Prescreening for Intracranial Hemorrhage on CT Head Scans with an AI-Based Radiology Workflow Triage Tool: An Accuracy Study

Chodakiewitz YG*, Maya MM, Pressman BD
Cedars Sinai Medical Center, Department of Imaging, Los Angeles, CA, USA

ABSTRACT

**Purpose:** To evaluate a recently FDA-approved AI-based radiology workflow triage device for accuracy in identification of intracranial hemorrhage (ICH) on prescreening of real-world CT brain scans.

**Method/Materials:** An AI-based device (“the algorithm”) for ICH detection on CT brain scans was studied at our institution; the algorithm was developed by outside company Aidoc (Tel Aviv, Israel). A retrospective dataset of 533 non-contrast head CT scans was collected from our large urban tertiary academic medical center. Following convention for studies evaluating sensitivities and specificities of imaging computer-aided detection and diagnosis devices, a prevalence-enriched dataset was utilized such that a 50% prevalence of intracranial hemorrhage was obtained. The algorithm was run on the dataset. Cases flagged by the algorithm as positive for ICH were defined as “positive”, and the rest as “negative”. The results were compared to the ground truth, determined by neuroradiologist review of the dataset. Sensitivity and specificity were calculated. Additionally, Negative-Predictive-Value (NPV) and Positive-Predictive-Value (PPV) calculations were made from the prevalence-enriched study data, which enable lower and upper threshold estimates for real-world NPV and PPV, respectively. Metrics were analyzed using a two-sided, exact binomial, 95% confidence-interval.

**Results:** Algorithm sensitivity was 96.2% (CI: 93.2%-98.2%); specificity was 93.3% (CI: 89.6-96.0%). Estimated real-world NPV was determined as at least 96.2% (CI: 93.2%-97.9%), and an estimated upper threshold for PPV was estimated as 93.4% (CI: 90.1%-95.7%).

**Conclusion:** The tested device detects intracranial hemorrhage with high sensitivity and specificity. These findings support the potential utility of using the device to autonomously surveil radiology worklists for studies containing critical findings, triage a busy workflow and ultimately improve patient care in clinically time-sensitive cases.

**Keywords:** Intracranial hemorrhage; CT; Artificial intelligence; Deep learning; Machine learning; Radiology.

INTRODUCTION

Radiology plays a pivotal role in fast and accurate diagnosis of medical conditions with the timely reporting of critical diagnostic results remaining one of the National Patient Safety Goals identified by the Joint Commission [1,2].

However, despite the volume of diagnostic imaging continuously increasing, the number of radiologists remains fairly stagnant [3]. The average number of medical images per minute requiring interpretation increased nearly sevenfold per radiologist from 1999 to 2010, and radiologists consequently struggle to keep up with the increasing workload without compromising quality and accuracy in diagnosis [4]. Specifically, delayed radiological reporting is a noteworthy problem, potentially leading to poorer patient outcomes [5-8].

Currently, except for some parameters of urgency (e.g. STAT), radiologists evaluate cases in a “First In, First Out” (FIFO) methodology, rather than through a triage-optimized queuing workflow system. As a result, scans of patients potentially suffering from a life-threatening condition might not be...
addressed within a critical time window. Previous studies have suggested that a computer-assisted triage system to prioritize radiological case review can potentially benefit radiological workflow. However, available prioritization/triage methods have yet to show a meaningful clinical impact [9-11].

A computer-aided automated evaluation system of incoming studies for acute or critical findings may help optimize the triaging of a busy radiology workflow. If such a system could accurately detect critical findings, it might be used to alert radiologists to studies that warrant more rapid interpretation; such a system offers significant potential to mitigate delays to critical diagnoses. However, prior to any such clinical implementation, such a system first needs to demonstrate sufficient accuracy.

The present study, therefore, seeks to investigate this fundamental accuracy question. Intracranial hemorrhage (ICH) was selected as a model critical-finding on which to focus. A recently developed AI-based radiology workflow triage tool ("the device") which was developed by an outside medical industry company, Aidoc (Tel Aviv, Israel) was investigated for this purpose. The device is a deep-learning-based algorithm trained to detect ICH on CT head scans and notify the radiologist of relevant scans for further evaluation. The device was approved by the FDA for use as a radiology workflow triage tool and appears to be the first device cleared for such a purpose [12,13]. The present study was designed to test the device’s ICH detection accuracy.

