Real-World Performance of Deep-Learning-based Automated Detection System for Intracranial Hemorrhage

Sehyo Yune, MD, MPH, MBA, Massachusetts General Hospital; Hyunkwang Lee, MS; Stuart R. Pomerantz, MD; Javier M. Romero, MD; Shahmir Kamalian, MD; Ramon G. Gonzalez, MD; Michael H. Lev, MD; Synho Do, PhD

Introduction
Most of currently published deep learning studies in medical image analysis report their performance using carefully selected data. To use such tools in the clinical practice, however, it is critical to know how they work with the real-world data. Here, we evaluated the applicability of our ICH detection system in the clinical setting by comparing the model performance on the real-world cases to the performance on the selected dataset.

Methods
We previously trained and validated the deep learning system for ICH detection using a total of 904 cases of 5mm, non-contrast head CT scans - 625 cases with ICH and 279 cases without ICH, labeled by six board-certified neuroradiologists. An additional, non-overlapping set of 200 cases (100 ICH-positive and 100 ICH-negative). Both sets excluded cases with history of brain surgery, intracranial tumor, intracranial device placement, skull fracture, or cerebral infarct. For performance evaluation in the real-world setting, all non-contrast head CT scans acquired at a single emergency department for three months from September to November 2017 were obtained. Collected were 2,606 consecutive cases, including 163 cases with ICH identified based on clinical report. Two sub-specialty trained neuroradiologists reviewed the result to describe the real-world performance of the deep-learning model.

Results
Area under the receiver operating curve (AUC) was 0.993 for detecting the presence of ICH on the 200 selected cases with sensitivity of 98.0% and specificity of 95.0%. The same model achieved AUC of 0.834 on the 2,606 real-world cases with sensitivity of 87.1%, specificity of 58.3%, positive predictive value of 11.8%, and negative predictive value of 98.5% at the high sensitivity operating point. Review of all 21 false-negative cases by radiologists revealed that these cases included 2 acute bleeding cases, 11 equivocal cases, and 8 cases of report-hedging with which no acute bleeding was found from the images. Review of randomly selected 204 false-positive cases found 7 cases of acute bleeding not clearly described in the reports, 110 cases with motion artifacts, 91 cases of falsely detected subdural hemorrhage on interhemispheric fissure or tentorium, 27 cases of calcification, and 105 other pathologies such as extracranial hematoma, brain mass, and encephalomalacia.

Discussion
The performance of deep-learning based systems should be evaluated on the real-world data before being used in the clinical practice to assist clinicians in interpreting the automated output. As we developed this model as the safety tool for inexperienced radiologists and non-radiologist professionals, we pre-defined the threshold for ICH detection to have high sensitivity. When applied to the real-world data, the negative predictive value was excellent at the cost of positive predictive value and specificity. By understanding the nature of incorrect prediction in the real-world data, the model can be improved through constant feedback from expert radiologists.

Conclusion
The performance of deep-learning-based ICH detection model was significantly different when tested on real-world data compared to when tested on the selected data that excluded potentially confusing cases. However, the model showed high negative predictive value in the real-world data, which was the primary intention when designing the tool.

Keywords
deep learning; machine learning; artificial intelligence; clinical application; intracranial hemorrhage
(a) Result on the 200 in-house test dataset

(b) Result on the real-world test dataset