng Kung University, Tainan, Taiwan; Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee.

 Anesthesiology 2022; 137:704–15. DOI: 10.1097/ALN.0000000000004378

at a high position may lead to air leaks or injury to the vocal cords and possibly increase the risk of accidental, unplanned extubation. Conversely, a mainstem bronchus intubation can result in hyperinflation of the intubated lung with subsequent pneumothorax and atelectasis of the nonventilated lung. Physical examination alone is unreliable for assessing the depth of ETT insertion.[1–4] In the neutral neck position, the optimal position of the ETT tip within the trachea is 3 to 7 cm above the carina, which is the reference point of the proper ETT position on portable chest radiographs.[5] It is recommended to evaluate the ETT position using a portable chest radiograph immediately after endotracheal intubation.[1,2,4,6–8] Given that radiologists are not always available at any time to read portable radiographs in the intensive care unit (ICU), timely interpretation of postintubation chest radiographs by critical care clinicians may improve the process of early decision-making.

The portable supine chest radiograph allows valuable information to be obtained without the risk of patient transport in the ICU.[9,10] Compared with standard standing chest radiographs, the quality is inconsistent due to higher image noise, because portable supine chest radiographs are obtained without an antiscatter grid.[10,11] The existence of the medical devices required for critical care (*e.g.*, nasogastric tubes, pacemaker wires, and electrocardiogram cables) and anatomical structures (*e.g.*, such as sternum, heart, and spines) may interfere with the reading of a portable chest radiograph to identify the precise ETT tip and carina locations. Nearly 40% of ICU patients are mechanically ventilated.[12] Thus, an algorithm designed to detect the ETT tip and carina on portable chest radiographs may help identify a suboptimal ETT position, reduce associated complications, and improve the ICU workflow.

With recent advances in image processing, artificial intelligence and deep learning have been gradually introduced into respiratory medicine and critical care.[13,14] Although some of these studies applied artificial intelligence and deep learning to recognize different pathologies (*e.g.*, malignancy) on standard chest radiographs and computed tomography,[13,15–18] only two reports in the literature have demonstrated the approaches and algorithm performance in identifying ETT malposition on portable supine chest radiographs.[19,20] These deep learning solutions are trained with image classification (categorization) on the basis of the entire image. However, an approach using image classification without labeling the objects (*i.e.*, the ETT and carina) on chest radiographs cannot localize the ETT and carina and is unlikely to accurately estimate the distance in between (*i.e.*, the ETT–carina distance), potentially limiting its application and reliability in clinical settings.

In the study presented here, we developed a deep learning–based automatic detection algorithm detecting the ETT tip and carina on portable supine chest radiographs to measure the ETT–carina distance using pixel-level segmentation labels. This study investigated the hypothesis that the algorithm might be more accurate than frontline critical care clinicians in ETT tip detection, carina detection, and ETT–carina distance measurement.

Materials and Methods

Training Datasets

The entire study protocol was approved by the Institutional Review Board of National Cheng Kung University Hospital (A-ER-108-305; Tainan, Taiwan). The study was conducted in the National Cheng Kung University Hospital, a 1,300-bed medical center that offers first-line and tertiary referral services for 1.8 million people in southern Taiwan. In this study, 1,870 de-identified portable supine chest radiographs of 1,870 intubated adult patients receiving surgical ICU care between 2015 and 2018 were randomly retrieved from the imaging database in the Department of Radiology. Patient consents were waived by the Institutional Review Board. The images had been de-identified before we received the files, and thus the patient demographics were not provided. The files were exported in the Digital Imaging and Communications in Medicine format. The length and width of these images ranged from 2,517 to 3,032 pixels, including 1,279 images sized at $2,517 \times 3,032$ pixels, 538 images sized at $3,032 \times 2,517$ pixels, and 53 images sized variously between $2,517 \times 2,517$ and $3,032 \times 3,032$ pixels. The image files were split into 4 folds. The number of images was estimated and 4-fold cross-validation was used based on our internal pilot study results. During the internal pilot experiment, we used approximately 400 images and 4-fold cross-validation to evaluate the performance of the algorithm. Based on the pilot study results, we chose the 4-fold cross-validation strategy and estimated the number of chest radiographs to be approximately 1,800.

Using a self-developed image annotation software, two board-certified intensivists (Drs. Huang and Lai; 14 and 9 yr of experience as intensivists, respectively) read the chest radiographs together at the same time and in the same location and conducted manual labeling as the ground truth annotations. The distal ETT end was labeled by the quadrangle constituted by P_1 to P_4, and the tracheal bifurcation was labeled by the polygon formed by P_5 to P_{13} (fig. 1A). If the distal ETT end or tracheal bifurcation could not be reliably identified at the discretion of two intensivists, the chest radiographs were eliminated from further processing. The numbers of chest radiographs eliminated from the 4 folds were 6, 6, 9, and 7, respectively. Thus, the chest radiograph numbers became 462, 461, 458, and 461 in the 4 folds, making a total of 1,842 images from 1,842 patients included in the training and cross-validation datasets. Across the 4 folds, a cranially misplaced ETT (*i.e.*, the ETT tip more than 7 cm above the carina) was found in 101, 109, 91, and 99 chest radiographs, respectively; a caudally misplaced ETT (*i.e.*, the ETT tip less than 3 cm above the carina) was found in 38, 45, 31, and 40 chest radiographs, respectively.