MATERIALS AND METHODS

The present study was granted IRB exemption by our institution. There was no outside financial support from this study from Aidoc or other industry sources. Aidoc did provide non-financial support in the form of assistance with technical deployment of the device and assistance with technical analysis of study data, which was always under the overall supervision of our institution’s imaging department. One of the authors sits on the medical advisory board of Aidoc as an unpaid member, and the other two members have no relevant disclosures.

The study involved 4 phases: (1) case selection; (2) device image processing; (3) “ground truth” determination; and (4) endpoint analysis.

Case selection proceeded as follows. An experimenter was designated as “case selector.” The case selector retrospectively reviewed consecutive CT brain reports beginning from January 2017 performed at our institution, a large urban tertiary academic medical center.

The case selector included or excluded cases solely based on information contained within the reports; the case selector was blinded to the associated images. Case selection continued consecutively until pre-determined quantities of each case type (ICH positive/negative) were collected (see section 2.4 for sample size rationale). A prevalence-enriched sample population was collected, following the convention for evaluation of imaging computer-aided detection and diagnosis devices, which does not introduce mathematical bias for sensitivity or specificity determinations [14].

Once the case-set was collected, de-identified case images were uploaded to a secure cloud-based server (see section 2.3 for de-identification protocol). The device subsequently processed each case’s images in the cloud. The device evaluated each case for imaging findings consistent with ICH, scoring each case as either positive or negative for ICH.

Separately, each case was assigned to one of two other experimenters (attending neuroradiologists), designated as “reviewers”, who independently reviewed each case blinded to the device’s score from phase 2 and provided their own positive or negative score. This reviewer score defined the “ground truth” for each case.

After both device scores and ground truths were obtained for all cases, the study was closed and study endpoints were calculated (see section 2.2). Primary endpoints were the quantification of device sensitivity and specificity.

Selection of study population

Case selection for the trial was performed on consecutive cases by the “Case selector” based on the following criteria:

Inclusion Criteria:
• Patients ≥ 18 years of age
• Non-enhanced CT (NECT)
• Head protocol

Endpoints

Each case was scored separately by the device and a reviewer. Data was subsequently arranged in the standard 2 × 2 table (Table 1). Both device sensitivity and specificity were calculated relative to the ground truth.

Table 1: Standard 2 × 2 table representation of study primary endpoints for sensitivity and specificity calculations.

<table>
<thead>
<tr>
<th></th>
<th>Ground truth score-ICH positive</th>
<th>Ground truth score-ICH negative</th>
<th>Totals (AI Screening device)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI Screening device score-ICH positive</td>
<td>255 (True Positive)</td>
<td>18 (False Positive)</td>
<td>273</td>
</tr>
<tr>
<td>AI Screening device score-ICH negative</td>
<td>10 (False Negative)</td>
<td>250 (True Negative)</td>
<td>260</td>
</tr>
<tr>
<td>Totals (Ground truth)</td>
<td>265</td>
<td>268</td>
<td>533</td>
</tr>
</tbody>
</table>
Data de-identification

The DICOM cases were de-identified per the safe harbor de-identification protocol defined by HIPAA. All 18 fields defined by the protocol were removed from the cases, and a unique study ID was assigned to each case.

Statistical methods and determination of sample size

Statistical and analytical plans: The data was summarized in tables listing count and percentage for categorical data. Age of the patients was summarized using the mean and standard deviation.

Determination of sample size: A sample size of 220 positives and 220 negatives provides 90% power to demonstrate that sensitivity and specificity each are greater than 89%, assuming that true sensitivity and specificity are 95%. Sample size computations were done using the Exact Binomial method for comparing one sample to a proportion.

For the purposes of the case set collection and achieving the targeted quantities of case types, we assumed that the case selector’s presumption of case positivity/negativity would closely correlate with the ground truth. However, given that case selection and ground truth determination involved two different methodologies that were also blinded to each other, a margin of error was anticipated. To buffer this risk, targeted sample size was increased for phase 1; a final minimum of 265 cases of each type was therefore targeted. Eventually, a total of 533 total cases were included (265 presumed positives, and 268 presumed negatives).

