EBioMedicine

Published by THE LANCET

Review

Looking beyond the hype: Applied AI and machine learning in translational medicine

Tzen S. Toh [a], Frank Dondelinger [b], Dennis Wang [c,d,*]

[a] The Medical School, University of Sheffield, Sheffield, UK
[b] Lancaster Medical School, Furness College, Lancaster University, Bailrigg, Lancaster, UK
[c] NIHR Sheffield Biomedical Research Centre, University of Sheffield, Sheffield, UK
[d] Department of Computer Science, University of Sheffield, Sheffield, UK

ARTICLE INFO

ABSTRACT

Big data problems are becoming more prevalent for laboratory scientists who look to make clinical impact. A large part of this is due to increased computing power, in parallel with new technologies for high quality data generation. Both new and old techniques of artificial intelligence (AI) and machine learning (ML) can now help increase the success of translational studies in three areas: drug discovery, imaging, and genomic medicine. However, ML technologies do not come without their limitations and shortcomings. Current technical limitations and other limitations including governance, reproducibility, and interpretation will be discussed in this article. Overcoming these limitations will enable ML methods to be more powerful for discovery and reduce ambiguity within translational medicine, allowing data-informed decision-making to deliver the next generation of diagnostics and therapeutics to patients quicker, at lowered costs, and at scale.

Contents

1. Introduction

Artificial intelligence (AI) and machine learning (ML) technologies have begun to change how we deliver healthcare. These disruptive

* Corresponding author at: NIHR Sheffield BRC, University of Sheffield, 385a Glossop Road, Sheffield S10 2HQ, UK.
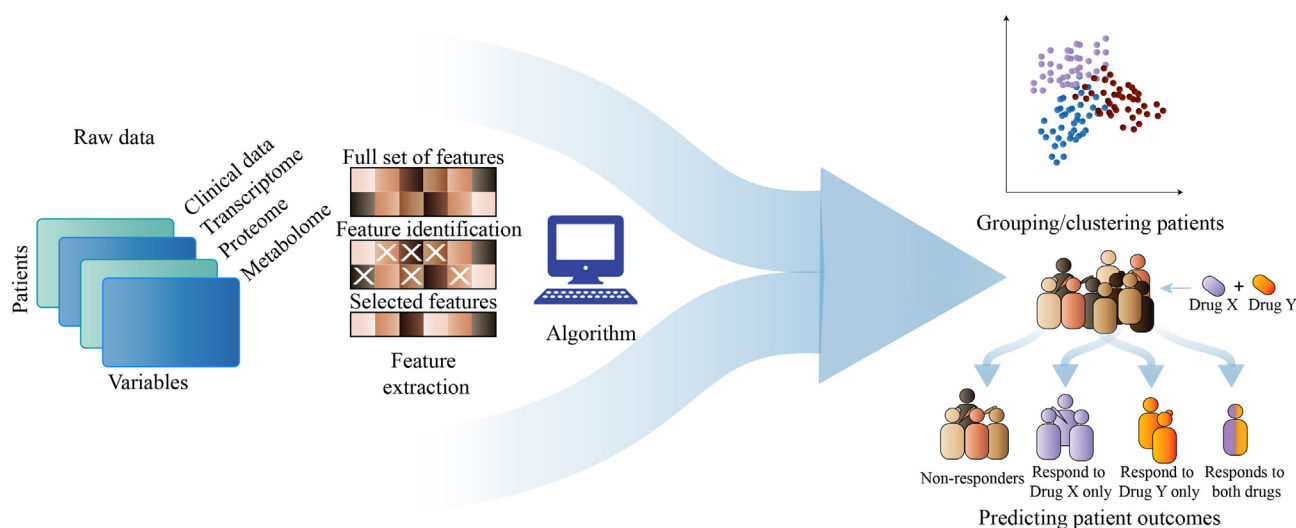   E-mail address: dennis.wang@sheffield.ac.uk (D. Wang).

Fig. 1. Process of AI/ML in translational medicine. A number of high-throughput assays generate data from many patient samples. Datasets are then structured into machine-readable format and potentially important variables are identified using an ML algorithm. The algorithm will learn relationships between the variables and may perform intelligent tasks such as grouping patients or predicting their outcomes.

computational methodologies are able to transform the way clinicians help patients in the clinic and also enhance the way scientists develop new treatments and diagnostics in the laboratories [1–3]. Now that biomedical scientists in academia and biotechnology industries frequently use high-throughput technologies such as parallelised sequencing, microscopy imaging, and compound screening, there has been a rapid expansion in the volume and quality of laboratory data being generated. Together with advancing techniques in ML, we are increasingly more capable of deriving biological insight from this "big data" to understand the mechanisms of disease, identify new therapeutic strategies, and improve diagnostic tools for clinical application.

The term AI is often used these days to denote ML techniques, so it is worth defining both terms. Here, we adopt the US Food and Drug Administration (FDA) definition, which describes AI as "the science and engineering of making intelligent machines", while ML is "an artificial intelligence technique that can be used to design and train software algorithms to learn from and act on data" [4]. It follows that all ML techniques are AI techniques, but not all AI techniques are ML techniques. In this article, we will mainly be concerned with ML techniques, as these are most relevant in the context of translational medicine. Both simple and advanced ML methods may enhance the capability of translational scientists who use them to develop new treatments and diagnostics for healthcare. Generally speaking, there are two main types of ML approaches we will focus on; 1) unsupervised learning of data to find previously unknown patterns or grouping labels for the patient samples; and 2) supervised learning from data with labels in order to make predictions on new samples (Fig. 1). We will describe specific ML methods that have been heavily promoted or hyped and those that are less well known but can also learn from a variety of large data sets generated in the laboratory and can be used to perform intelligent tasks that are difficult for human scientists. Specific technical challenges with applying each ML method (Table 1) will be discussed and all technical terms used in this review are defined in the Supplementary Appendix.

2. Drug discovery

Despite advances in technology and improved understanding of human biology, the process of drug discovery remains capital-intensive, tedious, and lengthy. The total cost of bringing a compound from the bench to bedside has been increasing over time, with the bulk of costs coming from Phase 3 clinical trials. The clinical success probability of bringing a drug through Phase 1 to approval was

estimated to be 11.83%, which shows the inherently risky nature of drug development [5].

