

Machine learning in translation

In machine learning applied to healthcare, challenges with the data stand between feasibility testing and clinically robust deployments.

Last year, a team from Google Health [found out](#) just how difficult it is to move machine-learning algorithms from early validation to clinical feasibility to real-world deployment. During prospective assessment in several clinics in Thailand of the feasibility and performance of a deep-learning system for the detection of diabetic retinopathy¹, the team documented, through interviews with nurses and technicians on the ground, the many sociological and environmental factors that affect the system's performance, the implementation workflow around it, and the experience of the patients. In particular, the system classified a larger-than-expected proportion of the retinal images as 'ungradable', owing to blurring or darkening caused by poor lighting during the eye-screening check-up.

To perform robustly in real-world settings, machine-learning solutions must therefore consider, from early design to deployment, how clinical staff and the patients interact with the technology, and how the technology integrates with patient-care workflows. However, even algorithms well designed for eventual clinical testing and deployment can stumble on their way to eventual clinical impact, owing to challenges related to the availability of high-quality, annotated and structured data for algorithm training and validation, as well as of data that are representative of the real-world conditions in which the algorithms will be deployed.

Foremost, large and high-quality clinical datasets that are well annotated are difficult to come by, especially in countries where operational and regulatory silos

raise barriers to data sharing and sustain inefficiencies in collaborative care and care delivery (including most countries that are members of the Organisation for Economic Co-operation and Development²). Necessary privacy restrictions and considerations of intellectual property also constrain the data that can be shared and accessed. Moreover, the data, codes and hyperparameters used to build, optimize and test machine-learning models are often not thoroughly reported or accessible³. In this regard, the recently developed artificial intelligence extensions of the Consolidated Standards of Reporting Trials (CONSORT)⁴, Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)⁵, and Standards for Reporting of Diagnostic Accuracy Studies (STARD)⁶ guidelines, and for medical imaging the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) best-practice guide⁷, should help increase the [credibility and reproducibility](#) of machine-learning studies.

The scarcity of data can be compensated for algorithmically, via generative adversarial networks (GANs) that create synthetic data. GANs are deep-learning architectures that mimic a zero-sum game for training a generator network so that it creates new datasets that mirror the data distribution of a real dataset by 'competing' with a discriminator network that determines whether the generated data belong to the real dataset. Datasets generated by GANs can also sidestep patient-privacy limitations to sharing. However, as [discussed](#) by Faisal Mahmood and colleagues in a Comment in this issue, the use of synthetic data calls for regulatory standards, as synthetic data

can also be exploited to circumvent patient privacy (by using the trained GANs to 'recover' samples from the training dataset), can amplify biases, and are particularly prone to authenticity issues.

GANs can also be used alongside federated learning — a technique for the decentralized training of machine-learning models. In federated learning, the data stay siloed while the model is trained locally (for example, at each healthcare institution or in each device at the 'edge' of the 'federated' data network). Federated learning and GANs can increase the performance and robustness of machine-learning algorithms by augmenting the training and validation datasets in size, diversity (by gathering data from a more diverse set of institutions, geographies, patient populations, or devices), and quality (by transforming datasets across data distributions and even data types, as in the generation of volumetric tomographic X-ray images from a single projection view⁸ and of computed-tomography images from magnetic resonance images⁹). In fact, as [shown](#) by Hadi Shafiee and co-authors in this issue of *Nature Biomedical Engineering*, adversarial learning (a variation of the GAN strategy lacking a generator and with a discriminator that is lax about shifts in the data distribution) allows for the effective training of machine-learning models from low-quality images (such as noisy photos from smartphone-based systems) that can be cheaply acquired and are largely unannotated. Additionally, algorithms for dimensionality reduction, such as the general and data-driven approach [reported](#) by Md Tauhidul Islam and Lei Xing in this