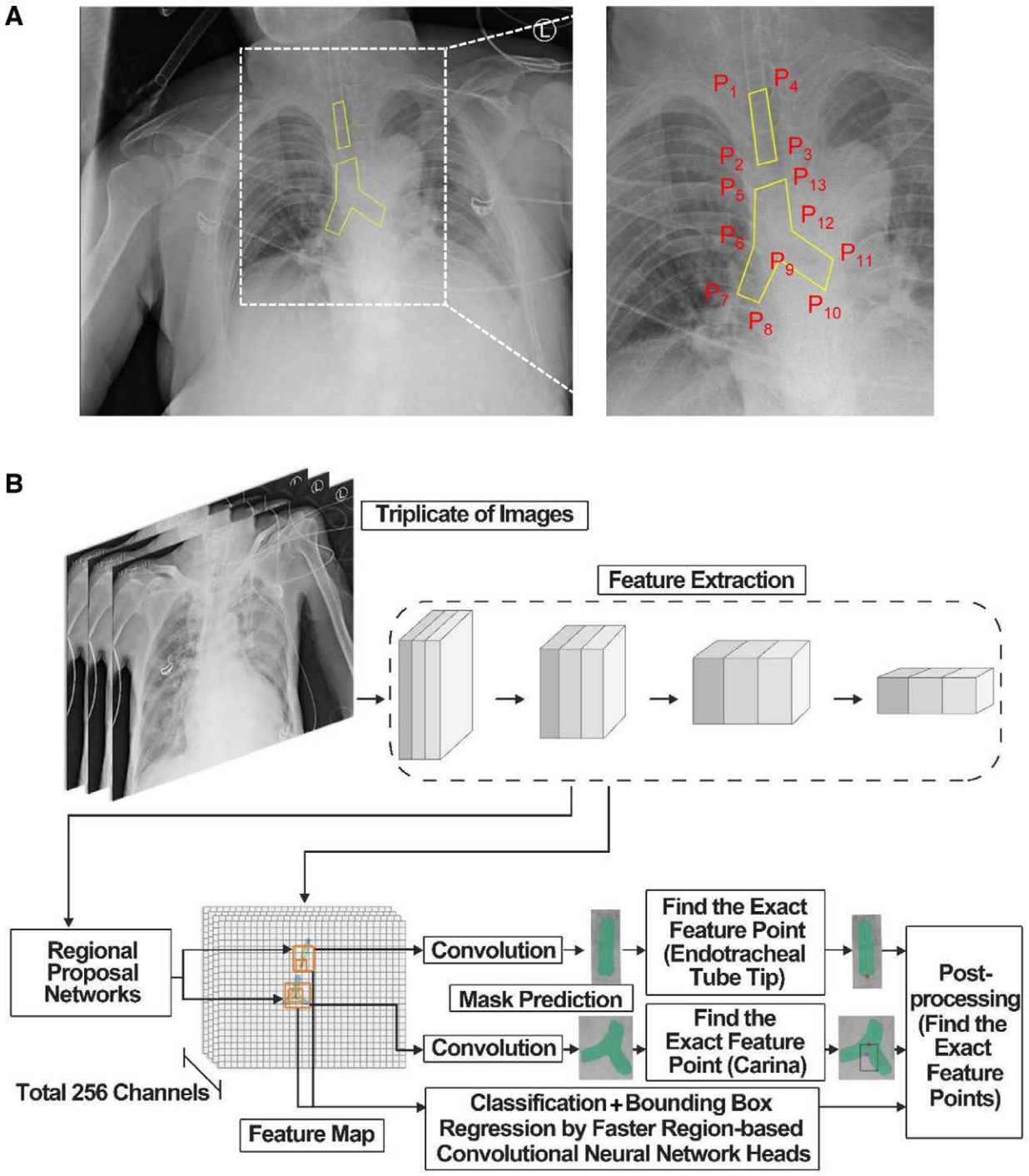

Fig. 1. Portable supine chest radiograph labeling and architecture of the deep learning–based algorithm. (*A*) Four points (P_1 to P_4) are used to label the distal endotracheal tube (ETT) end, and nine points (P_5 to P_{13}) are used to label the tracheal bifurcation as the ground truth. (*B*) The deep learning–based algorithm includes several steps, including feature extraction, mask prediction, classification, and bounding box regression. Because the architecture is originally designed to process the *red*, *green*, and *blue* color channels of the input images, triplicates of images are used as input to fit the three channels. The masks of the distal ETT end and tracheal bifurcation and the bounding boxes of the ETT tip and carina serve as supplements of each other to obtain the ETT tip and carina locations on chest radiographs.

Architecture of the Deep Learning–based Automatic Detection Algorithm

The ETT tip, defined as the midpoint between P_2 and P_3, and the carina (P_9) were selected as the feature points. In addition to the 13 points (P_1 to P_{13}) that labeled the distal ETT end and tracheal bifurcation, two ground-truth bounding boxes (used to define the location of the target objects; 48 × 48 pixels) with the ETT tip and carina at the center of each ground-truth bounding box were annotated.[21–23] The detection algorithm aimed to find the masks of the distal ETT end and tracheal bifurcation and the detected bounding boxes of the ETT tip and carina on portable supine chest radiographs (fig. 1B).

The mask region–based convolutional neural network (Mask R-CNN) has been known for its effectiveness in object recognition and instance segmentation.[21] In this study, a mask region–based convolutional neural network was trained to detect the distal ETT end and tracheal bifurcation masks through pixel-level segmentation of the two items. The mask region–based convolutional neural network algorithm for feature extraction was composed of 50-layer ResNeXt networks[24] as the backbone architecture with a recently proposed feature pyramid network.[22] During the inference step, only the masks with the maximal score for each class were preserved. A rule-based feature extraction method was performed to identify the feature points (*i.e.*, the ETT tip and carina) as the postprocessing procedure, which was developed based on the preliminary evaluation of the algorithm performance. To obtain the exact locations of the ETT tip and carina, the masks and detected bounding boxes localized by the mask region–based convolutional neural network were used to supplement each other. The ETT tip location was preferentially determined based on the detected bounding box center. Alternatively, the lowest point of the distal ETT end mask was accepted as the ETT tip location when the detected bounding box could not be identified on chest radiographs. Regarding the carina, the detected bounding box center was the preferred carina location. However, if the detected bounding box center was ≥100 pixels (13.9 mm) away from the feature point obtained from the tracheal bifurcation mask, the mask result was preferred as the carina location. The final detected locations of the distal ETT end and tracheal bifurcation were displayed as overlays on images. A supplemental method section (Supplemental Digital Content 1, http://links.lww.com/ALN/C918) is available to explain the architecture more thoroughly. The architecture of the deep learning–based algorithm and the rules in the postprocessing procedure were consistent during the training process of the four models. The ETT–carina distance was converted from pixels to millimeters using the pixel size of 0.139 mm based on the Digital Imaging and Communications in Medicine image data.

Validation Datasets

The validation steps included internal 4-fold cross-validation and external validation. Of the 4 folds, a single fold was retained as the validation dataset for testing the model, and the remaining 3 folds were used as the training datasets. For example, the first model was trained using the second, third, and fourth folds and tested using the first fold—each of the 4 folds serves as the validation dataset exactly once during 4-fold cross-validation. The external dataset was collected from intubated patients transferred from 12 neighboring urban hospitals between 2018 and 2019, whose images had been uploaded into the imaging database on patient admission. Overall, 216 de-identified chest radiographs were retrieved as the external validation dataset from the imaging database in our Department of Radiology.

