Journal for ImmunoTherapy of Cancer

Just how transformative will AI/ML be for immuno-oncology?

Daniel Bottomly, Shannon McWeeney

To cite: Bottomly D, McWeeney S. Just how transformative will AI/ML be for immuno-oncology? *Journal for ImmunoTherapy of Cancer* 2024;**12**:e007841. doi:10.1136/jitc-2023-007841

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/jitc-2023-007841).

Accepted 15 January 2024

Knight Cancer Institute, Oregon Health and Science University, Portland, Oregon, USA

Correspondence to
Dr Shannon McWeeney;
mcweeney@ohsu.edu

ABSTRACT

Immuno-oncology involves the study of approaches which harness the patient's immune system to fight malignancies. Immuno-oncology, as with every other biomedical and clinical research field as well as clinical operations, is in the midst of technological revolutions, which vastly increase the amount of available data. Recent advances in artificial intelligence and machine learning (AI/ML) have received much attention in terms of their potential to harness available data to improve insights and outcomes in many areas including immuno-oncology. In this review, we discuss important aspects to consider when evaluating the potential impact of AI/ML applications in the clinic. We highlight four clinical/biomedical challenges relevant to immuno-oncology and how they may be able to be addressed by the latest advancements in AI/ML. These challenges include (1) efficiency in clinical workflows, (2) curation of high-quality image data, (3) finding, extracting and synthesizing text knowledge as well as addressing, and (4) small cohort size in immunotherapeutic evaluation cohorts. Finally, we outline how advancements in reinforcement and federated learning, as well as the development of best practices for ethical and unbiased data generation, are likely to drive future innovations.

INTRODUCTION

With the development of several breakthrough immunotherapies including anti-PD-1/PD-L1 and anti-CTLA-4 immune checkpoint inhibitors (ICIs) reviewed in Ribas and Wolchok[1] as well as chimeric antigen receptor (CAR) T cells reviewed in June *et al*,[2] immuno-oncology (IO) has been established as a promising framework for the development of cancer therapeutics.[3] However, a number of challenges still remain.[4] For instance, only a subset of patients with cancer diagnoses that would be otherwise terminal see durable clinical benefit from a given immunotherapy. This has led to increased interest in development of personalized biomarkers to ensure rational development of clinical trials. Of those patients who do respond, as is also the case with chemotherapy and targeted inhibitors, they can eventually develop resistance (as reviewed in [5]). As a means of circumventing resistance, much focus is currently placed on the development of ICI combinations, reviewed in Sanmamed *et al*[6] as well as coupling an ICI with cytokines (or corresponding blocking inhibitors or antibodies) to fine tune the desired immune response as reviewed in Berraondo *et al*.[7] In addition to resistance, the occurrence of adverse events[8] including observations of "hyperprogressive" disease[9] presents a challenge to the development of clinically viable therapeutics. It is increasingly recognized that the complexities in immunotherapeutic treatment are in part due to immunoediting, reviewed in Gubin and Vesely.[10] A major tenant of the immunoediting hypothesis describes how the immune system can be modulated by tumors to maintain a stable disease state or escape, allowing progression. These interactions between tumor-intrinsic pathways, tumor microenvironment (TME) and immune mechanisms produce substantial heterogeneity in tumor dynamics and therefore patient response. At the same time advances in computational power and storage as well as in the theory and practice of machine learning (ML) have accelerated our ability to gain insights from large multimodal datasets.[11] ML approaches traditionally used for biomarker discovery or immune-therapeutic response have been reviewed previously.[12–15] This review focuses on recent advancements in artificial intelligence (AI) and ML, namely state-of-the-art neural network architectures and related applications. We note that there are numerous reviews that have focused on validation of IO targets and development of effective immunotherapies.[12–17] Therefore, we have focused on other areas where AI/ML holds promise to improve IO in clinical care.

AI and ML are often used interchangeably and while related are still distinct concepts. AI as a field was originally started in the 1950s[18] and focuses on computational approaches to mimic the ways that humans think in order to perform complex tasks. ML is a subfield of AI[19] that uses algorithms trained on data to produce models that can perform complex tasks. We note that often use of the term AI is colloquially associated with a specific type

