



# BMJ Open Do comprehensive deep learning algorithms suffer from hidden stratification? A retrospective study on pneumothorax detection in chest radiography

Jarrel Seah <sup>1,2</sup>, Cyril Tang,<sup>2</sup> Quinlan D Buchlak,<sup>2,3</sup> Michael Robert Milne <sup>2,2</sup>, Xavier Holt,<sup>2</sup> Hassan Ahmad,<sup>2</sup> John Lambert,<sup>2</sup> Nazanin Esmaili,<sup>3,4</sup> Luke Oakden-Rayner,<sup>5</sup> Peter Brotchie,<sup>2,6</sup> Catherine M Jones<sup>2,7</sup>

**To cite:** Seah J, Tang C, Buchlak QD, *et al.* Do comprehensive deep learning algorithms suffer from hidden stratification? A retrospective study on pneumothorax detection in chest radiography. *BMJ Open* 2021;**11**:e053024. doi:10.1136/bmjopen-2021-053024

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-053024>).

Received 01 May 2021

Accepted 11 November 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Dr Jarrel Seah;  
jarrel.seah@annalise.ai

## ABSTRACT

**Objectives** To evaluate the ability of a commercially available comprehensive chest radiography deep convolutional neural network (DCNN) to detect simple and tension pneumothorax, as stratified by the following subgroups: the presence of an intercostal drain; rib, clavicular, scapular or humeral fractures or rib resections; subcutaneous emphysema and erect versus non-erect positioning. The hypothesis was that performance would not differ significantly in each of these subgroups when compared with the overall test dataset.

**Design** A retrospective case–control study was undertaken.

**Setting** Community radiology clinics and hospitals in Australia and the USA.

**Participants** A test dataset of 2557 chest radiography studies was ground-truthed by three subspecialty thoracic radiologists for the presence of simple or tension pneumothorax as well as each subgroup other than positioning. Radiograph positioning was derived from radiographer annotations on the images.

**Outcome measures** DCNN performance for detecting simple and tension pneumothorax was evaluated over the entire test set, as well as within each subgroup, using the area under the receiver operating characteristic curve (AUC). A difference in AUC of more than 0.05 was considered clinically significant.

**Results** When compared with the overall test set, performance of the DCNN for detecting simple and tension pneumothorax was statistically non-inferior in all subgroups. The DCNN had an AUC of 0.981 (0.976–0.986) for detecting simple pneumothorax and 0.997 (0.995–0.999) for detecting tension pneumothorax.

**Conclusions** Hidden stratification has significant implications for potential failures of deep learning when applied in clinical practice. This study demonstrated that a comprehensively trained DCNN can be resilient to hidden stratification in several clinically meaningful subgroups in detecting pneumothorax.

## INTRODUCTION

Deep convolutional neural networks (DCNNs) are state-of-the-art for various

## Strengths and limitations of this study

- This study analysed a comprehensively trained deep learning algorithm that can detect up to 124 clinical findings on a chest radiograph.
- Strengths are that it not only evaluated the algorithm's ability to detect pneumothorax but also the clinically significant groups of pneumothorax—simple and tension pneumothorax.
- Additionally, the study evaluated the algorithm's ability to detect pneumothorax in specific, clinically salient subgroups.
- A large dataset of 2557 chest radiographs with a robust ground truth was evaluated.
- The study was limited by its retrospective nature, which necessitates further prospective, external validation studies.

image classification and processing tasks<sup>1 2</sup> In the medical imaging and artificial intelligence (AI) literature, these have frequently claimed to have near-human or even super-human performance in a variety of classification tasks performed by radiologists.<sup>3 4</sup> However, recent concerns have been raised about the translation of such results to clinical practice.<sup>5</sup> Most deep learning models in medical imaging are developed to detect specific findings or a group of similar findings, and as such performance is typically reported using a summary metric such as the area under the receiver operating characteristic curve (AUC). This can hide the performance of the models on clinically distinct and meaningful subgroups within these single findings, a phenomenon described as hidden stratification by Oakden-Rayner *et al.*<sup>6</sup> This work has shown that the algorithms trained to detect pneumothorax on chest X-ray are