RESULTS

Our studied dataset involved CT brain cases from a large urban tertiary academic medical center obtained over a 3-month long period (January-March 2017). Demographics of our studied population include age and gender. The mean age of patients whose scans were reviewed in the study was 62.4 years, with standard deviation of 21.3 years; the youngest patient was 21 years old, and the oldest was 97 years old. The gender distribution was 55.5% male (N=296), and 44.5% female (N=237). Race distribution data within the study patient population was unavailable.

A total of 1500 radiological reports from the 3 month period were reviewed by the case selector in the first phase of the study. From that review, cases were consecutively selected for inclusion into the study based on the criteria and methodology described above in the materials/methods section, until the pre-specified quantities of each case type were attained. At that point, the case-selection phase was closed and the remaining cases were automatically excluded. A total of 533 cases were included, such that 265 cases were “suspected positive” and 268 were “suspected negative”.

The 533 cases were then scored separately by both the device and reviewer-neuroradiologists, as detailed above in the materials/methods section. The device processed and scored cases in under 5 minutes.

Of the total 533 cases, neuroradiologist review scored 265 cases as positive for ICH (“Ground truth score-ICH positive”); the remaining 268 cases were negative for ICH (“Ground truth score-ICH negative”). Of the ICH positive cases, 255 cases were correctly flagged by the device (True-Positive Cases). The device failed to flag 10 of the ICH positive cases (False-negative cases). The device incorrectly flagged 18 of the ICH negative cases (False-Positive cases). The device correctly did not flag 250 of the ICH negative cases (True-Negative cases) (Figure 1 and Table 1).

Of the total 533 cases, neuroradiologist review scored 265 cases as positive for ICH (“Ground truth score-ICH positive”); the remaining 268 cases were negative for ICH (“Ground truth score-ICH negative”). Of the ICH positive cases, 255 cases were correctly flagged by the device (True-Positive Cases). The device failed to flag 10 of the ICH positive cases (False-negative cases). The device incorrectly flagged 18 of the ICH negative cases (False-Positive cases). The device correctly did not flag 250 of the ICH negative cases (True-Negative cases) (Figure 1 and Table 1).

Figure 1. Top left, “True Positive” example case, with Al-annotation (arrow within circle) of suspected ICH finding. Top right, “True Negative” example case. Bottom left, “False Positive” example, with Al-annotation of Al-suspected ICH finding determined not to represent hemorrhage during ground truth determination. While this case was not “normal”, as there is evidence of ischemic infarction in the left parietal lobe suggested by sulcal effacement and loss of grey-white matter differentiation, there was no hemorrhage present. Bottom right, example case of “False Negative” with arrowhead indicating a small area of ICH, which was not flagged by the device.

The device therefore demonstrated a case sensitivity of 96.2% (CI: 93.2%-98.2%) and a case specificity of 93.3% (CI: 89.6%-96.0%) (Table 2).

Table 2: Sensitivity and specificity with associated two-sided 95% confidence limits.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Lower confidence limit</th>
<th>Upper confidence limit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>96.20%</td>
<td>93.20%</td>
<td>98.20%</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Because we used a prevalence-enriched sample and the true prevalence of ICH is unknown, a calculation of Positive-Predictive Value (PPV) and Negative Predictive Value (NPV) for ICH could not be performed. However, based on the assumption that the true prevalence is much lower than the artificially high prevalence of ICH in our sample (which was 49.7%), these data permit estimating an upper threshold for true PPV and a lower threshold for true NPV. The results thus yield a minimum NPV of 96.2% (CI: 93.2%-97.9%) and a maximum PPV of 93.4% (CI: 90.1%-95.7%).

DISCUSSION

Physicians are increasingly relying on imaging for diagnosis and guidance of treatment. Consequently, radiologist workload has grown significantly in recent years, resulting in a constant tension between efficiency and quality [4,15]. Increasing workload also results in a significant worklist backlog such that one of the main reasons for delayed communication of critical imaging findings is the delay between image and radiologist review [16].

Currently, radiologists lack technological tools to assist in prioritization of their workflow based on image data content rather than external image characteristics (e.g. acquisition time, origin in hospital/ER/outpatient, ‘STAT’ order). Regarding ICH specifically, there is no established tool at present which screens head CTs for ICH prior to the radiologist’s interpretation in order to indicate those which may need prioritization.

Interestingly, a recent study suggested that time to the radiological interpretation of acute neurological events can be importantly decreased by employing an AI surveillance and triage system [11]. However, in that study, the authors point out that their studied system’s diagnostic accuracy was relatively low, potentially limiting its clinical relevance.

