Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs

Ju Gang Nam, MD • Sunggyun Park, PhD* • Eui Jin Hwang, MD • Jong Hyuk Lee, MD • Kwang-Nam Jin, MD, PhD • Kun Young Lim, MD, PhD • Thienkai Huy Vu, MD, PhD • Jae Ho Sohn, MD • Sangheum Hwang, PhD • Jin Mo Goo, MD, PhD • Chang Min Park, MD, PhD

The title of our first article is, “Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs.” This paper is a large collaboration: six institutions from Seoul, Korea are represented, one author from industry, and others from UC San Francisco.

Keeping track of useful developments in artificial intelligence is critical. As we go into the details, I will explain why this work a very serious effort in medical AI.

First the topic, detecting nodules on chest x-ray. Not novel by itself. There needs to be something unique for us to pay attention. The problems of chest x-ray are well known. Sensitivity is only moderate, about 20% of lung cancers are missed by humans on chest x-rays – but the nodules are obvious in retrospect when pointed out. There have been many prior efforts for computer-aided detection of nodules on chest x-ray. All are worse than a first-year radiology resident. The problem is false positives – the computer-aided diagnosis, or CAD system flags an abnormality on the film, requiring our time and attention to eventually dismiss the finding as negative. Prior CAD systems detected between 1 and 3 false positive findings per chest x-ray. A high number of false positives were the only way for conventional CAD systems to achieve adequate sensitivity to detect real findings.

Purpose: To develop a deep learning system to detect malignant pulmonary nodules on chest x-rays, and compare the performance to expert thoracic radiologists, general radiologists, and non-radiologists.

Methods: Lots of depth in this article:

Step 1: More than 43,000 chest x-rays were used to train the AI. About one-third of these had nodules, two-thirds were normal. The authors enlisted 29 readers at various institutions to label those training chest x-rays. More than 40,000 x-rays is a lot – but is it enough?

In the November podcast, we spoke about an AI from Stanford University also with the aim to interpret chest x-rays. But the Stanford team aimed to classify the frontal chest x-ray as simply normal or abnormal. Normal or abnormal is vague. Abnormal would be anything, a pneumonia, cardiomegaly, healed rib fracture, or lung cancer. The current study from Korea aimed to detect only malignant nodules.

The Stanford group had 200,000 frontal chest x-rays. But, they found the performance did not get much better after 20,000 films. Based on the Korea research, I think that number will not be enough. The Korea group used 43,000 films. As I will detail below, 43,000 was not enough by itself to train for all the patterns of malignant nodules. So that’s the first take-home message: more is better, despite the Stanford results. The reason why: think about the most unusual lung cancers you have ever detected. You saw it only on the lateral view in some odd location. Or the film had a little bit unusual contrast and you were able to see below the diaphragm through the liver dome. Your most unusual cancers ever, have occurred only a few times in your entire career. But to train an AI, it’s likely that dozens or hundreds of all of the unusual cases put together in your entire hospital would be needed.

My conclusion: once we get an AI for chest x-rays, we will need to keep it learning. Alexa or Siri have that advantage: their recognition and learning gets better over time.

AI will need to continue to learn from humans. If I call a nodule, the AI is supposed to learn. Therein lies the problem for the chest x-ray. If a human calls a nodule and requests a CT for follow up, only about one-third of cases are true positives. If we don’t actively teach the AI with good follow-up data, it will also read only one-third of cases as true positives.

Back to the research study. After the Korean researchers trained the AI, they did an internal validation on a set of 300 normal and 300 abnormal...
x-rays. After that, they did external validation based on 700 x-rays from 4 external hospitals – 3 in Korea, and 1 in the U.S.

The authors were extremely careful in training the AI. They had 29 readers doing labels for the 43,000 x-rays. But they did not necessarily believe each reader. They only called a nodule positive if they had a strong consensus of the 5 readers. They labeled up to 5 nodules per x-ray. If the nodule was indeterminate, they required a CT.

Results: Ok, how did it work? Remember the AUC curve. The AUC value, area under the receiver operating curve, goes from 0 to 1. It is a measure of diagnostic performance that combines sensitivity and specificity all in one number.

AUC of 0.5 is by chance, 1 is perfect. Given 2 images, 1 normal and 1 abnormal, AUC is the probability that the abnormal image is likely to be identified as abnormal. In this study: the AUC value of the AI was 0.96, or 96% on the internal validation dataset. At the 4 external hospitals, the AUC varied from 0.92 to 0.99.