Observer Performance Test

Eleven healthcare workers in the ICU, including two senior ICU nurse practitioners, two postgraduate year residents, five surgical residents, and two board-certified intensivists, participated in the observer performance test after consents were obtained. In Taiwan, postgraduate year residents participate in a generalized training program, and the surgical residency comes after 2 yr of postgraduate year residency training. Each clinician independently reviewed the first fold of the original dataset and labeled the ETT tip and carina on each chest radiograph. To ensure the labeling quality, these clinicians were temporarily exempted from clinical work and received standardized hands-on training before using the annotation software. The performance of each clinician was compared with that of the first model.

Performance Metrics and Statistical Analysis

The performance metrics measured consisted of the accuracy of ETT tip detection, carina detection, and ETT–carina distance measurement. As shown in Supplemental Digital Content 2 fig. S1 (http://links.lww.com/ALN/C919), the accuracy of ETT tip and carina detection was evaluated using the detection error between the detected location and ground truth location. Likewise, the accuracy of ETT–carina distance measurement was assessed using the measurement error between the estimated distance and ground truth distance. With reference from a previous study,[20] these performance metrics were further classified in terms of the errors from the ground truth within 5 mm, 10 mm, 15 mm, and beyond. In addition, whether the algorithm can detect a cranially misplaced ETT (*i.e.*, the ETT tip more than 7 cm above the carina) and a caudally misplaced ETT (*i.e.*, the ETT tip less than 3 cm above the carina) was evaluated. The overall performance of the algorithm in internal 4-fold cross-validation and external validation was calculated by pooling the results of the four individual models. During the observer performance test, the 462 chest radiographs in the first fold were used to compare the performance

(*i.e.*, the distribution of detection and measurement errors and proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error from the ground truth) of the algorithm and clinicians.

A data analysis and statistical plan were written after the data were accessed. Statistical analyses were performed using SPSS Statistics for Windows, Version 17.0 (SPSS Inc., USA). A P value < 0.05 was considered statistically significant. Categorical variables were expressed as percentage (number), whereas continuous variables were expressed as median (interquartile range). For categorical variables, independent samples (*i.e.*, comparisons in internal validation and internal validation results *versus* external validation results) were compared using the chi-square test and dependent samples (*i.e.,* comparisons in external validation and observer performance tests) using the McNemar test, with the P value of multiple comparisons adjusted by the Bonferroni correction. For comparisons of continuous variables between two independent groups (*i.e.*, internal validation results *versus* external validation results), the Mann–Whitney U test was used. For comparisons of continuous variables among multiple groups, independent samples (*i.e.*, comparisons in internal validation) were compared using the Kruskal-Wallis test, and dependent samples (*i.e.*, comparisons in external validation and observer performance tests) were compared using the Friedman test, followed by the *post hoc* analysis using Dunn's test.

Results

The overall performance of the deep learning–based automatic detection algorithm is summarized in table 1. During internal 4-fold cross-validation, the median error (interquartile range) and overall proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error of the deep learning–based algorithm were 2.8 (1.6 to 4.9) mm and 75.1%, 92.5%, and 96.4% in ETT tip detection, 3.6 (2.1 to 5.5) mm and 68.8%, 91.5%, and 95.6% in carina detection, and 3.9 (1.8 to 7.1) mm and 60.4%, 84.2%, and 92.8% in ETT–carina distance measurement, respectively. Among the four individual models, the performance (*i.e.*, the median error [interquartile range] and proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error from the ground truth) in ETT tip detection, carina detection, and ETT–carina distance measurement during internal 4-fold cross-validation was not significantly different (Supplemental Digital Content 3 table S1, http://links.lww.com/ALN/C920). During external validation, the median error (interquartile range) and overall proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error of the deep learning–based algorithm were 3.0 (1.7 to 5.3) mm and 72.6%, 90.4%, and 95.3% in ETT tip detection, 3.5 (2.0 to 5.9) mm and 67.8%, 89.2%, and 95.9% in carina detection, and 4.2 (1.7 to 7.8) mm and 57.6%, 83.2%, and 92.6% in ETT–carina distance measurement, respectively. Compared with the performance in internal cross-validation, the overall proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error and median error (interquartile range) from the ground truth in the three performance metrics were not significantly different, except a slight decline of median error (interquartile range) in ETT tip detection (2.8 [1.6 to 4.9] mm in internal cross-validation *versus* 3.0 [1.7 to 5.3] mm in external validation, $P = 0.046$). Thus, similar results were obtained during validation using the external dataset from neighboring hospitals. Among the four individual models, the accuracy of the three performance metrics was not significantly different during external validation (Supplemental Digital Content 4 table S2, http://links.lww.com/ALN/C921). For each model, the performance in ETT tip detection, carina detection, and ETT–carina distance measurement obtained during external validation (Supplemental Digital Content 4 table S2, http://links.lww.com/ALN/C921) was not significantly different from those obtained during internal

Table 1. Overall Performance of the Deep Learning–based Automatic Detection Algorithm

Validation*	Metric†	Median (Interquartile Range), mm	≤ 5 mm	≤ 10 mm	≤ 15 mm
Internal 4-fold cross-validation (n = 1,842)	ETT tip detection	2.8 (1.6–4.9)	75.1% (1,383)	92.5% (1,703)	96.4% (1,775)
	Carina detection	3.6 (2.1–5.5)	68.8% (1,268)	91.5% (1,686)	95.6% (1,760)
	ETT–carina distance	3.9 (1.8–7.1)	60.4% (1,112)	84.2% (1,551)	92.8% (1,709)
External validation (n = 864)‡	ETT tip detection	3.0 (1.7–5.3)§	72.6% (627)	90.4% (781)	95.3% (823)
	Carina detection	3.5 (2.0–5.9)	67.8% (586)	89.2% (771)	95.9% (829)
	ETT–carina distance	4.2 (1.7–7.8)	57.6% (498)	83.2% (719)	92.6% (800)

Error from the ground truth in ETT tip and carina detection and ETT–carina distance during internal 4-fold cross-validation (1,842 images, 1,842 patients) and external validation (216 images, 216 patients).

*The overall performance was calculated by pooling the results of the four individual models. †Data are expressed as percentage (number) unless otherwise indicated. §A statistically significant difference exists when compared with the result in internal 4-fold cross-validation ($P = 0.046$, Mann-Whitney U test). ‡The overall sample size in external validation is obtained from 216 images sampled four times by each individual model (n = 216 × 4 = 864).

ETT, endotracheal tube.

