Sensitivity and specificity of artificial intelligence with Microsoft Azure in detecting pneumothorax in emergency department: A pilot study

Kwok Hung Alastair Lai and Shu Kai Ma

Abstract

Background: Artificial intelligence is becoming an increasingly important tool in different medical fields. This article aims to evaluate the sensitivity and specificity of artificial intelligence trained with Microsoft Azure in detecting pneumothorax.

Methods: A supervised learning artificial intelligence is trained with a collection of X-ray images of pneumothorax from National Institutes of Health chest X-ray dataset online. A subset of the image dataset focused on pneumothorax is used in training. Two artificial intelligence programs are trained with different numbers of training images. After the training, a collection of pneumothorax X-ray images from patient attending emergency department is retrieved through the Clinical Data Analysis & Reporting System. In total, 115 pneumothorax patients and 60 normal inpatients are recruited. The pneumothorax chest X-ray and the resolution chest X-ray of the above patient group and a collection of normal chest X-ray from inpatients without pneumothorax will be retrieved, and these three sets of images will then undergo testing by artificial intelligence programs to give a probability of being a pneumothorax X-ray.

Results: The sensitivity of artificial intelligence-one is 33.04%, and the specificity is at least 61.74%. The sensitivity of artificial intelligence-two is 46.09%, and the specificity is at least 71.30%. The dramatic improvement of 46.09% in sensitivity and improvement of 15.48% in specificity by addition of around 1000 X-ray images is encouraging. The mean improvement of AI-two over AI-one is 19.7% increase in probability difference.

Conclusions: We should not rely on artificial intelligence in diagnosing pneumothorax X-ray solely by our models and more training should be expected to explore its full function.

Keywords
Pneumothorax, chest X-ray, artificial intelligence, Microsoft Azure, NIH CXR

Introduction

Artificial intelligence (AI) has been playing an increasing role in different aspects of human industry since its emergence. A variety of human industries have become increasingly dependent on the promising technology it provides including the medical field.1,2 AI surveillance programs may help radiologists prioritize work lists by identifying...
suspicious or positive cases for early review. At the same time, there are several technical challenges like complex and heterogeneous datasets, noisy medical datasets, and explaining their output to users.

An AI can be defined as a system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation. Machine learning is a subset of AI, which can be termed as an application of AI. The machine learning process consists of different algorithms, which can be broadly classified into supervised, semisupervised, and unsupervised learnings. Supervised learning in the context of AI, and machine learning can be defined as a type of system in which both input and desired output data are provided so that the aim is to provide a learning algorithm with known quantities to support future judgments. With an increasing database of images of different diseases, a supervised learning can be utilized to train a computer program for a desired pattern recognition and hence aiding a clinical decision-making process.

This article serves as a pilot study on the role of AI in the diagnosis of pneumothorax in emergency department. Pneumothorax is a common condition encountered in emergency department, and the diagnosis is radiologically prompted by clinical presentation. In one study performed by multiple network architects with 13,392 chest X-rays (CXR), internal testing regardless of pneumothorax size demonstrated a sensitivity of 0.55 and a specificity of 0.9, and the external testing on the National Institutes of Health (NIH) dataset showed performance decline with sensitivity 0.28–0.49 and specificity 0.85–0.97. Another study performed under full 26-layer convolutional neural network with 1596 CXR showed sensitivity and specificity of 61.1% and 93.0% in detecting pneumothorax 3 h after percutaneous transthoracic needle biopsy (PTNB) for pulmonary lesions. In another multicenter diagnostic cohort study, a commercial AI platform is used and the algorithm exhibited lower sensitivity (77.3% vs 81.8–97.7%) and higher specificity (97.6% vs 81.7–96.0%) than the radiologists.

**Methods**

**Study design and subjects**

We have utilized the available AI resource online and trained a program from scratch. The machine learning was a supervised learning based on publicly released NIH CXR dataset. The NIH CXR dataset consists of 112,120 X-ray images with disease labels from 30,805 individuals. A subset of the labeled images was used in the training of the AI (see Figure 1). The machine learning algorithm was based on the Microsoft Azure AI service.

The training data set was in the PNG file format with size 1024×1024. No preprocessing is performed on the image. Two AI programs were trained as a supervised classification. AI-one was trained with the first 1028 CXR labeled with pneumothorax and the first 1023 normal CXR from the training data set. AI-two was trained with the first 2187 pneumothorax CXR and the first 2192 normal CXR, with a probability of 50% as cutoff value for a positive prediction.

