The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms - How We Do It

Vidur Mahajan, Vasanthakumar Venugopal, Saumya Gaur,
Salil Gupta, Murali Murugavel, and Harsh Mahajan

Centre for Advanced Research in Imaging, Neurosciences & Genomics (CARING),
Mahajan Imaging, New Delhi, INDIA

Abstract

There is a plethora of Artificial Intelligence (AI) tools that are being developed around the world aiming at either speeding up or improving the accuracy of radiologists. It is essential for radiologists to work with the developers of such algorithms to determine true clinical utility and risks associated with these algorithms. We present a framework, called an Algorithmic Audit, for working with the developers of such algorithms to test and improve the performance of the algorithms. The framework includes concepts of true independent validation on data that the algorithm has not seen before, curating datasets for such testing, deep examination of false positives and false negatives (to examine implications of such errors) and real-world deployment and testing of algorithms.

Introduction

Artificial Intelligence (AI), today, refers to the ability of machines to perform ‘human-like’ tasks. In the context of radiology, and for purposes of this paper, we restrict AI-based tools to those that replicate the visual and cognitive part of the radiology workflow, i.e. detection of findings and interpretation of those findings. There are several research groups, comprising of academic centres, large healthcare companies and startups, focused on developing Artificial Intelligence (AI) based tools to automate these tasks. We classify all such groups into the category of “Vendors”. These tools are either in the domain of image acquisition, post-processing, triaging or radiology report creation.

We present, based on our experience of working with more than 20 such vendors, an ‘Algorithmic Audit’ methodology which can be used by radiologists to evaluate the performance of an AI algorithm, and share relevant feedback with vendors in order to help them improve the performance of the algorithms. The proposed method includes preparing datasets especially focused on validation of algorithms, examining failed cases post testing of AI on these validation images, and a series of other steps which are described further in more detail.

Prerequisites to the Algorithmic Audit

The Algorithmic Audit thrives on two essential components – data availability and clinical expertise. The availability of heterogeneous large datasets is essential for conducting of an Algorithmic Audit since testing on rare, but clinically relevant cases is paramount to determine how the algorithm performs in such situations. Clinical expertise is required to derive insights about the functioning of an algorithm – in many instances, given the black-box nature[1] of many AI algorithms, it is very hard to determine why an algorithm may have failed in a particular instance. In such cases, the ability of subspecialist radiologists to analyse failed cases and frame hypothesis around failure, is critical.

Truly Independent Validation

Lack of generalizability of AI algorithms is a major hindrance to the adoption of such tools in clinical practice. Kim et al found that as of 17th August 2018, only 6% of the 516 studies published on radiology AI did external or truly independent validation[2], an essential component of the Algorithmic Audit. To understand the nuance behind true independent validation, one must understand how a deep learning algorithm is developed and validated. We take the example of MRNet, an algorithm to automatically detect tears of the anterior cruciate ligament on knee MRIs, developed by Rajpurkar et al[3]. The authors first obtained a dataset of 1,370 knee MRIs along with their corresponding reports. These 1,370 MRIs were broken into 3 separate datasets – one training set of 1,130 MRIs, one tuning set of 120 MRIs and one validation set of another 120 MRIs. Note that all three sets are subsets of the larger dataset of 1,370 MRIs all of which have been obtained from the same hospital – Stanford University Hospital. The training set was used to create the deep learning algorithm, the tuning set used to fine-tune the parameters of the deep learning algorithm and finally the validation set was used to determine the performance of the algorithm. The algorithm was also tested on an external validation dataset from Stajduhar et al[4] of 917 MRIs, from a different country altogether. MRNet gave an impressive Area Under Curve (AUC) of 0.96 on the Stanford Hospital validation dataset, which dropped to 0.82 on the Stajduhar et al dataset – the truly independent dataset, thereby demonstrating the problem of generalizability of AI algorithms across datasets. The problem of generalisability is well documented in research but as such no concrete metric exists to measure both generalisability, or the lack of it.[5,6]

Another example of true independent validation comes from Chilamkurthy et al[7] where a deep learning algorithm that automatically detects critical findings in head CT scans was validated on a dataset from multiple hospitals from where no data was used to train the algorithm. This was one of the first peer-reviewed studies of an AI algorithm giving area-under curve (AUC) of more than .90 for all its findings on an external test set, comprising of data obtained from clinical sites outside of where the algorithm was trained on.