After identifying a target, key steps of designing a suitable chemical compound may be broken down into two processes, the retrosynthetic process and the formulation of a well-motivated hypothesis for new lead compound generation (*de novo* design) [6–9]. Retrosynthesis involves planning the synthesis of small organic molecules where target molecules are transformed recurrently by a search tree 'working backwards' into simpler precursors until a set of known molecules are gained. Each retrosynthetic step involves the selection of the most promising transformations from many known transformations by chemists, but there is no guarantee that the reaction will proceed as expected [9]. *De novo* design involves exploring compound libraries, estimated to contain >10^{30} molecules. Due to the vast size of compound libraries, automated screening of selected compounds with the required properties and possibility of activity presents an opportunity to apply ML to navigate these libraries [7,8]. Here, we review ML methods that help to design, synthesise, and prioritise new compounds before clinical testing.

2.1. Designing chemical compounds

Traditional computer-aided retrosynthesis is slow and does not provide results of high quality. Advances in ML are shifting the generation of chemistry rules from hand-coded heuristics, to autonomous systems that take advantage of a vast volume of available chemistry in assisting chemical synthesis. Chemists have been able to train deep neural networks (DNNs) (Fig. 2) on all available reactions published in organic chemistry to create a system that is thirty times quicker than traditional computer-aided methods [9].

New chemicals based on pharmacologically active natural compounds have been generated through a deep recurrent neural network (RNN) (Fig. 2a). A small set of known bioactive templates was sufficient to generate isofunctional new chemical entities [6]. Using a deep learning model based on RNNs to generate synthetic data shows efficacy in three applications of *de novo* drug design: generating libraries for high-throughput screening, hit-to-lead optimisation, and fragment-based hit discovery. Unlike traditional approaches for generating candidate compounds, this approach does not require similarity searching or external scoring and new molecular structures are generated immediately, which is useful for *in situ* real-time molecular modeling [7].

De novo drug design with desired properties may also be achieved with deep reinforcement learning. Unlike standard deep learning, where the neural network is trained once on a set of training data, in

Table 1

Recent ML tools and applications in various aspects of translational medicine with the key results and challenges faced by each application of ML.

Category	Application	ML technique(s)	Key result(s) or Advantages(s)	Challenge(s)
Drug discovery	Designing chemical compounds (retrosynthetic process)	Deep neural networks (DNNs) and Monte Carlo tree search [9] [70]	30× quicker than traditional computer-aided methods	1. Scarcity of training data 2. Stronger, but slower-reasoning, algorithms should be developed for this application
	Designing chemical compounds (*de novo* drug design)	Deep recurrent neural network (RNN) [6]	Generate isofunctional, new chemical entities	Appropriate predicted bioactivity which has been validated is required
		Generative deep learning (based on RNNs) [7]	Does not require similarity searching or external scoring and new molecular structures are generated immediately	User has to make a decision on when training should be stopped
		Reinforcement Learning for Structural Evolution (ReLeaSE - 2 DNNs, generative and predictive) [8]	Simpler to use compared to traditional methods	Only available for a single-task regime - development to extend to optimise several target properties together is required
	Drug screening	Random Forest and ChemVec [11]	Highest accuracy when compared to 3 other algorithms	1. Improve feature representation using deep learning 2. Experimental validation required
Imaging	Cell microscopy and histopathology	Bayesian matrix factorisation method, Macau [25]	Predictive performance comparable with that of DNNs	1. Current results for this method are based on a single HTI screen 2. Requires an adequate sized library of compound for training the model
		Gradient Boosting [27]	Reduction of disturbances to the cells, making sample preparation quicker and cheaper	Deep learning techniques should be tried to improve the model
	Defining relationships between morphology and genomic features	Inception v3 (based on convolutional neural networks) [22]	Capable of distinguishing between 3 types of histopathological images, predicting mutational status of 6 genes	Current data may not fully represent the heterogeneity of tissues
Genomic medicine	Biomarker discovery	Elastic net regression [33]	Identification of *BRAF* and *NRAS* mutations in cell lines, were among the top predictors of drug sensitivity for a MEK inhibitor	Technique does not allow for the comparison between drugs
		Unsupervised hierarchical clustering (part of ACME analysis) [30]	Identified associations between *BRAF* mutant cell lines of the skin lineage being sensitive to the MEK inhibitor	1. Distance metric and linkage criteria must be specified 2. Does not scale well
		Spectral clustering by Similarity Network Fusion (SNF) [34]	Identification of new tumour subtypes by utilising mRNA and methylation signatures	Prospective studies required to determine accuracy
	Integrating different modalities of data	iCluster [44]	Identified potentially novel subtypes of breast and lung cancers on top of subgroups characterised by concordant DNA copy number alterations and gene expression in an automated way	Only focuses on array data
		Kernel Learning Integrative Clustering (KLIC) [46]	Compared to Cluster-Of-Cluster Analysis (COCA), KLIC adds more detailed information about data from each dataset into the last clustering step and is able to merge datasets having various levels of noise, giving more weight to more significant ones	Only tested on simulated datasets
		Spectral clustering by SNF [43]	Identification of new medulloblastoma subtypes	1. Larger cohort size for validation 2. Current analysis of samples is bulk analysis
		Affinity Network Fusion (ANF) and semi-supervised learning [47]	Performs similarly or better when compared to SNF, less computationally demanding, generalises better	Results on four cancer types only (known disease types) and not yet validated on additional experimental data
		Clusternomics [48]	Outperforms existing methods and derived clusters with clinical meaning and significant differences in survival outcomes when tested on real-world data [44,49–51]	Comparison of performance to other methods on real-world data

reinforcement learning, the neural network continues learning with the help of a critic, *i.e.* a function that evaluates the utility of the latest output of the neural network. ReLeaSE (Reinforcement Learning for Structural Evolution), integrates two DNNs; one for generating novel chemical libraries, and one for evaluating their utility based on quantitative structure-property relationships (QSPRs). Reinforcement learning allows the first DNN to keep improving the generated libraries based on feedback from the second. This method uses chemical structures represented by SMILES (simplified molecular-input line-entry system) strings only, rather than calculating chemical descriptors. This differentiates ReLeaSE from traditional methods, making it simpler to use [8].

