Radiology

What Can We Learn from the RSNA Pediatric Bone Age Machine Learning Challenge?

Eliot L. Siegel, MD

From the Departments of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, 22 S Greene St, Baltimore, MD 21201; and VA Maryland Healthcare System, Baltimore, Md. Received November 19, 2018; accepted November 19, 2018. **Address correspondence to** the author (e-mail: *esiegel@umaryland.edu*).

Conflicts of interest are listed at the end of this article.

See also the article by Halabi et al in this issue.

Radiology 2019; 290:504–505 • https://doi.org/10.1148/radiol.2018182657 • Content code: IN •

Franklin D. Roosevelt said "Competition has been shown to be useful up to a certain point and no further, but cooperation, which is the thing we must strive for today, begins where competition leaves off."

Competition is an extraordinarily powerful driving force for innovation, improvement, and success. As far back as 776 BC, the Olympic Games have inspired athletes to ever greater levels of performance. Evolutionary biologists believe that competitiveness has coevolved in humans along with the struggle for survival (1).

The machine learning community in particular has taken advantage of the power of competition by issuing challenges, such as the annual ImageNet Large Scale Visual Recognition Challenge, which uses the more than 14 million hand-annotated color photographs from the ImageNet database. Kaggle, a company recently acquired by Google, has sponsored hundreds of machine learning competitions involving thousands of participants who frequently publish their winning strategies in academic journals and are awarded prize money in return for "a worldwide, perpetual, irrevocable and royalty-free license."

The Radiological Society of North America (RSNA) machine learning committee launched the RSNA Pediatric Bone Age Machine Learning Challenge and publicly recognized the teams that created the best-performing algorithms at the 2017 RSNA Annual Meeting. In this issue of *Radiology*, Halabi et al (2) describe the competition and its aims to demonstrate an application of machine learning in medical imaging, to promote "collaboration to catalyze artificial intelligence model creation," and to "identify innovators in medical imaging." The RSNA challenge used medical images, clinical reports, and the evaluation of four pediatric radiologists from data already collected in a study by Larson et al (3) that was designed to demonstrate the application of deep learning to predict the bone age of a child by using a radiograph of the hand.

An annotated training data set consisting of 12 611 pediatric hand radiographs was made available to competitors. The annotation included an expert consensus estimate of bone age and sex. Most contestants used a specific type of machine learning known as *deep learning*, which uses a convolutional neural network. The training data set was supplemented by an annotated validation data set of 1425 images that was designated to allow competitors to assess the accuracy of the algorithms that they developed from the training images. Finally, a test data set of 200 images without annotation was used to compare the accuracy of the algorithms developed by each team. There were 48 unique teams for a total of 105 contestants.

Ground truth in the study (3) was established with inputs from four radiologists, including a second interpretation by one of the radiologists 1 year later, as well as the original report. All of these interpretations used the Greulich and Pyle atlas matching method (4) and were combined by using a weighting factor that took into account the estimated relative accuracy of the radiologists, as judged by the other interpretations.

The individual performance of pediatric radiologists was judged by determining their mean absolute difference (MAD) from a weighted consensus score, with MADs ranging from 5 to 7 months for the four readers. In comparison, the 10 best machine learning teams achieved MADs with a narrow range of 4.265–4.907 months, with the three best entries separated by a remarkably thin margin of 3.5 days (2). All of the top 10 entries outperformed the machine learning model described in the original study (3) from which the data sets were derived and which had an MAD of slightly more than 6 months. When the results from the second- and fourth-place teams were combined, MAD improved to 4.00.

There were multiple common themes in the machine learning approaches used by the top five challenge winners, and all except the fourth-place team used deep learning. One theme was the use of data augmentation, in which data sets are increased in size by adding variants of images to the data set. These include flipping the images horizontally, vertically, or obliquely; applying image filters; and adding noise. Another theme was preprocessing, which could include a task to break up the hand images into subcomponents, such as fingers, metacarpals, and joint spaces. Finally, many teams created multiple algorithms that they teamed together to achieve a higher accuracy than could be achieved with any one algorithm alone. Multiple algorithms in combination in machine learning are termed *ensembles*.

