Performance of an artificial intelligence tool with real-time clinical workflow integration – Detection of intracranial hemorrhage and pulmonary embolism

Nico Buls¹,¹, Nina Watté¹, Koenraad Nieboer, Bart Ilsen, Johan de Mey

Department of Radiology, Universitair Ziekenhuis Brussel (UZ Brussel), Vrije Universiteit Brussel (VUB), Laarbeeklaan 101, 1090 Brussels, Belgium

**Introduction**

Acute pathologies require early detection with prompt communication of critical findings to ensure adequate clinical management. Intracranial hemorrhage (ICH) and pulmonary embolism (PE) are two of such frequent life-threatening pathologies, with significant morbidity and mortality, where misdiagnosis can lead to adverse outcome [1-3]. A non-contrast head CT scan is essential to confirm diagnosis and risk stratification of ICH, while contrast enhanced Computed Tomography Pulmonary Angiography (CTPA) is a standard scan for detecting and locating PE [3,4].

Advances in CT technology have led to the improvement of image quality and reduction of radiation dose, which allows the diagnosis of more subtle lesions. However, the increasing volume in number of examinations and images per examination, can have a disproportionate effect on the radiologist’s work stream. McDonald et al., calculated in their study on the influence of technological advancements of cross-sectional imaging on the radiology workflow, that a radiologist analyses an average of one image every three seconds [5]. This time-intensive encumbrance on the practicing radiologist, can accrue an increase in false negative results and misdiagnosis [6-8]. Real-time double reading by a peer is often done, which has been proved to aid in lowering the prevalence of misdiagnosis, however it is very labor-intensive. In addition, retrospective peer reviewing of cases does not immediately improve the patient’s clinical outcome, especially not in an acute setting [9-11].

Given the potential adverse outcome in case of misdiagnosis of ICH or PE, the increasing radiology workload, the constant development of new advanced computed tomography techniques and nowadays pandemics that effect our health care system, artificial intelligence (AI) technologies can assist radiologists and serve as a real-time clinical adjunct to diagnose ICH and PE. Using convolutional neural networks (CNN) based on deep learning, AI algorithms are becoming accessible, which can detect those life-threatening lesions [12-14]. AI technologies have multiple potential roles such as quality assurance and productivity enhancement. However, certain roles within specific pathologies have not yet been fully investigated. Implementing an AI tool during a real-time radiology work stream, has the potential to react earlier and/or even notice lesions that can be easily overlooked by a radiologist [13,15-18].

Much research in recent years has focused on such AI solutions, which has indicated a sensitivity of 0.95, specificity of 0.99, negative predictive value (NPV) of 0.98 and positive predictive value (PPV) of 0.98 with an overall accuracy of 0.98 for ICH detection. Rao et al. applied a AI solution to negative-by-report ICH cases. They found a false-negative rate of radiologists for ICH detection at 1.6%, and thus the technology could serve by minimizing false negatives [19]. Weikert et al. found a high degree of diagnostic accuracy for PE detection on CTPAs, and a balanced sensitivity and specificity of 0.93 and 0.96 [18]. Also with PE, sensitivity, specificity, positive and negative predictive values and accuracy compared with gold standard senior radiologists were reported, 0.85, 0.97, 0.85, 0.97 and 0.95 respectively [18].

The purpose of this study was to assess the performance of a commercially available AI tool as a second reader in detecting ICH and PE in a diverse clinical setting (e.g., emergency, routine, inpatient, and outpatient) in real-time assessment, by evaluation of the number processed studies by the AI tool and calculation of the diagnostic performance respectively.

**Materials and methods**

This retrospective study was conducted with approval by our local institutional medical ethics committee with a waiver of informed consent. Our study was bipartite and focused on two acute pathologies: Intra-cranial hemorrhage (ICH) and pulmonary embolism (PE). Each subdivision consisted of 4 stages: (1) Dataset collection; (2) Image data
processing by an automated AI tool; (3) Quality control and discrepancy revision by registered radiologists with certificate of added qualification in neuroradiology or thorax radiology; and (4) diagnostic performance analysis.

Dataset collection

Random case collection was performed from a consecutive database of patients referred to our radiology department for an non-contrast head CT or CTPA, blinded to clinical data regarding antecedents, diagnosis, therapy or outcome. Both for the brain and lung study, patients under the age of 18 were excluded. CT exams were pseudo-anonymized, retaining solely an identification code to link each report to its respective study. Control CTs were excluded, resulting in eliminating duplicate exams and a final cohort of unique patients and scans.