A major bottleneck for using DNNs for computer-aided retrosynthesis in the synthesis of natural products is the scarcity of training data for this application. Natural product synthesis could possibly be solved with stronger, but slower-reasoning, algorithms for

inventing reactions [9]. Improvements in algorithm development should come hand-in-hand with high quality training data to further improve this method.

For methods in computational *de novo* drug design, there is still a need to score and rank the generated designs. Meaningful selection of compounds from a set of new generated molecules require an appropriate predicted bioactivity which has been validated [6]. While a deep learning model based on RNNs may circumvent this limitation, available active ligands for parameter optimisation are still required. Fine-tuning of this model over a certain number of epochs is required to avoid the generation of duplicates. However, the user has to make a decision on when training should be stopped, suggesting that objective rules to fine-tune the model will be required to further improve this method [7]. Based on the example above, the ReLeaSE method currently only allows for a single-task regime,
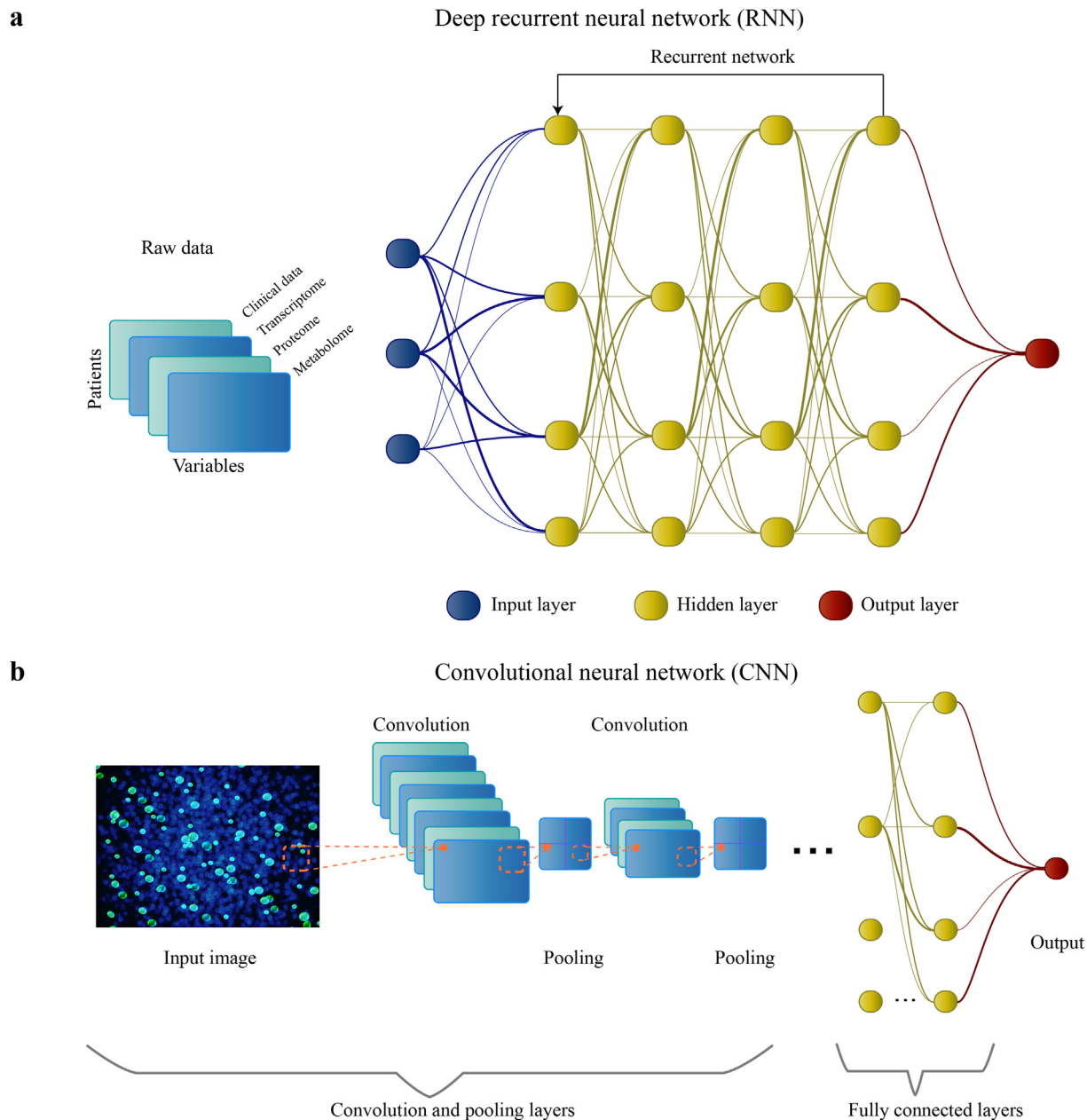
Fig. 2. Summary of conventional deep neural networks (DNNs) and their architectures. DNNs are powerful algorithms based loosely on the human brain. Each node ('neuron') in the neural network can only perform simple calculations, but DNNs work by connecting the nodes to form a number of layers, where each layer performs a calculation based on the output of the previous layer. In this way, the DNN can perform tasks that are much more complex than what a single node could achieve. The 'deep' in DNNs refers to the fact that the number of layers can be very large. In some cases, DNNs outperform other ML algorithms due to their capability of finding patterns in vast amounts of unstructured data. As the number of nodes and layers implies a large number of parameters that need to be learned, DNNs often require large amounts of data to achieve top performance. (a) A type of DNN, called a deep recurrent neural network (RNN), consists of an input layer, numerous hidden layers corresponding to each feature of the input data, and an output layer. The input will consist of a dataset in machine-readable format, and information is propagated forward through the layers to provide an output. The thickness of the lines indicates the weight of the connections entering and leaving the nodes. RNNs use feedback loops throughout the process of computation. RNNs are particularly useful for modeling a sequence of data (eg. designing chemical compounds) because the feedback loops allow the network to effectively retain a `memory` of inputs it has seen previously [7,8]. (b) Inspired by the visual cortex, convolutional neural networks (CNNs) allow for more effective processing of the complexity of a given input (eg. raw images); this is done by transforming them into simpler forms, without the loss of important features for a prediction. In the first convolution layer, filters are applied to the input to create low-level convoluted features (eg. edges of images). Convolution can be thought as matrix multiplication between the filter (kernel) with subsets of the input data where the filter is overlapping. Next, the pooling layer reduces the sizes of convoluted features and declutters it to select important features for training the model. The number of convolution and pooling layers is dependent on the complexity of the data; more computation is required as the number of layers increases. The final layers consist of one or more fully connected layers (to perform computations on the extracted features) and the output layer (usually the result of the prediction or classification).