There are many additional lessons from the RSNA Machine Learning Challenge. An important point that is usually ignored by the media and often by authors of machine learning articles is the fact that the radiologists typically are given a subtly different task than the deep learning or artificial intelligence algorithm. The clinically oriented task for the radiologists who set the reference

standard was to simply determine the best match for a given radiograph by using the atlas. This differs from the task for the machine learning algorithm, which was to guess the expected mean weighted score of the radiologists for a given radiograph. For example, if a radiologist knows that Alice the radiologist usually guesses high, he might change the prediction of the mean score in that direction. However, the radiologists are not asked to predict the MAD and are thus at a disadvantage when they are compared with the machine learning algorithms. Another challenge that has been well described in the literature (4) is the substantial difference in bone age estimation between the atlas matching method of Greulich and Pyle and other methods, such as the Tanner-Whitehouse method. This is partly due to the fact that Greulich and Pyle studied American children of high socioeconomic status in the 1940s, while Tanner and Whitehouse studied Scottish children of low socioeconomic status in the 1950s. It is probable that a database derived from an ethnically diverse population in 2019 of children whose disease likely manifests earlier in puberty would also differ greatly from those older less diverse databases. This raises the question of the optimal approach for a data set and validation for the development of a commercial version of the bone age software for use in the United States and other parts of the world.

The extremely small margin between performance of the five best methods raises several interesting questions: Have we reached the limitations of the granularity of the atlas method itself for bone age or have we reached a plateau in deep learning in diagnostic imaging where incremental advances in the technique provide minimal improvement on performance? Or, as Somers asks in the MIT Technology Review, is artificial intelligence riding a one-trick pony (5)? Will methods emerge in the next 5–20 years that would be able to substantially improve on the performance from the data set used in the RSNA challenge? Should we aggressively encourage the development of novel approaches to create more effective and efficient algorithms from images? Will revolutionary breakthroughs in hardware and software, such as quantum computing, which holds the promise of speeding up computation by many orders of magnitude, have any effect on image analysis?

In conclusion, the inaugural RSNA Machine Learning Challenge was successful in meeting its goals to demonstrate machine learning in medical imaging, catalyze model creation, and recognize innovators. The selection of pediatric bone age determination worked very well on multiple levels to achieve those aims and to demonstrate the potential to advance research in machine learning by sharing a common data set and goal. This task is particularly well suited to machine learning because of the relatively well-defined nature of the quantitative assessment of bone age and relative consistency and simplicity of the anatomy of the digital radiographs of the hand. Consequently, the use of machine learning to determine bone age of a pediatric hand radiograph has been the subject of hundreds of research papers over the past 20 years. It is also a relatively tedious, repetitive, and time-consuming job from a clinical perspective that makes it a good candidate for clinical implementation. Finally, and most importantly, it provides a compelling example of the research potential associated with the sharing of raw data in a published article with the research community to allow novel and innovative ideas and incremental improvements. Underscoring this was the improved performance achieved when the authors combined the second-place (deep learning) and fourth-place (conventional machine learning) teams that resulted in an accuracy that surpassed that of the first-place team. This sharing of raw data is commonly described in other specialty journals, such as the *Journal of the Optical Society of America*, which encourages authors to upload data sets to their portal as supplemental materials. This could serve as an excellent model for the diagnostic imaging community's journals and their authors and could provide a way to make the transition from competitions, such as those described by Halabi et al, to an enduring culture in which researchers facilitate innovation and creativity in their colleagues by sharing their work. As President Roosevelt aptly observed, cooperation begins where competition leaves off.

Disclosures of Conflicts of Interest: **E.L.S.** disclosed no relevant relationships.

References

1. Nowak MA. Five rules for the evolution of cooperation. Science 2006;314(5805):1560–1563.
2. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology 2019;290:498–503.
3. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287(1):313–322.
4. Bull RK, Edwards PD, Kemp PM, Fry S, Hughes IA. Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. Arch Dis Child 1999;81(2):172–173.
5. Somers J. Is AI riding a one-trick pony? MIT Technology Review. September 29, 2017.