***Prompt:*** *Now that you have been exposed to some of the challenges in applying machine learning to Cancer research problems we would like you to write a 2000 word essay on how advances in machine learning and precision medicine are likely to impact Cancer care over the next ten years. Layout how you see this field developing and what are the most important problems to be solved*.

Since the war against cancer was declared in 1970s, we, researchers and clinicians, made significant progress in understanding cellular and molecular mechanisms of this disease and in developing effective therapies for certain types of cancers. However, even in the best cases, it is a game of numbers and percentages. In other words, over 40 years of intense research has failed to produce a treatment that can reliably cure 100% of cancers of a specific tissue origin. This because every single cancer is a unique disease with unique molecular, cellular, physiological, and epidemiological features. Even within a single tumor, there is inherent heterogeneity in the cells that make up the tumor, due to the evolutionary aspect of tumor development. This heterogeneity in cancers is why only some patients respond to drug treatments while others do not, or some cancers rapidly spread to the body, and some do not. It is clear to researchers and clinicians that successful treatment of cancer requires accurate prognostic assessment, and selection of the most effective treatment plan based on the features of an individual’s tumor – the concept that lies at the heart of precision medicine.

Historically, physicians have been performing this task based on years of experience and clinical trial results. But with the increasing availability of high-throughput cancer data, computational approaches such as machine learning and artificial intelligence are starting to be employed in predicting disease risk, prognosis, and response to treatment. In this essay, I will be discussing how machine learning is likely to alter the day-to-day practices in the clinic and at the bench in the near future, with the appreciation of the challenges lying ahead.

The holy grail of precision medicine is to be able to say “Patient X will respond to treatment Y because of Z”. Being able to make the connections between X and Y and Z is not always easy in cancer because of the aforementioned heterogeneity and takes years of research. This is where the machine learning methods will be most widely applied to, ten years from now. Supervised learning methods are already being used in academia and industry to develop “gene signatures” that can predict prognosis, metastatic potential, or recurrence of a patient’s cancer.(1) Unsupervised learning methods has led to identification of subtypes of cancers that behave differently and have different prognostic outcomes. (2) Results of these machine learning-driven approaches have already changed the clinical practice where physicians, now, follow different treatment strategies based on a patient’s tumor subtype and predicted outcome. (3)

Industry has also adopted machine learning approaches to build predictive models using genomic data. The greatest example of this is the genetic test Oncotype DX (developed by Genomic Health) which can predict response to chemotherapy and the chances of recurrence in early stage breast cancers. Another example is the company 23andMe which offers genetic testing product that can tell you your susceptibility for 10 diseases (though cancer is not one of them). All of these tests come with the warning saying that they are not 100% accurate. Clinicians think that these tests can never be 100% accurate because they are solely looking at genetic-level data. I anticipate that in ten years more accurate tests will be produced by building predictive models for cancer outcome using not only genomic data, but also transcriptomic, epidemiological, and even geospatial data. Companies will be selling these products in the form of tests to be used at home, or in the form of a hardware or software to be used in the clinic.

Another field where machine learning will have an impact is drug repurposing. Drug repurposing is the idea that a drug that has been FDA-approved to be used to treat a certain disease (or a certain type of cancer, in our case) can be repurposed to treat another disease that is of a similar molecular make-up. An example of this is the case of Pfizer’s Viagra, which has initially been developed to treat cardiovascular diseases, but then have been repurposed as an erectile dysfunction drug. Examples of drug repurposing in cancer, however, has so far been more coincidental or through mechanistic similarities between two cases. Metformin, for example, is a drug used against Type 2 diabetes. Researchers realized that patients treated with Metformin for their diabetes had lower rates of cancer, suggesting Metformin as an anti-cancer drug. Metformin’s efficacy in treating cancer is being vigorously tested right now.

Machine learning methods can be used to reposition drugs within the cancer domain more systemically and more globally. A predictive model can be trained based on the similarities between drugs, or the transcriptomic changes they induce in tumors (4). One can also predict a potential repurposing based on similarities in the molecular profiles of responder tumors. Using machine learning methods this way will facilitate identification of new indications of previously approved drugs. The advantage of drug repurposing is that repurposed drugs are more likely to succeed in clinical trials (at least in the early phases) because their toxicity and side-effect profiles are already known.

Image recognition is another type of machine learning that will impact cancer care in the near future. Most of the diagnostic tests used in the clinic today are image-based tools, where the patients undergo some sort of imaging procedure(e.g. X-ray, MRI, endoscopy, mammography) to detect and type a tumor. The images taken for the diagnostic purposes, then, are evaluated by a pathologist. I anticipate that, in ten years, most of image-based diagnosis will be performed computationally using models trained on hundreds of thousands of images across multiple cancer types. The premise is that machine learning and deep learning models can identify tumors in tissues more accurately and with higher sensitivity than a pathologist can with bare eyes. Potentially, image-based models can go beyond identifying tumors, but also subtyping them, or even identifying features that correspond to prognosis, or drug response. For example, cancer researchers know that if a tumor is infiltrated by macrophages, the prognosis is usually worse since these macrophages (called tumor-associated macrophages) have been shown to help tumor cells survive, grow, metastasize, and resist therapy. If a tumor has a lot of cytotoxic T cell infiltration, on the other hand, immunotherapy is much more likely to work because it activates these T cells to fight off the cancer. If an image recognition model can identify macrophages or T cells within a tumor, then it would also have predictive/prognostic value in addition to diagnostic value. However, such approach may require a multi-class classification if a single model was to be developed.

