Turning the crank for machine learning: ease, at what expense?

Excitement around the transformative potential of machine learning in health care belies a reliance on deep technical expertise that leaves this technology in the hands of the few. Typically, a practitioner of machine learning undertakes numerous tasks in the process of training and testing a model for classification. The process requires substantial technical knowledge and—perhaps somewhat incongruously—is often both highly detailed and loosely defined. In *The Lancet Digital Health*, Livia Faes, Siegfried Wagner, and colleagues report on their experience of using a service that creates an abstraction from the training and testing process, enabling a professional with no coding experience to build a model that might once have been out of reach. In the study, the authors train models for classifying disease in medical images using Cloud AutoML, a service that requires minimal technical knowledge. Discriminative performance is compared with values previously reported in the academic literature for matching tasks. The study does not claim to introduce new methods for machine learning, but it does highlight the potential of an easily accessible service that could be used by health-care providers. The work is compelling because it suggests that expert-level results in image classification are now achievable by anyone with cursory training.

We cautiously share the authors’ optimism that removing obstacles to algorithmic modelling will lead to improvements in patient care, but the risks of bypassing mathematical, statistical, and programming expertise must be emphasised. The use of machine learning methods without in-depth knowledge can result in misleading or outright erroneous results that would cause harm if used to guide the delivery of care. A reliance on simple performance metrics alone does not allow the practitioner to interpret other aspects of model development. For example, a model could demonstrate racial bias by yielding differing results for different subpopulations. Often the data itself introduces its own modelling challenges. Artifacts could cause a model to learn spurious rules, for example, such as the skin cancer algorithm that associates suspicious lesions with the surgical skin markers that surround them. Data quality might drift over time (for example, with changing equipment or operators), confounding an analysis that fails to account for these changes.

The models demonstrated in this study perform well on benchmark tasks, but they fail to generalise to external data. In reviewing the performance issues, the authors acknowledge their limited ability to audit the models or the data. Since the costs of misclassification can be high, this lack of transparency is concerning. Considerations of bias and technical rigour should be dominant considerations in health care, an issue that we hope will be addressed as the service develops. The “sharp contrast of the model’s discriminative performance” when moving from the “internal” to the “external” testing dermatology dataset leads the authors to conclude that a “small data” approach might be the ultimate use case for automated deep learning software. The approach would involve researchers and clinicians training models within their own institutions for “a specific geographical patient population that a given clinic might encounter”. While the simplicity of this approach is appealing, caution is clearly needed. What is the fate of the patient from outside this geographic population, previously unseen by the model?

We welcome efforts to improve the ease of developing models in health care, but it is important that governance, ethics, and technical oversight are allowed to keep up. Faes, Wagner and colleagues conclude that regulatory guidelines are needed for both medical deep learning and clinical implementation of these models before they might be used in clinical practice. We absolutely agree. A wider discussion about the ethics of training and deploying machine learning models in routine clinical practice—involving multiple disciplines spanning the clinical and computing worlds—must ensue. For now, machine learning in health care should remain collaborative, with experts from across disciplines working together. Meaningful machine learning for health is not just about turning a crank, but it requires the careful and thoughtful application of analytical techniques.

This online publication has been corrected. The corrected version first appeared at thelancet.com/digital-health on September 16, 2019.

See Articles page e232
*Tom J Pollard, Irene Chen, Jenna Wiens, Steven Horng, Danny Wong, Marzyeh Ghassemi, Heather Mattie, Emily Lindemer, Trishan Panch
Massachusetts Institute of Technology, Cambridge, MA 02139, USA (THP, IC); Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA (JW); Beth Israel Deaconess Medical Center, Boston, MA, USA (SH); Guy’s and St Thomas’ Hospitals NHS Foundation Trust, London, UK (DW); University of Toronto, Vector Institute, Toronto, ON, Canada (MG); Harvard T H Chan School of Public Health, Harvard University, Cambridge, MA, USA (HM); and Wellframe, Boston, MA, USA (EL, TP)

tpollard@mit.edu

TJP reports non-financial support from Google Cloud, outside the submitted work. SH reports grants from Philips Research, grants from Doris Dukes Charitable Foundation, outside the submitted work. MG worked previously for Verily, a subsidiary of Alphabet Inc and is supported in part by Microsoft Research. The other authors declare no competing interests.

Copyright © 2019 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

1 Faes L, Wagner SK, Fuet DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digital Health 2019; 1: 232–42.

2 Chen I, Johansson FD, Sontag D. Why is my classifier discriminatory? ArXiv 2018. https://arxiv.org/pdf/1805.12002.pdf (preprint).

3 Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IV, Ranganath R. Practical guidance on artificial intelligence for health-care data. Lancet Digital Health 2019; 1: e157–59.

4 Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. NPJ Digit Med 2019; 2: 77.

5 Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 2019; published online Aug 14. DOI:10.1001/jamadermatol.2019.1735.

6 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nature Med 2019; published online Aug 19. https://doi.org/10.1038/s41591-019-0548-6.