Machine Learning in Medicine

THE AUTHORS REPLY: Alexander and Kolodny question our conclusion that many providers have stopped prescribing opioids altogether. We reached this conclusion on the basis of the stability of the provider network during the sample period. Furthermore, our conclusion remains valid in analysis of data from the same providers over time.

Although we agree (and note in the article) that we did not have sufficient clinical information to determine whether a given prescription was appropriate, CDC guidelines help to identify prescriptions that are more likely to be inappropriate. We did not imply that all other prescriptions were appropriate.

Prescription opioids have played an important role in the opioid crisis. However, we believe that the causes are complex and that other factors, such as heroin and fentanyl, have also played a role.

Finally, we agree that the serious risks associated with opioids occur with use as directed, although the risks increase with higher doses and durations of opioid treatment.[1,2] At the same time, there are medical contexts in which opioid use may be indicated (especially when preferred therapies have failed), although recent evidence suggests that the list is growing ever smaller.[3]

Wenjia Zhu, Ph.D.
Harvard Medical School
Boston, MA

Tisamarie B. Sherry, M.D., Ph.D.
RAND
Boston, MA

Nicole Maestas, M.P.P., Ph.D.
Harvard Medical School
Boston, MA
maestas@hcp.med.harvard.edu

Since publication of their article, the authors report no further potential conflict of interest.

1. Barnett ML, Olenski AR, Jena AB. Opioid-prescribing patterns of emergency physicians and risk of long-term use. N Engl J Med 2017;376:663-73.
2. Brat GA, Agniel D, Beam A, et al. Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study. BMJ 2018;360:j5790.
3. Krebs EE, Gravely A, Nugent S, et al. Effect of opioid vs nonopioid medications on pain-related function in patients with chronic back pain or hip or knee osteoarthritis pain: the SPACE randomized clinical trial. JAMA 2018;319:872-82.
DOI: 10.1056/NEJMc1905100

Machine Learning in Medicine

TO THE EDITOR: Rajkomar and colleagues (April 4 issue)[1] summarize the advantage of machine learning for medical predictive analytics over traditional statistical methods. We agree that there is no clear distinction between the two types of algorithms but find the discussion of their differences to be caricatural. They argue that use of statistical algorithms would be limited to simple problems based on a limited set of curated and standardized predictors. For complicated problems that involve a large number of noisy and heterogeneous predictors, machine learning would be preferred. Machine learning indeed requires large sample sizes, but it is unclear how this will yield accurate predictions regarding highly noisy data, such as electronic health records (EHRs). Sample size does not solve fundamental data problems. On the contrary, machine learning may not outperform traditional statistical models when the "signal-to-noise" ratio is low.[2-4] We therefore need a better understanding of when different algorithms have maximal value. We call for external validation studies by independent researchers in order to understand model generalizability to new data and different environments. Although such studies are scant,[5] they can inform society on the strengths and weaknesses of medical predictive analytics.

Ben Van Calster, Ph.D.
KU Leuven
Leuven, Belgium
ben.vancalster@kuleuven.be

Laure Wynants, Ph.D.
Maastricht University
Maastricht, the Netherlands

No potential conflict of interest relevant to this letter was reported.

1. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-58.
2. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17:2501-8.
3. Hand DJ. Classifier technology and the illusion of progress. Stat Sci 2006;1:1-14.
4. Goldstein BA, Pomann GM, Winkelmayer WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. Stat Med 2017;36:2750-63.
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
DOI: 10.1056/NEJMc1906060

TO THE EDITOR: The article by Rajkomar and colleagues provides a thorough overview of machine

learning and its potential to transform health care. They provide impressive examples of machine-learning algorithms that correctly identify abnormal imaging findings with an accuracy similar to that of a specialist physician.[1] Image-based problems lend themselves well to prediction because the totality of the data is contained within the individual pixels. However, there are important caveats to applying neural networks.

Neural networks can be fooled by small alterations in image orientation or positioning.[2] Alcorn and colleagues showed how a deep neural network can achieve near-perfect accuracy on correctly identifying basic objects (e.g., a fire truck), but when those same objects are shown with slight changes in positioning, it incorrectly identifies the object and, disturbingly, does so with a high level of confidence.[2] Azulay and Weiss found that changes in even a few pixels can drastically affect the ability of neural networks to correctly identify images.[3] Quality control and preprocessing of images will be critical in the real world, where seemingly insignificant changes could have substantial implications for patient care.

Mike Fralick, M.D.
Errol Colak, M.D.
Muhammad Mamdani, Pharm.D.

Li Ka Shing Centre for Healthcare Analytics Research and Training
Toronto, ON, Canada
mike.fralick@mail.utoronto.ca

No potential conflict of interest relevant to this letter was reported.

1. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392:2388-96.
2. Alcorn MA, Li Q, Gong Z, et al. Strike (with) a pose: neural networks are easily fooled by strange poses of familiar objects. Ithaca, NY: Cornell University arXiv, April 18, 2019 (https://arxiv.org/abs/1811.11553).
3. Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations? Ithaca, NY: Cornell University arXiv, February 18, 2019 (https://arxiv.org/abs/1805.12177).