After training of the model, a retrospective retrieval of the patient group through the Clinical Data Analysis and Reporting System (CDARS) was done (see Figure 2).

CDARS is a computer system implemented in public hospitals in Hong Kong, which helps to facilitate the retrieval of clinical data captured from different operational systems. It was utilized for retrieval of subjects, and their images will be exported as a JPG file with appropriate size to be saved in the storage server. All the patients attending the Accident and Emergency Department of our hospital from 9 July 2014 to 9 June 2019 were retrieved. The diagnosis code (ICD9) for data search included spontaneous tension pneumothorax (512.0), iatrogenic pneumothorax (512.2), other spontaneous pneumothorax (512.8), traumatic pneumothorax without mention of open-wound into thorax (860.0), and traumatic pneumothorax with open-wound into thorax (860.4). There was no age limitation of the patient recruitment. The diagnosis of pneumothorax was made upon admission by emergency physicians. It was further confirmed with in-hospital record.

One hundred sixty cases were retrieved. Forty-five cases were excluded. Twenty-one cases were excluded due to repeated entry in the CDARS, seven were excluded as the X-ray turn out to be without pneumothorax, six were excluded as no resolution X-ray are available, six were excluded as the patients succumbed during the same admission, three were excluded as the patients were discharged against medical advice before resolution of pneumothorax, and two were excluded as the patients were found to have...
pneumothorax on CT for other disease condition and the pneumothorax were not observable on CXR. A total of 115 cases underwent testing by the trained AI. Another 60 inpatients without pneumothorax were randomly selected and underwent a similar AI prediction process, and their specificity was calculated.

Data collection
For the interval validation, after training the models, a precision and recall graph was constructed with different cut-off thresholds for positive finding for the training sample.

For the external validation, we have chosen 50% as the threshold for a positive finding. Each X-ray was tested and it was labeled as pneumothorax if the predicted probability was greater than 50%; otherwise, it was labeled normal.

Statistical analysis
Statistical analyses were performed using the IBM SPSS subscription. During the training of the AIs, the probability threshold of 50% was chosen as the cutoff value for detecting pneumothorax. Precision and recall are two evaluation metrics in machine learning. Precision is defined as the fraction of identified classifications that are correct, for example, true positive/(true positive + false positive). Recall is defined as the fraction of actual classifications that are correctly identified, for example, true positive/(true positive + false negative).

A precision and recall figure was obtained for the interval validation and the area under curve (AUC) will be calculated. For external validation, the X-ray images were tested for the possibility of being pneumothorax. The sensitivity and specificity were calculated. The receiver operating characteristic (ROC) curve and the AUC were calculated. To evaluate the effectiveness of additional training on the AI programmes, we performed a paired sample t-test. For each of the AIs, the change in the predicted probability in the pneumothorax XR and the resolution XR was calculated (e.g. predicted probability in pneumothorax XR – predicted probability in resolution XR). The directional change was taken into consideration. A change would be positive if the AIs were performing what we expected, e.g. the AI programmes would give a

Figure 2. Flow diagram of subject recruitment.
higher probability in the pneumothorax XR. A paired-sample t-test was then performed on the changes in probability of the two AIs. Different from trained physicians who identify pneumothorax by visual detection for pleural edge and radio-lucency at peripheral indicating the absence of lung markings, when training the AI, each X-ray is broken down into each pixel level and trained with the Microsoft Azure algorithm. Different parts of the CXR, for example, ribs, heart, mediastinal contour, lung markings, and so on, will contribute differently to the final predicted probability. If there is a statistically significant improvement in the predicted probability difference, it may indicate the image features we physicians focus most, for example, the absence of lung markings at the peripheral, contribute much lower weighting in the AI model being built.

Results

AI training result

After training, the AI-one and AI-two have an overall precision (with recall of the same value under our chosen threshold) of 80.3% and 84.5%, respectively. It means when applying the trained AIs back to the training images, among the positive prediction by the AI-one, 80.3% (and 84.5%
for AI-two) of those prediction is correct. And 80.3% of those labeled images in the training set can be correctly identified. The AUC of PR graph on AI-one is 0.89 and AI-two is 0.91. It shows it has a good interval validation (Figure 6(a)).

**Prediction**

Figures 3–5 show the predicted probability of each X-ray test with our models.

For the patients with pneumothorax, the sensitivity for AI-one in detecting pneumothorax is 33.04% (38/115), and for AI-two, it is 46.09% (53/115). The relative improvement (46.09%−33.04%/33.04%=39.5%) in sensitivity is contributed by the addition of 1000 images.