often affected by hidden stratification, performing well in summary across an entire test dataset but performing worse in the subset of pneumothorax patients without the presence of an accompanying intercostal drain. As a pneumothorax is frequently treated with the insertion of an intercostal drain,<sup>7</sup> many chest radiography training datasets labelled for pneumothorax, including well-known public datasets such as the National Institutes of Health CXR14 dataset<sup>8</sup> demonstrate a strong correlation between pneumothorax and intercostal drains. Due to this correlation and the absence of explicit labels that distinguish intercostal drains as a separate finding, DCNNs trained on these datasets frequently erroneously rely on the presence of intercostal drains to identify pneumothoraces, a process which has been called ‘shortcut learning’ or ‘unintended cue learning’.<sup>9</sup> Reliance on these unintended cues can lead to reduced performance when the model is evaluated on the subset of cases without intercostal drains. This example is particularly dangerous as intercostal drain insertion usually indicates that the pneumothorax has already been identified and treated. As such, an algorithm demonstrating good performance for pneumothorax detection overall may be masking poor performance within the most clinically relevant subgroups, namely those yet to be treated and who would most benefit from prompt diagnosis.

Other clinically meaningful pneumothorax subgroups include patients with subcutaneous emphysema; patients with acute rib, clavicular, scapular or humeral fractures or rib resections and patients with semierect or supine (referred to as non-erect) positioning. While not an exhaustive list, these subgroups contain features that correlate with pneumothorax that a DCNN may erroneously rely on. For instance, subcutaneous emphysema may be benign but is often associated with pneumomediastinum and pneumothorax.<sup>10</sup> Patients with trauma with rib and other skeletal fractures often have associated pneumothorax or haemopneumothorax.<sup>11</sup> Additionally, trauma bay patients are often imaged in the supine or semierect position due to the acute nature of their injuries. The positioning of the patient also alters the visibility of pneumothoraces as well as their radiological appearances.<sup>12</sup> Pneumothorax is a common postoperative complication following thoracic surgery, of which rib resection is a common indicator.<sup>13</sup>

Labelling these subgroups and evaluating the performance of a DCNN on each one has been described as schema completion.<sup>6</sup> Evaluating the performance of DCNNs in such a way can help answer critical questions about the true clinical utility of such AI-driven computer-aided diagnosis tools.

We hypothesised that a DCNN trained comprehensively to detect multiple findings, including some of these subgroups, would demonstrate non-inferior performance as measured by AUC for both simple and tension pneumothorax in these subgroups when compared with the overall test dataset.

## METHODS

### DCNN software

A commercially available DCNN-based computer-aided diagnosis algorithm (Annalise CXR V.1.2, annalise.ai, Sydney, Australia) was evaluated. This algorithm has been trained to detect 124 clinical findings on chest radiography<sup>14</sup> and is publicly available at <https://cxrdemo.annalise.ai>. This algorithm indicates if each of the findings are present, as well as provides a numerical score indicating its confidence that the finding is present. This algorithm consists of several convolutional neural networks based on the EfficientNet architecture using the Keras library with Tensorflow V.2.1.

### Patient and public involvement

Patients and public were not involved in the design, conduct or reporting of this study.

### Study design and dataset

This project’s test dataset was obtained retrospectively from a more extensive study by (Seah *et al*) titled ‘Radiologist chest X-ray diagnostic accuracy performance when augmented by a comprehensive deep learning model: a multireader multicase study’, previously undertaken to validate the DCNN algorithm, which describes the case selection and participant flow. A reanalysis of the performance data from that study was conducted to test the hypothesis that the DCNN algorithm is resilient to hidden stratification. The chest radiographs for the test dataset were retrospectively obtained from two sources: a large dataset from a private radiology clinic in Australia and the publicly available Medical Information Mart for Intensive Care CXR (MIMIC-CXR) dataset.<sup>15</sup> These radiographs were not used in the DCNN training dataset, and there was no overlap between patients in the training and testing cohorts. Each study comprised multiple images from a single patient. The test dataset’s inclusion criteria were age >16 years; and studies that contained at least one frontal image. Additional frontal or lateral images acquired within each study were also used. Studies were in DICOM format and were deidentified. The original resolution and bit-depth was preserved. For the original study, cases were selected to comprise a wide variety of pathology that included each of the subgroups analysed in this investigation, as well as non-pneumothorax related pathologies such as lung nodules or airspace opacities. **Table 1** presents the breakdown of the number of cases with and without simple and tension pneumothorax in the overall dataset and in each subgroup.

