Machine learning in cancer diagnostics

In 1959, the phrase “machine learning” (ML) was first used by Arthur Samuel, a computer scientist who worked at IBM at the time and later became a Professor at Stanford University (CA, USA). ML defines the ability of a machine to learn and predict future events and outcomes based on large datasets. The research field of ML started to flourish in the 1980s and 1990s when artificial network connectivity and computational power improved, and digitalised information became more available. At that time, ML started to separate itself from the research field of artificial intelligence and shifted its focus to solving problems of a more practical nature. From there, it was not long before its potential in medical science was realised and scientists started to explore its utility in various medical specialties such as radiology, pathology, mental health, and cardiology. ML in healthcare is intended to improve medical data interpretation and thus expedite outcomes based on large datasets. The whole research process led to the development of novel computational tools for stratification, grading, and prognostication of patients with the goal of improving patient care.

The Gleason score is one of the oldest and most widely used grading systems for prostate cancer relying on histology slides of tissue biopsies. It was developed in 1966 and stratifies prostate cancer on the basis of tumoural architectural patterns. Pathologists evaluate the tissue visually, and based on the scoring, various treatment options are recommended. This method is subjective and suffers from high interobserver variability and discordance. In a study published June, 2019, in npj Digital Medicine, Martin Stumper (AI and Data Science, MI, USA), Craig Mermel (Google AI Healthcare, CA, USA), and collaborators developed a deep-learning system (DLS) for the Gleason scoring based on whole tissue images. The DLS achieved a diagnostic accuracy of 0.7 (on a scale from 0.5 [random] to 1 [100% correct]), which is an impressive result when compared with the diagnosis of 29 expert pathologists, the accuracy of which was only 0.61. Furthermore, follow-up data confirmed a better patient risk stratification by the DLS. Overall, this study demonstrates how ML can improve well established standards such as the Gleason scoring by eliminating subjective evaluation by the human eye, thus yielding to more precise prognostication.

Microsatellite instability (MSI) is a main driver in various cancers and can predict responses to immunotherapy. In the clinic, testing for MSI is time and resource consuming, which precludes its use in routine practice. In a 2019 study published in Nature Medicine, Tom Luedde (University Hospital RWTH Aachen, Germany), and collaborators developed a convolutional neural network directly from histology images to predict MSI in patients’ tissues. In patients with gastrointestinal cancer, the network yielded a maximal accuracy of 0.84, and in endometrial cancers it achieved an accuracy of 0.75. Both performances exceeded previously published predictions based on molecular markers.

Accurate drug-response prediction, especially before treatment, is essential for the stratification of patients to guarantee the best treatment option for each patient and will help mapping patients to the best therapy possible. Ultimately, accurate drug-response prediction can also help to identify patients who will not benefit from the therapy and spare these patients from unnecessary and potentially harmful treatments. Implementing ML can improve such strategies with retrospective datasets from clinical trials. Several studies were presented at the 2019 American Society of Clinical Oncology Annual Meeting (Chicago, IL, USA) on this topic; especially exciting was the study presented by Jame Abraham (Cleveland Clinic, OH, USA) and collaborators. They analysed the response to HER2-targeted neoadjuvant chemotherapy in patients with HER2-positive breast cancer by using pre-treatment MRI images. The neural network was able to predict responses with a maximal accuracy of 0.93 based on the MRI datasets, whereas the multivariate clinical model only achieved an accuracy of 0.67.

All these studies show how ML could improve diagnostic performance and prediction accuracy in clinically relevant patient cohorts; yet, only a small percentage of hospitals use ML routinely in clinical practice. ML is making huge progress in improving prediction accuracies, but such algorithms are barely used in the clinic (source: hospital survey by the Healthcare Information and Management System Society). Furthermore, even if the prediction power is highly improved, this information still needs to be interpreted and applied in a meaningful way. This is an active discussion in the field and it is worth considering why there is such great discrepancy. A recent quote from Eric Topol (Scripps Research, CA, USA) summarises the issue in a humorous way: “Computer scientist: “My deep learning algorithm has an AUC of 0.99!”—Doctor: “But does it help patients?”.” At the 18th MIT Annual Koch Institute Summer Symposium on Machine Learning and Cancer (Boston, MA, USA) in June, 2019, leading scientists in the field came together to present their work and to discuss how to implement ML into routine practice to improve daily patient care. One of the main challenges that ML is facing is that the workflow designed by researchers is not easy to integrate into the electronic data and health-record systems used in hospitals. Most ML approaches generated in the laboratory do not have direct clinical implementation as a primary goal. An intermediate step between the scientist and the clinician seems to be needed. Regina Barzilay and Connie Lehman at Massachusetts General Hospital (MGH, MA, USA) approached this problem directly and implemented a model for predicting the risk of developing breast cancer based on mammography images. With this ML approach they calculated the clinical risk and could improve the accuracy from 0.66 of the currently used Tyrer-Cuzick risk model to 0.88 for the 1-year risk assessment. Additionally, one major improvement in this study was the implementation of the ML results directly into the clinical workflow at MGH.
mammograms are acquired, the images are processed, and the assessment is directly integrated into the medical report where it can be reviewed by the clinician (published in Radiology, January, 2019).

Another hurdle for the implementation of diagnostic ML tools in the clinic lies in the difficulty of evaluating and selecting the algorithm to use. ML approaches and publications have dramatically increased in number, with several studies and various algorithms approaching the same clinical question. A systematic comparison of ML tools developed for the same clinical application should be performed to prioritise the most robust approach. This was the focus of a study by Harald Kittler (Medical University of Vienna, Austria) and collaborators published in The Lancet Oncology in June 2019, which compared the performance of 139 algorithms developed by 77 different ML labs to classify pigmented skin lesions based on dermatoscopic images. These algorithms performed with varying prediction power, with some even favouring human expert opinion. The top three outperformed the human reader by achieving the correct diagnosis with increased accuracy. More cross-validation studies such as this one will be needed to evaluate the performance of ML networks.

EBioMedicine is devoted to supporting high-quality translational scientific studies that have direct implications for clinical practice. In the field of ML for cancer diagnostics, our role is to facilitate progress for the development of better models to improve health care. This role entails upholding the best scientific practice, which uses validation cohorts or independent testing and comparison to already existing diagnostic standards. We will ask our authors to outline the strategy and workflow for clinical implementation of the developed ML tools. With this, EBioMedicine will support ML to play a more important role in clinical practice.