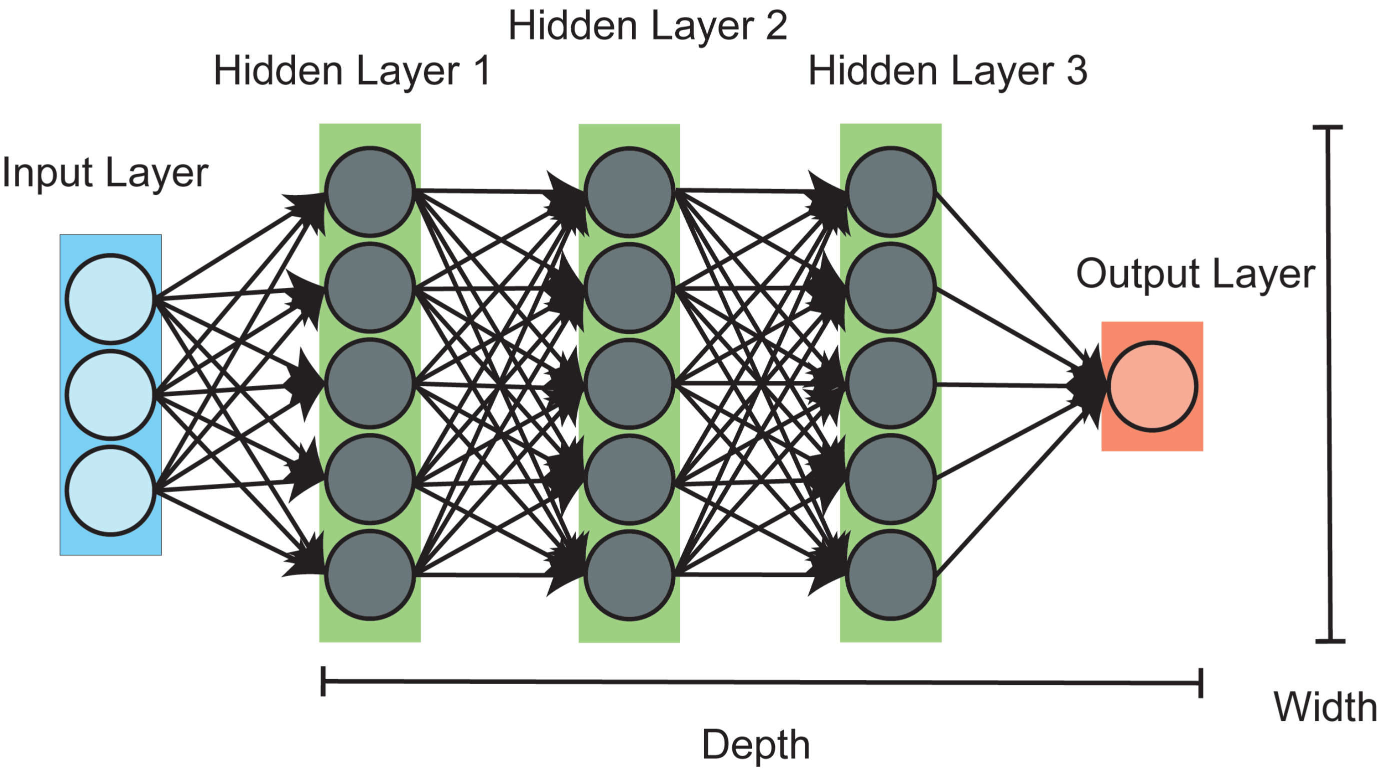

Figure 1 Architecture diagram of a simple feedforward deep neural network is depicted. Circles represent nodes/units with lines and arrows indicating information flow from the input layer to the output layer. The intermediate layers are termed hidden layers. Depth is defined as the number of non-input layers and width as the number of nodes in a layer.

of ML model known as a neural network. In particular, so-called deep neural networks (DNNs) have been shown to approach human level performance on a number of tasks.[20] Traditionally, neural networks are conceptualized as having fully connected layers of nodes starting at the input with intermediary "hidden" layers terminating at the output layer (termed a feedforward network; figure 1). Not counting the input layer which is specified by the data, the depth is typically considered to be the number of remaining layers, while the width is measured as the number of nodes in a given layer. Although modern neural network layers can be composed of different types of components and connections, DNNs by definition all have multiple hidden layers. The power of DNNs is that they are thought to be able to represent complex relationships by encoding more general information in the early layers followed by increasing levels of abstraction.[19] Arguably, to date, the most potentially impactful of these DNN technologies is the generative pretrained transformer version 4 (GPT-4) large language model (LLM) which has been shown to score respectably on a number of accreditation and academic exams.[21] The primary interface to these models is a prompt-based question-and-answer framework tuned by reinforcement learning from human feedback (RLHF).[22] In addition to the adaptation of GPT-4 into an "AI chatbot for medicine",[23] many of the concepts that make the GPT model successful can also be applied as a whole or in part to the study of IO (as well as other clinical research fields) as we discuss below.

It is critical that we can discern reality from the hype when it comes to AI/ML. Hype can distort our understanding of any new technology which makes it difficult to understand where it is best used. The Gartner hype cycle[24] provides a graphical representation to represent the maturity, adoption, and social application of specific technologies. Recognizing that the time frame for AI/ML is not recent, the hype cycle can allow us to assess expectations versus reality for specific subsets of AI/ML such as DNNs or LLM. While there have been numerous criticisms of the hype cycle,[25] it serves as a useful reminder that popularity of a technology may not often coincide with its maturity, leading to premature applications or misuse. The implications of this in the healthcare setting are far-reaching beyond economic impact. Therefore, we need to consider "how soon is now", that is, which approaches will be realized in the near, mid and long term? In addition to the maturity of the technology, we need to also consider organization maturity and readiness for the development, deployment and maintenance of AI/ML. The term "Technical Debt", originally defined in software engineering, has been used to describe tradeoffs in the long-term costs of adopting AI/ML systems.[26] Quickly deploying an AI/ML model often results in increased future expenses to maintain performance and stability when faced with changes to the underlying code, data format/distributions as well as computing or inference environment. To address these challenges, machine learning operations (MLOps) was devised as an

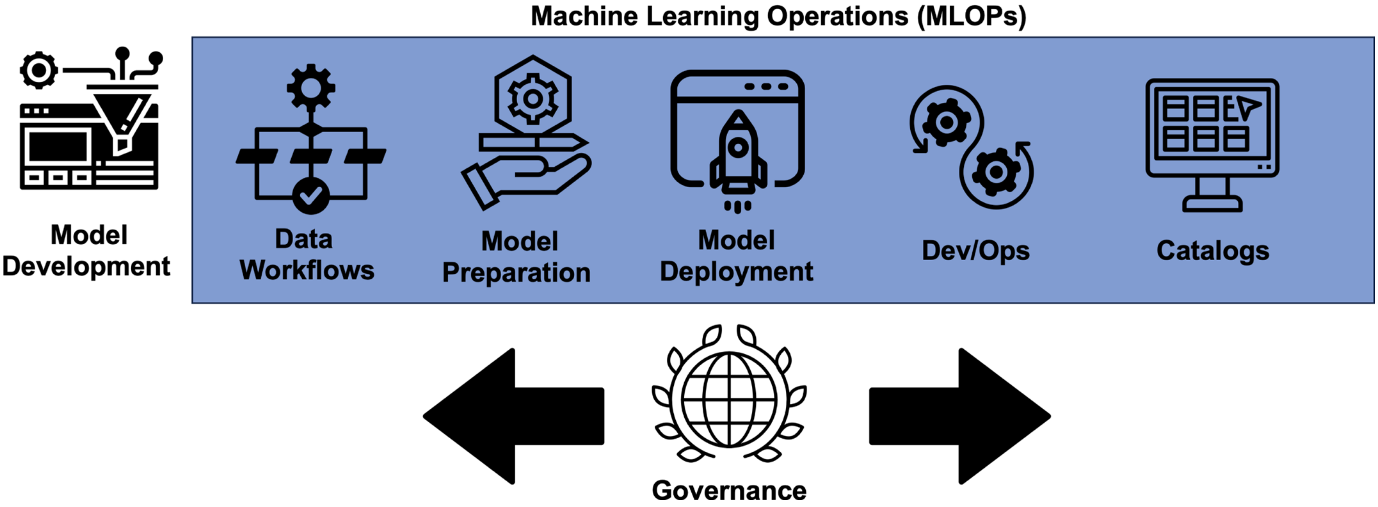

Figure 2 An overview of the main components of the field of machine learning operations (MLOps) is depicted. Model deployment can arguably be considered outside of MLOps as the resulting models are research grade and may never be deployed in a production environment. For deploying production models, first care must be taken to ensure reproducibility of the data and computational workflows. To prepare the model in question, it first needs to be refined to minimize resources and maximize evaluation performance. Only then is the model deployed with maintenance and updates occurring using best practices formed originally from software development operations (DevOps). Active monitoring of the model given potential drifts in the data over time is also critical. Production grade models can additionally be registered in model catalogs simplifying access to alternative model versions. Of importance is that this entire process is subject to institutional governance to ensure that resulting production systems meet ethical, privacy, security, performance and fairness standards.

organization-level paradigm drawing on and extending best practices for deploying complex software systems. However, fully using MLOps requires substantial expertize, infrastructure and resources (figure 2) and is still in its infancy.[27] In addition, governance of AI has been more reactive with many institutions still grappling with the complex issues around best practices such as evaluation, deployment and drift.[28] Finally, in addition to training and workforce development, professional development and guidance are needed to support key roles (eg, C-level executive roles) responsible for oversight of implementation/operations.