A total of 500 consecutive non-contrast CT exams of the head performed over 31 days from September 1, 2019 until October 1, 2019, were included. This consecutive case collection varies considerably in terms of neurologic pathology signs such as hemorrhage, mass effect, hydrocephalus, suspected acute infarct, encephalomalacia or no evidence of intracranial disease. Cases also differ markedly in hemorrhage age and attenuation on CT (respectively hypo- and hyperattenuating), hemorrhage size and location (epidural, subdural, subarachnoid and intraparenchymal). Scans with movement artifacts, or beam hardening artifacts, were included as well to represent routine standard practice.

Secondly, we considered 500 consecutive CTPA scans performed between July 1, 2019 and February 1, 2020. This data set consists of pulmonary emboli (central, segmental and subsegmental), common diseases such as interstitial pneumonias, acute respiratory distress syndrome, sarcoidosis, lymphangitic carcinomatosis, cardiogenic pulmonary edema and normal findings. Scans containing moderate breathing images with arrows pointed where the pathology is situated. The AI report is seamlessly integrated into the clinical workflow, with the results being automatically added to the CT study. The typical time between the CT acquisition to the notification (AI results available in PACS) varies between 3 and 7 min for ICH studies and 5–9 min for PE studies.

Diagnostic performance and discrepancy review

From the 500 consecutive head CT exams and CTPA exams that were presented to the AI tool, we registered the number of studies that were sent back with an AI report. Secondly, we evaluated its diagnostic performance by comparison to expert reviewing. The original clinical radiology report after consensus review by 3 neuro-radiologists for ICH and 3 thorax radiologists PE, was considered as gold standard. Six board-certified radiologists participated in the consensus review with each 5 up to 15 year of experience in reading unenhanced head CT and CTPA studies. The reviewers had access to prior and future studies, and were able to see clinical history and reports to diagnose. AI results were classified into true positive, false positive, true negative and false negative.

**Table 1**

Scanner models, scan- and reconstruction parameters and radiation doses used for the two indications. Doses represent median values with 95% confidence intervals between brackets.

<table>
<thead>
<tr>
<th>Scanner model (indications)</th>
<th>Number of cases</th>
<th>Tube voltage (kV)</th>
<th>Single collimation width (mm)</th>
<th>Reconstructed slice thickness (mm)</th>
<th>Reconstruction method</th>
<th>CTDIvol (mGy)</th>
<th>DLP (mGy.cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intracranial hemorrhage (ICH)</td>
<td>500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE Revolution</td>
<td>212 (42%)</td>
<td>DECT</td>
<td>0.625</td>
<td>0.625</td>
<td>DLR-M</td>
<td>35.8</td>
<td>724 (698–751)</td>
</tr>
<tr>
<td>GE Discovery 750HD</td>
<td>173 (35%)</td>
<td>120</td>
<td>0.625</td>
<td>0.625</td>
<td>ASiR 30%</td>
<td>38.4</td>
<td>756 (725–787)</td>
</tr>
<tr>
<td>Philips iCT</td>
<td>109 (22%)</td>
<td>100</td>
<td>0.625</td>
<td>0.8</td>
<td>iDose3</td>
<td>29.9</td>
<td>598 (581–614)</td>
</tr>
<tr>
<td>Siemens Somatom AS40</td>
<td>6 (1%)</td>
<td>120</td>
<td>0.6</td>
<td>0.6</td>
<td>FBP</td>
<td>60.8</td>
<td>1006 (990–1023)</td>
</tr>
<tr>
<td>Pulmonary embolism (PE)</td>
<td>500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE Revolution</td>
<td>282 (56%)</td>
<td>DECT</td>
<td>0.625</td>
<td>0.625</td>
<td>ASiR-V 70%</td>
<td>6.4 (5.9–6.8)</td>
<td>229 (211–248)</td>
</tr>
<tr>
<td>GE Discovery 750HD</td>
<td>203 (41%)</td>
<td>100-120</td>
<td>0.625</td>
<td>0.625</td>
<td>ASiR 30%</td>
<td>11.8</td>
<td>411 (405–478)</td>
</tr>
<tr>
<td>Philips iCT</td>
<td>15 (3%)</td>
<td>100-120</td>
<td>0.625</td>
<td>0.9</td>
<td>iDose3</td>
<td>4.1 (3.8–4.4)</td>
<td>174 (163–183)</td>
</tr>
</tbody>
</table>
negative cases. True-positive (TP) cases contained hemorrhage or embolism detected by the AI tool, and subsequently confirmed by the consensus reviewers. True-negatives (TN) consisted of exams without ICH nor PE according to the AI tool and reviewers. False-positives (FP) were defined as cases that were flagged positive by the AI tool but found out to be negative. False-negatives (FN) were defined as cases that were classified by the AI tool as negative but decided to be positive for ICH/PE by consensus review. We quantified the diagnostic performance of the AI algorithm in ICH and PE detection by calculating the sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV) and accuracy. The concordance between the AI tool and consensus review on each pathology, was calculated using percentage of agreement and Cohen’s k statistic.