### Ground truthing

Each study was evaluated by three subspecialist thoracic radiologists, including one of the authors (CMJ), who had each undertaken dedicated chest imaging fellowships. Each of these radiologists was trained on a specific set of definitions for simple and tension pneumothorax, as well as each of the subgroups, as defined in online supplemental appendix A. In particular, tension pneumothorax

**Table 1** Demographics of the test dataset

Patients	2286
Studies	2568
Images	4568
Sex	29% male 28% female 43% unknown*
Age	74 years (SD 15 years)*
View position	28% Posteroanterior (PA) 33% Anteroposterior (AP) 31% Lateral (LAT) 8% other

\*MIMIC-CXR does not provide sex or age information and hence data for this is incomplete.

was defined as pneumothorax with mediastinal shift towards the contralateral lung. For the ‘no fractures’ subgroup, specialist radiologists were instructed to label acute rib, humerus, clavicular, spinal and scapular fractures, as well as the presence of any rib resections. The presence of any of these fractures or rib resection disqualified the study from the ‘no fractures’ subgroup. Radiologists independently assessed each study, with access to the patient’s past and future imaging, clinical reports, as well as any CT chest reports if available, and identified if each finding was absent or present within that study. The consensus for each finding for each triple-read study was obtained using the Dawid-Skene consensus algorithm,<sup>16</sup> which considers the relative accuracies of each labeller for each finding. This was performed to mitigate variability and resolve discrepancies. Additionally, a radiology registrar (JS) reviewed the radiographer annotations on each image to identify if it was erect, semierect or supine. Studies with an accompanying lateral were considered erect. Where such annotations were not present, the positioning of the patient was estimated by considering indicators on the image such as the presence of air fluid levels or arm positioning. All annotations were performed on an in-house web-browser-based labelling tool capable of displaying DICOM images.

### Statistical analysis

For both simple and tension pneumothorax, the AUC was calculated, which is a commonly used metric of interest in the assessment of diagnostic classification tests.<sup>17</sup> To obtain the performance in a subgroup, the sample was filtered to retain only patients from that subgroup before recalculating the AUC. The difference in AUC between the full test set and each subgroup was bootstrapped to obtain a Bonferroni adjusted 95% CI. A difference of greater than 0.05 in AUC was considered clinically significant, therefore, if the lower bound of the CI of the delta exceeded  $-0.05$ , the performance in that subgroup was considered statistically non-inferior. This is a commonly

**Table 2** Number of studies with simple or tension pneumothorax within the testing set as well as throughout each subgroup

Subgroup	Simple pneumothorax	Tension pneumothorax	Total
Total	166	49	2557
Non-erect	64	19	595
No subcutaneous emphysema	96	44	2409
No intercostal drain	65	39	2277
No fractures	130	45	2017

chosen non-inferiority margin in diagnostic radiology AUC analysis.<sup>18</sup> A  $p < 0.00625$ , adjusted for eight hypotheses tested, was considered statistically significant. Analyses were conducted using Excel 2016 as well as custom Python scripts and the scipy,<sup>19</sup> scikitlearn,<sup>20</sup> nltk,<sup>21</sup> gensim<sup>22</sup> and keras<sup>23</sup> packages.