The remainder of the paper highlights four clinical/biomedical challenges relevant to IO, as well as other clinical research fields, and how current AI/ML methodologies may help in the near/mid-term timeframe (figure 3). Finally, we offer thoughts about key innovations and practices necessary for long-term impact.

Challenges of efficiency in clinical workflows

Although much attention is given to the application of DNNs and other advanced ML systems in biomedical and clinical research, clinical practice and administration stand to benefit as well[29] and are likely a near-term beneficiary of advancements in AI/ML. For example, AI/ML medical scribes can automate documentation, improving accuracy, reducing physician burden, and enhancing patient care.

Another promising area is the ability to increase patient engagement and adherence to treatment. Approaches to increasing patient engagement have been systematically reviewed previously.[30] Access to LLM chatbots as well as mobile apps are newer ways that can provide a patient greater access to their health information.[14] A recent study found similar accuracy between humans and ChatGPT in answering questions related to clinical genetics. However, it was noted that ChatGPT displayed poorer performance if critical thought was required.[31] One of the strengths of ChatGPT as well as other LLMs is the ability to rapidly summarize prompted information. This is an ability which could be readily adapted to generating discharge summaries[32] and is seeing rapid uptake in medical transcription.[23] However, as with the other more critical diagnostic and treatment uses, use of specific LLMs for these purposes would have to be rigorously validated.[33] The successful application and adoption of DNN powered applications have great potential for lowering the cost of delivering healthcare. A recent study found that utilization of DNNs for treatment in particular could potentially result in large cost savings for hospitals.[34]

Another challenge for healthcare systems is the management of adverse drug reactions, estimated to cost US$30.1 billion annually in the USA alone.[35] Data on adverse events are gathered as part of "pharmacovigilance" efforts and AL/ML approaches are considered to be a tool to help with case processing and assessment or as part of "human-in-the-loop" systems.[36] AI/ML systems designed to predict adverse events ahead of time such as the adverse events atlas[37] are under development and have the potential to improve patient care as well as realize significant cost savings.

Adverse event detection extends beyond the pharmacological use cases to other aspects of patient risk. An area of clear interest is AI/ML prediction of point-of-care risk stratification which has received tremendous attention

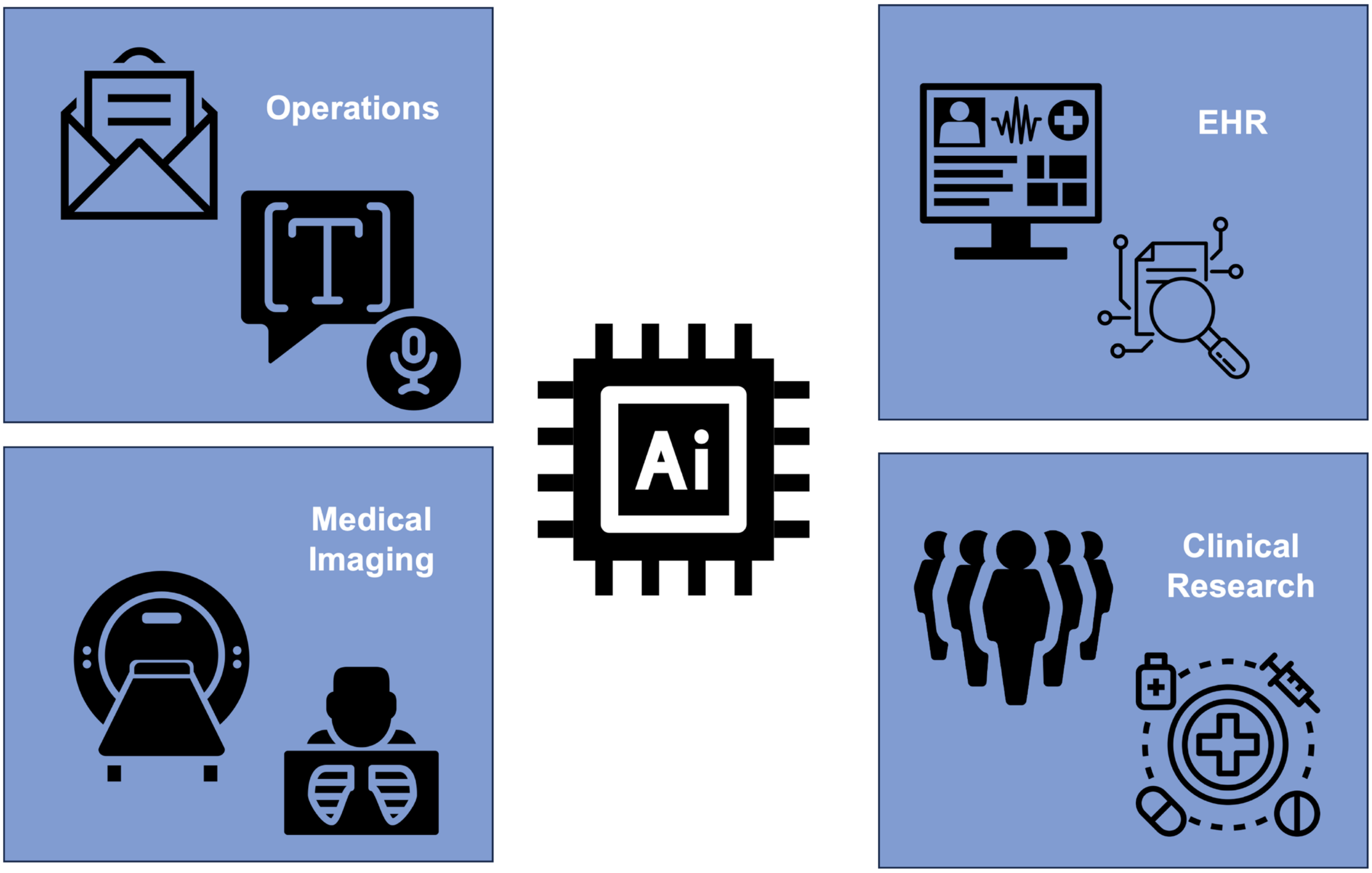

Figure 3 Current advancements in AI/ML technology stand to have the greatest impact in four areas relevant to immune-oncology and clinical research as a whole. The field of clinical operations and medical imaging are likely poised to see near term benefit from AI/ML, while natural language processing applications for electronic health records (EHR) and research are likely to see mid-term benefits. Clinical research due to relative sparsity of data, on the other hand may only see mid-term to long-term benefits from AI/ML. AI/ML, artificial intelligence and machine learning.

