International Journal of Convergence in Healthcare

IJCIH
Converging Healthcare & Technology

Published by
IJCIH & Pratyaksh Medicare LLP

www.ijcih.com

Harmonizing Reporting and Identification of Lesions in Chest X-rays - Issues and their Implications for Development of an AI Tool

Kshitij Agarwal[1], Reetika Malik Yadav[2], Manoranjan Pattanaik[3], Vikram Vohra[4], Atul Tayade[5], Manjula Singh[6]

[1]Respiratory Medicine, University College of Medical Sciences (University of Delhi) & Guru Teg Bahadur Hospital, Dilshad Garden, Delhi, [2]ICMR-National Institute of Immuno-Haematology, Mumbai, [3]Shreeram Chand Banga Medical College & Hospital, Cuttack, [4]National Institute of Tuberculosis and Respiratory Diseases, Delhi, [5]Mahatma Gandhi Institute of Medical Sciences, Wardha, [6]Indian Council of Medical Research (ICMR), Delhi

Abstract

This commentary addresses the critical challenges in harmonizing the reporting and identification of lesions in chest X-rays (CXRs), particularly in the context of developing artificial intelligence (AI) tools for radiological interpretation. Despite the significant advancements in medical imaging, the error rate in clinical practice remains alarmingly high, with millions of misinterpretations occurring annually. AI models have shown promise in interpreting CXRs, yet they struggle with assessing cardiomegaly, hilar abnormalities, and diaphragm positioning in view of the subjective decision various radiologists/physicians which affects the training of the AI tool. This paper highlights four primary issues: the complexities of accurately diagnosing cardiomegaly due to projection variations and anatomical confounders; the challenges in recognizing hilar abnormalities due to the intricate anatomy and variability among individuals; and the difficulties in assessing diaphragm position and shape, which can be influenced by various physiological factors or anatomical variations. The authors advocate for the establishment of standardized uniform objective criteria for abnormality identification in CXRs, which would enhance the accuracy of AI models and improve clinical diagnosis. By fostering collaboration among radiologists/physicians and AI developers, the goal is to create a uniform criteria for abnormality detection which would help in development of more reliable diagnostic tool would minimize errors and maximize patient safety.

Keywords: *Artificial intelligence (AI), chest x-rays (CxRs), radiological interpretation, diagnostic accuracy, standardized criteria.*

Commentary

In the world of modern medicine, it is hard to come by a scenario where radiological examination does not form one of the pillars of clinical management. With no less than one billion performed annually world-over, the role of roentgenograms, colloquially called X-rays, can hardly be highlighted enough. Typically, the rate of errors in clinical practice is around 4%, which translates to a mind-blowing 40 million errors in interpretation per year leaving radiologists vulnerable to medical negligence lawsuits[1,2]. In 1949, Garland reported 33.3% errors in interpretation of positive films through group consensus and 8% intra-reader variation, which in spite of the strides that medical technology has taken over the last 76 years, have remained surprisingly unchanged, underscoring the limitations of the human eye and brain interpreting them[3,4].

Recently, artificial intelligence (AI) based deep neural training models, albeit in a primeval form, have shown

Corresponding Author:
Dr. Manjula Singh
Scientist F, Division of Delivery Research, Indian Council of Medical Research, V Ramalingaswami Bhawan, Ansari Nagar, New Delhi
Phone: +919868245793; Fax No:011-26588896
e-mail: drmanjulasb@gmail.com

some promise in reliably interpreting chest X-rays (CXRs). Current AI models report sensitivity ranging from 67% to 98%; and specificity ranging from 84% to 95% (3-6) for triaging chest x-ray abnormalities in optimized settings where homogenous, high quality CXRs curated/annotated by specialists are provided to the model[5–8] .

However, these models too seem to fare subpar when challenged with tasks requiring subjective decision making such as determining heart size, hilar, and diaphragmatic abnormalities, which, it may be noted, is additional to the traditional 'problem areas' of the CXR such as the apices, retrocardiac and subdiaphragmatic areas. In present manuscript, an attempt is made to highlight such challenges faced by the present AI-based learning models during their practical application in the real-world scenario and to suggest remedial measures for the same through harmonizing the process of abnormality identification. We focused on primarily four issues related to cardiomegaly, hilum, hyperinflation and diaphragm.