Additionally, the FDA recently approved marketability of a computer-aided triage software for detection of ischemic stroke and direct notification to neurovascular specialists upon software detection of an abnormality, bypassing the typical delays associated with radiologist interpretation of CT images prior to involving the neurovascular specialist [17].

In the present study, we evaluated the accuracy of an AI-based device in the detection of ICH on non-contrast head CTs, with the eventual goal being the optimization of radiology workflow by prioritizing for radiologist review cases with the critical finding of suspected ICH. Overall, the device demonstrated a high-level of accuracy, with 96.2% sensitivity and 93.3% specificity, supporting the device’s potential effectiveness in radiological workflow triage. Additionally, although prevalence was artificially increased in our dataset, our results imply an NPV of at least 96.2% for ICH detection; this is because the true prevalence of ICH is assumed to be much lower under real-world conditions, which implies an even higher true NPV. The calculated lower threshold for NPV remains high, offering a strong degree of confidence and reassurance that when a study is not flagged by the device for ICH that it is indeed negative for ICH.

In contrast to the estimated NPV, the estimate of PPV is likely without much meaning, offering only an upper threshold for the true PPV; the true PPV may lie anywhere from 0% up to the calculated upper limit of 93.4%. Therefore, in the absence of the known true prevalence of ICH at our institution and/or in the greater population, our inability to more precisely characterize the true PPV remains a limitation of this study.

We further note several additional limitations regarding the present study. For one, while more than one neuroradiologist participated in establishing the study’s ground truth, each individual case was reviewed by a single neuroradiologist. It is therefore conceivable that the ground truth for an individual case could be flawed, particularly in cases with subtle or questionable findings. Blinded double-reading for the establishment of ground truth may alleviate this limitation in future studies.

Additionally, the present study focuses exclusively on ICH detection. While ICH is a common and critical diagnosis, many other such diagnoses exist [18]. It remains unclear whether we can extrapolate algorithm performance in ICH to predict performance for other diagnoses, therefore placing a limitation of our results regarding the generalizability of AI-based detection of imaging findings. Furthermore, because current machine learning technology requires a separate algorithm to detect any given pathology, a comprehensive triage tool covering multiple commonly-encountered diagnoses would necessitate a suite of algorithms deployed in parallel [11,19]. The incorporation of additional tools within a comprehensive triage system and its seamless integration into the radiologist’s workflow remains a challenge. Further development and investigations are ongoing, and no determination regarding the value of such a system in the radiologist workflow can yet be made.

The present study also considered the detection accuracy only regarding the general category of intracranial hemorrhage, without specific attention to whether there is differential detection accuracy based on the subtype of hemorrhage, for instance, epidural versus subdural versus subarachnoid versus parenchymal versus intraventricular. This differential accuracy question may be further elucidated in subsequent studies may elucidate this question of differential accuracy [20-24].

Finally, while this study demonstrates excellent algorithm performance in detecting ICH, it does not evaluate whether triage/prioritization yields any downstream benefits such as decreased report turnaround-time or more rapid communication of a critical finding. As suggested by Titano et al. [11], we assume such downstream benefits would be significant. However, additional studies will be required to formally test this assumption.

CONCLUSION

To conclude, our findings indicate that the AI-based device can accurately screen CT head scans for ICH. We anticipate that the device can be deployed to prioritize cases with critical findings as...
they arise on a radiologist’s worklist. We furthermore assume that such prioritization could translate to downstream improvements in result notification, and ultimately more prompt treatment, potentially leading to improved clinical outcomes.

Additional ancillary benefits include raising the reviewer’s index of suspicion when reviewing scans with subtle or borderline findings flagged as potentially positive. Furthermore, it is conceivable that on a workflow level, by prompting a sense of urgency on an interpreting radiologist regarding a flagged case, or reassurance in negative cases, the device may improve overall reader efficiency. This is a question we are actively investigating in another ongoing study, where preliminary data supports that hypothesis.

Finally, the device’s very high NPV could be applied in scenarios in which a radiologist is not immediately available to review a scan, but where a rapid result with high confidence is required. For example, in the setting of ischemic stroke, where thrombolysis is under consideration. Further study is warranted to assess these potential clinical benefits.

REFERENCES