In addition, the expert reviewers performed a detailed discrepancy analysis of the false-positive and false-negative cases in order to identify the reason for miss classification by AI.

Results

Table 2 summarizes the AI performance results for detecting ICH and PE. From the 500 presented cases, the AI algorithm could process 77.6% (388 of 500) for ICH evaluation during real-time radiology work stream. No ICH evaluation was performed for 112 studies. There was a difference in process rate between scanners models ranging from 84.9% (GE Discovery HD) to 53.8% (Philips iCT). All 6 Siemens cases were rejected for processing. From all 388 processed studies, the AI tool flagged 31 (7.9%) exams as having ICH. Expert review flagged 37 (9.5%) hemorrhages. Substantial agreement (kappa-value of 0.65) between AI and expert reading was observed. The performance for ICH showed 0.84 sensitivity and 0.94 specificity. The negative and positive predictive values were 0.98 and 0.61 respectively. The AI tool failed to label 1.7% (6 of 337) cases, that were agreed to have hemorrhage after review by three subspecialists (false negative by the AI tool). Those six cases summed in Table 3, conclude two discrete subarachnoid hemorrhages, two subdural hemorrhages and two pachyenchymal hemorrhages. Twenty positive results out of 37 flagged exams by the AI tool, were labelled as false-positive findings after consensus review, which the reviewers assigned to falcine or basal ganglia calcifications (9/20 cases), beam hardening artifacts (8/20 cases) and hyperdense dural sinuses (3/20), shown in Table 3.

The AI technology created a report for 448 (89.6%) consecutive CTPA’s. Similar to ICH, the process rate for the Philips iCT was lower (13.3%), compared to the GE scanners (90.6% and 92.5%). The sensitivity and specificity and accuracy were 0.73, 0.95 and 0.90 respectively. The kappa value for agreement between AI and expert reading was 0.73, indicating a substantial concordance. The expert readers detected 82 cases positive for PE. The AI system did not identify 19 of these 82 PE cases (false negative by AI). Nine of these patients, had chronic pulmonary embolisms. Six cases had masquerading artifacts. In three cases, an underlying pathology had concealed the present emboli. In 1 patient, a superimposing vein was the cause of the missed pulmonary embolism. 17 out of 19 misdiaagnosed patients had subsegmental and segmental PE. Chronic known central emboli were missed in 1 patient. Lobar emboli were misdiagnosed by AI in 1 patient during delayed scan phase.

24.4% of studies were found to be false-positive findings by the AI solution. According the consensus reviewers, six FP cases were due to contrast agent-related flow artifacts, beam hardening artifacts and breathing artifacts. Another six false positive cases were due to the masking effect of associated pathologies (such as infiltrate, metastasis, pleural effusion, atelectasis and fibrosis) or superposition anatomy (e.g. pulmonary vein, lympha node, hilar soft tissue, bronchus, azygos vein or pulmonary artery bifurcation). Withal, 7 out of 18 patients had a false positive diagnosis caused by a combination by the aforementioned factors.

Examples of intracranial hemorrhages and pulmonary embolism with AI detection are shown in Figs. 1-4.