Cardiomegaly is classically identified by a cardiothoracic (CT) ratio (ratio of the maximal outward cardiac dimensions on either side from the midline to the maximal thoracic diameter measured from the inner border of the ribs at the superior diaphragmatic margin) of more than 1:2 or 0.5. However, certain caveats apply to this rule. Firstly, this ratio is defined for the postero-anterior (PA) projection of the CXR; the antero-posterior (AP) projection inherently magnifies the cardiac dimensions by 20% and thus using the same mathematical logic for the CT ratio tends to over-diagnose cardiomegaly erroneously. Furthermore, images not taken in full inspiration, particularly those taken at the end of expiration, lead to more prominent-appearing pulmonary vessels, and heart which can again be misinterpreted as cardiomegaly or heart failure with this fixed ratio[9]. Similarly, mispositioning the patient in the PA view gives rise to issues like rotation which is practically a CXR taken in an oblique projection wherein the heart apex appears to rotate away or towards the screen on the median axis such that the heart appears larger if the right shoulder rotates forward and smaller if reverse, and thus the CT ratio may again be erroneously under- or overestimated respectively[10]. Moreover, presence of anatomically contiguous paracardiac lesions in the mediastinum or lung could make heart borders indistinguishable resulting in an erred CT ratio, that is if one can be calculated at all[11] All these features lead to subjectivity and often resulting in contradictory reporting for cardiomegaly by different experts. While it is imperative that the model be trained through a fixed numerical value to diagnose cardiomegaly to have a uniformity, dealing with the subjectivity arising due to the issues mentioned above that is required without affecting impact on clinical judgement. However, the model's word should be taken with a pinch of salt in light of the above, if the subjectivity is not tackled. One of the ways is to have the reporting based on the CT ratio, the criteria once fixed should be adhered to for reporting of cardiomegaly, and then the other features can be looked and commented upon accordingly. Therefore, the model first needs to be trained to calculate CT ratio and then trained to adjust for these aberrations which may not only induce apparent cardiomegaly but also falsely miss the latter. Therefore, once flagged as cardiomegaly, the model should sequentially look for the projection, centering or rotation and extent of inspiration.

While it may be difficult to predict with certainty, the AP and PA differ in that the AP view tends to project the clavicles and ribs more horizontally consequently the anterior ends of the latter appear more radio dense than the posterior ends, the scapulae tend to overshadow the lung fields bilaterally and heart appears magnified. Indeed, AI-models have been successful in distinguishing PA from AP with 95% accuracy[12]. The concept of centering is built on the right and left halves of the body being symmetrical around the median vertical axis which on a CXR is represented by the spinous processes of the vertebral bodies, therefore, a well-centered CXR can be inferred objectively if the medial ends of both the clavicles placed equidistant from the spinous processes[9], provided the patient is standing straight, which in-turn can be assessed by having the acromion processes of both sides lying in the same horizontal line. Inspiration on the other hand, is widely considered satisfactory if the dome of the right diaphragm lies level with the posterior aspect of the 10th rib in the mid-clavicular line[9]. Finally, recognition of ancillary lesions confounding the heart borders forms the heart and soul of the AI-model built to read CXR, and it is needless to say the discriminability for these lesions is essential for the model. It would be appropriate to consider each of these cases and decide on its label while comparing with the actual disease conditions which may help to harmonize the decisions of physicians/radiologists in such cases thus making it easy for the AI developers.