## RESULTS

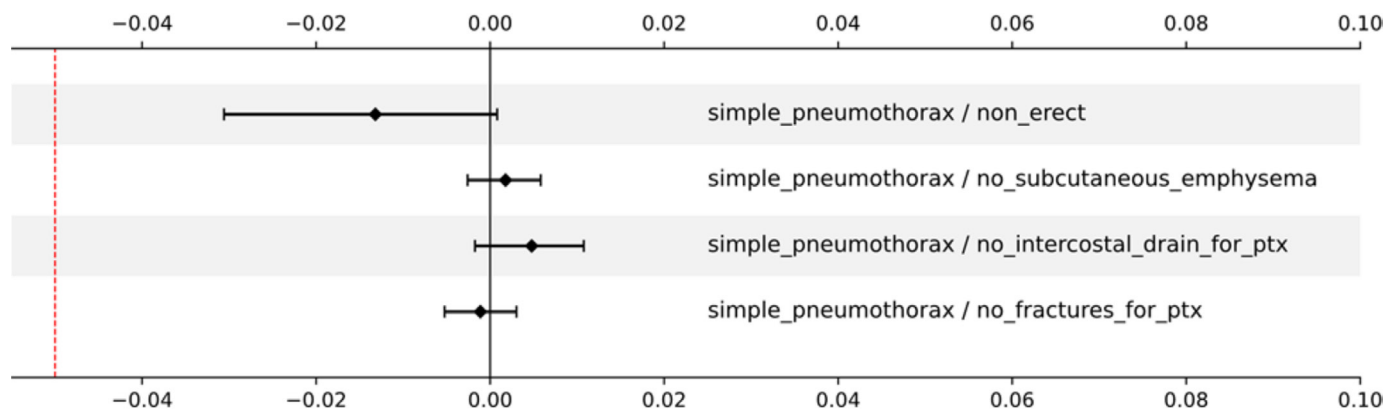
### Population characteristics

A total of 2568 studies from 2286 patients were included, comprising 4568 images. Forty-three per cent of cases from the test dataset were drawn from the MIMIC-CXR dataset and 57% were drawn from the private Australian radiology practice dataset. Table 1 presents the demographic and imaging characteristics of the test dataset. MIMIC-CXR does not provide age and sex data.

Eleven studies were deemed unsuitable by the DCNN and hence were excluded from analysis. None of these 11 studies were labelled by the ground-truthers as positive for simple or tension pneumothorax. There were 162 cases of simple pneumothorax and 49 cases of tension pneumothorax. Most cases of pneumothorax were found on erect chest radiographs. Sixty simple pneumothoraces and 11 tension pneumothoraces were seen on non-erect chest radiographs. Table 2 presents the number of cases with simple or tension pneumothorax in the entire test dataset, as well as within each subgroup. The complete co-occurrence matrix of each of the subgroups is presented in online supplemental appendix B.

### AUC performance

Figures 1 and 2 present the difference in AUC within each subgroup as compared with the overall test dataset. For simple pneumothorax, the lower bound of the adjusted 95% CI of the AUC delta exceeds  $-0.05$  and the upper bound exceeds 0 in all subgroups, indicating that performance in those subgroups was statistically non-inferior to the overall test dataset. For tension pneumothorax, the ‘no fractures’ and ‘non-erect’ subgroups were statistically non-inferior to the overall test dataset, however the ‘no subcutaneous emphysema’ and ‘no intercostal drain’ subgroups both demonstrated lower bounds of the



**Figure 1** Difference in AUC for detecting simple pneumothorax in the test dataset versus each specific subgroup with adjusted 95% CI. AUC, area under the receiver operating characteristic curve.

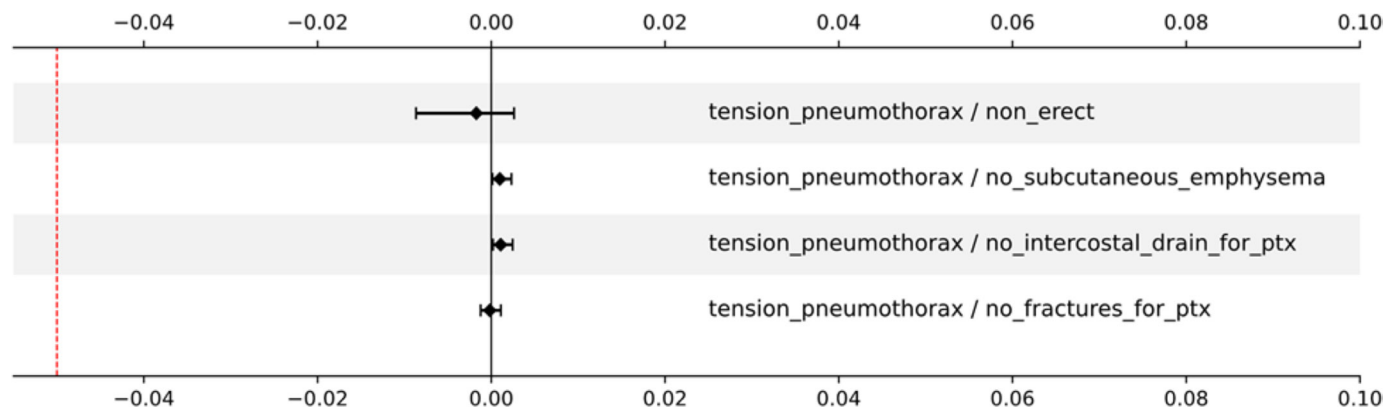
adjusted 95% CIs of the AUC delta exceeding 0, meaning these two subgroups demonstrated significantly superior performance. Table 3 presents the raw AUC values for the DCNN's performance and table 4 presents AUC deltas.

