Towards trustable machine learning

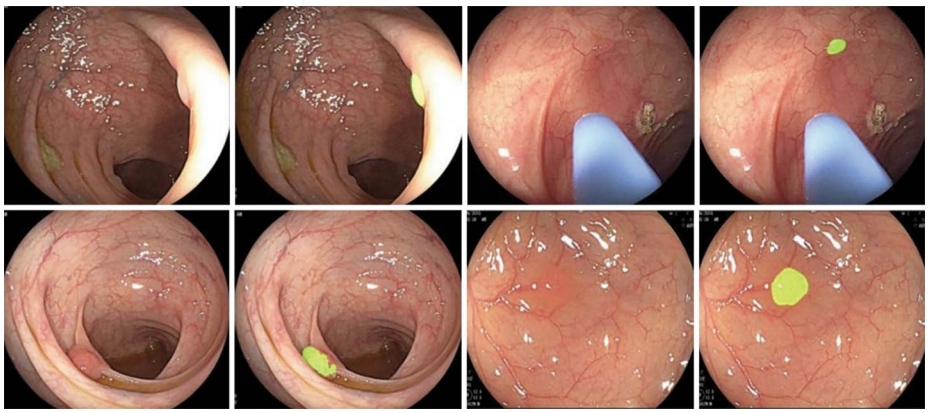
Clinical implementations of machine learning that are accurate, robust and interpretable will eventually gain the trust of healthcare providers and patients.

Reports of machine-learning algorithms that perform similarly or better than experts at specific tasks are increasingly less likely to elicit surprise. Provided that sufficiently large curated datasets and computer power are available, properly trained algorithms can outperform qualified physicians in most well-defined tasks that largely rely on pattern recognition — such as classifying medical images or complex patient data, or diagnosing disease risk from a defined set of biomarkers.

Yet excellent performance with previously curated datasets is insufficient as a validation threshold when the algorithms are intended to help physicians and patients in the complex world of healthcare, where data can be heterogeneous and where edge cases cannot be easily identified beforehand. In clinical use, the stakes are much higher. The performance of the software should be robust across the target patient population, its implementation ought to ensure proper use, and its analyses or predictions should provide context that aids interpretability.

Provided that the right choice of machine-learning model and of training and tuning processes are made (there are standardized and increasingly automated techniques that can help with these tasks), robustness in algorithm performance largely depends on the size and quality of the training and validation datasets. Any biases in the data with respect to the true representative population (such as biases in the patient cohort, in patient geographical spread, genetic background or disease incidence, in data labelling, or in data-acquisition technology or data-curation methodology), will transfer to the outcomes of the machine-learning algorithm. Because large and well-curated datasets such as those in the UK Biobank are not commonplace, performance pay-offs will accrue to organizations that are able to acquire large-scale medical data or efficiently curate and annotate previously acquired data (from medical electronic records, for example); all the more so because processes of data acquisition, curation and annotation can actually be sped up via machine learning.

The clinical implementation of machine-learning software should be informed by the needs of the healthcare providers and their patients. Especially in healthcare systems with stretched budgets and in departments with physicians and nurses with a real risk of burn out, proper use of artificial intelligence should ideally lead to faster or improved care (for example through improved patient–physician interactions), and to a reduction in overall costs. For instance, in some hospitals, machine learning is being tested to maximize the use of operating rooms (the most expensive real estate in hospitals) by estimating the duration of procedures and by anticipating demand.



Real-time detection of polyps during colonoscopy, via machine learning. Credit: Figure adapted from the Supplementary Information for the Article by Xiaogang Liu and colleagues, Springer Nature Ltd.

Where machine-learning software is to affect patient care directly, algorithm interpretability will be essential. Yet addressing the well-known 'black box' problem (one main critique to IBM's Watson for Oncology) by using computational techniques that 'peer' into the innards of the algorithms will not be sufficient in many cases. The interpretability of an algorithm's predictions can be skewed by the quality of the data used for the training of the algorithm, the characteristics and quality of the input data specific to the patient being assessed, and local medical standards and practice. Efforts to address these challenges are only starting. For example, a recent study from a collaboration between DeepMind and Moofields Eye Hospital in London reported that a machine-learning algorithm that makes referral recommendations on dozens of retinal diseases on the basis of optical coherence tomography scans highlighted structures in the scans that could lead to ambiguous interpretation. And an earlier study by Google Research showed that deep-learning models for predicting cardiovascular risk factors from photographs of the retina could indicate which anatomical features (such as the optic disc or blood vessels) the algorithm used to generate the predictions.

Two additional research studies, included in this issue, demonstrate the importance of algorithm interpretability and of clinically minded algorithm implementation. Su-In Lee and colleagues trained a gradient-boosting machine-learning model (an ensemble model that 'boosts' a set of weak prediction learners into becoming a strong learner) on minute-by-minute data from the electronic medical records of over 50,000 surgeries to predict the risk of unexpected hypoxaemia (abnormally low levels of oxygen in blood) during surgery under anaesthesia and, crucially, to provide quantitative measures of the contribution of the many risk factors (such as changes in the level of blood oxygen saturation, in ventilation rate and in the exhalation of previously administered anaesthetics). In real surgeries, the software's interpretable predictions and risk warnings could increase the safety of patients under general anaesthesia by strengthening preventive assistance and improving surgical workflows. In another study, Xiaogang Liu and collaborators demonstrate that a deep-learning algorithm (a convolutional neural network, which emulates how the visual

cortex responds to visual stimuli) trained with colonoscopy images from thousands of patients and colonoscopy videos from over 100 patients retrospectively detects polyps in newly acquired clinical colonoscopies in real time (pictured). The software (partially) tags detected polyps also under difficult visualization conditions (such as when a polyp is partially obscured by a fold, or when the illumination conditions are poor), raising the attention of the endoscopist, who can then guide the colonoscope around the tagged area.

As noted by Isaac Kohane and colleagues in a Review Article in this issue, implementing artificial intelligence in healthcare environments is never straightforward. Designing an algorithm for implementation in the clinic involves matters well beyond algorithm performance: from unintended consequences — such as increased workloads, disrupted workflows and patient–physician engagement and alert fatigue — to regulatory, reimbursement and data-privacy considerations. Yet none of these will matter if algorithm design neglects the need for trust. In this respect, opening up algorithms to interpretation is a necessary first step. ❐

Published online: 10 October 2018
https://doi.org/10.1038/s41551-018-0315-x