Likewise, abnormalities at the hilum pose another challenge to the AI-model. Although the radiological hilum is the rather simple junction of the upper lobe vein and lower lobe artery, is a fact well-known to the human interpreter of the CXR, it is the latter who also understands that the anatomical hilum is more complex complex - comprising of the bronchi, lymphatics and lymph nodes, and bronchial arteries, overlapping each other and communicating with mediastinal structures such that visualization of these structures is severely limited by the 2 dimensional character of CXR[9]. Also, it would not be an overstatement that the hila of no two individuals appear the same. The present AI models on the other hand, are not yet adept at discriminating differing radiodensities created by the differing spatial orientation of different lesions, let alone the aetiology. It is not uncommon for the AI-models to stay mute on hilar abnormalities or, more commonly, to mark the normal hilar vessels as abnormal lesions given the slightest anatomical variations, tortuosity or other conformational changes in shape, caliber and radiodensities of the structures. While radiological signs such as the dense hilum, hilum convergence and hilum overlay are well-known aide-mémoire to the clinician/radiologist in determining the site and thus possible aetiology of the lesion[13]. As it stands today, the AI model lacks the faculty to take note of the same and given the wide gamut of aetiological possibilities, many of which may be sinister in their outcome, if missed. A case of lung cancer, for example, may have a hilar, subcarinal, or paratracheal lymph node as the sole discernible radiological anomaly on the CXR each of which would sequentially upstage the stage of the cancer thereby drastically changing the treatment protocol and finally, outcome[14]. On one hand, while it is imperative that the AI-model be trained to recognize the spatial conformation and differences in radiodensities of the various normal hilar structures, on the other, it is advisable that the model be trained to demarcate any abnormalities falling outside of this ‘normal’ pattern and preferably err on the side of a false positive diagnosis instead of a missed identification, since the stakes associated with a missed diagnosis of a true hilar abnormality would be too high - in terms of morbidity, mortality as well as financial costs. Thus, it would be appropriate to decide the features that would be considered normal including the normal variations, in terms of its size, shape, densities, location and also decide criteria for abnormal lesion or feature that would represent pathological finding or suggestive of a diseased condition to be marked as abnormal. Such approach will need a large number of normal and abnormal X-rays (confirmed disease as per gold standard criteria) across various age groups and gender before arriving at a consensus. This might look difficult but is possible and this will not only bring harmonization in reporting of X-rays but also help in development of AI tool as a diagnostic test. A practical approach in this regard should be to train the AI model to recognize the following attributes of the hilum - shape, density, size, and contour of the pulmonary arteries - both, individually as well as relative to the contralateral side[13].

The diaphragm is another structure that poses yet another challenge to interpretation for the AI-model. The placement of the diaphragm in the body as the partition between the thorax and abdomen is sui generis, which causes the former being amenable to reflect anatomical variations and pathological processes affecting structures on either side of it, in addition to its own intrinsic processes. In general, the diaphragm appears as a smooth, regular, thin-lined dome at the interface of the lower border of the lungs with the abdominal viscera lying, at full inspiration, below the posterior aspect of the 10th rib (corresponding to the anterior end of the 6th rib in the mid-clavicular line) on the right and an intercostal space lower on the left [15]. While ‘spot diagnoses’ in the vicinity of the diaphragm such as most causes of elevated hemidiaphragm, pleural collections, lobar consolidations, and masses are rarely missed by the AI model, subtle processes such as apparent depression of the diaphragm test the performance of the model in real time. Classically, the diaphragm is considered ‘low-lying’ if the dome is found lying lower than the anterior end of the 7th rib in the mid-clavicular line[16], a level below the 6th rib has also been proposed as part of the diagnostic criteria to diagnose hyperinflation[15]. Therefore, this debate needs to be settled before drafting an objective training algorithm for AI-models. On the other hand, the diaphragm is labelled ‘flat’ if the maximum height of the dome from an imaginary line linking the ipsilateral costophrenic and cardiophrenic angles is shorter than 1.5 centimetres[16]. It is not uncommon to encounter situations where the diaphragm may be low-lying without objective flattening as in the case of a trained athlete or a tall person or conversely, flatter as in the case of a more stocky or obese build. Practically, to make the radiological diagnosis of hyperinflation, the presence of both attributes of the diaphragm is required; other signs such as lung length of >30 cm, a small-sized, pulled-up or tubular heart, everted diaphragm and obtuse costophrenic angles, saber sheath trachea aid in identifying hyperinflation but are present inconsistently[15, 16]. While dome height correlates well with severity of hyperinflation, the plain CXR has poor sensitivity for mild-to-moderate emphysema and could miss nearly 60% cases[18,19]. Therefore, the algorithm to diagnose the same should first ascertain the position of the respective domes of the diaphragm relative to the ribs and then identify the degree of flattening as described above and finally, identify confounding supra- and/or infra-diaphragmatic factors that might be confounding the appearance or height of the domes which could be indicated by identifying irregularities, and humps and bumps in the regular outline of the diaphragm. Thus, it would be appropriate to first decide on the lung length taking into consideration the level of ribs, that would be considered for hyperinflation and then various other conditions including the mid-expiration films etc., it can be corelated with other features like shape of diaphragm before deciding on the criteria for hyperinflation or abnormal diaphragm. Once the criteria is decided, it should be adhered to objectively and then used to train the AI models.