prompting future development to optimise several target properties together. This is required for drug discovery where the compound would need to be optimised based on characteristics such as potency, selectivity, solubility, and drug similarity [8]. AstraZeneca were able to use RNNs to expand their portfolio by optimising the bioactivity, pharmacokinetic properties, and solubility of their new compounds [10].

2.2. Drug screening

Drug repositioning involves evaluating a library of approved drugs or drugs in trials for application in a new disease indication [11]. Pharmaceutical companies maintain and assess millions of unique compounds as potential treatments for a variety of diseases [12,13]. The high number of permutations for comparing compounds across disease

indications pose a significant computational and experimental challenge, and even more so for combination treatments [14,15]. The chances of predicting the probability of compounds being repositioned may be improved by the integration of molecular and chemical information from the compound and diseases [16].

Feature engineering can be applied to chemical information to identify important features needed for prioritising drugs for clinical investigation. From RepurposeDB and DrugBank, chemical descriptors of the small molecules (*i.e.* quantitative structure activity relationship parameters, electronic, topological, geometrical, constitutional and hybrid) were integrated by algorithms like ChemVec [17,18]. The combination of features from different molecular features determines a framework representing the chemical space of drug molecules used in this study [11]. To fully extend the utility of this method, the authors have proposed to include datasets from PubChem in the future to improve feature representation using deep learning [11]. While the proposed direction for improvement would require large volumes of data, good quality data is often equally or more critical to ensure optimal performance.

After the compilation of molecular features with ChemVec, Random Forest (RF) was utilised to predict the suitability of small molecules as candidates for drug repositioning. It was suggested that RF had the highest accuracy rate when compared to other algorithms including Naïve Bayes, Support Vector Machines (SVM), and the Ensemble Model (combining the stated 3 models) [11]. Studies that benchmark different algorithms for single and paired treatments have highlighted the need to include genomic information [19,20]. Although high accuracy is reported in these studies, there has been limited clinical validation of the ML predicted treatments partly because due to their translatability from *in vitro* to *in vivo*. Nevertheless, the predicted genetic biomarkers for treatment response, such as those for drugs targeting the AKT protein, have been shown to be useful in the clinical setting.

3. Imaging

Opportunities are already available to apply ML to assist researchers in cell image analysis with the increasing availability of high quality volumes of cell imaging data. This is particularly evident in cell microscopy and histopathology. Unraveling disease heterogeneity by improving cellular profiling of certain morphology characteristics is becoming more feasible with the help of ML. Here, we focus on promising ML techniques that can be applied to aid the classification of samples and even extract new disease features which evade the human eye.

3.1. Cell microscopy and histopathology

Diagnosis by histopathology is fraught with inconsistency in the interpretation of slides by pathologists [21]. Traditional techniques like the microscopic examination of a specimen are limited and do not allow for the detection of specific genomic driver mutations and patterns within the subcellular structure of cells [22,23]. Biomarker detection in the clinical laboratories also typically requires specific reagents, specialised equipment, and/or complicated laboratory techniques, which might disturb the original state of biological specimens [24]. This presents an opportunity to leverage advances in ML for label-free cell-based diagnostics to highlight disease states that cannot be identified by humans alone.

High-throughput imaging (HTI) allows researchers to resolve cell morphology by microscopy. It is frequently used to screen compounds based on alterations to cell morphology by a particular drug. An untrained eye may not fully leverage the entirety of morphological data from images, leaving behind a relatively large untapped reservoir of biological insight [25]. Rich data from a HTI assay could be repurposed for the prediction of biological activity of drugs in other assays, including predicting those targeting a different pathway or biological process. A

Bayesian matrix factorisation method was revealed to yield a predictive performance comparable with that of DNNs. Both methods can successfully repurpose the glucocorticoid HTI assay to predict activity of >30 unrelated protein targets [25]. This approach suggests the justification of repurposing high-throughput imaging assays for the drug discovery process.

Current results for this method are based on a single HTI screen but future studies that fuse data across multiple HTI screens could be more capable for prediction. This method relies on a supervised ML method which requires an adequate sized library of compounds to train the model. Convolutional neural networks (CNNs) (Fig. 2b) could potentially predict activity directly from raw images, eliminating the need to extract features from each cell [26].

Unstained white blood cells (WBCs) were classified using a label-free approach with imaging flow cytometry and various ML algorithms (AdaBoost, Gradient Boosting, K-Nearest Neighbors, RF, and SVM) [27]. Imaging flow cytometry combines the sensitivity and high-content morphology of digital microscopy with the high-throughput and statistical power of flow cytometry. Among those algorithms, the Gradient Boosting algorithm is best able to classify the WBCs into their subtypes. This approach demonstrates the reduction of disturbances to the cells, and makes sample preparation quicker and cheaper, leading to potential application of label-free identification of cells in the clinic [24].

In the future, applying other ML techniques such as deep learning may help to improve the model. Deep learning algorithms are potentially better suited for discerning cellular images as they allow learning of features directly from the vast amounts of raw data. Using unsupervised ML to cluster subpopulations of cells could aid interpretation of similarity among clustered cells [27]. Researchers from Google AI have been benchmarking CNNs for routine Gleason scoring of prostate cancer cases and comparing them to actual classifications by qualified pathologists [28]. The algorithm's diagnostic accuracy of 0.70 was far superior than the mean accuracy of 0.61 among pathologists, so we may see a significant improvement in the clinical diagnosis of common cancers in the near future.

