Circulation

EDITORIAL

Trusting Magic

Interpretability of Predictions From Machine Learning Algorithms

Articles, see p 1274 and p 1287

Michael A. Rosenberg, MD

In his third of 3 laws, the late science fiction writer Sir Arthur Clarke once surmised that "any sufficiently advanced technology is indistinguishable from magic."[1] To those unfamiliar with the methods of machine learning (ML), it seems almost magical to see a computer interpret an ECG—a technique that, for a human, requires many years of training, as well as continued practice—in less than a second. Yet, computers are the furthest thing from magical. Computers are machines—machines that execute a script exactly as it is written by a human. They are machines that do not get bored, hungry, distracted, or sick and are more than capable of doing the same thing over and over again. One requires little imagination to see how a computer that is trained to replicate human skills in medical diagnosis could be useful in today's fast-paced, high-volume world of health care. Yet, before we place our lives and livelihoods in the hands of a computer algorithm, how do we know we can trust it? What is the process for obtaining trust in a model?

In this week's issue of *Circulation,* 2 exciting papers are presented that describe the magic of ML applied to ECG interpretation. Giudicessi et al developed and verified an algorithm to predict the QT interval at the same accuracy of an expert-interpreted 12-lead ECG using only 2 leads.[2] Raghunath et al developed and tested a model capable of predicting 1-year incident atrial fibrillation (AF) as well as AF-related strokes using a sinus rhythm ECG.[3] Both approaches use a ML method called deep learning[4] that involves multilayered convolutional neural networks applied to the pixels of an image to predict the outcome. Both report impressive accuracy of predictions on held-out testing samples, and both offer enormous potential to improve the efficiency of health care delivery. However, the 2 papers approach the application of ML to ECG interpretation from nearly opposite directions: one seeks to predict a well-known ECG parameter of future risk with greater accuracy, and the other attempts to predict future risk agnostic to any previously described ECG parameters—a contrast worth examining in more detail.

Although Giudicessi et al describe it as an example of artificial intelligence, the model the authors developed and reported performs quite a straightforward task: predicting a measurement, the heart rate–corrected QT interval, which is already provided by most computer-generated ECG programs. The caveat, of course, is that this model provides an accurate measure that is based on only 2 ECG leads, rather than requiring a 12-lead ECG and concordance with expert human interpretation. This innovation is powerful in that many wearable and mobile telemetry monitoring systems are based on 1- or 2-lead measurements, and thus integration with this algorithm could allow remote monitoring of patients at risk for QT prolongation. It is important to note that the authors demonstrated that the algorithm was most accurate at identifying QTc intervals >500 ms, where

Key Words: Editorials ■ atrial fibrillation ■ electrocardiography ■ machine learning ■ sudden cardiac death ■ Torsades de pointes

https://www.ahajournals.org/journal/circ

risk of a lethal arrhythmia such as torsade de pointes is greatest, indicating that the custom loss function applied to train the model—a composite of linear regression and binned integer classification—was well-crafted to perform the task at hand. However, to state that this ML algorithm is a clear demonstration of the power of deep learning perhaps ignores the potential for even greater insights. The QT interval itself is not an outcome, but rather a biomarker for which perturbation, especially prolongation, is a well-known risk factor for risk of torsade de pointes and sudden cardiac death. Although an algorithm that can predict the QT interval accurately is of great use in clinical management, even more enlightening is an algorithm that can predict torsade de pointes itself. Would such an algorithm confirm that QT length itself was most important, or perhaps identify a specific T wave morphology that can predict increased risk independent of duration? Or perhaps we are asking too much of a computer algorithm to uncover deeper biology, and we should be satisfied with the fact that this program will provide accurate QT interval measurements at a scale no human could achieve.

In contrast, the approach taken by Raghunath et al was to predict AF, an outcome for which, unlike the QT interval and torsade de pointes, there is no well-established prognostic ECG metric. Leveraging the power of deep learning, the authors demonstrated that a model applied to 12-lead ECG images was capable of predicting future AF within a year, with estimates on par with 2 of the strongest clinical risk factors for AF, age and sex. The authors take the intriguing approach of examining the predicted risk of AF as one would study a novel biomarker—demonstrating the relative hazard above and below a dichotomized level of prediction, modeling the time-to-event using Kaplan–Meier curves, and examining interactions with other known clinical risk factors. The results are indeed impressive, except for 1 problem: We have no idea what the model is using to make these predictions. The model could be identifying some novel pattern in the P wave that reflects abnormal atrial signal propagation that increases risk of AF…or it could be picking up something as trivial as a lower resting heart rate, which could reflect medication already being used to treat previously undiagnosed or unrecognized AF. Interpretability is a challenging topic in model development, and one can argue that even in the inductive realm of basic science, biological mechanisms have a way of adapting to the stochastic results of clinical trials—or, put another way, the human brain is a black box and we seem to be fine trusting its decisions. However, it is not a given that we must accept that deep learning models are impenetrable black boxes; there are existing methods to examine the parts of convolutional neural networks that are contributing to a given prediction. For example, saliency maps allow one to directly map the weights of the network back onto an image, and visualize the part of the image that was used in the prediction.[5] Ablation approaches, such as Local Interpretable Model-agnostic Explanations,[6] systematically remove parts of the model and examine the effect on predictive accuracy in order to identify which information is behind the prediction. No method is perfect, but if we are proposing to use a prediction model to dictate empirical use of anticoagulation, we would like to have some understanding of why that prediction is being made.