and already provided key use cases on the challenges of implementation. External validation of a proprietary sepsis prediction model which had been implemented in hundreds of hospitals found that the model predicted the onset of sepsis with an area under the curve of 0.63, which is substantially worse than the performance reported by the vendor, highlighting the need for external validation before adoption.[38] With regard to oncology and immunology, applications have included quality improvement by identification of oncology patients at risk for a near-term emergency room visit,[39] prediction of mortality in immunotherapy patients[40] and early identification of patients with who could realize durable clinical benefit from PD-(L)1 blockade-based ICI treatment and chemotherapy,[41] among others. Best practices for implementing risk stratification models in the clinic are actively evolving—one common theme is that use of a "plug and play" model is fraught with issues given concerns about clinical relevance of predictions, the lack of considerations for how it will be integrated into workflows, as well as the need for training and change management.[42–45] Evaluation of the models must be intuitive to clinicians and ideally focused on metrics related to quality of care and patient outcomes rather than technical performance of the model.[46] As with all clinical decision models, we must also address ethical concerns such as bias and fairness as it relates to both the models and the data the models were trained on. While we view many aspects of AI/ML improving efficiency in clinical workflows as candidates for near term benefit, for the reasons above, the risk stratification approaches are likely more near to mid-term.

For academic medical centers and cancer centers, the impact of the savings due to increased efficiency from the near-term applications could potentially be used not only for expansion of new clinics, but also implementation of immunotherapeutic trials and other clinical and translational research innovation.

Challenge of curation of high-quality image data

The advent of digital pathology, reviewed in Baxi *et al*,[47] has led to an increasing reliance on AI/ML methodology and is a near term beneficiary of advancements in AI/ML. The importance of digital pathology for IO should not be overlooked (see Morad *et al*[48] and references therein). The automation of tissue slides in digital pathology, in conjunction with the rapid advances in multiplex imaging technology, sets the stage for extensive characterization of the TME.

Relevant for digital pathology, we provide additional background on ML. Traditionally, ML approaches are classified into two main types: supervised and unsupervised.[49] Supervised learning requires the presence of a high-quality "label", often a classification or outcome. In the case of clinical research, this typically requires either collection of requisite data and evaluation of patient response, say to a given immunotherapeutic, by clinicians or access to existing collections of well-curated and annotated images. The data collection and evaluation process can be time-consuming and laborious and often limits the utility of retrospective data. Unsupervised learning on the other hand does not require prelabeled data, making it more appealing for discovery in clinical research such as for new biomarkers. However, datasets used for biomarker discovery in particular tend to have large numbers of correlated features. In addition, genomics data are also often impacted by unknown technical factors, reviewed in Leek *et al*.[50] These issues make true unsupervised biomarker discovery especially challenging regarding validation and interpretability. However, there are other approaches to model training such as reinforcement learning (discussed in the Future Impact and Concluding Remarks section) and hybrid approaches of supervised and unsupervised learning which provide a bridge between a priori knowledge and discovery.[19]

In digital pathology, while high quality classification/outcome/annotation information exists, often there are large stores of unannotated images are also available. Semisupervised approaches seek to leverage both characterized and uncharacterized data to improve model performance. Related to semisupervised learning is self-supervision reviewed in Krishnan *et al*.[51] This type of training formulates a supervised learning problem based on a secondary outcome that can be derived from the data itself. Although this conceivably can be done in many ways, two common approaches exist: contrastive learning and generative (masked) learning.[51] Contrastive learning first develops a model that can distinguish, for example, whether two images are similar or different. In this manner, it captures important features for distinguishing images without explicit labeling. Masked learning on the other hand removes a portion of the data, says a word in a sentence, and requires the model to predict the word. The process guides the model to infer approximations of the underlying context.

Another approach to dealing with limited curated data for algorithm development is to instead generate more high-quality data in silico. The ability to generate new realistic data and content is a powerful and controversial development in AI/ML. These so-called generative neural network architectures are commonly used for image, speech and text generation.[52–54] The fundamental type of generative neural network is the Generative Adversarial Network (GAN).[55] GANs consist of two interlocked models, a generator, which models the data distribution producing a synthesized result and a discriminator which tries to distinguish the synthesized version from the training data. GANs have been developed to help with digital pathology, analyzing histological images of tumor samples both in terms of generating new high-quality images but also in learning latent representations in an unsupervised manner.[56] Additional uses in pathology include preprocessing, color normalization, virtual staining, image enhancement, removal of ink marks and augmentation) as well as other types of analyses including nuclei detection, segmentation and domain adaptation.[57] Generative neural network approaches have also been developed for the analysis of single-cell data in general[58] including prediction of immune response from a baseline measurement.[59] Of high relevance to IO are spatial transcriptomics technologies that combine gene expression (including single-cell) and imaging.[60] Similar to imaging, there is much interest in the development of AI/ML analysis approaches. For instance, the recently proposed GraphST method leverages graph neural networks refined using contrastive learning.[61]