Ok, just a little more math, but worth reviewing. You have heard many times in these podcasts and seen in many papers that we prefer to see the AUC values, the overall measure of diagnostic performance. One reason we like it very much: the AUC curve gives us the entire set of sensitivity and specificity values. If you have a first-year resident, they miss a lot of findings. But when they do see a fracture, it’s pretty definite. Mathematically: their sensitivity for finding a fracture is low – but their specificity is very high. Let’s give the first-year resident an AUC value is 0.6 – better than chance, which was 0.5.

The second- or third-year resident may be the opposite: they start to find ALL of the abnormalities. Actually, they find too many; they over call – their sensitivity is very high, but the specificity is low. The AUC is still only 0.6; they get penalized for over calling lesions with the AUC.

Now let’s take the world’s best radiologist: she reads a chest x-ray as well as humanly possible. Of course, the chest x-ray is far from perfect compared to a CT. The world’s best chest x-ray reader compared to CT has an AUC only of perhaps 0.75. It’s not the reader: it’s the chest x-ray technology. In this study, the AI had a score of about 0.96 compared to the human readers, not compared to CT.

One more term in case you have not heard of it: JAFROC (J-A-F-R-O-C) figure of merit. The authors report this number as well. It also goes from 0 to 1. JAFROC stands for Jackknife alternative free-response receiver operating characteristic, quite a mouthful. The problem: if you read a chest film, you do not know how many abnormalities there are. There could be 1, there could be 20. You also do not know where the abnormalities are. I show you an x-ray, ask you to mark any lesions. Let’s say you put a mark on the x-ray within 0.5 cm of the real lesion. I would say you correctly found the lesion. But for another real finding, you put a mark 2 cm from the real lesion. In that case, I doubt you were marking the right finding. JAFROC takes that into account: you have to correctly mark all the lesions and you have to mark their locations. So pay attention to the JAFROC numbers: they will almost always be lower than the AUC values. The JAFROC numbers allow us to see differences in performance better than the AUC – by taking location into account.

So far I discussed the AI reading all by itself. But how did AI compare to human readers? For 15 of 18 readers, the AI added value: there was significantly better performance when the AI was the second reader.

What about those false positives from older CAD systems – I mentioned that as the most frustrating part of prior nodule detection systems. For non-radiologists – such as pulmonary physicians: when the AI suggested a nodule, the human incorrectly overruled the AI 37% of the time. But the most experienced thoracic radiologists: they incorrectly overruled the AI almost 70% of the time! Why did the best readers ignore the AI? The authors suggest 2 reasons.

#1: The AI missed essentially all the nodules below the diaphragm – nodules that overlapped with the dome of the liver. In retrospect, the authors recognized they did not correctly train the AI for those lesions. Also, the AI missed most nodules less than 1 cm: the AI got 11% of those nodules. Humans did much better: 40% of nodules less than 1 cm. Remember the AI still performed at or better than most of the humans overall: as a group, humans were erratic for nodules between 1 and 3 cm.

#2 Second reason for incorrectly overruling the AI: radiologists had prior experience with prior nodule CAD systems. Those CAD systems had many false positives: 1 to 3 false positives for every chest x-ray. But the new AI was much better: only had 1 false positive for every 3 x-rays. The expert radiologists had not calibrated to the better performance of the new AI, they failed to realize it was really working.

Conclusion: The depth of the research in this article is impressive. A large number of x-rays, 29 individuals labeling the images, 18 readers of the x-rays, external validation from 4 institutions, extensive statistical analysis, a tremendous amount of work, nice results.

But the authors are also realistic: they realized they still need more training for the AI. The AI needed to learn the lateral chest x-ray. The AI needed to learn nodules below the diaphragm. And the AI needed to learn about nodules less than 1 cm.

One interesting remark from the authors: Compared to a human, the AUC was almost perfect, AUC score 0.96 compared to the maximum score of 1.0. But as I mentioned: the world’s best radiologist has an AUC for chest x-ray that is much worse compared to CT scan. The authors speculate: perhaps they should skip the humans, instead train the AI to see findings invisible on chest x-ray for humans, but that are present on CT.
The authors used an AI called ResNET 18. You can download a pre-trained version of ResNET 18. Out of the box, ResNET 18 can recognize common objects such as a computer keyboard, mouse, a pencil, many animals, up to 1,000 different objects. When researchers use a pre-trained network to interpret medical images, there can be an advantage, called transfer learning. The research paper of the year in this journal from Thomas Jefferson University used a network with transfer learning to diagnose pulmonary tuberculosis on chest x-rays.

