AJNR

Reply:

P.D. Chang and D.S. Chow

AJNR Am J Neuroradiol published online 29 November 2018
http://www.ajnr.org/content/early/2018/11/29/ajnr.A5913

Published November 29, 2018 as 10.3174/ajnr.A5913

REPLY:

Thank you very much for your keen and insightful comments. We furthermore congratulate your team on both the incredibly large dataset and impressive results across various head CT findings in the reference that has been provided.[1]

Looking back on our experimental data, we saw no difference in algorithm performance between the training dataset and each respective cross-validation fold across all hemorrhage sizes. Instead of overfitting, the slight relative drop in algorithm performance for small-volume (<5 mL) hemorrhage likely relates to the inherent difficulty in identifying subtle CT findings, as well as a degree of interpreter subjectivity in differentiating microhemorrhage from punctate high-density mimics (Fig 1*C* in our original article), especially without corresponding comparison studies, advanced imaging, or clinical history that may otherwise be available in routine practice. Furthermore, no statistically significant differences in performance were noted between cross-validation and test datasets, while acknowledging the overall low number of punctate ($n = 4$, <0.1 mL) and small ($n = 11$, <5 mL) test set hemorrhages.

However, while no significant overfitting was observed in our internal dataset, we agree that generalization of deep-learning algorithms remains an unsolved challenge for the Artificial Intelligence (AI) medical imaging community. To some extent, this relates to difficultly in the curation of large, diverse datasets shared among multiple institutions; in the United States, a number of logistic barriers and concerns for robust patient anonymization are key bottlenecks. To this end, we applaud the impressive curation effort and open-source release of data in the provided reference.[1]

However, a large dataset alone does not guarantee generalizability. For true clinical relevance and widespread adoption, an AI tool must be flexible enough to generalize across use cases and clinical contexts. For example, the referenced dataset and corresponding trained algorithms[1] do not include any postoperative CT scans, patients with postsurgical changes or hardware, or pediatric patients. While this exclusion may make sense in certain clinical contexts (eg, community hospitals or outpatient clinics), these exclusion criteria account for a significant population at most large academic centers in the United States; algorithms trained using such a dataset may thus fail to generalize against hardware streak artifacts or other high-density mimics that are commonly seen in such a setting. Conversely, an algorithm optimized for high disease prevalence and rare entities seen at an academic center may produce too many false-positives in a more routine, healthy population.

This issue of generalizability and a number of other key practical considerations remain key unsolved problems that must be addressed before the potential of medical deep learning is realized on a large scale. To this end, we look forward to working alongside your capable team and the radiology deep-learning community across the world to identify solutions to these problems and together build the next generation of AI-enabled tools.

http://dx.doi.org/10.3174/ajnr.A5913

REFERENCE

1. Chilamkurthy S, Ghosh R, Tanamala S, et al. **Development and validation of deep learning algorithms for detection of critical findings in head CT scans.** March 13, 2018. https://arxiv.org/pdf/1803.05854.pdf. Accessed August 5, 2018

P.D. Chang
D.S. Chow
Division of Neuroradiology
Department of Radiological Sciences
Center for Artificial Intelligence in Diagnostic Medicine
University of California, Irvine Health System
Irvine, California