Such results make the case for true independent external validation of AI algorithms before putting them to use in clinical environments. Since the performance of AI algorithms is typically much better on data sourced from sites it was trained on, as a radiologist or a radiology manager wanting to work with vendors to develop or validate AI algorithms, it is important to think through whether one would like to contribute data for the development of algorithms, or would one like to use their data exclusively for validation. Developing and validating an algorithm on data from the same clinical site may not give the true picture of the performance of algorithms.

Making Data Usable

Depending on what position one takes, with respect to assisting with developing and validating AI algorithms, the next step is to ‘convert’ one’s data from crude unstructured data, to refined data with very strong ‘ground truth’. Consider the validation study of a high-sensitivity deep learning algorithm (by Singh et al[8]) that classifies chest X-Rays into normal and abnormal[9]. The algorithm was not trained on images from the test sites and hence all the test data was 'unseen' by the algorithm. To test the performance of the algorithm, a dataset of 430 chest X-Rays from different radiology clinics, comprising of different X-Ray machines (CR, DR and retrofit DR), was assembled. Subsequently, the 430 scans were read by multiple radiologists to develop consensus around whether the findings truly exist in the scans – the ground truth for validating the AI. Once ground truth was established, it was a matter of running the algorithm on the X-Rays and comparing the results of the AI to the said ground truth. The algorithm delivered a sensitivity of 97% in this true independent validation setting giving confidence to use the algorithm in the real world. The process of defining ground truth is a painstaking one but as far as conducting validation studies is concerned, it is the most important piece – our group is now working on datasets that have objective ground truth, for example, chest X-Rays with corresponding chest CT scans, or CT scans with corresponding biopsies.

Choosing the right data mix

Once a test dataset with corresponding ground truth has been assembled, it is important to determine the mix of cases required to aptly validate the AI algorithm. Generally, there are two types of algorithmic errors one is looking for – false positives and false negatives. False positives are cases which the AI calls out to be positive, but they in fact are negative (for the finding under question). To check for an algorithms false positive rate, it is important to have a dataset that comprises heavily of cases without many positives – this gives the validator a chance to see how frequently the AI calls a truly negative case as positive and is only possible when there are a high number of negatives. The same is true for false negatives, which are cases where in fact there is a finding but the algorithm misses it. For this, a dataset comprising heavily of positive cases is required to determine whether (and to what extent) AI misses positive cases. This is especially important in today's scenario when AI is being pegged as a tool for either triaging[10] or for automatically identifying normal cases[9,11] with a high degree of confidence.

Deep examination of false positives and negatives

Now that a dataset has been assembled, the algorithm under validation has been run on the test data, and results of the algorithm have been compared to ground truth, the most important part of the algorithmic audit begins – a deep examination of false positives and negatives. There is no easy way to do this – a radiologist and a data scientist need to sit together for hours at end and read every case where the AI failed. There are two independent approaches to this, and both are equally important. First, one must examine the implications of AI going wrong – for example, a chest X-ray algorithm that misses broncho-pulmonary markings is significantly less dangerous than one that misses massive pneumothoraces, even if the summary statistics in the former may be better than the latter.

Second, one must attempt to examine reasons for AI failure. For example, our group helped GE Healthcare build an algorithm to automatically detect pneumothorax[12] on chest X-Rays done on a portable X-Ray system[13]. This algorithm used rich annotation information, i.e. pixel level annotation of pneumothorax done by multiple radiologists, to drive very high algorithm performance[14]. During validation of this algorithm, we were provided several false positive and false negative scans, along with Class Activation Maps or heat maps of areas on the chest X-ray the algorithm gave 'importance' to while determining whether pneumothorax was present in the scans. We examined each false positive and false negative in an attempt to develop hypothesis around *why* the algorithm made the decision regarding presence of pneumothorax in the image. Eventually, patterns such as cases with mach bands and chest tubes falsely being labelled as pneumothorax, emerged. Once these patterns were confirmed on more test images, the algorithm was retrained to ignore these findings, and results of the algorithm improved to a great degree. Such insight can only be obtained by diving deep into failures and attempting to decipher why an algorithm fails. It is important to note that such insight can only be derived by deep collaboration between data scientists and radiologists and links back to the ready availability of clinical expertise as a prerequisite to successful collaboration.