## DISCUSSION

We evaluated the clinical performance of a commercially available DCNN algorithm for the detection of simple and tension pneumothorax in a retrospective study with a large test dataset, representative of real-world clinical conditions. Of note, 33% of images in the test dataset were AP images, reflecting the inpatient and emergent nature of the studies. The test dataset was obtained from a wide range of Australian radiology sites as well as the MIMIC-CXR dataset, improving the generalisability of the results. When compared with existing algorithms<sup>24 25</sup> the DCNN algorithm demonstrated comparable or higher AUC for the detection of pneumothorax. The DCNN algorithm performed better in detecting tension pneumothorax when compared with simple pneumothorax, presumably due to the larger size and conspicuity of this type of pneumothorax, along with other associated features like mediastinal shift. The DCNN algorithm appeared to be resilient to hidden stratification for the four tested subgroups, with statistically non-inferior performance in six of the

eight subgroups tested, with the remaining two demonstrating borderline statistically superior performance. The superior performance of the tension pneumothorax subgroups with no intercostal drain or no subcutaneous emphysema was likely due to limitations with the testing dataset itself, which contained 49 tension pneumothoraces in total. Because of the relatively low number of cases, a small number of strongly confidently positive or negative cases unique to any one of these subgroups would likely influence the distribution of AUC deltas and may explain the counterintuitive result observed. Out of all subgroups, the DCNN algorithm demonstrated the greatest decrease in AUC when tested on the non-erect subgroup, although the result remained statistically non-inferior.

This DCNN algorithm appears to be resilient to hidden stratification as it has been trained on a comprehensively labelled dataset with 124 findings, including subgroups. The fact that the non-erect subgroup demonstrates the greatest decrease in AUC is circumstantial evidence that comprehensive labelling is beneficial as 'non-erect' was the only subgroup examined in this study that was not part of the 124 findings explicitly labelled during model training. Another likely contributing factor is that non-erect pneumothoraces are less conspicuous and indeed may not be visible on supine chest radiographs at all.



**Figure 2** Difference in AUC for detecting tension pneumothorax in the test dataset versus each specific subgroup with adjusted 95% CI. AUC, area under the receiver operating characteristic curve.

**Table 3** AUC values with 95% CI (non-adjusted) for the DCNN's performance on simple and tension pneumothorax in the test dataset as well as in specific subgroups

	AUC
Simple pneumothorax	0.981 (0.976–0.986)
No subcutaneous emphysema	0.983 (0.977–0.989)
No fractures	0.979 (0.972–0.986)
No intercostal drain	0.986 (0.979–0.992)
Non erect	0.968 (0.954–0.980)
Tension pneumothorax	0.997 (0.995–0.999)
No subcutaneous emphysema	0.998 (0.997–0.999)
No fractures	0.997 (0.995–0.999)
No intercostal drain	0.998 (0.997–0.999)
Non erect	0.995 (0.990–0.999)

AUC, area under the receiver operating characteristic curve; DCNN, deep convolutional neural network.

As computer-aided diagnosis and clinical decision support software becomes more prevalent in clinical practice,<sup>26</sup> it is likely that clinically meaningful failures will stem from hidden stratification, or more specifically, the lack of evaluation of clinically relevant subclasses.<sup>6</sup> Therefore, explicit evaluation of such clinically relevant subclasses is critical to responsible clinical decision support research, and that it is the domain of clinicians to define these subclasses and resist the temptation of oversimplifying performance into a single metric such as AUC for broad disease categories. While recent literature has highlighted this problem, it has been identified and cautioned against for years,<sup>27</sup> and unfortunately ignored in most clinical deep learning research. This is one of the