The authors of our current paper used an untrained version of ResNET 18, believing that dogs and cats had little to do with cerebral aneurysms on MIP images. ResNET was released in 2015 to solve a problem with convolutional neural networks: as the number of layers of neural connections increased, the performance of the algorithm would plateau, or actually decrease. In ResNET, the neural connections are allowed to skip layers in the training phase. Those neural layers can be added back later, after the AI has enough knowledge to know more image features. ResNet 18 has 18 neural layers.

Results: ResNET 18 found 91% of aneurysms in the internal data. When tested on completely external data, the results were similar, 93%, detecting 74 of 80 aneurysms.

Remember that an AI does not really diagnose an aneurysm. It predicts the probability that an aneurysm is present. In order to get a sensitivity of more than 90%, the AI also found many possible aneurysms. For us, that means a lot of false positives. I spoke of this in the last article.

After a while, humans get tired of false positives suggested by CAD software. We starting thinking of the CAD as having the ability of a second- or third-year resident. Able to find many abnormalities, but not knowing the significance of those findings.

On the other hand, in the case of aneurysms, the authors decided at the outset that they would use the AI to read the MRI after the radiologist. As a double check.

In that context: compared to the radiologists: the AI found 5% more aneurysms than the humans found. There were 31 additional aneurysms in 649 cases. In the external test set, the AI found 13% more aneurysms not found by the humans.

Finally: the authors looked at the results of ResNET as a function of MRI manufacturer, hospital, location, and size of the aneurysm. Overall, the AI was very robust: relatively similar performance that was independent of manufacturer or MRI field strength. That was impressive and usually a pitfall of most other AI research.

Interesting, ResNET was really good at finding very small aneurysms less than 3mm. But some large aneurysms more than 10 mm were missed. The authors looked at those cases: the larger aneurysms had flow abnormalities. Humans understood that the flow in large aneurysms would be uneven, not uniformly bright. But the AI did not understand this. Also, some aneurysms in the vertebral artery regions were missed. In both cases, the authors could provide training to ResNET for those kinds of aneurysms.

Conclusion: Impressive, robust performance over multiple field strengths, multiple institutions, multiple manufacturers for detection of cerebral aneurysms. By far the largest study of this type – still, the authors acknowledge they need even more training.

ResNET found between 5 and 13% more aneurysms than humans. But the AI also missed some aneurysms that a human would easily detect and functioned like a second-year resident with many false positives. So on an optimistic note, let me quote directly from these very matter-of-fact authors from Japan: “Our results suggest that radiologists and the present algorithm with deep learning technique are complementary to each other.” That means we can co-exist, at least for now.

For a nice perspective, Dr. Adam Flanders at Thomas Jefferson has an editorial to go along with the paper. He asks the question: if state-of-the-art MRI and humans already have sensitivity in the upper 90th percentile to detect aneurysms, do we need an AI? He finds these results to be beneficial for small aneurysms less than 3 mm that may otherwise escape the eye of the busy clinician. I believe that was exactly the aim the authors had in mind.

Spot and Diffuse Signs: Quantitative Markers of Intracranial Hematoma Expansion at Dual-Energy CT

Radiology 2019; 290:179–186

Can Ozan Tan, PhD • Stephanie Lam, MD • Daan Kuppers, BS • Rick H. J. Bergmans, BS • Bimal Kumar Parameswaran, MD • Reza Forghani, MD • Ranliang Hu, MD • Laleh Daftari Besheli, MD • Joshua N. Goldstein, MD, PhD • James Thrall, MD • Michael Lev, MD • Javier M. Romero, MD • Rajiv Gupta, MD, PhD

Our next article is also neuro. The title is, “Spot and Diffuse Signs: Quantitative Markers of Intracranial Hematoma Expansion on Dual-Energy CT.” The senior author is Dr. Rajiv Gupta from MGH in Boston. Co-authors are from McGill, University of Twente in Holland, and Emory in Atlanta.

The topic is to predict which patients do poorly after intracerebral bleed detected on CT. Expansion of hematoma is reported in about one-third of patients within 24 hours. For patients on anticoagulation, 50% of hematoma may expand. Early hematoma expansion is strongly associated with neurologic deterioration, worse outcome, and mortality.

The standard of reference for predicting hematoma expansion is called the spot sign at CT angiography. On delayed images, the spot sign is one or more spots of bright contrast enhancement within a hematoma. The spot sign is due to active contrast extravasation. The specificity for the spot sign is extremely good, 80-93%. But pooled data from a 14-center meta-analysis shows the sensitivity is poor, about the toss of a coin, at 53%.