Challenge of finding, extracting and synthesizing knowledge

Electronic health records (EHRs) have become an increasingly important source of real-world data (RWD). In turn, RWD can enhance the development of IO therapeutics and companion diagnostics as well as provide critical real-world evidence for regulatory approval.[62] Algorithmic and technical innovations coupled with access to huge quantities of text data from the internet enabled the creation of LLMs such as Bidirectional Encoder Representations from Transformers (BERT)[63] and the GPT.[64] These models can have billions of parameters, for instance, the previous generation GPT-3 had 175 billion,[65] which encode the capacity for highly sophisticated pattern recognition and generalization. Because of their potential they are considered to be "foundation models"[66] with utility in many different tasks. In the case of GPT-3, it was observed that "few-shot learning", in this case applying in-line text examples for the system, produced performance on par with many state-of-the-art systems that had been specially tuned using labeled training data.[65] This feature has been explored in the clinical domain.[67–69] Additionally, LLMs specific to science and biomedical tasks have been created which may be better suited for technical use-cases.[70–72] In fact, it was observed that using only biomedical texts in pretraining provided a benefit over models that included other less relevant sources of information.[73] In addition to the general clinical applications reviewed in,[23] of most relevance to IO research, is the ability to retrieve and summarize existing research. One of the challenges facing widespread adoption of methods in this area is the observation that LLMs can "hallucinate" producing realistic sounding content that is factually incorrect. As mentioned on their homepage,[74] the prevalence of hallucination led to one scientific LLM endeavor, Galactica,[70] removing its demo from the web. This is due to these models encoding knowledge as a function of their learnt weights—not necessarily facts.[75] Ongoing work is exploring how to incorporate prior knowledge in the form of graphs which may be able to help address this issue.[76]

The ChatGPT model in particular is incredibly powerful, however, it can still produce inaccurate results and therefore caution is needed. For instance, in online supplemental figure S1A, an interaction is shown where ChatGPT is asked to list approved immunotherapies, their indications, companion diagnostics and references. Although, it left blank most of the indications, generally the companion diagnostic information seemed accurate with the exception of CD19 expression for Axicabtagene ciloleucel CAR-T therapy. One possible source of ambiguity is that Axicabtagene ciloleucel is an anti-CD19 CAR-T therapy, although having a positive expression of CD19 is not a prerequisite for treatment.[77] To attempt to gain further insight into what chatGPT classifies as a companion diagnostic, it was further asked to define a companion diagnostic. The chatGPT system described the companion diagnostic as "a test or assay used alongside a specific drug to identify patients who will benefit from the treatment" (online supplemental figure S1C). This highlights that there are still nuances not recognized by the LLMs. One way to combat hallucinations as well as other ambiguous responses is to ask the model to check itself for accuracy.[23] Finally, we provided the above definition of companion diagnostic and asked it to confirm if CD19 is considered a companion diagnostic for Axicabtagene ciloleucel. Again, the system answered in the affirmative justifying its reasoning (online supplemental figure S1D). This further emphasizes the need for conversational interactions with the system as opposed to strict question and answer. Similarly, in online supplemental figure S1B, when asked to verify the references, the system did indicate that the references were placeholders. Hopefully in the future, if correct references cannot be given these fields can be left blank or at least clarified to the user as part of the response text to avoid potential issues.

LLMs can also be used to extract unstructured or semi-structured information found in EHRs as is reviewed in Fu *et al*.[78] This is highly relevant to IO as a means to gather data that would otherwise be laborious to achieve, for example, tumor-infiltrating lymphocyte classification.[79] Additionally, EHR information can be queried to evaluate patients with respect to clinical trial eligibility. In a recent adaption of this approach, using CT-BERT, a version of BERT fine-tuned on ClinicalTrials.gov data[80] an AI/ML framework was proposed to assess the generalizability of a given clinical trial with respect to its eligibility criteria.[81] The recently developed LLM GatorTron focused not only on computational phenotyping/cohort characterization but also on its potential use in pharmacovigilance.[82] One of the current challenges with application of ChatGPT to EHR mining tasks concerns patient privacy. This is because currently interactions with ChatGPT are transmitted over the internet.[83] Additionally, if limited to basic (non-identifying) diagnostic information, the performance of ChatGPT degrades considerably.[83] These and related issues have influenced the development of LLMs derived specifically from deidentified clinical notes as well as other relevant sources.[82] Despite general advances in LLM capabilities, the robust application to biomedical and clinical text is likely a mid-term beneficiary of AI/ML advancements. We note that implementation of the largest scale LLMs is often beyond the resource of individual cancer centers so optimization and adaptation of existing models will likely be key.

Challenge of small cohort size in immunotherapeutic evaluation cohorts

One of the challenges with developing treatment strategies or evaluating biomarkers for personalized therapy is that most clinical patient cohorts tend to be small and heterogeneous. This more often than not leads to the lack of novel discovery and/or generalizability.

An often-pursued approach especially in biomarker studies is to develop models first in larger cohorts such as The Cancer Genome Atlas (TCGA)[84] and then apply them to more focused studies. For instance, using a curated list of 29 functional gene expression signatures describing TME processes, 4 signatures related to melanoma were derived using TCGA that could predict response to ICI in smaller independent studies.[85]

The concept of "pretraining" larger DNN models is commonly used in other fields, often with the goal of leveraging these models (termed sources) to help solve other problems similar in form to the original task (termed targets), a procedure referred to as transfer learning.[86] With a large pretrained model in hand, transfer learning can be implemented using a number of different strategies. This approach relies on the assumption that earlier layers in a large DNN have learnt sufficiently general patterns to be useful to other tasks. Learnt weights within the source DNN model are either "frozen" or used as the starting point for additional fine-tuning with the target dataset as reviewed in the study of medical image classification.[87] In addition to imaging, with the advent of single-cell transcriptomics, large quantities of publicly available data can be collected to pretrain models for the purpose of transfer learning. One variation of this is annotation of single-cell experiments relative to existing datasets.[88 89] As larger single-cell atlases are generated such as the human cell landscape,[90] related approaches can be used to map single-cell datasets onto the atlases to facilitate analyses.[91]

Based on the success of models such as GPT and BERT, there is much interest in developing large pretrained transformer models to support transfer learning in many aspects of clinical research including IO. The transformer architecture was originally developed in the context of sequence (eg, language) modeling using DNNs.[92] At its core is the concept of self-attention which was originally devised as a means to reduce computational complexity, increase parallelizability and to improve detection of long-range dependences in these sequence models.[92] However, importantly, at the same time it also provides a means of model interpretation. Interestingly, in addition to the utilization of the pretrained transformer architecture in state-of-the-art LLMs, models have also been devised to pretrain on available large biomedical datasets. Using publicly available single-cell data, pretrained transformer models have been devised to capture network dynamics such as Geneformer[93] as well as begin to build more foundational generative models such as scGPT[94] and tGPT.[95] The tGPT model in particular showed promise for differentiating immunotherapeutic treatment outcomes in urothelial carcinoma leveraging available bulk RNASeq data. In addition to single-cell data, models are being developed that utilize clinical and mutational data for ICI treatment outcome prediction such as the Clinical Transformer.[96] Given challenges with relevant dataset accrual and potential issues in bias, it is likely this is a mid-term to long-term beneficiary of AI/ML in the clinic.