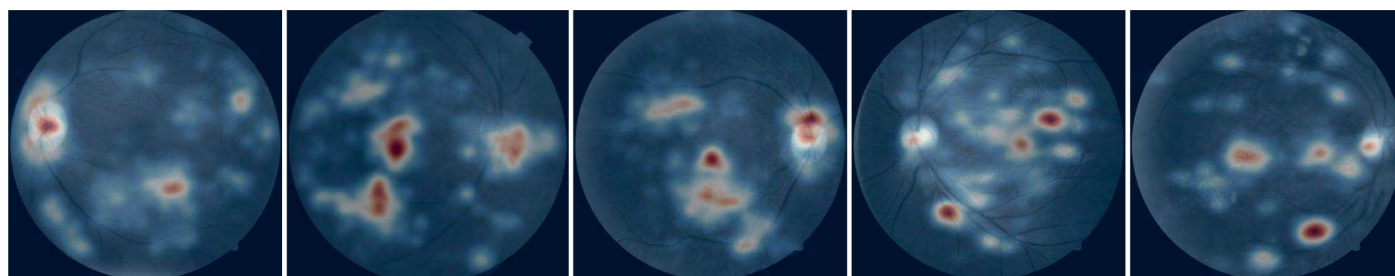


Fig. 1 | Retinal images with overlaid saliency maps indicating areas relevant to the prediction, by machine-learning models, of chronic kidney disease (early, advanced and severe; three leftmost images) and type-2 diabetes (with or without diabetic retinopathy; two rightmost images). Images reproduced with permission from the [Article](#) by Kang Zhang and co-authors.


issue, can be used to denoise data and to project data onto a lower number of dimensions to aid interpretability and visualization.

When big data are available yet cannot be efficiently processed, interpretable weakly supervised deep-learning methods leveraging attention-based learning (a technique for enhancing the most relevant parts in datasets) can be used to more efficiently process big data, as Faisal Mahmood and colleagues [demonstrate](#) with a model for the classification of whole-slide images annotated only with slide-level labels (rather than labels for pixels, patches, or regions of interest in the images). When abundant raw datasets are available — for example, videos of medical ultrasound — and can be processed, using all available raw data (rather than hand-crafted measurements derived from the data) to train machine-learning models can substantially improve their performance, as Brandon Fornwalt and colleagues [show](#) for the prediction of all-cause mortality within one year from echocardiography videos.

However, clinical data are often heterogeneous (in quality, completeness, or acquisition frequency, for example)

and unstructured (in format), and hence raw datasets usually need to be ‘cleaned up’ and processed before they can be fed to machine-learning systems. This is exemplified by an [Article](#) by Guangyu Wang and colleagues describing a modular deep-learning pipeline for the automated processing of chest X-ray images to identify and discriminate viral, non-viral and COVID-19 pneumonia and to assess disease severity. The pipeline included steps for the detection of anatomical landmarks, image registration, and the segmentation of lung lesions; and it was validated, retrospectively and prospectively, across patient cohorts and geographies.

Prospective assessment of the performance of machine-learning models is a necessary step before their actual deployment for routine clinical use. Model performance can be deteriorated (on deployment, but also over time) by myriad factors — in particular, changes in the incidence of the disease and in patient demographics, and alterations in clinical workflows and data-acquisition equipment — that affect the distribution of the data fed into the models. Such ‘domain shifts’ can be accounted for via regular auditing

and by model re-training and updating. Population-based prospective assessment is also a test of whether the performance of models can be generalized across settings, such as smartphone-based photography and multi-ethnic patient cohorts, as Kang Zhang and colleagues [show](#) for the detection of chronic kidney disease and type-2 diabetes with deep-learning algorithms trained with large datasets of retinal images (Fig. 1) acquired with standard fundus cameras. The performance of the algorithms will now need to be tested for robustness in real-world settings and eventually monitored after deployment. 

Published online: 15 June 2021
<https://doi.org/10.1038/s41551-021-00758-1>

References

1. Gulshan, V. et al. *JAMA* **316**, 2402–2410 (2016).
2. Oederkirk, J. OECD Health Working Papers, No. 127 <https://doi.org/10.1787/55d24b5d-en> (2021).
3. Roberts, M. et al. *Nat. Mach. Intell.* **3**, 199–217 (2021).
4. Liu, X. et al. *Nat. Med.* **26**, 1364–1374 (2020).
5. Rivera, S. C. et al. *Nat. Med.* **26**, 1351–1363 (2020).
6. Sounderajah, V. et al. *Nat. Med.* **26**, 807–808 (2020).
7. Mongan, J., Moy, L. & Khan, C. E. *Radiol. Artif. Intell.* **2**, e200029 (2020).
8. Shen, L., Zhao, W. & Xing, L. *Nat. Biomed. Eng.* **3**, 880–888 (2019).
9. Nie, D. et al. *IEEE Trans. Biomed. Eng.* **65**, 2720–2730 (2018).