Table 2. Detection of a Cranially Misplaced Endotracheal Tube (*i.e.*, the Endotracheal Tube Tip more than 7 cm above the Carina) by the Deep Learning–based Automatic Detection Algorithm during Internal 4-Fold Cross-validation (1,842 Images, 1,842 Patients) and External Validation (216 Images, 216 Patients)

Validation*	Model†	Total Number of Images	Number of Images with the ETT tip > 7 cm above the Carina	Correctly Detected by the Algorithm (Sensitivity)	Number of Images with the ETT Tip ≤ 7 cm above the Carina	Correctly Detected by the Algorithm (Specificity)
Internal 4-fold cross-validation	1st	462	101	76.2% (77)	361	95.0% (343)
	2nd	461	109	82.6% (90)	352	94.9% (352)
	3rd	458	91	73.6% (67)	367	94.8% (348)
	4th	461	99	75.8% (75)	362	96.1% (348)
	Overall	1,842	400	77.3% (309)	1,442	95.2% (1,373)
External validation‡	1st	216	52	71.2% (37)	164	95.1% (156)
	2nd	216	52	73.1% (38)	164	93.3% (153)
	3rd	216	52	73.1% (38)	164	96.3% (158)
	4th	216	52	71.2% (37)	164	96.3% (158)
	Overall	864	208	72.1% (150)	656	95.3% (625)

*The overall performance was calculated by pooling the results of the four individual models. †Data are expressed as number or percentage (number). ‡The overall sample size in external validation is obtained from 216 images sampled 4 times by each individual model (n = 216 × 4 = 864).

ETT, endotracheal tube.

cross-validation (Supplemental Digital Content 3 table S1, http://links.lww.com/ALN/C920).

Whether the deep learning–based algorithm can detect a cranially or caudally misplaced ETT was also evaluated. For chest radiographs with a cranially misplaced ETT (table 2), the sensitivity and specificity of the algorithm were 77.3% and 95.2% during internal 4-fold cross-validation and 72.1% and 95.3% during external validation, respectively. For chest radiographs with a caudally misplaced ETT (table 3), the sensitivity and specificity of the algorithm were 70.8% and 96.4% during internal 4-fold cross-validation and 69.3% and 96.6% during external validation, respectively.

During the observer performance test, the median error (interquartile range) of the algorithm in ETT tip detection was 2.6 (1.6 to 4.8) mm (fig. 2), significantly superior to that of six clinicians. The sensitivities of the algorithm at 5 mm, 10 mm, and 15 mm error from the ground truth or less were 77.1%, 92.9%, and 96.5%, respectively (table 4). Compared with the 11 clinicians, the algorithm had significantly higher sensitivities than 7, 3, and 0 clinicians at 5 mm, 10 mm, and 15 mm error or less. No clinician was more accurate than the algorithm in ETT tip detection.

For carina detection, the median error (interquartile range) of the algorithm (3.6 [2.1 to 5.9] mm) was significantly

Table 3. Detection of a Caudally Misplaced Endotracheal Tube (*i.e.*, the Endotracheal Tube Tip less than 3 cm above the Carina) by the Deep Learning–based Automatic Detection Algorithm during Internal 4-Fold Cross-validation (1,842 Images, 1,842 Patients) and External Validation (216 Images, 216 Patients)

Validation*	Model†	Total Number of Images	Number of Images with the ETT Tip < 3 cm above the Carina	Correctly Detected by the Algorithm (Sensitivity)	Number of Images with the ETT tip ≥ 3 cm above the Carina	Correctly Detected by the Algorithm (Specificity)
Internal 4-fold cross-validation	1st	462	38	62.2% (24)	424	96.9% (411)
	2nd	461	45	75.6% (34)	416	95.4% (397)
	3rd	458	31	64.5% (20)	427	95.8% (409)
	4th	461	40	77.5% (31)	421	97.6% (411)
	Overall	1,842	154	70.8% (109)	1,688	96.4% (1,628)
External validation‡	1st	216	22	63.6% (14)	194	97.4% (189)
	2nd	216	22	68.2% (15)	194	97.9% (190)
	3rd	216	22	63.6% (14)	194	93.8% (182)
	4th	216	22	81.8% (18)	194	97.4% (189)
	Overall	864	88	69.3% (61)	776	96.6% (750)

*The overall performance was calculated by pooling the results of the four individual models. †Data are expressed as number or percentage (number). ‡The overall sample size in external validation is obtained from 216 images sampled 4 times by each individual model (n = 216 × 4 = 864).

ETT, endotracheal tube.