Future impact and concluding remarks

Distinct from more traditional supervised learning described previously, reinforcement learning provides a mechanism through which models can be continuously trained. This is accomplished through the accumulation of rewards from chosen actions with the goal being to maximize the total received rewards in the long term. This approach, though currently used infrequently in biomedical research, has been implemented to predict drug sensitivity as well as optimizing chemotherapeutic and radiotherapy doses in retrospective and simulated clinical settings as reviewed in Eckardt *et al*.[97 98] Similar approaches are also being adapted to assess immunotherapeutic challenges such as achieving control of the balance of a patient's immune and tumor cells with respect to treatment.[99] Use of reinforcement learning has also been successfully used in the search for T cell receptor beta chain CDR3 sequences that have enhanced affinity for peptide sequences .[100] This has implications for adoptive T cell immunotherapy. One breakthrough in the application of reinforcement learning has been the successful use as a means to fine-tune the interaction with a given LLM through RLHF[101] as in ChatGPT. With the accelerated adoption of large DNNs in IO research variations on this approach have the potential to allow the research community as well as key stakeholders such as clinicians and other subject matter experts to be able to refine predictions of these models in an efficient manner.

One of the main challenges facing deployment of DNN systems in healthcare and biomedical research today is access to data. State-of-the-art DNN models with potentially billions of parameters need proportionally large datasets which may be more than a single institution or organization may have. At the same time concerns with privacy, security, computational constraints as well as intellectual property can limit use of a shared repository. As reviewed in Rieke *et al*,[102] the implementation of digital healthcare can use advances in federated learning to help address these concerns. Federated learning allows the sharing of DNN model optimization updates and parameters in order to learn a consensus model that performs better than models trained in isolation. This approach can further be leveraged in the biomedical sciences as reviewed in Kaissis *et al*.[103] For both healthcare and research, the use of federated learning alone is not guaranteed to address privacy and security concerns.

Aggregating data for secondary analysis/modeling can be challenging depending on the data types and sources. While AI/ML algorithms are often robust to individual issues (eg, different variable definitions/coding, mismatched distributions, diverse data types, missing data and class imbalance), biomedical data often is characterized as having many of these issues simultaneously.[104] AI/ML methods can be key as part of data cleaning and quality assurance (QA) / quality control (QC) to detect data anomalies and quality issues, as well as to facilitate mapping and transformation for integration. One such approach for automated concept mapping used DNNs to match data and coding at individual institutions to those of the observational medical outcomes partnership common data model.[105] Additionally, work carried out as part of the INCISIVE project provides a promising framework for unifying, harmonizing, and securely sharing scattered cancer-related data to aggregate the large datasets needed to develop and evaluate trustworthy AI models.[106]

While the amount of data is clearly a critical factor, issues with bias and fairness have reiterated that the greatest potential for impact is when the data are generated. This has led to efforts focused on developing the tools and approaches to support prospective data collection. For example, the National Institutes of Health (NIH) Common Fund Bridge2AI project[107] is focused on generating flagship biomedical data sets that are ethically sourced, well curated and accessible. The adoption of data generation best practices from these efforts in the key cancer consortia such as the Cancer Moonshot Immuno-oncology translational network[108] and others will provide invaluable data for realizing the potential of AI/ML in IO.

Acknowledgements The authors thank Guest Editor Dr. Alan Hutson and the Journal for ImmunoTherapy of Cancer for the invitation to contribute to this series.

Contributors DB and SM jointly developed the outline for the paper. DB wrote the initial draft. SM provided additional text and references, as well as final editing. All authors read and approved the final manuscript.

Funding This study was funded by NCI 2U54CA224019-05 and NIH Common Fund 3OT2OD032644-01S1.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iD