The authors identified a potential problem with the spot sign. The blood in the hematoma is already bright. So finding bright spots within a bright hematoma can be inaccurate. Dual-energy CT to the rescue. Dual-energy CT can separate the x-ray properties of hematoma from iodine extravasation.

Purpose: To compare dual-energy CT to single-energy CT spot sign to predict intracranial hematoma expansion at 48 hours.

Methods: The authors had 43 patients with various causes of intracranial hematoma on baseline CT. At 48-hour follow up CT, one-third of patients had expanding hematoma, defined as growing 3 cm3 larger at follow up compared to baseline. The authors did quantitative analysis of the dual-energy CT. The CT protocol included a CT angiogram and a delayed CT at 90 seconds after iodine injection.

After much analysis, there were only 2 main factors that predicted hematoma expansion. Factor #1: the total concentration of iodine in the hematoma. Factor #2: the iodine concentration in the brightest spot in the hematoma.

Next, the authors compared dual-energy to the conventional spot sign interpreted by radiologists blinded to patient outcomes. They evaluated a set of 65 patients to see if the new dual-energy formula held up.

Results: The conventional spot sign. Like prior literature, the sensitivity was only a little over 50%: in the MGH cohort, the sensitivity was 57%. But the specificity was higher, about 80% as expected.

For dual-energy CT: The sensitivity was much better: about 80%. The specificity was also slightly better at 88-89%.
Diagnosis of Pulmonary Hypertension with Cardiac MRI: Derivation and Validation of Regression Models

Radiology 2019; 290:61–68

Christopher S. Johns, FRCP • David G. Kiely, MBChB • Smitha Rajaram, MD • Catherine Hill, FRCR • Steven Thomas, FRCR, MSc • Kavitasagary Karunasagarar, FRCR • Pankaj Garg, PhD • Neil Hamilton, DPharm • Roshni Solanki • David A. Capener, MSc • Charles Elliot, MD • Ian Sabroe, PhD, FRCR • Athanasios Charalamopoulos, MRCP • Robin Condiffe, MD • James M. Wild, PhD • Andrew J. Swift, PhD

A datamining approach was used by the research team: they measured every possible cardiac MRI parameter that could potentially predict pulmonary hypertension. There were 30 parameters. The authors used a multivariable regression equation to create a mathematical formula to predict the presence or absence of pulmonary hypertension.

A main strength of the study is the large size. 300 patients were used to derive the regression equation. Then, a different set of 300 patients was used to test if the equation really worked.

Results: Fortunately for clinical simplicity, only 3 parameters were needed to predict pulmonary hypertension: the ratio of the mass of the right ventricle compared to the left ventricle, the area of the main pulmonary artery. The third parameter defines the abnormal shape of the left ventricle that results from the abnormal right heart pressure: that 3rd parameters is called the interventricular septal angle.

The sensitivity of the final 3 parameters was 92% to detect pulmonary hypertension greater than 25 mm Hg. The AUC value was 0.93, quite high. But specificity was only about 60%. The specificity was a little better if slow blood flow in the pulmonary artery was taken into account. But the images that showed slow blood flow were qualitative and likely not reproducible on different MRI scanners.

Conclusion: this is very nice work, on a huge cohort of more than 600 patients evaluated for pulmonary hypertension. It is the first and best study of noninvasive MRI to successfully identify the presence or absence of pulmonary hypertension.

Deep Learning–based Method for Fully Automatic Quantification of Left Ventricle Function from Cine MR Images: A Multivendor, Multicenter Study

Radiology 2019; 290:81–88

Qian Tao, PhD • Wenjun Yan, BSc • Yuanyuan Wang, PhD • Elisabeth H. M. Paiman, MD • Denis P. Shamonin, MSc • Pankaj Garg, MD, PhD • Sven Plein, MD, PhD • Lu Huang, MD, PhD • Liming Xia, MD, PhD • Marek Sramko, MD, PhD • Jarsolav Tintera, PhD • Albert de Roos, MD, PhD • Hildo J. Lamb, MD, PhD • Rob J. van der Geest, PhD

Next article, short title, “Deep Learning for Fully Automatic Quantification of Left Ventricle Function on MRI: A Multivendor, Multicenter Study.” The senior author is Dr. Rob van der Geest. Rob is from University of Leiden in the Netherlands. Rob has been working on quantitative cardiac analysis for about 30 years—probably longer than anyone in the world. Maybe he has a lot of motivation—his wife is a cardiologist, but in Holland I believe she does interventional cardiology rather than noninvasive MRI. Rob’s innovations keep getting more remarkable, but so do the treatments offered by interventional cardiology. Slow leakage of secondary vessels contribute to further delayed hematoma expansion. This slow leakage is represented by the total iodine content in the hematoma, without a discrete spot.