As a final point, it bears mention that the 2 approaches presented in this week's issue are applying the most basic form of ML called supervised learning, which in essence only predicts a label that has already been assigned. There are ML methods that are far more sophisticated and dynamic to the point of being nearly indistinguishable from magic. Unsupervised learning approaches, such as generative adversarial networks and variational autoencoders,[7,8] can learn the deeper structure of data, beyond what can be perceived by humans, in order to seamlessly modify a picture, change a person's voice, or create an entire database of pictures of nonexistent, synthetic people. Reinforcement learning algorithms,[9] which are arguably the true manifestation of artificial intelligence, can learn to identify the optimum action to take in any situation to achieve a long-term goal and thereby achieve suprahuman performance, which has been demonstrated in video games[10] and the board game *Go*.[11,12] It is certain that we have only skimmed the surface of what is possible using computers and large amounts of data to guide clinical care. There may be a future in which computers make clinical decisions, provide tailored explanation and education to patients, perform procedures, and follow up on vital data being generated in the patient's home 24 hours a day, 7 days a week, without fatigue, burnout, or error. Books have been written and companies have been founded on such promises. However, before we can step into that brave new world of artificial intelligence, we must identify ways to interpret, understand, and trust the magic behind these incredibly accurate and efficient prediction algorithms. In that respect, we are still at the very beginning.

ARTICLE INFORMATION

Correspondence

Michael Rosenberg, MD, University of Colorado Anschutz Medical Campus, 13001 E 17th Avenue, E5315, Aurora, CO 80045. Email michael.a.rosenberg@cuanschutz.edu

Affiliation

Cardiac Electrophysiology Section, Division of Cardiology; and Colorado Center for Personalized Medicine, Anschutz Medical Campus, University of Colorado School of Medicine, Aurora.

Circulation. 2021;143:1299–1301. DOI: 10.1161/CIRCULATIONAHA.121.053733

Sources of Funding

Dr Rosenberg is funded by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute (R01 HL146824, K23HL127296) and Google, Inc.

Disclosures

None.

REFERENCES

1. Clarke AC. Hazards of prophecy: the failure of imagination. *Profiles of the Future: An Enquiry into the Limits of the Possible.* Harper & Row; 1962.
2. Giudicessi JR, Schram M, Bos JM, Galloway CD, Shreibati JB, Johnson PW, Carter RE, Disrud LW, Kleiman R, Attia ZI, et al. Artificial intelligence-enabled assessment of the heart rate corrected QT interval using a mobile electrocardiogram device. *Circulation.* 2021;143:1274–1286. doi: 10.1161/CIRCULATIONAHA.120.050231
3. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, vanMaanen DP, Hartzel DN, Ruhl JA, Lagerman BF, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation–related stroke. *Circulation.* 2021;143:1287–1298. doi: 10.1161/CIRCULATIONAHA.120.047829
4. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436–444. doi: 10.1038/nature14539
5. Philbrick KA, Yoshida K, Inoue D, Akkus Z, Kline TL, Weston AD, Korfiatis P, Takahashi N, Erickson BJ. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. *AJR Am J Roentgenol.* 2018;211:1184–1193. doi: 10.2214/AJR.18.20331
6. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery; 2016:1135–1144. doi: 10.1145/2939672.2939778
7. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. In: *Advances in Neural Information Processing Systems (NIPS 2014).* Neural Information Processing Systems Foundation; 2014:2672–2680.
8. Kingma DP, Welling M. Auto-encoding variational Bayes. In: *Conference Proceedings: Papers Accepted to the International Conference on Learning Representations (ICLR) 2014.* Accessed December 12, 2019. https://arxiv.org/abs/1312.6114
9. Sutton RS, AG B. *Reinforcement Learning.* 2nd ed. Cambridge: MIT Press; 2018.
10. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. *Nature.* 2015;518:529–533. doi: 10.1038/nature14236
11. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. Mastering the game of Go without human knowledge. *Nature.* 2017;550:354–359. doi: 10.1038/nature24270
12. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. Mastering the game of Go with deep neural networks and tree search. *Nature.* 2016;529:484–489. doi: 10.1038/nature16961