Discussion

This two-folded study assessed the diagnostic performance of an AI algorithm to automatically rule out ICH on non-contrast head CTs on one hand, and PE diagnose via CTPAs on the other hand, with real-time clinical work flow integration. In a diverse real-time clinical setting, 77.6% (388/500) of consecutive head CT exams and 89.6% (448/500) CTPA exams could be automatically evaluated by AI. However, AI process rates increase to 84.6% (326/385) for head CT exams and 91.7% (445/485) for CTPA if we only consider the two GE scanners who represent the bulk of the studies. We did not assess the cause of failure to process nor the follow up of these patients, as this was not part of the study protocol. Possible causes can be hospital-network related or can be attributed to inadequate radiological quality due to, for example, increased noise, the presence of motion artifacts or metal artifacts. The few Siemens cases were rejected for AI processing (0/6) because they were not compliant with the AI-tool requirements (<64 slice CT).

With a specificity of 0.94 and a 0.98 negative predictive value, our ICH study is in line with prior work. Previous research with the same AI tool reported a specificity of 0.99 and a NPV of 0.98 for ICH detection. However, the sensitivity and PPV were 0.84 and 0.61 in our study, and remain moderate in comparison with previous studies, in which a sensitivity of 0.95 and 0.98 PPV was obtained [19]. Weikert et al., stated that prior research with a sensitivity above 0.85 accepted more false positive findings, which increase the amount of false positive cases, thereby increasing the radiologist’s workload and time to therapy initiation [18].

False positive cases were more frequent (54% or 20/37) than false negatives, and mostly due to falcine or basal ganglia calcifications, hyperdense dural sinuses and streak artifacts, these results are in line with prior work assessed by Roa et al., effortlessly recognized without

### Table 3

<table>
<thead>
<tr>
<th>False negative ICH cases by AI</th>
<th>False positive ICH cases by AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subarachnoid hemorrhages</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Subdural hemorrhages</td>
<td>33% (2/6)</td>
</tr>
<tr>
<td>Parenchymal hemorrhages</td>
<td>33% (2/6)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>False negative PE cases by AI</th>
<th>False positive PE cases by AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falcine and Basal ganglia calcifications</td>
<td>45% (9/20)</td>
</tr>
<tr>
<td>Beam hardening artefacts</td>
<td>40% (8/20)</td>
</tr>
<tr>
<td>Hyperdense dural sinuses</td>
<td>15% (3/20)</td>
</tr>
</tbody>
</table>
any difficulties by the original reporting radiologist [19]. An example of a false positive ICH case by AI (yellow arrow) is shown in Fig. 3, probably due to the presence of surrounded periventricular white matter hypo-attenuation.

False negatives ICH cases occurred in a limited number of cases (6/337) and were mainly seen in very small hemorrhages or follow-up exams within deteriorating patients. The most commonly false negative cases were sulcal subdural and subarachnoid hemorrhage, predominant in convex brain regions. However, we had one less subtle stroke hemorrhagic transformation (Fig. 2). Although, this missed finding could still be explained by very low HU densities within the lesion itself, luckily it was fast diagnosed by the original reporting radiologist. On one hand, the interpreting radiologist should scrutinize those critic brain regions very carefully. On the other hand, a missed subtle hemorrhage may be inconsequential, it needs to get your attention as well and may indicate further examination.

With high specificity and negative predictive value, the AI tool shows the potential to rule out ICH.

For our PE study we achieved a rather low sensitivity of 0.73 when compared to the study by Weikert et al., who reached a sensitivity level of 0.93 [18]. Firstly, this can be explained by a high prevalence of chronic emboli in our study population (9/19 cases). Secondly, 6 out of 19 patients had marked artifacts (movement, beam hardening and contrast agent-related flow artifacts). After consensus review, we also found 18 false positive cases flagged by the AI tool. According to the reviewers, also mainly due to artifacts, the other cases can be clarified by a combination of artifacts, masquerading pathologies (such as infiltrate, metastasis, pleural effusion, atelectasia, fibrosis) and superposing anatomy (e.g. pulmonary vein, lymph node, hilar soft tissue, bronchus, azygos vein, pulmonary artery bifurcation). Detailed analysis of false negative and false positive PE cases by AI are reported in Table 4. Examples of false negative and false positive PE cases by AI are shown in Fig. 4. Although false positive results can increase workload, these were again easily recognized by the expert readers. Also, an AI solution with a high false-negative rate can be more harmful, especially in outpatients [20-23]. Via Dual-energy CT with associated iodine maps for perfusion defect detection, the expert radiologists easily picked up false negative AI reports. In future updates, the AI tool should implement those iodine maps as well, to decrease false negative results [6,24-26]. Even though AI solution may potentially identify PE and thus assist radiologists, ultimately, we have to weigh the significance of such findings, which is for PE in most cases location-bound. Radiologists will give more significance to identifying central PE compared to missing a very small sub-segmental PE [18,26,27].