For the resolution CXR, the specificity of AI-one is 61.74% (71/115); for AI-two, it is 71.30% (82/115) with a 15.48% improvement (73.8%−66.7%)/66.7%.

For the normal inpatient CXR, the specificity of AI-one is 80% (48/60); for AI-two, it is 81.67% (49/60), with a 2% change.

### Table 1. Performance of AI-one in pneumothorax XR and resolution XR.

|       | 115 with PTHx | 115 resolution XR |
|-------|---------------|--------------------|
| Positive | 38 | 44 |
| Negative | 77 | 71 |
| Sensitivity = 38/ (38 + 77) = 33.04% (95% CI: 24.72%–42.51%) | Specificity = 71/ (44 + 71) = 61.74% (95% CI: 52.17%–70.51%) |

AI: artificial intelligence; CI: confidence interval; XR: X-ray.

### Table 2. Performance AI-one in pneumothorax XR and normal inpatient XR groups.

|       | 115 with PTHx | 60 normal XR |
|-------|---------------|--------------|
| Positive | 38 | 12 |
| Negative | 77 | 48 |
| Sensitivity = 38/ (38 + 77) = 33.04% (95% CI: 24.72%–42.51%) | Specificity = 48/ (12 + 48) = 80% (95% CI: 67.30%–88.81%) |

AI: artificial intelligence; CI: confidence interval; XR: X-ray.

### Table 3. Performance of AI-two in pneumothorax XR and resolution XR.

|       | 115 with PTHx | 115 resolution XR |
|-------|---------------|--------------------|
| Positive | 53 | 33 |
| Negative | 62 | 82 |
| Sensitivity = 53/ (53 + 62) = 46.09% (95% CI: 36.84%–55.60%) | Specificity = 82/ (33 + 82) = 71.3% (95% CI: 62.00%–79.16%) |

AI: artificial intelligence; CI: confidence interval; XR: X-ray.

### Table 4. Performance of AI-two in pneumothorax XR and normal inpatient XR.

|       | 115 with PTHx | 60 normal XR |
|-------|---------------|--------------|
| Positive | 53 | 11 |
| Negative | 62 | 49 |
| Sensitivity = 53/ (53 + 62) = 46.09% (95% CI: 36.84%–55.60%) | Specificity = 49/ (11 + 49) = 81.67% (95% CI: 69.15%–90.07%) |

AI: artificial intelligence; CI: confidence interval; XR: X-ray.

The AUC of ROC for AI-one is 0.50 (95% confidence interval (CI): 0.45–0.57), and that for AI-two is 0.62. (95% CI: 0.56–0.69) (Figure 6(b)). The p value = 0.02 for the change in AUC.

A paired sample t-test was performed on the change in the predicted probability in the pneumothorax XR and the resolution XR. A two-tailed test was performed on AI-one and AI-two. The mean of the change in probability of
AI-two minus AI-one is 19.7% (95% CI: 9.7%–29.7%). The two-tailed p value is 0.0002, t=3.90.

Discussion

In the above study, we have shown that a more well-trained AI with more training images can increase the sensitivity and specificity in detecting pneumothorax (see Tables 1–4).

To visualize the enhancement of the AI, we can construct a probability graph. In the above data, the predicted probability graph is plotted with the recruited subject in order. The data points appear to be chaotic. To enhance the visual presentation, we can rearrange the probability in ascending order.

If we overlay both graphs of AI-one and AI-two, we will get the following graphs (see Figure 7(a) to (c)). From the graph, we can see that the sensitivity and specificity in the pneumothorax CXR and resolution CXR are improved in AI-two. The systematic improvement results in a left shift of the graph, for example, the intersection of the horizontal line of a predicted probability of 50% with the curves is left-shifted, indicating that more cases are correctly identified. Yet, the improvement on the normal inpatient CXR is not visually significant.

The key finding is that with easily accessible and easy to use Microsoft Azure online resource, we can construct an AI with a good specificity. And the sensitivity can be improved to a great extent with addition of images. In our case, an addition of 1000 images results in a 39.5% improvement in sensitivity and a 15.48% improvement in specificity. The probability graph provides a visual representation of this improvement.