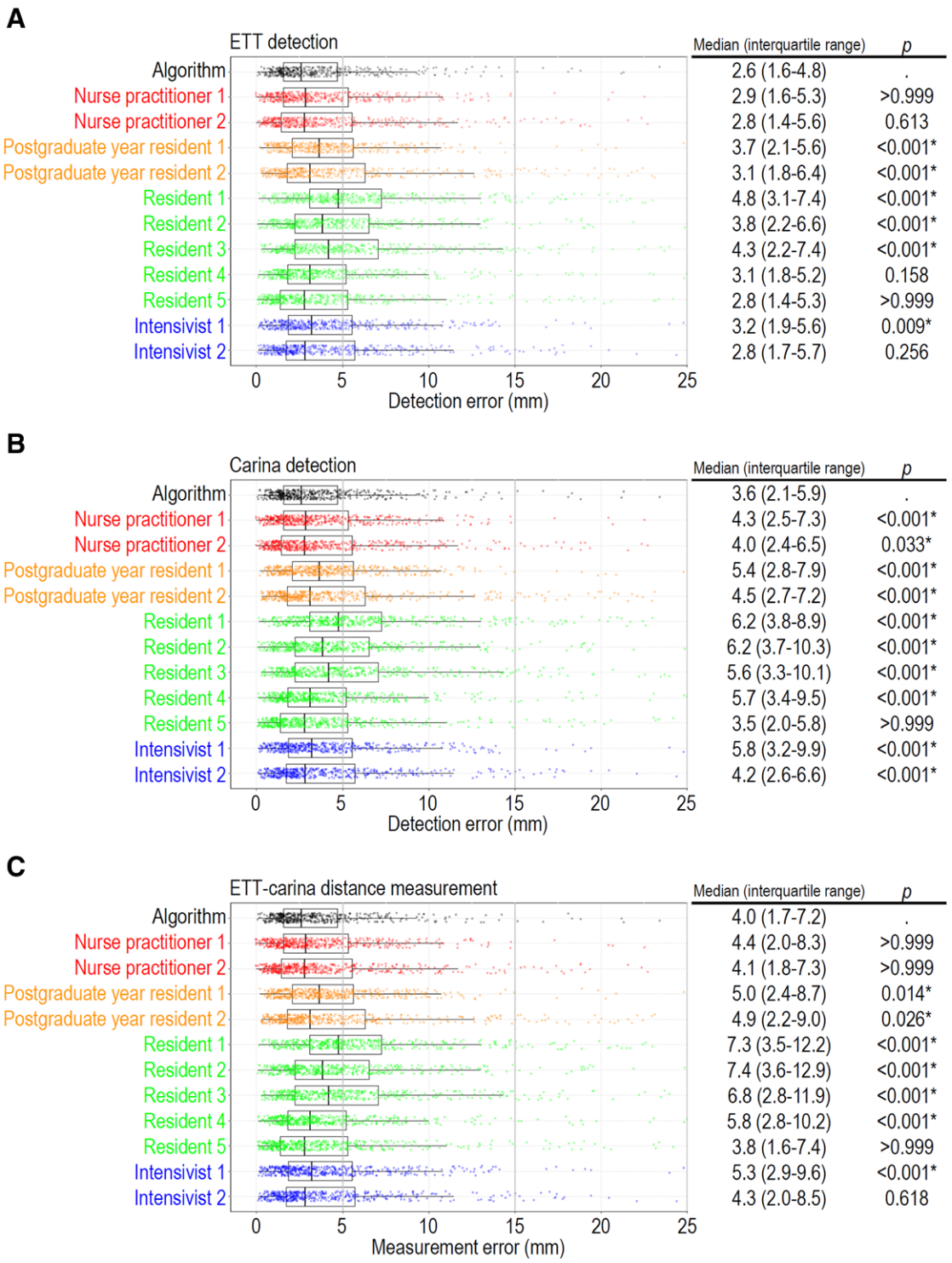

Fig. 2. Accuracy of endotracheal tube (ETT) tip detection, carina detection, and ETT–carina distance measurement in the observer performance test (462 images, 462 patients). (*A*) Distribution of detection error in ETT tip detection. (*B*) Distribution of detection error in carina detection. (*C*) Distribution of measurement error in ETT–carina distance measurement. *Signifies statistical significance when compared with the results of the algorithm.

Table 4. Endotracheal Tube Tip Detection in the Observer Performance Test (462 Images, 462 Patients)

	Detection Error from the Ground Truth					
Observer*†	**≤ 5 mm**	***P* Value‡**	**≤ 10 mm**	***P* Value‡**	**≤ 15 mm**	***P* Value‡**
Algorithm	77.1% (356)	—	92.9% (429)	—	96.5% (446)	—
Overall clinicians§	67.3% (3,411)	—	89.9% (4,561)	—	95.9% (4,865)	—
Nurse practitioner						
Nurse practitioner 1	73.2% (338)	0.139	93.7% (433)	0.652	98.3% (454)	0.134
Nurse practitioner 2	71.9% (332)	0.040	90.3% (417)	0.155	95.9% (443)	0.728
Postgraduate year resident‖						
Postgraduate year resident 1	68.2% (315)	0.001#	93.9% (434)	0.568	98.1% (453)	0.189
Postgraduate year resident 2	66.9% (309)	< 0.001#	85.7% (396)	< 0.001#	92.6% (428)	0.008
Surgical resident‖						
Resident 1	53.7% (248)	< 0.001#	84.0% (388)	< 0.001#	93.3% (431)	0.020
Resident 2	71.4% (330)	0.032	91.8% (424)	0.568	95.9% (443)	0.678
Resident 3	63.9% (295)	< 0.001#	89.0% (411)	0.330	96.3% (445)	> 0.999
Resident 4	58.7% (271)	< 0.001#	84.2% (389)	< 0.001#	93.1% (430)	0.020
Resident 5	72.9% (337)	0.113	92.6% (428)	> 0.999	96.8% (447)	> 0.999
Intensivist						
Intensivist 1	68.8% (318)	0.002#	91.6% (423)	0.461	97.4% (450)	0.523
Intensivist 2	68.8% (318)	0.001#	90.5% (418)	0.144	95.5% (441)	0.424

Comparisons of the deep learning–based automatic detection algorithm and clinicians in terms of the error from the ground truth within 5 mm, 10 mm, and 15 mm.

*Data are expressed as percentage (number). †Nurse practitioners 1 and 2 had 15 and 3 yr of intensive care unit experience; postgraduate year residents 1 and 2 were postgraduate year 1 and postgraduate year 2 residents; residents 1 and 2 were second-year surgical residents; residents 3, 4, and 5 were third-year surgical residents; intensivist 1 and 2 had 2 and 6 yr of intensive care unit experience. ‡Comparisons of the algorithm and clinicians were performed using the McNemar test. A *P* value < 0.005 (adjusted by the Bonferroni correction) was considered statistically significant. §The performance of overall clinicians was calculated by pooling the results of the 11 critical care clinicians. ‖Postgraduate year residents participate in a generalized training program, and the surgical residency comes after 2 yr of postgraduate year residency training. #Signifies statistical significance compared with the results of the algorithm.

superior to that of 10 clinicians (fig. 2). The sensitivities of the algorithm at the error of 5 mm, 10 mm, and 15 mm or less were 67.5%, 90.0%%, and 95.0%, respectively (table 5). Compared with the 11 clinicians, the algorithm was significantly more sensitive than 9, 6, and 4 clinicians at the error of 5 mm, 10 mm, and 15 mm or less. No clinician was significantly more accurate than the algorithm in carina detection.

The results of ETT–carina distance measurement of the algorithm and clinicians are shown in Supplemental Digital Content 5 fig. S2 (http://links.lww.com/ALN/C922). For ETT–carina distance measurement, the median error (interquartile range) of the algorithm (4.0 [1.7 to 7.2] mm) was significantly superior to that of 7 clinicians (fig. 2). Of the algorithm, the proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error from the ground truth were 59.3%, 84.4%, and 91.1%, respectively (table 6). In the comparisons with the 11 clinicians, the proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error of the algorithm were significantly higher than those of 5, 5, and 3 clinicians. No clinician was significantly more accurate than the algorithm in ETT–carina distance measurement.

Discussion

In the current study, we aimed to develop an algorithm to localize the ETT tip and carina on chest radiographs and estimate the ETT–carina distance. The performance of the algorithm was compared with that of clinicians in ETT tip detection, carina detection, and ETT–carina distance measurement in an observer performance test. Of note, the algorithm did perform better than some clinicians, and no clinician was more accurate than the algorithm in any comparison (regardless of the distribution of errors or proportions of chest radiographs within 5 mm, 10 mm, or 15-mm error). Thus, although the clinical effects remain to be determined, the deep learning–based algorithm might play a role to complement and augment the ability of critical care clinicians by offloading their routine duties and enabling them to focus on cognitively demanding tasks.