3.2. Defining relationships between morphology and genomic features

Physical differences between cancer cell types are often subtle, especially in poorly differentiated tumours. Advanced learning techniques for image analysis may help identify existing patterns in cell morphology that relate to biological processes [22,24]. A technique based on CNNs, Inception v3, was capable of distinguishing between three types of histopathological images (normal tissue, lung adenocarcinoma, and lung squamous cell carcinoma). CNNs were also capable of predicting the mutational status of six genes (*STK11, EGFR, FAT1, SETBP1, KRAS, TP53*) from lung adenocarcinoma whole-slide images [22].

As whole-slide images may not fully represent the heterogeneity of tissues, images containing more features would be required to train the model for improved performance. Continuing work to extend this method to other histological subtypes and consider features in the tumour microenvironment will better help researchers through a semi-automated approach [22]. In particular, Steele et al. were able to translate knowledge from histopathologists into rule-based algorithms that can measure immune cell infiltrate to improve the success of immunotherapy durvalumab in Phase 1/2 studies [29].

4. Genomic medicine

In order to ensure targeted treatments can be delivered to patients, biomarkers of disease are essential to select patients who might best respond to treatment. Recent technological and analytical advances in genomics have allowed the identification and interpretation of genetic variants for cancers spanning almost all protein-coding genes of the human genome [30,31]. These variants could potentially act as biomarkers of disease, leading to targeting the mechanisms contributing

to disease which are unique to a patient [32]. However, finding meaningful biological insights into drivers of disease progression from the sheer number of potential biomarkers arising from variants is challenging for researchers to achieve manually.

ML techniques have helped to systematically define biomarkers driving diseases [30,33,34]. Integration between experimental and clinical data is required to help understand new disease subtypes, but remains a challenge due to the heterogeneity in both experimental and clinical data. Clinical data is often messy, with missing data points proving to be a challenge to integrate into an ML model [35]. Understanding new subtypes remains key for the application of delivering precise treatments to patients, which will hopefully improve clinical outcomes.

4.1. Biomarker discovery

The need for precision medicine is evident within oncology; cancers are more frequently categorised and treated based on their molecular profile. For melanoma patients with unresectable or metastatic tumours, ipilimumab is currently a recommended treatment option [36]. As the heterogeneity of tumours is difficult to define solely by the naked eye, ML presents a data-driven approach to account for various markers that drive tumour growth, including those present in a patient's molecular profile. Regression and clustering methods have been used to identify biomarkers that predict drug sensitivity [30,33,34]. Elastic net regression has identified features that are known predictors of drug sensitivity. For example, models including *BRAF* and *NRAS* mutations were among the top predictors of drug sensitivity in cell lines for a MEK inhibitor drug [33]. Unsupervised hierarchical clustering of AUC drug sensitivity measurements was able to identify sensitivity of *BRAF* mutant skin cancer cell lines to the MEK inhibitor selumetinib. This finding is expected and validates a known vulnerability observed in patients [30,37].

A particularly powerful approach to conduct biomarker discovery when you have information about samples from multiple sources is to use networks (Figs. 3a–c). An individual similarity network of samples

Fig. 3. Multi-omics analysis of samples and biomarkers using networks. (a) Multi-omic datasets across a group of samples. (b) Samples are first connected based on their similarity within each dataset (c) Similarity networks based on different datasets are fused by updating each network repeatedly with data from other networks. (d) Final fused network can be analysed to identify biomarkers that describe the similarity between groups of samples. The colour of the edges show the data type contributing to the similarity.

is constructed for each data type using the Euclidean distance, where the network nodes are patients and edges are the pairwise similarities for patients. Networks for each omics data type are combined by iteratively updating edges on the basis of the information from the individual networks until convergence, with the final fused network being clustered using spectral clustering (Fig. 3d) [34]. The identification of tumour subtypes in the The Cancer Genome Atlas (TCGA) study was based on the Illumina Infinium DNA Methylation 450 BeadChip, mRNA expression, and integrated mRNA and DNA methylation (iRM) [31,34,38]. Other advanced methods have been proposed for biomarker discovery, but they do not truly identify common deviations between multiple data types before analysis for enriched biomarkers begins [30,33,39].

A particular instance of the network clustering approach, Similarity Network Fusion (SNF), uses spectral clustering. It has shown a high level of correlation between the mRNA and methylation signatures to tumour types identified histologically, suggesting the application of mRNA and methylation profiling in determining tumour subtypes. This could be especially useful in situations where tumour histology or morphology is ambiguous. Future studies are required to determine the accuracy of classifier development as a tool for clinical cancer diagnostics [34]. Methods that integrate both clinical and new multi-omic datasets represents the next step forward to develop more targeted therapeutics suitable for individual patients [40,41]. While the impact of these multi-omic platforms are at the beginning of the drug discovery pipeline, a handful of new candidate compounds, like BRD-7880, have come as a result of their predictions [42].

4.2. Integrating different modalities of data

Unsupervised clustering approaches have been applied to various applications and data types [34,43,44]. Previously, researchers applied 'manual integration' after separate analysis of individual data types, which typically requires domain expertise and is subject to inconsistencies. A joint latent variable model generates an integrated cluster assignment based on simultaneous inference from multi-omics data (aligning DNA copy number and gene expression data), revealing potential new subgroups in breast and lung cancer [44]. However, this approach, called iCluster, focuses only on array data.

Another integrative clustering method first clusters each dataset individually before forming a binary matrix encoding the cluster allocations of each observation in each dataset [45]. A key limitation of this method, the Cluster-Of-Clusters Analysis (COCA) method, is that the algorithm output gives equal weight to datasets with poor quality or with unrelated cluster structures due to clustering structures from each dataset being unweighted [46]. To address this, a new method was introduced and it was able to add more detailed information about the data from each dataset into the last clustering step. It was also able to merge datasets having various levels of noise, giving more weight to datasets with more signal [46].