DOI: 10.1056/NEJMc1906060

THE AUTHORS REPLY: Van Calster and Wynants raise common concerns about using machine-learning models rather than traditional statistical models. As emphasized in our review article, the selection of a model should be tailored to the clinical scenario, and we agree that it is often preferable to use a simple model if it is sufficiently accurate for a particular application. However, many clinical prediction tasks are hard both because many potential variables and interactions among them need to be modeled and because the data encountered in clinical care are inherently noisy (e.g., an elevated potassium level from a hemolyzed sample, entry of an erroneous vital sign, or idiosyncratic abbreviations in a medical note). Modern machine-learning models excel at handling those issues, and there are compelling findings that accurate predictions across multiple clinical domains can be obtained with unharmonized and uncurated EHR data.[1]

We agree that validation of models is important, and we would like to address a common confusion about this point. Validating a model using EHR data from one site on another site is possible if the data are harmonized. However, most EHR data are not harmonized, particularly when the full electronic record is considered, which includes local customizations, abbreviations, notes, and more. This makes harmonization a difficult and time-consuming task owing to a lack of technical and semantic interoperability. Because of decreasing costs of computing power, it may be preferable to retrain models at new hospital sites; this would allow models to incorporate site-specific signals and minimize the overhead for performing and maintaining harmonization.

Fralick et al. caution against using neural networks because some networks may be brittle to slight changes in the features in an input through positioning, pixel changes, or preprocessing. We agree that before models are deployed, they must be tested for robustness to the quality of data and images and for how they handle edge cases. After a model is deployed, we advocate for monitored deployment that systematically evaluates whether live data are consistent with the distribution of data seen during training, that model performance remains at expected levels, and, most important, that patient care is being improved by its use.[2]

Alvin Rajkomar, M.D.
Jeffrey Dean, Ph.D.

Google
Mountain View, CA

Isaac Kohane, M.D., Ph.D.

Harvard Medical School
Boston, MA
isaac_kohane@harvard.edu

Since publication of their article, the authors report no further potential conflict of interest.

1. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. May 8, 2018 (https://www.nature.com/articles/s41746-018-0029-1).
2. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169:866-72.
DOI: 10.1056/NEJMc1906060

More on Ivabradine in Tachycardia with Paraganglioma

TO THE EDITOR: Malaza et al. (March 28 issue)[1] describe the benefit of ivabradine administration in a patient with catecholamine-induced tachycardia and metastatic paraganglioma. However, further data on hemodynamic variables before and after ivabradine treatment were lacking. Previous investigations have attempted to detect the hemodynamic effects of ivabradine in counteracting the undesirable tachycardia that is related to exogenous catecholamines. Vitale and colleagues[2] found a significant heart-rate–lowering response to ivabradine, which was associated with increased cardiac index, stroke-volume index, and mixed venous oxygen saturation in a series of patients with sinus tachycardia that had been induced by inotropic agents after cardiac surgery. A subsequent single-center, prospective, nonrandomized analysis by Gallet et al.[3] showed a significant improvement in diastolic function and hemodynamic status in patients with refractory cardiogenic shock who had been treated with oral ivabradine during dobutamine infusion. The hemodynamic benefit of this drug combination was validated in a preclinical experimental animal model, in which intravenous ivabradine was found to reverse the chronotropic properties of dobutamine, thus resulting in prolonged diastolic filling time and restored stroke volume.[4] Larger, randomized, controlled trials are needed to confirm these results.

Riccardo Scagliola, M.D.
Claudio Brunelli, M.D.

IRCCS Ospedale Policlinico San Martino
Genoa, Italy
risca88@live.it

No potential conflict of interest relevant to this letter was reported.

1. Malaza G, Brofferio A, Lin F, Pacak K. Ivabradine in catecholamine-induced tachycardia in a patient with paraganglioma. N Engl J Med 2019;380:1284-6.
2. Vitale D, De Santis V, Guarracino F, Fontana A, Pellegrini F, Tritapepe L. Use of ivabradine in catecholamine-induced tachycardia after high-risk cardiac surgery. Clin Res Cardiol 2010;99:853-5.
3. Gallet R, Ternacle J, Damy T, et al. Hemodynamic effects of ivabradine in addition to dobutamine in patients with severe systolic dysfunction. Int J Cardiol 2014;176:450-5.
4. Bakkehaug JP, Naesheim T, Torgersen Engstad E, Kildal AB, Myrmel T, How OJ. Reversing dobutamine-induced tachycardia using ivabradine increases stroke volume with neutral effect on cardiac energetics in left ventricular post-ischaemia dysfunction. Acta Physiol (Oxf) 2016;218:78-88.
DOI: 10.1056/NEJMc1905730

THE AUTHORS REPLY: We agree with Scagliola and Brunelli that the hemodynamic benefits of ivabradine have been reported in multiple studies. However, we would like to note that the patient populations in previously published studies differ from our patient, who never had low cardiac output. Invasive cardiac monitoring was not obtained in our patient, because there were no clinical indications to do so; therefore, many of the variables discussed in the aforementioned studies were not available for our patient. Throughout the course of our patient's treatment, he was monitored with echocardiography, which revealed a stable stroke volume and cardiac index. Grade I to II diastolic dysfunction was present in our patient, but these changes did not correlate with the use of ivabradine. It is important to consider that our patient had severe hypertension during treatment, which often led to the use of multiple antihypertensive medications, with an inevitable effect on diastolic function. Further studies in large populations of patients appear to be warranted to better delineate the wide spectrum of hemodynamic benefits associated with the use of ivabradine.

Gelinemae Malaza, M.D.
Alessandra Brofferio, M.D.
Karel Pacak, M.D., Ph.D.

National Institutes of Health
Bethesda, MD
karel@mail.nih.gov

Since publication of their letter, the authors report no further potential conflict of interest.

DOI: 10.1056/NEJMc1905730