**Table 4** Difference in AUC values between each specific subgroup and the overall test dataset with 95% adjusted CI

	Delta AUC (subpopulation – full population)		
	Mean	Lower	Upper
Simple pneumothorax			
No subcutaneous emphysema	0.002	–0.003	0.006
No fractures	–0.001	–0.005	0.003
No intercostal drain	0.005	–0.002	0.011
Non erect	–0.013	–0.030	0.001
Tension pneumothorax			
No subcutaneous emphysema	0.001	0.000	0.002
No fractures	–0.000	–0.001	0.001
No intercostal drain	0.001	0.000	0.002
Non erect	–0.002	–0.009	0.003

AUC, area under the receiver operating characteristic curve.

risks to be mitigated when implementing computer-aided diagnosis tools at the bedside.<sup>28</sup>

### Limitations and future research

One notable limitation of this study is that the test dataset was drawn from the same population as the training dataset, and further research and external validation is required to verify these results. Another limitation is that radiologist readers may have missed subtle pneumothoraces, especially on supine patients,<sup>12</sup> although this was mitigated by the availability of future chest radiographs and reports as well as contemporaneous CT reports. Results may underestimate the true decrease in performance in the non-erect subgroup.

Opportunities for future research include providing paired chest CT images and radiographs to radiologists engaged in the ground-truth process to ensure that the ground-truth truly reflects the underlying pathology, as well as testing for resilience to other subgroups that were not available in this study. Future work is required to define the performance of the model in subgroups of tension pneumothorax over a larger number of cases than available in the test dataset to clarify whether superior performance is truly indicative of model behaviour, or simply an artefact of the test dataset. This would require obtaining data from sources and populations external to the training dataset and ensuring sufficiently large numbers of tension pneumothorax cases to verify results.

Additionally, this work was conducted as a retrospective analysis, which limits the generalisability of results to datasets that the DCNN algorithm has not seen before. Furthermore, as this is a reanalysis of previously acquired data, it may be underpowered to detect subtle differences in these subgroups. Therefore, additional prospective studies in different geographies, with a priori power analyses to determine adequate sample sizes, are required to see if similar performance is obtained in other populations. The resilience of the comprehensively trained DCNN algorithm to hidden stratification may also be due to additional factors, such as the already high baseline performance in identifying simple and tension pneumothorax. Further work is needed to explore the benefits of comprehensive labelling and training in findings that the DCNN algorithm does not perform as well on. One hypothesis is that as such findings may be more difficult to identify, the DCNN algorithm may rely on associated features or 'unintended cues' more, leading to worse hidden stratification.

### CONCLUSION

We have demonstrated that in a retrospective analysis a comprehensively trained DCNN algorithm can be resilient to hidden stratification when detecting simple and tension pneumothorax in clinically relevant subgroups. Further external validation and prospective study is needed to see if the benefits of a comprehensively trained model are generalisable in other settings.

### Author affiliations

- <sup>1</sup>Radiology, Alfred Health, Melbourne, Victoria, Australia  
<sup>2</sup>annalise.ai, Sydney, New South Wales, Australia  
<sup>3</sup>University of Notre Dame Australia, Sydney, New South Wales, Australia  
<sup>4</sup>University of Technology Sydney, Sydney, New South Wales, Australia  
<sup>5</sup>Australian Institute for Machine Learning, The University of Adelaide, Adelaide, South Australia, Australia  
<sup>6</sup>Radiology, St Vincent's Hospital Melbourne Pty Ltd, Fitzroy, Victoria, Australia  
<sup>7</sup>I-MED Radiology, Brisbane, Queensland, Australia

**Contributors** JS, CT, QDB, MRM, XH, HA, JL, NE, LO-R, PB and CMJ made substantial contributions to the planning, conception and design of the work, interpretation of the data, revision of the manuscript, approval of the final version, and agree to be accountable for all aspects of the work. Additionally, authors JS, CT and XH contributed to acquisition and analysis of the data. JS is the guarantor of the study.

**Funding** This work was supported by Annalise.ai. Award/grant number N/A.

**Competing interests** All authors have reviewed and approved this manuscript. Authors JS, CT, QDB, MRM, XH, HA, JL, PB and CMJ are employees of, or are seconded to, Annalise.ai. NE and LO-R have no interests to declare.