Image recognition models that can make tumor/no tumor classification can have implications in development of diagnostic devices. I can envision the release of a hand-held device, within the next ten years, that can make a benign/malignant call on images of skin moles. Similar approach can also be used to identify different types of skin rashes or infections. Such devices can even be used as self-diagnosis tools that a patient can use at home and avoid a visit to the hospital for a mole check. However, FDA-approval of these devices might take longer than their development.

Machine learning is likely to change the way scientists conduct basic cancer research as well. Particularly, studies that aim to test drug response and efficacy will soon be carried out mostly computationally, instead of experimental testing. Experimental science will assume the role of validating what is predicted by the drug response models. Thanks to databases like NCI60 and GDSC, now we have drug-response data from many drugs tested on many cell lines *in vitro*. In the near future, these data sets will only grow larger. As the next generation sequencing technologies get cheaper, we will be generating even more high-throughput labeled data. Within the next ten years, I anticipate similar databases to be built for drug combinations as well as for animal testing of drug efficacy. Soon we will be able to take hundreds of newly generated drugs, and predict their effect on cell growth or morphology without touching a pipette.

Similar approach will be implemented in clinical research as well. The concept of precision medicine has been pushing the boundaries of the way we test drugs in clinical trials. Current clinical trial system give minimal consideration to the molecular features of a tumor. Patients who are participating in the study usually get randomly assigned into either the treatment arm or the control arm. This usually yield low response rates because non-responders are assigned with responders at random. I expect this to change with the implementation of machine learning models that can predict whether a patient will respond to a drug *before* the patient is put in a group. This way the drug will be tested on people who are already more likely to respond to it. In other words the clinical trial would be validating what the model predicted in the first place.

These approaches are, indeed, being slowly implemented in the clinical trials. There is a new trial design called the “basket trial”, which aims to group cancer patients into a treatment a treatment arm based on the target mutation that their cancer harbors, regardless of the histology and tissue origin of the cancer (5). In these trials, a breast cancer patient can be put in the same treatment arm along with a lung cancer patient as long as both cancers have the same mutational aberration that the treatment is targeting. Examples of these trials are NCI-MATCH, CREATE, CUSTOM, Genentech MyPathway, and Novartis signature (discussed in ref. 5) This type of trial design aims to speed up drug approval process and increase response rates by targeting the right patient with the right kind of drug. However, the assignment still falls on single genetic feature, and not overall clustering of all the patients.

I think in the future, machine learning methods will be employed to more accurately identify cancer cases that are more similar to each other in their predicted response to a novel treatment in a given basket trial. The treatment arm assignments will be based on not only the single type of genetic aberration shared by the two different cases, but instead will be based on models that have been trained on multiple-level omic data (e.g. genomic, transcriptomic, epigenomic) which predict favorable response for a given drug.

Implications of machine learning and deep learning methods in cancer research and clinical care are numerous. There are, however, some challenges that need to be overcome for machine learning applications to reach their full potential.

The biggest challenge right now is data availability. All predictive models that influence treatment of a cancer patient will have to reach near 100% accuracy. Training deep learning models that can predict patient outcomes with high accuracy will require massive amount of data. However, most of cancer-related data today are collected as private property of a hospital or a company. There are roughly 1.7 million new patients diagnosed with cancer every year in the United States, but the size of publically available cancer data is limited to a few 1000s. This is also why majority of cutting-edge deep learning and artificial intelligence research is carried out in big companies that own the largest data sets in the world, like Google. Therefore, tremendous amount of effort must be made to increase the number of available data sets that are relevant to oncology.

Another challenge for implementation of deep learning in cancer care is the black-box aspect of these deep models. Referring back to the “Patient X will respond to treatment Y because of Z” dogma of precision medicine, current deep learning methods are impeccable in predicting the X-Y relationship, but they all fail at explaining Z. Justification and trust are important requirements for a predictive model that change the course of a patient’s journey to cure. If a model predicts the wrong response to a treatment, it is absolutely essential to know why that prediction failed. Explainable AI (XAI) approaches are promising learning methods that tackle this problem. Adoption of an explainable AI in the clinic would be a lot easier than a black-box AI since it helps build an intuition in the doctor and the patient about the decisions they are making. XAI can also facilitate basic research by helping generate hypotheses about the mechanism that leads to the prediction. A prediction like “Patient X will respond to Treatment Y because their tumor has high expression of gene Z” will generate an immediate mechanistic hypothesis that can be tested experimentally.

**REFERENCES**

1. Lee et al., *PLoS One*. 2013 Dec 11;8(12):e82125.
2. Perou et al. *Nature* 406, 747–752 (17 August 2000)
3. Yersal et al. *World J Clin Oncol.* 2014 Aug 10; 5(3): 412–424.
4. Sleire et al. *Pharmacol Res.* 2017 Oct;124:74-91. Epub 2017 Jul 13.
5. Cunanan et al. *Journal of Clinical Oncology* 35, no. 3 (January 2017) 271-273.