Far North and Far South Cases

As a clinical group conducting an audit of AI algorithms, another concept we describe is that of *Far North* and *Far South* cases. Most often, the output of an AI algorithm is a probability or abnormality score using which the decision of whether an abnormality is present or not is taken. Such a score is a measure of the 'certainty' with which an AI algorithm is giving a particular output. Typically, developers of algorithms determine a 'threshold' beyond which a finding is taken to be present, and below which a finding is determined to be not present.

These concepts are best understood using an example - let us take the hypothetical case of a deep learning algorithm for detection of fractures in an X-ray. Assume that the algorithm gives a probability estimate of whether a fracture is present in an X-Ray on a scale of 0% to 100%, with a threshold of 50% - a probability of >50% implies the presence of a fracture, and less than 50% means no fracture. Let us further assume we have 6 test images (1-6) out of which images 1, 2 and 3 are normal (no fracture), and images 4,5 and 6 have fractures. The AI algorithm gives a probability estimate of 25% for image 1, 60% for image 2, 90% for image 3, 75% for image 4, 40% for image 5 and 10% for image 6, implying that the algorithm is most 'certain' about the presence of a fracture in image 3, then image 4 and then image 2. Note that image 1 is hence a true negative and image 4 is a true positive, and rest of the images are situations where the algorithm was wrong based on the threshold of 50% for positivity.

Now, while one would not be wrong in clubbing images 2 and 3 into false positives, and images 5 and 6 into false negatives, it would be wrong to equate the 'extent of falseness'. The algorithm was 90% certain that image 3 has a fracture and 60% certain that image 2 has a fracture, whereas both do not have fractures. It is important to understand that the error made by the algorithm in image 3 was much 'worse' than image 2 since it is an error that cannot be fixed by simply adjusting the threshold of the probability. Similarly, algorithm was very certain that image 6 does not have a fracture (10% estimate) and fairly certain that image 5 does not have a fracture (40%), whereas both in reality had fractures, but again the mistake made by the algorithm in image 6 is much worse than image 5. Such cases are dubbed as Far North (high probability false positives) and Far South (low probability false negatives) by our group.

It is essential to monitor and test algorithms from point of view of identifying far north and far south cases because they demonstrate fundamental issues in algorithms' ability to understand an image and make a prediction. A radiologist, in contrast, is able to say 'I am not sure' and hence seek appropriate guidance in cases where help is needed. In order for AI to be a reliable companion to radiologists, the probability estimates given by AI need to be less divergent in terms of correlating with the ground truth.

Real World Deployment and Testing

The litmus test for any medical imaging AI algorithm is whether radiologists use it or not, how frequently radiologists need to change the findings of the algorithm and how long they take to 'trust' it. Unfortunately, there is hardly any published literature around real-world deployments and testing of algorithms. From our experience of deploying AI algorithms that automatically classify chest X-Rays into normal or abnormal and that automatically detect and characterize lung nodules on chest CT scans, we understood that it is important to present the findings of the AI algorithm in the most radiologist-friendly way possible. This insight comes from two main reasons – one, most radiologists use workstations which are 'locked', i.e. additional software cannot be loaded on them, limiting the ways in which AI results can be displayed t them. Second, and possibly more importantly, it is very difficult to toggle between viewers for specific cases, and such toggling negatively impacts adoption of new software and solutions. To counter this, we suggest either HL7[15] integration with the radiology reporting software, or, in situations where the AI generates an image / segmentation, simply 'writing' no raw dicom images like a 'screensaved' image.

One successful implementation strategy for AI that automatically classifies chest X-rays into normal or abnormal, in the workflow, is the following – first, chest X-rays are automatically pulled from PACS using DCM4CHEE[16] based tools every 2 minutes to process them in real-time; next, the AI algorithm is run on the chest X-Ray and if the AI is unable to find an abnormality in the image, a 'normal' report template is automatically sent to the reporting software using an HL7 message such that the radiologist simply has to approve the report after looking at the chest X-Ray. The radiologist has full freedom to change the report and such changes are measured in real time to determine accuracy of the algorithm, displayed through a real-time dashboard. Such prospective validation studies, akin to phase 4 clinical trials for drugs, will help in improving algorithms in the long term and are the only way to realistically measure the impact of such algorithms in real clinical practice.