Interestingly, while the AI tool processed more PE studies, our data showed the impact of the tool to be more sensitive for ICH (0.84 versus 0.73). Even so, the PE study scored a higher positive predictive value of 0.76 versus 0.61 for ICH. Similar results were achieved regarding specificity (0.94 for ICH and 0.94 for PE) and accuracy (0.93 for ICH and 0.98 for PE). Future studies with adjusted prevalence of each target pathology could provide more insight. To our knowledge, no study has evaluated the diagnostic accuracy of AI technology during real-time radiology work flow in detecting ICH and PE, using a consecutive dataset for each target pathology. Prior studies focused on prototype algorithms which limits their usefulness in clinical setting, while an important strength of this study is that the AI algorithm is commercially available, and has never previously been exposed to images from our department or our CT equipment [20]. Also, the AI technology used in our study was applied to data from two single-energy CTs and one dual-energy CT, suggesting the robustness of AI processing. Roa et al., studied more closely retrospective peer review systems to minimize false negatives in particular, whereby this AI tool could function as a real-time prospective, peer review for radiologists [19]. Earlier research evaluated on rather small data sets with a high percentage of positive cases or
even exclusively positive cases, which does not represent real-time clinical workflow and might influence diagnostic accuracy [17,27-29].

Following limitations of our study merit consideration. We included emergency, in and out patients which differ markedly in diseases and pathology signs. We did not study the underlying or associated pathologies nor characteristics (gender, age, etc.) of our study population. We do not know whether these may have influenced our results. Likewise, we did not calculate the prevalence of our target pathologies at our institution. However, there is a remarkable variety geographically and it is well known that prevalence has a strong influence on PPV and NPV. This can provoke an application site-dependent performance which could lead to future replicability issues. Nevertheless, testing on all consecutive head CTs and CTPAs during a vast time frame at our department, ensured a representative clinical reality regarding the ICH and PE distribution and even the positive and negative cases. All original radiology reports included in our study, were used for making clinical decisions at our institution. As aforementioned, artifacts, sloped and even postoperative studies were included to represent routine radiology work stream. These inclusion and exclusion criteria were established to approximate a standard practice data set and could partially explain our rather lower achieved sensitivity for ICH detection. Since our gold standard was the consensus review, we still may have missed undetected cases. Also, we did not assess the follow-up imaging of patients with missed ICH/PE. Another limitation is that we did not assess the reason
why some studies were not processed by the AI tool, nor did we consider the diagnostic accuracy stratified per scanner model. Future studies with a higher number of cases per scanner and a rejection analysis might provide interesting insights in the performance of AI tools in function of the scanner model and applied scan protocols including image quality and radiation dose. An additional limitation of this study relates to the retrospective methodology in a single center. Lastly, we did not assess any AI solution for the detection of intracranial pathology besides hemorrhage nor pulmonary pathology besides embolism.

Due to the fact that we evaluate the AI tool next to an original reporting radiologist (and not without human primary and secondary review), it is hard to assess the clinical impact of the findings detected by the AI tool. Currently, the tool is integrated into our PACS as an automated triage system with a pop-up window whenever the tool suspects the presence of ICH or PE. In this way, quality increases because abnormalities are brought to our attention immediately. The automated case prioritization ensures that the most urgent patient will be diagnosed first. We did not evaluate the time saving for each patient due to the automated triage system. Future investigation should focus on the added value of worklist prioritization, which will give more information about the clinical impact as well.

Conclusion

Our study demonstrates that, in a diverse clinical setting, an AI solution has the potential to assist radiologists and serve as a real-time clinical adjunct for non-contrast head CT scans to rule out ICH on one hand and for CTPAs to diagnose PE on the other hand. An important fraction of consecutive studies could not be analyzed (22% ICH studies and 10.4% PE studies), however these fractions were reduced to 15.4% for ICH and for CTPAs to diagnose PE on the other hand. An important retrospective methodology in a single center. Lastly, we did not assess an AI solution acting as an adjunct to current real-time radiology workflow as a second reader of non-contrast CT studies. SPIE Medical Imaging, 2019, Proceedings Volume 10949, Medical Imaging 2019: Image Processing; 109493.

References