Shannon McWeeney http://orcid.org/0000-0001-8333-6607

REFERENCES

1 Ribas A, Wolchok JD. Cancer Immunotherapy using checkpoint blockade. *Science* 2018;359:1350–5.
2 June CH, O'Connor RS, Kawalekar OU, *et al*. CAR T cell Immunotherapy for human cancer. *Science* 2018;359:1361–5.
3 Hoos A, Britten CM. The Immuno-oncology framework. *OncoImmunology* 2012;1:334–9.
4 Hegde PS, Chen DS. Top 10 challenges in cancer immunotherapy. *Immunity* 2020;52:17–35.
5 Sharma P, Hu-Lieskovan S, Wargo JA, *et al*. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell* 2017;168:707–23.
6 Sanmamed MF, Berraondo P, Rodriguez-Ruiz ME, *et al*. Charting Roadmaps towards novel and safe synergistic immunotherapy combinations. *Nat Cancer* 2022;3:665–80.
7 Berraondo P, Sanmamed MF, Ochoa MC, *et al*. Cytokines in clinical cancer immunotherapy. *Br J Cancer* 2019;120:6–15.
8 Vaishnav R, Arslan W, Mehta DG. Unforeseen consequences of cancer immunotherapy. *Int J Mol Immuno Oncol* 2018;3:20.
9 Adashek JJ, Kato S, Ferrara R, *et al*. Hyperprogression and immune checkpoint inhibitors: Hype or progress *Oncologist* 2020;25:94–8.
10 Gubin MM, Vesely MD. Cancer immunoediting in the era of immuno-oncology. *Clin Cancer Res* 2022;28:3917–28.
11 Capobianco E. High-dimensional role of AI and machine learning in cancer research. *Br J Cancer* 2022;126:523–32.
12 Yang Y, Zhao Y, Liu X, *et al*. Artificial intelligence for prediction of response to cancer immunotherapy. *Semin Cancer Biol* 2022;87:137–47.
13 Gao Q, Yang L, Lu M, *et al*. The artificial intelligence and machine learning in lung cancer immunotherapy. *J Hematol Oncol* 2023;16:55.
14 Damane BP, Mkhize-Kwitshana ZL, Kgokolo MC, *et al*. Applying artificial intelligence prediction tools for advancing precision oncology in immunotherapy: future perspectives in personalized care. In: Dlamini Z, ed. *Artificial Intelligence and Precision Oncology: Bridging Cancer Research and Clinical Decision Support*. Cham: Springer Nature Switzerland, 2023: 239–58.
15 Li T, Li Y, Zhu X, *et al*. Artificial intelligence in cancer immunotherapy: applications in neoantigen recognition, antibody design and immunotherapy response prediction. *Semin Cancer Biol* 2023;91:50–69.
16 Kang CY, Duarte SE, Kim HS, *et al*. Artificial intelligence-based radiomics in the era of Immuno-oncology. *Oncologist* 2022;27:e471–83.
17 Bilal M, Nimir M, Snead D, *et al*. Role of AI and Digital pathology for colorectal immuno-oncology. *Br J Cancer* 2023;128:3–11.
18 Brunette ES, Flemmer RC, Flemmer CL. A review of artificial intelligence. 2009 4th International Conference on Autonomous Robots and Agents; Wellington. IEEE, 2009
19 Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press, 2016.
20 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
21 OpenAI. GPT-4 technical report [arXiv [cs.CL]]. 2023. Available: http://arxiv.org/abs/2303.08774
22 Christiano PF, Leike J, Brown T, *et al*. Deep reinforcement learning from human preferences. *Adv Neural Inf Process Syst [Internet]* 2017:30. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html
23 Lee P, Bubeck S, Petro J. Limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:1233–9.
24 Gartner. Gartner Hype cycle. Available: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle [Accessed 01 Aug 2023].
25 Steinert M, Leifer L. Scrutinizing Gartner's Hype cycle approach. In: *PICMET 2010 Technology Management For Global Economic Growth*. 2010: 1–13.
26 Sculley D, Holt G, Golovin D, *et al*. Hidden technical debt in machine learning systems. *Adv Neural Inf Process Syst* 2015;28.
27 John MM, Olsson HH, Bosch J. Towards Mlops: A framework and maturity model. 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA); Palermo, Italy. IEEE, 2021
28 Butcher J, Beridze I. What is the state of artificial intelligence governance globally. *RUSI J* 2019;164:88–96.
29 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
30 Bombard Y, Baker GR, Orlando E, *et al*. Engaging patients to improve quality of care: a systematic review. *Implement Sci* 2018;13:98.
31 Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. *Eur J Hum Genet* 2023.
32 Patel SB, Lam K. ChatGPT: the future of discharge summaries. *Lancet Digit Health* 2023;5:e107–8.
33 Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. *Nat Med* 2023;29:1930–40.
34 Khanna NN, Maindarkar MA, Viswanathan V, *et al*. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. *Healthcare (Basel)* 2022;10:2493.
35 Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother* 2013;4:S73–7.
36 Ball R, Dal Pan G. "Artificial intelligence" for pharmacovigilance: ready for prime time. *Drug Saf* 2022;45:429–38.
37 Kucukosmanoglu A, Scoarta S, Wijnands T, *et al*. Abstract 6312: the adverse events Atlas, towards a strategy to predict synergistic adverse events of combination therapies. *Cancer Res* 2022;82:6312.
38 Wong A, Otles E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065–70.
39 Coombs L, Orlando A, Wang X, *et al*. A machine learning framework supporting prospective clinical decisions applied to risk prediction in oncology. *NPJ Digit Med* 2022;5:117.
40 Wu Y, Zhu W, Wang J, *et al*. Using machine learning for mortality prediction and risk stratification in atezolizumab-treated cancer patients: integrative analysis of eight clinical trials. *Cancer Med* 2023;12:3744–57.
41 Zhou Z, Guo W, Liu D, *et al*. Multiparameter prediction model of immune checkpoint inhibitors combined with chemotherapy for non-small cell lung cancer based on support vector machine learning. *Sci Rep* 2023;13:4469.
42 Harris AH. Path from predictive analytics to improved patient outcomes: a framework to guide use, implementation, and evaluation of accurate surgical predictive models. *Ann Surg* 2017;265:461–3.
43 Amarasingham R, Patzer RE, Huesch M, *et al*. Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs* 2014;33:1148–54.
44 He J, Baxter SL, Xu J, *et al*. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30–6.
45 Baxter SL, Bass JS, Sitapati AM. Barriers to implementing an artificial intelligence model for unplanned readmissions. *ACI Open* 2020;4:e108–13.
46 Kelly CJ, Karthikesalingam A, Suleyman M, *et al*. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.
47 Baxi V, Edwards R, Montalto M, *et al*. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod Pathol* 2022;35:23–32.
48 Morad G, Helmink BA, Sharma P, *et al*. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell* 2021;184:5309–37.
49 Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning*. New York: Springer,
50 Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;11:733–9.
51 Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng* 2022;6:1346–52.
52 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [arXiv [cs.LG]]. 2016. Available: http://arxiv.org/abs/1511.06434
53 Zhang Y, Gan Z, Carin L. Generating text via adversarial training. NIPS workshop on Adversarial Training; 2016. 21–32.
54 Wali A, Alamgir Z, Karim S, *et al*. Generative adversarial networks for speech processing: a review. *Comput Speech Lang* 2022;72:101308.

55 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, et al., eds. *Advances in Neural Information Processing Systems*. Curran Associates, Inc, 2014. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

56 Quiros AC, Murray-Smith R, Yuan K. Pathologygan: learning deep representations of cancer tissue [arXiv [eess.IV]]. 2021. Available: http://arxiv.org/abs/1907.02644

57 Jose L, Liu S, Russo C, *et al*. Generative adversarial networks in digital pathology and histopathological image processing. *J Pathol Inform* 2021;12:43.

58 Lopez R, Regier J, Cole MB, *et al*. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;15:1053–8.

59 Fallahzadeh R, Bidoki NH, Stelzer IA, *et al*. In-Silico generation of high-dimensional immune response data in patients using a deep neural network. *Cytometry A* 2023;103:392–404.

60 Nerurkar SN, Goh D, Cheung CCL, *et al*. Transcriptional spatial profiling of cancer tissues in the era of immunotherapy: the potential and promise. *Cancers (Basel)* 2020;12:2572.

61 Long Y, Ang KS, Li M, *et al*. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nat Commun* 2023;14:1155.

62 O'Donnell JC, Le TK, Dobrin R, *et al*. Evolving use of real-world evidence in the regulatory process: a focus on Immuno-oncology treatment and outcomes. *Future Oncol* 2021;17:333–47.

63 Devlin J, Chang M-W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding [arXiv [cs.CL]]. 2018. Available: http://arxiv.org/abs/1810.04805

64 Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training; 2018.

65 Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;33:1877–901.