Another real-world deployment strategy for AI algorithms is an AI-enabled retrospective quality audit. In this quality audit, several hundreds or thousands of chest X-rays that are deemed 'normal' based on their previous radiology reports, are read by an AI algorithm at once. Subsequently, abnormality scores (probability of each X-ray having an abnormality) are determined and images having high abnormalities are re-read by a 2nd arbitration radiologist. It is highly likely that the AI finds 'missed' findings which either the radiologists did not pick up on, or did not feel the need to comment on – either way, the AI algorithm gives a low-risk method to pick mistakes and improve the quality of radiology reporting in general.

Improving the algorithm

The outcome of an algorithmic audit should ideally be solutions or suggestions using which vendors can improve the functioning of the algorithm. There are two possible solutions, in our experience. The first possible solution may be to retrain the model using false positives and false negatives found during the audit, especially the far north and far south cases. Such retraining can also be done locally, i.e. at the deployment site using a technique called federated learning[17] where data does not need to leave the premises of the hospital/healthcare provider. The second possible solution is to use 'Dynamic Thresholds'. A dynamic threshold refers to a threshold value (for determining presence or absence of an abnormality, as discussed earlier) that changes based on the given clinical context. Although not published in literature, preliminary work done by our group demonstrates the ability of dynamic thresholds to significantly reduce the error rate of algorithms by adding a layer of 'clinical sense' into the output of AI algorithms. It is intuitive and can be explained using an example - a patient who has come for a screening chest X-Ray for a routine health check should have a very high threshold for detecting pneumothorax, as opposed to one who is in the intensive care unit with a falling oxygen saturation. Adding a simple checkbox to determine the patient's clinical context (screening, trauma, fever, cough, intensive care, etc.) can aid in improving AI's accuracy and making it more accessible.

In conclusion, one should keep the following in mind while working with vendors to validate algorithms:

- Develop strong ground truth - whether it is by individual or group consensus or by finding objective measures such as chest CT for chest X-Rays;
- Datasets should ideally be from sites the algorithm has not been trained on;
- Develop a good test dataset – case mix of positives and negatives is very important and depends on the use-case of AI under test;
- Focus on the cases that failed - what the implications of failure, and build hypothesis around why AI might have failed;
- Focus on Far North and Far South cases – these will limit trust that doctors can bestow on AI;
- Aim for real-world deployment and testing;
- Stay in constant communication with developers of AI – help them determine how algorithms can be improved.

References

1. Handelman, G. S. *et al.* Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *Am. J. Roentgenol.* **212**, 38–43 (2019).
2. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean J. Radiol.* **20**, 405 (2019).
3. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Med.* **15**, e1002699 (2018).
4. Štajduhar, I., Mamula, M., Miletić, D. & Ünal, G. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput. Methods Programs Biomed.* **140**, 151–164 (2017).
5. Prevedello, L. M. *et al.* Challenges Related to Artificial Intelligence Research in Medical Imaging and the Importance of Image Analysis Competitions. *Radiol. Artif. Intell.* **1**, e180031 (2019).
6. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
7. Chilamkurthy, S. *et al.* Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet* **392**, 2388–2396 (2018).
8. Singh, R. *et al.* Deep learning in chest radiography: Detection of findings and presence of change. *PLOS ONE* **13**, e0204155 (2018).
9. Venugopal, V. K., Mahajan, V. & Mahajan, H. Automated classification of chest x-rays as normal/abnormal using a high sensitivity deep learning algorithm. in (2019).
10. Annarumma, M. *et al.* Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* **291**, 196–202 (2019).
11. Rodriguez-Ruiz, A. *et al.* Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur. Radiol.* **29**, 4825–4832 (2019).

12. Taylor, A. G., 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.* **15**, e1002697 (2018).
13. AI-embedded X-Ray system could help speed up detection of a collapsed lung. (2018).
14. Zhao, Q., Zhang, M., Gopal, A., Venugopal, V. K. & Mahajan, V. Pneumothorax Detection and Localization in X-Ray Images Given Richer Annotation Information. in
15. Health Level Seven International.
16. Open Source Clinical Image and Object Management.
17. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J. & Bakas, S. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (eds. Crimi, A. et al.) **11384**, 92–104 (Springer International Publishing, 2019).