It is worth noting that all the above mentioned features may be the reasons causing the discrepancies in X-ray reporting across various radiologists/physicians and it would be appropriate to discuss and decide on the criteria which then can be validated against the confirmed disease (confirmed by reference standard tests) as gold standard which will not only help harmonizing the lesion identification but also increase the accuracy of X-ray reporting as diagnostic tool which in turn will not only simplify development of AI tool but also make AI more explainable by reducing the error.

Presently available AI based models designed to interpret CXRs therefore still have a long way to go before they can be applied to public use with confidence for which they first need to be trained enough on an algorithm drafted on a holistic diagnostic approach. This approach requires uniformly accepted objective criteria for identification of abnormality for annotation, along with the possible underlying aetiology and respective diagnoses associated with various thoracic structures. The potential repercussions of a missed identification, and diagnoses thereof can be tackled using an algorithm wherein other factors like symptoms etc. can be considered.

As we move forward in an era wherein AI is gradually expanding its role in all areas, it is essential that we all come together and develop objective criteria to harmonize the anomaly detection using human intelligence for betterment of public health before it is taken over by Artificial Intelligence.

References

1. Waite, S. *et al.* Interpretive Error in Radiology. *American Journal of Roentgenology* **208**, 739–749 (2017).
2. Bruno, M. A., Walker, E. A. & Abujudeh, H. H. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *RadioGraphics* **35**, 1668–1676 (2015).
3. Kim, Y. W. & Mansfield, L. T. Fool Me Twice: Delayed Diagnoses in Radiology With Emphasis on Perpetuated Errors. *American Journal of Roentgenology* **202**, 465–470 (2014).
4. Garland, L. H. On the Scientific Evaluation of Diagnostic Procedures. *Radiology* **52**, 309–328 (1949).
5. Annarumma, M. *et al.* Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* **291**, 196–202 (2019).
6. Albahli, S. & Ahmad Hassan Yar, G. N. AI-driven deep convolutional neural networks for chest X-ray pathology identification. *J Xray Sci Technol* **30**, 365–376 (2022).
7. Hwang, E. J. *et al.* Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open* **2**, e191095 (2019).
8. Anderson, P. G. *et al.* Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. *Sci Rep* **14**, 25151 (2024).
9. Kelly, B. *The Chest Radiograph*. www.ums.ac.uk.
10. Wilson, A. G. The Chest Radiograph in Heart Disease. *Medicine* **30**, 18–26 (2002).
11. Felson, B. & Felson, H. Localization of Intrathoracic Lesions by Means of the Postero-Anterior Roentgenogram. *Radiology* **55**, 363–374 (1950).
12. Hosch, R., Kroll, L., Nensa, F. & Koitka, S. Differentiation Between Anteroposterior and Posteroanterior Chest X-Ray View Position With Convolutional Neural Networks. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren* **193**, 168–176 (2021).
13. Jash, D., Maji, A., Patra, A. & Sarkar, S. Approach to unequal hilum on chest X-ray. *The Journal of Association of Chest Physicians* **1**, 32 (2013).
14. Shah, P. K. *et al.* Missed Non–Small Cell Lung Cancer: Radiographic Findings of Potentially Resectable Lesions Evident Only in Retrospect. *Radiology* **226**, 235–241 (2003).
15. Proschek, P. & Vogl, T. J. Chest and Mediastinum. in *Diagnostic and Interventional Radiology* 479–587 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2016). doi:10.1007/978-3-662-44037-7_19.
16. Shaker, S. B., Dirksen, A., Bach, K. S. & Mortensen, J. Imaging in Chronic Obstructive Pulmonary Disease. *COPD: Journal of Chronic Obstructive Pulmonary Disease* **4**, 143–161 (2007).
17. Proschek, P. & Vogl, T. J. Chest and Mediastinum. in *Diagnostic and Interventional Radiology* 479–587 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2016). doi:10.1007/978-3-662-44037-7_19.
18. Shiraishi, M. *et al.* Diaphragm dome height on chest radiography as a predictor of dynamic lung hyperinflation in COPD. *ERJ Open Res* **9**, 00079–02023 (2023).
19. Thurlbeck, W. & Simon, G. Radiographic appearance of the chest in emphysema. *American Journal of Roentgenology* **130**, 429–440 (1978).