Using data integration with SNF and spectral clustering allowed integration of experimental and clinical data (763 primary frozen medulloblastoma samples), further enabling identification of 12 different subtypes of medulloblastoma. This was achieved after integrating somatic copy-number alterations and clinical data specific to each already known subtype of medulloblastoma. Integrative analysis of experimental and clinical data with this method further defines group 3 from group 4 medulloblastoma, which was not previously visible by analyses of individual data types [43]. A future study with a larger cohort size could assess the heterogeneity of the defined subgroups in greater detail. Bulk analysis of the samples are also another limitation to this current approach [43]. To further separate groups 3 and 4 medulloblastoma subgroups and potentially reveal subgroups with similar mechanisms and developmental origins, single-cell genomics can be applied to distinguish the full subclonal structure of the subgroup.
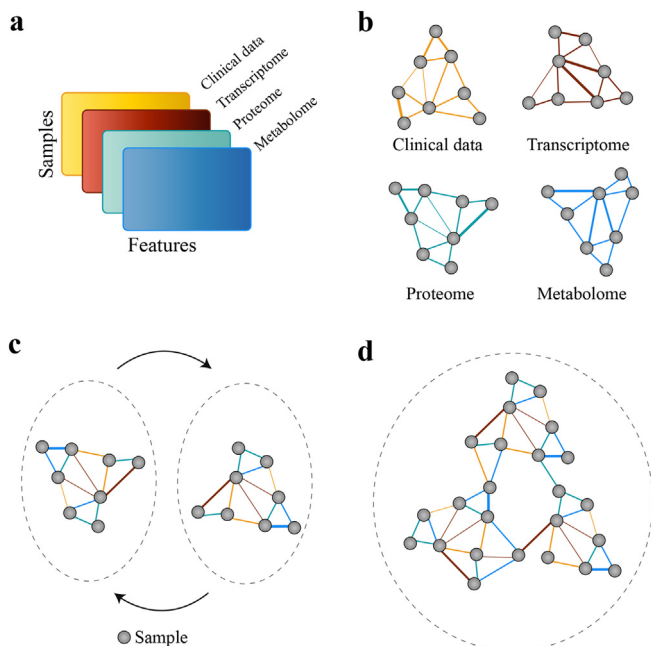
Further improvements in data aggregation allowed for the integration of multi-omic data for clustering cancer patients in combination with semi-supervised learning [47]. When benchmarked against a prior method like SNF, it performs similarly or better, is less computationally demanding, and generalises better. A semi-supervised model combining this method, called the Affinity Network Fusion (ANF) method and neural network for few-shot learning can achieve >90% accuracy on the test set with training <1% of the data, in several cases [47].

The methods described prior assume a common cluster structure across datasets. However, clusters do not always share a similar structure. Datasets with different cluster structures are considered distractions and this might result in missing biologically meaningful differences. A method was introduced as a Bayesian context-dependent model that identifies groups across datasets that do not share similarities in their cluster structure [48]. Using a Dirichlet mixture model, Clusternomics defines clustering at two levels: a local structure within each dataset and a global structure arising from combinations of dataset-specific clusters to model heterogeneous datasets without the same structure. It outperforms existing methods for clustering on a simulated dataset [44,49–51]. When tested on a breast cancer dataset from the TCGA (integrating gene expression, miRNA expression, DNA methylation, and proteomics), this method derived clusters with clinical meaning and significant differences in survival outcomes [31]. Evaluation upon lung and kidney cancer TCGA data showed clinically significant results and the ability of Clusternomics to scale [48].

While the integration of genomics and electronic health records has been shown to be effective for precision diagnostics of complex disorders, integrating laboratory and clinical data remains a challenge due to differences between preclinical models and patients [52]. Recently developed ML models can be used to transfer molecular predictors found in mouse models to humans [53,54]. More structured mathematical modeling rather than machine learning of *in vitro*, *in vivo*, and clinical data has been useful for identifying optimal doses that could be tested in patients [55]. Connecting laboratory results to clinical records can still be difficult with inconsistencies in nomenclature used by clinicians while ordering tests for a similar diagnosis. Better integrated information systems can not only accelerate healthcare AI but also significantly improve test utilisation practices, increase patient satisfaction, and improve health care processes [56].

5. Data collection, governance, reproducibility, and interpretability

There are general non-technical issues that are required to be addressed before mainstream application of ML within translational medicine takes place. Oftentimes, sensitive data is required to train ML algorithms. Access to data should be carefully regulated to ensure privacy without stifling innovation and technological advancement to improve outcomes [57]. A proposed scheme called privacy-preserving clinical decision with cloud support (PPCD) is an encouraging step in this direction [58].

Biases within the training datasets of ML algorithms need to be avoided to reduce the risk of failure of ML methods to generalise. Rethinking responsibility and accountability of individuals or organisations selecting datasets used to train ML algorithms are key to address this. Ethical frameworks should be developed by scientific committees and regulatory bodies to recognise and minimise the effect of biased models while guiding design choices to introduce systems that build trust, understanding, and maintaining individual privacy [57,59].

Reproducibility is another aspect that needs to be managed to ensure widespread adoption of ML in translational medicine. Caution should be exercised when drawing conclusions solely from large reservoirs of clinical data as it is often fraught with heterogeneity in quality [60]. Responsibly sharing data and code should be made requirements for authors alongside their publications to ensure trust in research findings. Reproducibility of results requires the software environment, source code, and raw data used during experiments [61].

Lastly, there is also a need to explain and easily interpret predictions of ML systems to implement them in clinical settings. Black-box algorithms may have good prediction accuracy but their predictions are difficult to interpret and are not actionable, hence limiting their clinical application [62]. Fortunately, new methods have been devised to allow researchers to interpret black-box algorithms to ensure their predictions are sensible [63–65].