**Patient consent for publication** Not applicable.

**Ethics approval** This project was approved by the University of Notre Dame Australia's human research ethics committee (2020-127S). All data were anonymised prior to use in this study.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available on reasonable request. The research team may make the model and radiologist performance datasets and the test dataset available to interested research partners with the goals of supporting the research community and making further collaborative contributions to the literature. Requests for access can be made through the Annalise.ai website (<https://annalise.ai/contact/>). The model is publicly available as a commercial software product (<https://annalise.ai/products/annalise-cxr/>). The free web-based demonstration can be accessed online (<https://cxrdemo.annalise.ai/>).

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

### ORCID iDs

Jarrel Seah <http://orcid.org/0000-0002-2305-7873>  
 Michael Robert Milne <http://orcid.org/0000-0003-2082-5723>

### REFERENCES

- Khan A, Sohail A, Zahoor U, *et al.* A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 2020;53:5455–516.
- Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput* 2017;29:2352–449.
- Rajpurkar P, Irvin J, Ball RL, *et al.* Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018;15:e1002686.
- Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. *Evol Intell* 2021;1:3.
- Aggarwal R, Sounderajah V, Martin G, *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit Med* 2021;4:1–23.
- Oakden-Rayner L, Dunnmon J, Carneiro G. *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging.* In: *ACM CHIL 2020 - Proceedings of the 2020 ACM conference on health, inference, and learning.* Association for Computing Machinery, Inc, 2020: 151–9.
- Kwiat M, Tarbox A, Seamon MJ, *et al.* Thoracostomy tubes: a comprehensive review of complications and related topics. *Int J Crit Illn Inj Sci* 2014;4:142.
- Wang X, Peng Y, Lu L. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on Weakly-Supervised classification and localization of common thorax diseases. Available: <https://uts.nlm.nih.gov/metathesaurus.html> [Accessed 18 Apr 2021].
- Geirhos R, Jacobsen J-H, Michaelis C, *et al.* Shortcut learning in deep neural networks. *Nat Mach Intell* 2020;2:665–73.
- Balaji SM. Subcutaneous emphysema. *J Maxillofac Oral Surg* 2015;14:515–7.
- Sirmali M, Türüt H, Topçu S, *et al.* A comprehensive analysis of traumatic rib fractures: morbidity, mortality and management. *Eur J Cardiothorac Surg* 2003;24:133–8.
- Omar HR, Abdelmalak H, Mangar D, *et al.* Occult pneumothorax, revisited. *J Trauma Manag Outcomes* 2010;4:12.
- Łochowski MP, Kozak J. Video-assisted thoracic surgery complications. *Wideochir Inne Tech Maloinwazyjne* 2014;9:495–500.
- Annalise.ai - Annalise CXR comprehensive medical imaging AI. Available: <https://annalise.ai/products/annalise-cxr/> [Accessed 18 Apr 2021].
- Johnson AEW, Pollard TJ, Berkowitz SJ, *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019;6:1–8.
- Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the em algorithm. *Appl Stat* 1979;28:20.
- Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5:1315–6.
- Obuchowski NA, Bullen JA. Statistical considerations for testing an AI algorithm used for prescreening lung CT images. *Contemp Clin Trials Commun* 2019;16:100434.
- Virtanen P, Gommers R, Oliphant TE, *et al.* {SciPy} 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 2020;17:261–72.
- Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
- Bird S, Klein E, Loper E. *Natural language processing with python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc, 2009.
- Rehurek R. Scalability of semantic analysis in natural language processing 2011.
- Chollet F. Keras, 2015. Available: <https://github.com/fchollet/keras>
- Gooßen A, Deshpande H, Harder T. Deep learning for pneumothorax detection and localization in chest radiographs. Available: <http://arxiv.org/abs/1907.07324> [Accessed 18 Apr 2021].
- Taylor AG, Mielke C, 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:e1002697.
- Buchlak QD, Esmaili N, Leveque J-C, *et al.* Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg Rev* 2020;43:1235–53.
- Webb GI, Ting KM. On the application of ROC analysis to predict classification performance under varying class distributions. *Mach Learn* 2005;58:25–32.
- Buchlak QD, Esmaili N, Leveque J-C, *et al.* Ethical thinking machines in surgery and the requirement for clinical leadership. *Am J Surg* 2020;220:1372–4.