Several study groups and companies have announced working on relevant projects. However, only Lakhani *et al.*[19,20] documented the details of their approaches and results in two studies. Their deep learning–based algorithms were trained using image classification, *i.e.*, category labeling for the entire image rather than annotation for specific objects. In their former study,[19] the authors found that the deep convolutional neural networks achieved a relatively poorer area under the curve of 0.81 in differentiating the low or normal position of the ETT. In the latter study,[20] the Inception V3 deep neural network was used to classify the ETT–carina distance. A total of 22,960 chest radiographs were classified into 12 categories, including bronchial

Table 5. Carina Detection in the Observer Performance Test (462 Images, 462 Patients): Comparisons of the Deep Learning–based Automatic Detection Algorithm and Clinicians in Terms of the Error from the Ground Truth within 5 mm, 10 mm, and 15 mm

	Detection Error from the Ground Truth					
Observer*†	**≤ 5 mm**	***P* Value‡**	**≤ 10 mm**	***P* Value‡**	**≤ 15 mm**	***P* Value‡**
Algorithm	67.5% (312)	—	90.0% (416)	—	95.0% (439)	—
Overall clinicians§	50.7% (2,570)	—	82.8% (4,197)	—	92.4% (4,687)	—
Nurse practitioner						
Nurse practitioner 1	56.1% (259)	< 0.001‖	88.1% (407)	0.380	95.2% (440)	> 0.999
Nurse practitioner 2	65.2% (301)	0.419	88.5% (409)	0.494	96.8% (447)	0.215
Postgraduate year resident#						
Postgraduate year resident 1	47.6% (220)	< 0.001‖	83.1% (384)	0.002‖	93.7% (433)	0.461
Postgraduate year resident 2	54.3% (251)	< 0.001‖	84.8% (392)	0.009	92.2% (426)	0.085
Surgical resident#						
Resident 1	37.2% (172)	< 0.001‖	79.4% (367)	< 0.001‖	90.5% (418)	0.008
Resident 2	69.3% (320)	0.275	92.2% (426)	0.237	97.2% (449)	0.110
Resident 3	39.8% (184)	< 0.001‖	74.5% (344)	< 0.001‖	87.4% (404)	< 0.001‖
Resident 4	42.0% (194)	< 0.001‖	74.9% (346)	< 0.001‖	88.1% (407)	< 0.001‖
Resident 5	42.9% (198)	< 0.001‖	76.6% (354)	< 0.001‖	88.7% (410)	0.001‖
Intensivist						
Intensivist 1	42.4% (196)	< 0.001‖	76.4% (353)	< 0.001‖	87.2% (403)	< 0.001‖
Intensivist 2	59.5% (275)	0.003‖	89.8% (415)	> 0.999	97.4% (450)	0.071

*Data are expressed as percentage (number). †Nurse practitioners 1 and 2 had 15 and 3 yr of intensive care unit experience; postgraduate year residents 1 and 2 were postgraduate year 1 and postgraduate year 2 residents; residents 1 and 2 were second-year surgical residents; residents 3, 4, and 5 were third-year surgical residents; intensivist 1 and 2 had 2 and 6 yr of intensive care unit experience. ‡Comparisons of the algorithm and clinicians were performed using the McNemar test. A *P* value < 0.005 (adjusted by the Bonferroni correction) was considered statistically significant. §The performance of overall clinicians was calculated by pooling the results of the 11 critical care clinicians. ‖Signifies statistical significance compared with the results of the algorithm. #Postgraduate year residents participate in a generalized training program, and the surgical residency comes after 2 yr of postgraduate year residency training.

insertion, distance from the carina at 1.0-cm intervals up to 10 cm (0.0 to 0.9 cm, 1.0 to 1.9 cm, …, 9.0 to 9.9 cm), and 10 cm or greater. The mean differences between the algorithm and radiologists in ETT–carina distance were 0.69 ± 0.70 cm on the internal test dataset and 0.63 ± 0.55 cm on the external test dataset, with both intraclass correlation coefficients greater than 0.8. On the internal test images, the algorithm was 66.5% sensitive and 99.2% specific in detecting ETT–carina distance greater than 7 cm and 95.0% sensitive and 91.8% specific in detecting ETT–carina distance less than 3 cm, respectively. Although Lakhani's work is more sensitive in detecting a caudally misplaced ETT, our algorithm performs slightly better in detecting a cranially misplaced ETT. However, as acknowledged by the authors, an approach using such "weak" labeling needed substantially more training data. More important, the algorithms were trained through image classification (*i.e.*, low or normal position of the ETT in the former study and 12 numerical categories of ETT–carina distance in the latter study). No accurate object annotation for the ETT and carina was made on training dataset images so that the deep learning solutions classified only the low or normal position of the ETT or ETT–carina distance. Therefore, if the clinicians have any suspicion of the ETT–carina distance reported by the algorithm, no localization information of the ETT and carina can be provided.

In the current study, we aimed to improve model explainability (*i.e.*, transparency) using a deep learning–based object detection algorithm instead of image classification. The deep learning–based algorithm learned how to localize the ETT tip and carina on chest radiographs to estimate the ETT–carina distance. Although pixel-level segmentation labeling performed by two board-certified intensivists together was a time-consuming and labor-intensive task, the image annotations using 4 and 9 points each provided abundant information to recognize the distal ETT end and tracheobronchial tree on chest radiographs, substantially reducing the number of images in training datasets. In addition, complementary application of the mask and bounding box results may enhance ETT tip and carina detection and consequently contribute to the accuracy of ETT–carina distance measurement. The deep learning–based algorithm, trained using the bounding boxes denoting the ETT tip and carina locations and pixel-level segmentation of the distal ETT end and tracheal bifurcation, exhibited robustness in ETT–carina distance measurement during internal cross-validation and external validation. In addition, the overlays, which localize the distal ETT end and tracheal bifurcation on images, can help users perceive the ETT tip in relation to the carina (fig. 3), especially when a disagreement exists between the interpretation of clinicians and the detection of the algorithm.