The strength of this study is that Microsoft Azure is an easy-to-use tool, and with the availability of free CXR collection over the Internet, the study is highly repeatable. The availability of NIH X-ray can provide more training images. The results show a possible improvement with further addition and selection of training images. The limitation is that this study is a retrospective study and not randomized. The diagnosis of pneumothorax is being made on admission by an emergency physician and from the in-hospital admission notes, which most if not all is made by an internal medicine physician. A randomized double-blinded control should be performed with a radiological report as a reference standard. Another weakness is that the resolution X-ray is...
obtained at different time intervals due to the different recovery periods and outpatient follow-up times. The AUC for the ROC for our better model is 0.62. This model is not good enough to put into clinical use. More training and refinement are expected to be carried out.

**Improvement**

To enhance the performance of AI, we suggest the following measures.

It is shown that for the above two AI programs, both sensitivity and specificity can be improved by increasing the number of training samples.

Second, when scrutinizing the training images, it is found that although the training images are labeled pneumothorax solely, there is certain heterogeneity with diversified features not being labeled. For example, some XRIs contain foreign objects like endotracheal tube or ECG leads, some with distorted anatomy and hardly visually differentiable pneumothorax and some with poor orientations and XR taking techniques. Those features are not desirable and may cause confusion in training of the AI. For improvement, we can refine our training images to suit the common scenario encountered in the emergency setting to better match the expected patient sample, for example, relatively young age, without foreign objects, without obvious lung pathologies, and proper XR techniques.

Third, the above AI trainings are used as a classification algorithm. Each training image is labeled whether there is pneumothorax or not. We propose that with another algorithm, for example, object detection training, apart from providing the AI whether the XR has pneumothorax, we also provide the location information to the AI, and theoretically, it shall enhance the performance by putting more weighting on this feature.

Fourth, we can collaborate with the radiological department so that a local collection of XR is available and we can eliminate the potential demographic confounding factors as NIH is from the United States.

Finally, the above study is a retrospective case–control study. A future study with double-blinded randomized trial should be conducted to ensure more evidence is in place.

**Conclusion**

This study showed that a self-trained AI with online resource and publicly available XR has low sensitivity but a relatively good specificity. However, the performance can be improved by adding more training images. We have also suggested ways to improve the AI. At this stage, AI cannot replace a physician and radiologist. However, with future studies, the AI may act as an aide as a screening tool in daily practice as it comes with great promises to solve the problem of volume versus applicability of information in science.9

**Acknowledgements**

All authors are acknowledged for their contributions to the article.

**Author contributions**

All authors had full access to the data, contributed to the study, approved the final version for publication, and were responsible for its accuracy and integrity. Concept or design: K.H.A.L. and S.K.M. Data acquisition: K.H.A.L. Analysis or interpretation of data: K.H.A.L.

Drafting of the article: K.H.A.L. and S.K.M. Critical revision for important intellectual content: S.K.M.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**Availability of data and materials**

1. X-ray images of pneumothorax from National Institutes of Health (NIH) CXR dataset online.
2. CDARDS retrieval of hospital X-ray images.

**Informed consent**

Written informed consent was not necessary because no patient data have been included in the article.

**Ethical approval**

This study was approved by the Kowloon East research ethics committee (Ref.: KC/KE-19-0219/ER-2).

**Human rights**

Human right is not applicable for this study.

**ORCID iD**

Kwok Hung Alastair LAI https://orcid.org/0000-0002-1470-9665

**Supplemental material**

Supplemental material for this article is available online.

**References**

1. Buch VH, Ahmed I and Maruthappu M. Artificial intelligence in medicine: current trends and future possibilities. *Br J Gen Pract* 2018; 68(668): 143–144.
2. Hosny A, Parmar C, Quakenbush J, et al. Artificial intelligence in radiology. *Nat Rev Cancer* 2018; 18(8): 500–510.
3. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges,
pitfalls, and criteria for success. J Am Coll Radiol 2018; 15(3 Pt. B): 504–508.

4. Altman RB. Artificial intelligence (AI) systems for interpreting complex medical datasets. Clin Pharmacol Ther 2017; 101(5): 585–586.

5. Taylor AG, Mielke C and Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. PLoS Med 2018; 15(11): e1002697.

6. Park S, Lee SM, Kim N, et al. Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. Eur Radiol 2019; 29(10): 5341–5348.

7. Hwang EJ, Hong JH, Lee KH, et al. Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. Eur Radiol. Epub ahead of print 11 March 2020. DOI: 10.1007/s00330-020-06771-3.

8. Wang X, Peng Y, Lu L, et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE CVPR, 2017, https://arxiv.org/abs/1705.02315

9. El-Hassoun O, Maruscakova L, Valaskova Z, et al. Artificial intelligence in service of medicine. Bratisl Lek Listy 2019; 120(3): 218–222.