6. Outstanding questions

Most of the methods described in this review unfortunately involve many steps and some of these may be beyond the abilities of the non-expert. In order for ML systems to be better integrated in a laboratory and clinical setting, a simpler end-to-end approach for applying ML needs to be available to biomedical scientists. One recent trend in AI is the development of automated machine learning (AutoML) techniques. AutoML seeks to make the process of developing an ML model more accessible by automating the difficult tasks of defining the structure and type of the model and selecting the hyperparameters [66]. Examples of AutoML frameworks include open-source frameworks such as TPOT and Auto-Keras, and industrial software suites that offer pre-trained models, such as Google's Cloud AutoML [67–69]. While this development would appear to be a boon in that it allows clinicians with limited ML knowledge to apply ML models to their datasets, we were not successful in identifying any current applications in translational medicine. Care should also be taken when applying these techniques, as automation may miss some of the subtleties and particular biases of medical data.

Since the goal of translational research is to see methods tested in the laboratory setting used in the clinic, computational methods will need to undergo the scrutiny of healthcare regulators. In the US, the FDA has recently proposed a new regulatory framework for medical devices (including software) that deals with the changes brought about by software employing AI and ML techniques. In particular, they propose that rather than having to review the software each time it changes, the premarket submission to the FDA would include a "predetermined change control plan" which describes the type of changes that could be made to the algorithm and how the risk to patients would be managed. We believe that this is a beneficial step that represents a reasonable trade-off between the burden of legislation and the potential risks to patients, and hope that other regulatory bodies will follow suit.

7. Conclusions

Throughout this review, we discussed key applications of ML in translational medicine. We also emphasised key bottlenecks and limitations of each application, as well as prerequisites needed to overcome them before these technologies may be applied effectively in research or clinical settings. Although the application of ML to translational medicine is met with great enthusiasm, we should collectively exercise caution to prevent premature rollout and the harm they might cause to patients secondary to any faults in these technologies.

New laboratory technologies applied to translational medicine must undergo sharing of results in peer-reviewed publications, rigorous testing, and real-world validation - ML algorithms applied in this domain should be held to the same high standards. While the temptation to apply various types of computation to large volumes of data grows, it is still of utmost importance to first consider the underlying biological question and the data required to answer it instead of blindly applying ML techniques to any dataset. Nevertheless, advancements in ML will bring truly objective methods to address the ambiguity observed within translational medicine, allowing for more robust, data-driven decision making to bring the next generation of diagnostic tools and therapeutics to patients.

Search strategy and selection criteria

Articles for this review were identified *via* PubMed and Google Scholar using the search terms "machine learning", "AI", "artificial intelligence", "deep learning", "deep neural networks", "de novo drug design", "drug development", "cell microscopy", "feature extraction", "histopathology", "biomarker development", "electronic health records", "data collection", "governance", and "reproducibility" on top of references from relevant articles. We also identified articles from the names of prominent investigators in the field. Articles published in English from 2016 onwards were included in the review. Exceptions to this include data integration methods (iCluster, Bayesian Consensus Clustering, Multiple Dataset Integration, SNF), datasets (TCGA, CTRP v2, DrugBank 4.0, GDSC), research articles to better give context and key statistics to the described challenges, and clinical trial findings. We included reference to four non-peer reviewed research articles, in *bioRxiv* and *arXiv*.

Author contributions

TST and DW designed the review outline, wrote the review, did the literature search, designed the figures, and carried out data extraction and interpretation. FD provided critical reviews, wrote the review, and interpreted the data. All authors read and approved the final version of the manuscript.

Declaration of competing interests

The authors declare no conflict of interest.

Acknowledgements

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.08.027.

References

[1] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24–9.

[2] Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886.

[3] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6.

[4] Artificial Intelligence and Machine Learning in Software as a Medical Device. US food and drug administration. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device#whatis; 2019. (accessed July 19, 2019).

[5] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 2016;47:20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

[6] Merk D, Grisoni F, Friedrich L, Schneider G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. Commun Chem 2018;1. https://doi.org/10.1038/s42004-018-0068-1.

[7] Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G. Generative recurrent networks for de novo drug design. Mol Inform 2018;37:1700111. https://doi.org/10.1002/minf.201700111.

[8] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci Adv 2018;4:eaap7885. https://doi.org/10.1126/sciadv.aap7885.

[9] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555:604–10.

[10] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Chem 2017;9:48.

[11] Shameer K, Johnson K, Glicksberg BS, Hodos R, Readhead B, Tomlinson MS, et al. Prioritizing small molecule as candidates for drug repositioning using machine learning; 2018. https://doi.org/10.1101/331975.

[12] Kogej T, Blomberg N, Greasley PJ, Mundt S, Vainio MJ, Schamberger J, et al. Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer pharma AG case. Drug Discov Today 2013;18:1014–24. https://doi.org/10.1016/j.drudis.2012.10.011.

[13] Engels MFM, Gibbs AC, Jaeger EP, Verbinnen D, Lobanov VS, Agrafiotis DK. A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. J Chem Inf Model 2006;46:2651–60.

[14] Schamberger J, Grimm M, Steinmeyer A, Hillisch A. Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. Drug Discov Today 2011;16:636–41.

[15] Di Veroli GY, Fornari C, Wang D, Mollard S, Bramhall JL, Richards FM, et al. Combenefit: an interactive platform for the analysis and visualization of drug combinations. Bioinformatics 2016;32:2866–8.

[16] Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Chem 2013;5:30.

[17] Shameer K, Glicksberg BS, Hodos R, Johnson KW, Badgeley MA, Readhead B, et al. Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. Brief Bioinform 2018;19:656–78. https://doi.org/10.1093/bib/bbw136.

[18] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014;42:D1091–7.

[19] Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 2014;32:1202–12.

[20] Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat Commun 2019;10:2674.

[21] Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.

[22] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67. https://doi.org/10.1038/s41591-018-0177-5.

[23] Ozaki Y, Yamada H, Kikuchi H, Hirotsu A, Murakami T, Matsumoto T, et al. Label-free classification of cells based on supervised machine learning of subcellular structures. PLoS One 2019;14:e0211347.

[24] Doan M, Carpenter AE. Leveraging machine vision in cell-based diagnostics to do more with less. Nat Mater 2019;18:414–8. https://doi.org/10.1038/s41563-019-0339-y.