Table 6. Endotracheal Tube–Carina Distance Measurement in the Observer Performance Test (462 Images, 462 Patients): Comparisons of the Deep Learning–based Automatic Detection Algorithm and Clinicians in Terms of the Error from the Ground Truth within 5 mm, 10 mm, and 15 mm

	Measurement Error from the Ground Truth					
Observer*†	**≤ 5 mm**	***P* Value‡**	**≤ 10 mm**	***P* Value‡**	**≤ 15 mm**	***P* Value‡**
Algorithm	59.3% (274)	—	84.4% (390)	—	91.1% (421)	—
Overall clinicians§	49.0% (2,487)	—	76.9% (3,901)	—	88.9% (4,508)	—
Nurse practitioner						
Nurse practitioner 1	56.5% (261)	0.397	82.0% (379)	0.351	92.6% (428)	0.464
Nurse practitioner 2	59.5% (275)	> 0.999	87.2% (403)	0.208	94.4% (436)	0.063
Postgraduate year resident‖						
Postgraduate year resident 1	50.6% (234)	0.010	80.5% (372)	0.127	93.1% (430)	0.314
Postgraduate year resident 2	51.3% (237)	0.011	79.2% (366)	0.031	88.3% (408)	0.160
Surgical resident‖						
Resident 1	37.2% (172)	< 0.001#	66.5% (307)	< 0.001#	83.1% (384)	< 0.001#
Resident 2	61.0% (282)	0.614	84.8% (392)	0.920	94.4% (436)	0.044
Resident 3	34.4% (159)	< 0.001#	63.0% (291)	< 0.001#	79.4% (367)	< 0.001#
Resident 4	41.1% (190)	< 0.001#	68.4% (316)	< 0.001#	84.2% (389)	0.002#
Resident 5	42.9% (198)	< 0.001#	75.3% (348)	0.001#	87.7% (405)	0.101
Intensivist						
Intensivist 1	46.8% (216)	< 0.001#	76.0% (351)	0.001#	87.2% (403)	0.050
Intensivist 2	56.9% (263)	0.462	81.4% (376)	0.223	91.3% (422)	> 0.999

*Data are expressed as percentage (number). †Nurse practitioners 1 and 2 had 15 and 3 yr of intensive care unit experience; postgraduate year residents 1 and 2 were postgraduate year 1 and postgraduate year 2 residents; residents 1 and 2 were second-year surgical residents; residents 3, 4, and 5 were third-year surgical residents; intensivist 1 and 2 had 2 and 6 yr of intensive care unit experience. ‡Comparisons of the algorithm and clinicians were performed using the McNemar test. A *P* value < 0.005 (adjusted by the Bonferroni correction) was considered statistically significant. §The performance of overall clinicians was calculated by pooling the results of the 11 critical care clinicians. ‖Postgraduate year residents participate in a generalized training program, and the surgical residency comes after 2 yr of postgraduate year residency training. #Signifies statistical significance compared with the results of the algorithm.

For the deep learning–based automatic detection algorithm, ETT tip and carina detection was accurate to within a 10-mm error from the ground truth in ~90% of images and within 15-mm error in ≈95% of images. In addition, ETT–carina distance measurement was accurate to within 10-mm error in ≈85% of images and within 15-mm error in ≈90% of images. More important, the performance of the deep learning–based algorithm was consistent in internal 4-fold cross-validation and external validation. We compared the performance of the deep learning–based algorithm with that of a diverse group of 11 critical care clinicians. In terms of the median error (interquartile range) from the ground truth, the algorithm performed better than 6, 10, and 7 clinicians in ETT tip detection, carina detection, and ETT–carina distance measurement, respectively. The algorithm was superior to 7, 3, and 0, 9, 6, and 4, and 5, 5, and 3 clinicians regarding the proportions of chest radiographs within 5 mm, 10 mm, and 15 mm error in ETT tip detection, carina detection, and ETT–carina distance measurement. The algorithm outperformed clinicians in many comparisons, particularly when a lower error (*i.e.*, 5 mm) from the ground truth was allowed. No clinician was significantly more accurate than the algorithm in terms of the sensitivities within 5 mm, 10 mm, and 15 mm error or median error (interquartile range) from the ground truth. These findings suggested that the deep learning–based automatic detection algorithm can match or even outperform frontline critical care clinicians in measuring the ETT–carina distance. Whether clinical use of the algorithm might reduce complications associated with ETT malposition and improve the ICU workflow warrants further investigation.

The current study has some limitations. First, in the observer performance test, only the performance of the first model was compared with that of the clinicians. The performance of the four individual models was not significantly different during internal 4-fold cross-validation and external validation. However, it is not equivalent to or in place of comparing the other three individual models with clinicians. Second, the possibility of overfitting cannot be avoided considering that a rule-based feature extraction method was used as the postprocessing procedure in identifying the ETT tip and carina. Finally, the performance of our algorithm cannot be compared comprehensively with previous works. The algorithms presented in previous studies were trained using image classification,[19,20] and thus the area under the curve and intraclass correlation coefficients are used as evaluation metrics. However, in a detection task like our work, using the area under the curve or intraclass correlation coefficients to evaluate the algorithm

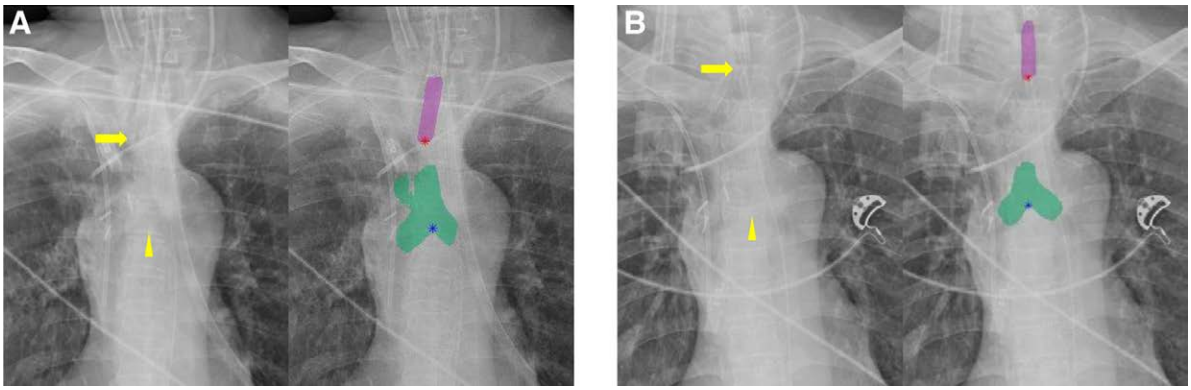

Fig. 3. Endotracheal tube (ETT) tip and carina detection by the deep learning–based algorithm. The *yellow* arrow indicates the ETT tip, and the *yellow arrowhead* indicates the carina on chest radiographs. The overlays of the deep learning–based algorithm may help users perceive the ETT tip in relation to the carina on images. (*A*) A properly placed ETT. (*B*) Interval cephalic migration of the ETT.

performance tends to discretize continuous variables, leading to a loss of information. Also, evaluation using different testing datasets could result in biased comparisons. A standard regarding the performance evaluation for relevant studies remains lacking. Thus, conducting an observer performance test using the same dataset may be a more feasible and direct approach to identify whether the algorithm works before further validation.