Limitations of the research: the authors did not account for blood pressure, anticoagulant therapy, and clinical variables. That will be left to a future project.
different physicians, technologists, and MRI centers is much greater, perhaps 10 to 20%. Radiologists who focus on ventricular function speak about drawing circles around the heart. My research lab did this for almost 15 years for about 10,000 research patients. Each patient had about 800 images to evaluate. The conventional software analysis is completely tedious and brute force by drawing circles at the inner and outer contours of the heart. About 16 million circles drawn in my research lab over 15 years.

Now things are changing with AI. One of my doctoral students had a few extra months after finishing his PhD. He did part of his PhD research at the NIH and then completed the work with Dr. Alison Nobel at Oxford. Her lab invents novel techniques for software analysis of the heart, based on CT, MRI, and echocardiography. After 15 years of conventional work, our newly-trained PhD student developed an AI method to automatically analyze the entire heart — including the atria and both ventricles — in about 3 months. The current article is also about an AI for the heart.

Purpose: develop deep learning for automated quantification of left ventricular structure and function from MRI.

Four-dimensional Flow MRI as a Marker for Risk Stratification of Gastroesophageal Varices in Patients with Liver Cirrhosis

Radiology 2019; 290:101–107

Utaroh Motosugi, MD, PhD • Alejandro Roldán-Alzate, PhD • Peter Bannas, MD • Adnan Said, MD • Sean Kelly, MD • Ryan Zea, MS • Oliver Wieben, PhD • Scott B. Reeder, MD, PhD

Last topic, also related to blood flow. The short title: “Four-dimensional Flow MRI for Risk Stratification of Gastroesophageal Varices in Patients with Liver Cirrhosis.” The senior author is Dr. Scott Reeder, University of Wisconsin, Madison.

Background: Patients with cirrhosis who have hepatic venous pressure more than 12 mm Hg are likely to have varices. But the potential for variceal bleed depends on the type of collaterals that they develop. Collaterals that develop around the esophagus are mostly likely to bleed, but other collaterals serve to decompress the liver via splenorenal shunts or from the paraumbilical veins leading back to the systemic veins.

Endoscopy can assess the risk of esophageal varices for rupture. High-risk esophageal varices that rupture tend to have red patches on the mucosal surface that can be directly visualized. Other varices have fibrin-platelet plugs that are seen on the mucosa. These high-risk varices can undergo sclerotherapy to prevent rupture and bleed.

Purpose: determine if MRI can identify high-risk esophageal varices using 4D Flow images.

Methods: 4D flow uses radial imaging with phase contrast technique. The velocity encoding gradient, or VENC, is set to a low value of about 30 cm per second to detect slow, venous flow. About 10,000 radial projections through k-space are acquired in 10 minutes. A respiratory bellows is used to minimize motion.

Results: This was a proof-of-concept study in 23 patients. The authors looked at factors on the 4D flow MRI that correlated with high risk features of varices at endoscopy.

There were 2 primary MRI features that were identified. The first risk factor was flow that was detected in the azygos vein of more than 0.1 liters per minute. Flow in the azygos vein is near zero in the normal individual. The other factor was to assess the portal vein. The total portal venous flow should be approximately the total of flow in the superior mesenteric vein plus the splenic vein. If splenic flow plus SMV flow do not add up to portal flow, this means that there is reversed flow likely occurring in the coronary vein. The coronary vein, also known as the left gastric vein, has reversed flow in patients with portal hypertension and is the major blood supply to the esophageal collaterals.

Conclusion: This is sophisticated MRI analysis, not quite ready for the clinic except in a few specialized centers. Exquisite movies are provided online, showing the elegance of the technique. In an example patient with large varices, the coronary veins that should drain into the portal vein instead clearly show reversed flow: blood is shunted from the mesenteric venous system to the coronary veins and then to the abundant esophageal varices. The main advantage of the 4D flow method is a single acquisition. Conventional 2D phase contrast images would be nearly impossible to use. The technique requires the technologist to place cross sections perpendicular to the vessel lumen. But venous collaterals are tortuous and have unpredictable patterns. 4D flow cine MRI solves this problem with a single 3D acquisition. The method is becoming more automated and a commercial release is already available by a major vendor.

That concludes this week’s articles. I hope these podcasts were helpful to you. Until next time, this is Dr. David Bluemke for the journal Radiology. I hope you have a good rest of your week.