[25] Simm J, Klambauer G, Arany A, Steijaert M, Wegner JK, Gustin E, et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. Cell Chem Biol 2018;25 611–8.e3.

[26] Hofmarcher M, Rumetshofer E, Clevert D-A, Hochreiter S, Klambauer G. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. J Chem Inf Model 2019;59:1163–71.

[27] Nassar M, Doan M, Filby A, Wolkenhauer O, Fogg DK, Piasecka J, et al. Label-free identification of white blood cells using machine learning. Cytometry A 2019. https://doi.org/10.1002/cyto.a.23794.

[28] Nagpal K, Foote D, Liu Y, Chen P-HC, Wulczyn E, Tan F, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med 2019;2:48.

[29] Steele KE, Tan TH, Korn R, Dacosta K, Brown C, Kuziora M, et al. Measuring multiple parameters of CD8+ tumor-infiltrating lymphocytes in human cancers by image analysis. J Immunother Cancer 2018;6:20.

[30] Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov 2015;5:1210–23.

[31] The cancer genome atlas. National Cancer Institute; 2018 https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, Accessed date: 24 June 2019.

[32] Dugger SA, Platt A, Goldstein DB. Drug development in the era of precision medicine. Nat Rev Drug Discov 2017;17:183.

[33] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.

[34] Mamatjan Y, Agnihotri S, Goldenberg A, Tonge P, Mansouri S, Zadeh G, et al. Molecular signatures for tumor classification: an analysis of the cancer genome atlas data. J Mol Diagn 2017;19:881–91.

[35] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. Npj Digit Med 2018;1:18.

[36] Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med 2010;363:711–23.

[37] Catalanotti F, Solit DB, Pulitzer MP, Berger MF, Scott SN, Iyriboz T, et al. Phase II trial of MEK inhibitor selumetinib (AZD6244, ARRY-142886) in patients with BRAFV600E/K-mutated melanoma. Clin Cancer Res 2013;19:2257–64.

[38] Infinium HumanMethylation450K BeadChip documentation. http://emea.support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/documentation.html; 2019, Accessed date: 29 July 2019.

[39] Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570–5.

[40] Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald 3rd ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 2019;569:503–8.

[41] Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature 2019;568:511–6. https://doi.org/10.1038/s41586-019-1103-9.

[42] Yu C, Mannan AM, Yvone GM, Ross KN, Zhang Y-L, Marton MA, et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. Nat Biotechnol 2016;34:419–23.

[43] Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral heterogeneity within medulloblastoma subgroups. Cancer Cell 2017;31 737–54.e6.

[44] Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer sub-type analysis. Bioinformatics 2009;25:2906–12.

[45] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.

[46] Cabassi A, Kirk PDW. Multiple kernel learning for integrative consensus clustering of genomic datasets; 2019.

[47] Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods 2018;145:16–24.

[48] Gabasova E, Reid J, Wernisch L. Clusternomics: integrative context-dependent clustering for heterogeneous datasets. PLoS Comput Biol 2017;13:e1005781.

[49] Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics 2013;29:2610–6.

[50] Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 2012;28:3290–7.

[51] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11:333–7.

[52] Li J, Pan C, Zhang S, Spin JM, Deng A, Leung LLK, et al. Decoding the genomics of abdominal aortic aneurysm. Cell 2018;174. https://doi.org/10.1016/j.cell.2018.07.021 1361–72.e10.

[53] Normand R, Du W, Briller M, Gaujoux R, Starosvetsky E, Ziv-Kenet A, et al. Found in translation: a machine learning model for mouse-to-human inference. Nat Methods 2018;15:1067–73.

[54] Mourragui S, Loog M, van de Wiel MA, Reinders MJT, Wessels LFA. PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors; 2019. https://doi.org/10.1101/536797.

[55] Buil-Bruna N, López-Picazo J-M, Martín-Algarra S, Trocóniz IF. Bringing model-based prediction to oncology clinical practice: a review of pharmacometrics principles and applications. Oncologist 2016;21:220–32.

[56] Aziz HA, Alshekhabokbakr HM. Health informatics tools to improve utilization of laboratory tests. Lab Med 2017;48:e30–5.

[57] Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018;15:e1002689.

[58] Ma H, Guo X, Ping Y, Wang B, Yang Y, Zhang Z, et al. PPCD: privacy-preserving clinical decision with cloud support. PLoS One 2019;14:e0217349.

[59] Dignum V. Ethics in artificial intelligence: introduction to the special issue. Ethics Inf Technol 2018;20:1–3.

[60] Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. Npj Digit Med 2019;2:2.

[61] Celi LA, Citi L, Ghassemi M, Pollard TJ. The PLOS ONE collection on machine learning in health and biomedicine: towards open code and open data. PLoS One 2019;14:e0210232.

[62] Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749–60. https://doi.org/10.1038/s41551-018-0304-0.

[63] Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Proc 2018;73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011.

[64] Olah C, Mordvintsev A, Schubert L. Feature visualization. Distill 2017;2. https://doi.org/10.23915/distill.00007.

[65] Bau D, Zhou B, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations IEEE Conference on Computer Vision and Pattern Recognition (CVPR) ; 2017. https://doi.org/10.1109/cvpr.2017.354.

[66] Yao Q, Wang M, Chen Y, Dai W, Yi-Qi H, Yu-Feng L, et al. Taking human out of learning applications: A survey on automated machine learning; 2018.

[67] Olson RS, Moore JH. TPOT: a tree-based pipeline optimization tool for automating machine learning. Auto Mach Learn 2019:151–60. https://doi.org/10.1007/978-3-030-05318-5_8.

[68] Jin H, Song Q, Hu X. Auto-keras: An efficient neural architecture search system; 2018.

[69] Cloud AutoML – Custom Machine Learning Models | Google Cloud. Google Cloud; 2019 https://cloud.google.com/automl/ (accessed July 28, 2019).

[70] Szymkuć S, Gajewska EP, Klucznik T, Molga K, Dittwald P, Startek M, et al. Computer-assisted synthetic planning: the end of the beginning. Angew Chem Int Ed Engl 2016;55:5904–37.