In summary, we have developed a deep learning–based automatic detection algorithm detecting the ETT tip and carina on portable supine chest radiographs to measure the ETT–carina distance. Our study demonstrates that the deep learning–based algorithm is comparable or even superior to frontline critical care clinicians in detecting the ETT tip and carina and measuring the ETT–carina distance.

Acknowledgments

The authors appreciate Cheng-Shih Lai, B.S., at the Department of Radiology, National Cheng Kung University Hospital, Taiwan, for his excellent technical support and Kai-Wen Li, M.S., at the Department of Nursing, National Cheng Kung University Hospital, Taiwan, for her laborious contribution to this work.

Research Support

Support for article research was provided from the Ministry of Science and Technology, Executive Yuan, Taiwan (MOST 109-2634-F-006-023) and from National Cheng Kung University Hospital, Tainan, Taiwan (NCKUH-10901003).

Competing Interests

Dr. Lai received support for article research from the Ministry of Science and Technology, Executive Yuan, Taiwan (MOST 109-2634-F-006-023) and from National Cheng Kung University Hospital, Tainan, Taiwan (NCKUH-10901003). The other authors declare no competing interests.

Correspondence

Address correspondence to Dr. Lai, Department of Surgery, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, No. 138, Sheng-Li Rd, Tainan 70403, Taiwan. d303878@mail.hosp.ncku.edu.tw.

Supplemental Digital Content

Supplemental Digital Content 1: Supplemental method, http://links.lww.com/ALN/C918
Supplemental Digital Content 2: Fig. S1, http://links.lww.com/ALN/C919
Supplemental Digital Content 3: Table S1, http://links.lww.com/ALN/C920
Supplemental Digital Content 4: Table S2, http://links.lww.com/ALN/C921
Supplemental Digital Content 5: Fig. S2, http://links.lww.com/ALN/C922

References

1. Brown CA III, Bair AE, Pallin DJ, Walls RM; NEAR III Investigators: Techniques, success, and adverse events of emergency department adult intubations. Ann Emerg Med 2015; 65:363–70.e1
2. Ono Y, Kakamu T, Kikuchi H, Mori Y, Watanabe Y, Shinohara K: Expert-performed endotracheal

intubation-related complications in trauma patients: Incidence, possible risk factors, and outcomes in the prehospital setting and emergency department. Emerg Med Int 2018; 2018:5649476
3. Sitzwohl C, Langheinrich A, Schober A, Krafft P, Sessler DI, Herkner H, Gonano C, Weinstabl C, Kettner SC: Endobronchial intubation detected by insertion depth of endotracheal tube, bilateral auscultation, or observation of chest movements: randomised trial. BMJ 2010; 341:c5943
4. Brunel W, Coleman DL, Schwartz DE, Peper E, Cohen NH: Assessment of routine chest roentgenograms and the physical Examination to confirm endotracheal tube position. Chest 1989; 96:1043–5
5. Goodman LR, Conrardy PA, Laing F, Singer MM: Radiographic evaluation of endotracheal tube position. AJR Am J Roentgenol 1976; 127:433–4
6. Lotano R, Gerber D, Aseron C, Santarelli R, Pratter M: Utility of postintubation chest radiographs in the intensive care unit. Crit Care 2000; 4:50–3
7. Bentz MR, Primack SL: Intensive care unit imaging. Clin Chest Med 2015; 36:219–34, viii
8. Amorosa JK, Bramwit MP, Mohammed TL, Reddy GP, Brown K, Dyer DS, Ginsburg ME, Heitkamp DE, Jeudy J, Kirsch J, MacMahon H, Ravenel JG, Saleh AG, Shah RD: ACR appropriateness criteria routine chest radiographs in intensive care unit patients. J Am Coll Radiol 2013; 10:170–4
9. Hobbs DL: Chest radiography for radiologic technologists. Radiol Technol 2007; 78:494–516; quiz 7–9
10. Schaefer-Prokop C, Neitzel U, Venema HW, Uffmann M, Prokop M: Digital chest radiography: an update on modern technology, dose containment and control of image quality. Eur Radiol 2008; 18:1818–30
11. Wiener MD, Garay SM, Leitman BS, Wiener DN, Ravin CE: Imaging of the intensive care unit patient. Clin Chest Med 1991; 12:169–98
12. Wunsch H, Wagner J, Herlim M, Chong DH, Kramer AA, Halpern SD: ICU occupancy and mechanical ventilator use in the United States. Crit Care Med 2013; 41:2712–9
13. Gonem S, Janssens W, Das N, Topalovic M: Applications of artificial intelligence and machine learning in respiratory medicine. Thorax 2020; 75:695–701
14. Gutierrez G: Artificial intelligence in the intensive care unit. Crit Care 2020; 24:101
15. Massion PP, Antic S, Ather S, Arteta C, Brabec J, Chen H, Declerck J, Dufek D, Hickes W, Kadir T, Kunst J, Landman BA, Munden RF, Novotny P, Peschl H, Pickup LC, Santos C, Smith GT, Talwar A, Gleeson F: Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. Am J Respir Crit Care Med 2020; 202:241–9
16. González G, Ash SY, Vegas-Sánchez-Ferrero G, Onieva J, Rahaghi FN, Ross JC, Díaz A, San José Estépar R, Washko GR; COPDGene and ECLIPSE Investigators: Disease staging and prognosis in smokers using deep learning in chest computed tomography. Am J Respir Crit Care Med 2018; 197:193–203
17. Nam JG, Kim M, Park J, Hwang EJ, Lee JH, Hong JH, Goo JM, Park CM: Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. Eur Respir J 2021; 57:2003061
18. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, Goo JM, Aum J, Yim JJ, Park CM; Deep Learning-Based Automatic Detection Algorithm Development and Evaluation Group: Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 2019; 69:739–47
19. Lakhani P: Deep convolutional neural networks for endotracheal tube position and x-ray image classification: Challenges and opportunities. J Digit Imaging 2017; 30:460–8
20. Lakhani P, Flanders A, Gorniak R: Endotracheal tube position assessment on chest radiographs using deep learning. Radiol Artif Intell 2021; 3:e200026
21. He K, Gkioxari G, Dollar P, Girshick R: Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 2020; 42:386–97
22. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S: Feature pyramid networks for object detection. Proc IEEE Conf Comput Vis Pattern Recognit 2017:2117–25
23. Ren S, He K, Girshick R, Sun J: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017; 39:1137–49
24. Xie S, Girshick R, Dollár P, Tu Z, He K: Aggregated residual transformations for deep neural networks. Proc IEEE Conf Comput Vis Pattern Recognit 2017:1492–500