66 Bommasani R, Hudson DA, Adeli E, *et al*. On the opportunities and risks of foundation models [arXiv [cs.LG]]. 2021. Available: http://arxiv.org/abs/2108.07258

67 Liévin V, Hother CE, Winther O. Can large language models reason about medical questions? [arXiv [cs.CL]]. 2023. Available: http://arxiv.org/abs/2207.08143

68 Agrawal M, Hegselmann S, Lang H, *et al*. Large language models are few-shot clinical information extractors [arXiv [cs.CL]]. 2022. Available: http://arxiv.org/abs/2205.12689

69 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge [arXiv [cs.CL]]. 2022. Available: http://arxiv.org/abs/2212.13138

70 Taylor R, Kardas M, Cucurull G, *et al*. Galactica: a large language model for science [arXiv [cs.CL]]. 2022. Available: http://arxiv.org/abs/2211.09085

71 Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36:1234–40.

72 Luo R, Sun L, Xia Y, *et al*. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;23:bbac409.

73 Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare* 2022;3:1–23.

74 Meta AI. Galactica Demo. n.d. Available: https://galactica.org/

75 AlKhamissi B, Li M, Celikyilmaz A, *et al*. A review on language models as knowledge bases. *arXiv* 2022.

76 Pan S, Luo L, Wang Y, *et al*. Unifying large language models and knowledge graphs: a roadmap. *arXiv* 2023.

77 Jacobson CA, Farooq U, Ghobadi A. Axicabtagene Ciloleucel, an anti-CD19 Chimeric antigen receptor T-cell therapy for relapsed or refractory large B-cell lymphoma: practical implications for the community oncologist. *Oncologist* 2020;25:e138–46.

78 Fu S, Wen A, Liu H. Clinical natural language processing in secondary use of EHR for research. In: Richesson RL, Andrews JE, Fultz Hollis K, eds. *Clinical Research Informatics*. Cham: Springer International Publishing, 2023: 433–51.

79 Yang J, Lian JW, Chin Y-PH, *et al*. Assessing the prognostic significance of tumor-infiltrating lymphocytes in patients with Melanoma using pathologic features identified by natural language processing. *JAMA Netw Open* 2021;4:e2126337.

80 Liu X, Hersch GL, Khalil I, *et al*. Clinical trial information extraction with BERT. 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI); Victoria, BC, Canada. IEEE, 2021

81 Liu X, Shi C, Deore U, *et al*. A Scalable AI approach for clinical trial cohort optimization. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Springer International Publishing, 2021: 479–89.

82 Yang X, Chen A, PourNejatian N, *et al*. A large language model for electronic health records. *NPJ Digit Med* 2022;5:194.

83 Reese JT, Danis D, Caulfied JH, *et al*. On the limitations of large language models in clinical diagnosis. *Health Informatics* [Preprint] 2023.

84 The Cancer Genome Atlas program (TCGA). CCG - National Cancer Institute. 2022. Available: https://www.cancer.gov/ccg/research/genome-sequencing/tcga [Accessed 28 Jul 2023].

85 Bagaev A, Kotlov N, Nomie K, *et al*. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell* 2021;39:845–65.

86 Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010;22:1345–59.

87 Kim HE, Cosa-Linan A, Santhanam N, *et al*. Transfer learning for medical image classification: a literature review. *BMC Med Imaging* 2022;22:69.

88 Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;177:1888–1902.

89 Lieberman Y, Rokach L, Shay T. Classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* 2018;13:e0205499.

90 Han X, Zhou Z, Fei L, *et al*. Construction of a human cell landscape at single-cell level. *Nature* 2020;581:303–9.

91 Lotfollahi M, Naghipourfar M, Luecken MD, *et al*. Mapping single-cell data to reference Atlases by transfer learning. *Nat Biotechnol* 2022;40:121–30.

92 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Adv Neural Inf Process Syst* 2017:30. Available: https://proceedings.neurips.cc/paper/7181-attention-is-all

93 Theodoris CV, Xiao L, Chopra A, *et al*. Transfer learning enables predictions in network biology. *Nature* 2023;618:616–24.

94 Cui H, Wang C, Maan H, *et al*. scGPT: Towards building a foundation model for single-cell multi-omics using generative AI. *bioRxiv* [Preprint] 2023.

95 Shen H, Shen X, Hu J, *et al*. Generative pretraining from large-scale transcriptomes: implications for single-cell Deciphering and clinical translation. *bioRxiv* [Preprint] 2022.

96 Kipkogei E, Arango Argoty GA, Kagiampakis I, *et al*. Explainable transformer-based neural network for the prediction of survival outcomes in non-small cell lung cancer (NSCLC). *Oncology* [Preprint] 2021.

97 Eckardt J-N, Wendt K, Bornhäuser M, *et al*. Reinforcement learning for precision oncology. *Cancers (Basel)* 2021;13:4624.

98 Yang C-Y, Shiranthika C, Wang C-Y, *et al*. Reinforcement learning strategies in cancer chemotherapy treatments: a review. *Comput Methods Programs Biomed* 2023;229:107280.

99 Chen L, Zhang Y-W, Zhang S-C. Optimal drug dosage control strategy of immune systems using reinforcement learning. *IEEE Access* 2023;11:1269–79.

100 Chen Z, Min MR, Guo H, *et al*. T-cell receptor optimization with reinforcement learning and mutation polices for precision immunotherapy. Springer Nature Switzerland; 2023. 174–91.

101 Ziegler DM, Stiennon N, Wu J, *et al*. Fine-tuning language models from human preferences [arXiv [cs.CL]]. 2020. Available: http://arxiv.org/abs/1909.08593

102 Rieke N, Hancox J, Li W, *et al*. The future of digital health with federated learning. *NPJ Digit Med* 2020;3:119.

103 Kaissis GA, Makowski MR, Rückert D, *et al*. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell* 2020;2:305–11.

104 Martínez-García M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. *Front Med* 2021;8:784455.

105 Kang B, Yoon J, Kim HY, *et al*. Deep-learning-based automated terminology mapping in OMOP-CDM. *J Am Med Inform Assoc* 2021;28:1489–96.

106 Kosvyra A, Filos D, Fotopoulos D, *et al*. Towards data integration for AI in cancer research. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); IEEE, 2021

107 Bridge to artificial intelligence (Bridge2Ai). 2020. Available: https://commonfund.nih.gov/bridge2ai [Accessed 01 Aug 2023].

108 Annapragada A, Sikora A, Bollard C, *et al*. Cancer moonshot immuno-oncology translational network (IOTN): accelerating the clinical translation of basic discoveries for improving immunotherapy and immunoprevention of cancer. *J Immunother Cancer* 2020;8:e000796.