Review Article

# How data science and AI-based technologies impact genomics

Jing Lin, PhD[1], Kee Yuan Ngiam, FRCS[1,2]

[1]NUHS Corporate Office, National University Health System, [2]Department of Surgery, National University of Singapore, Singapore

## Abstract

Advancements in high-throughput sequencing have yielded vast amounts of genomic data, which are studied using genome-wide association study (GWAS)/phenome-wide association study (PheWAS) methods to identify associations between the genotype and phenotype. The associated findings have contributed to pharmacogenomics and improved clinical decision support at the point of care in many healthcare systems. However, the accumulation of genomic data from sequencing and clinical data from electronic health records (EHRs) poses significant challenges for data scientists. Following the rise of artificial intelligence (AI) technology such as machine learning and deep learning, an increasing number of GWAS/PheWAS studies have successfully leveraged this technology to overcome the aforementioned challenges. In this review, we focus on the application of data science and AI technology in three areas, including risk prediction and identification of causal single-nucleotide polymorphisms, EHR-based phenotyping and CRISPR guide RNA design. Additionally, we highlight a few emerging AI technologies, such as transfer learning and multi-view learning, which will or have started to benefit genomic studies.

**Keywords:** Artificial intelligence, deep learning, genome-wide association study, pharmacogenomics, phenome-wide association study

## INTRODUCTION

High-throughput sequencing technologies, such as microarrays and next-generation sequencing (NGS), enable hundreds of millions of DNA molecules to be sequenced at a time and become powerful tools in genomic studies. The genome-wide association study (GWAS) has been a popular method to apply to genomic data to investigate the association between genetic variants (e.g., single-nucleotide polymorphisms [SNPs]) and phenotypes (a particular trait or disease). GWAS is usually performed by comparing the DNA of participants having the phenotype with similar people without the phenotype (as controls).

One landmark publication in the history of GWAS, the largest GWAS ever conducted at the time of its publication in 2007, was presented by the Wellcome Trust Case Control Consortium and included 14,000 cases of seven common diseases (coronary heart disease, type 1 diabetes mellitus, type 2 diabetes mellitus, rheumatoid arthritis, Crohn's disease, bipolar disorder, and hypertension) and 3,000 shared controls.[1] Since then, many more GWAS studies have been performed, which led to the discovery of thousands of risk variants involved in multiple traits/diseases. According to the GWAS Catalog, a free online database that compiles GWAS published data, more than 250,000 SNP phenotype associations have been identified from more than 5,000 studies as of May 2021.[2]

The phenome-wide association study (PheWAS), a complementary approach to GWAS, has been developed with the emergence of clinical electronic health records (EHRs) with linked genetic data. The major difference between the design of GWAS and PheWAS is the direction of inference: GWAS typically focuses on a single disease or a small set of diseases and studies their association with many possible genetic variants, whereas in PheWAS, a single genetic variant is tested across a large number of different phenotypes. In 2010, the first PheWAS study was published by researchers from Vanderbilt University, who used EHR billing codes (International Classification of Diseases, 9th revision [ICD-9]) to define 776 case/control

Access this article online

Quick Response Code:

Website:
https://journals.lww.com/SMJ

DOI:
10.4103/singaporemedj.SMJ-2021-438

**Correspondence:** A/Prof Kee Yuan Ngiam,
Group Chief Technology Officer, NUHS Corporate Office,
National University Health System, 1E Kent Ridge Road, 119228, Singapore.
E-mail: kee_yuan_ngiam@nuhs.edu.sg

**Received:** 23 Nov 2021 **Accepted:** 26 Mar 2022 **Published:** 19 Jan 2023

**For reprints contact:** WKHLRPMedknow_reprints@wolterskluwer.com

**How to cite this article:** Lin J, Ngiam KY. How data science and AI-based technologies impact genomics. Singapore Med J 2023;64:59-66.

phenotypes and studied their association with five SNPs with known disease associations.[3] This study replicated >50% of previously reported SNP-disease associations and identified additional new associations. To date, the number of phenotypes defined in PheWAS studies has been expanded to more than 1,600 and can be based on other diagnosis codes (e.g., ICD-10) in addition to ICD-9. A combination of GWAS and PheWAS offers great potential to uncover pleiotropy and elucidate the genetic architecture of complex traits.[3]

GWAS/PheWAS has greatly contributed to pharmacogenomics, which is the study of how human genetic information impacts drug response, with the aim of improving efficacy and reducing side effects.[4] For example, several pharmacogenomic GWASs identified and confirmed that the genetic variants in VKORC1 and CYP2C9 are the principle genetic determinants of stable warfarin dosing.[5] Another GWAS for clopidogrel effect identified the association of the CYP2C19*2 variant with diminished platelet response to clopidogrel therapy.[6] Recently, clinical decision support (CDS) systems based on genetic testing of these identified SNPs have been, or are being, implemented into EHR systems to trigger clinical recommendations and/or alerts at the point of care in many healthcare systems (e.g., Mayo Clinic, Vanderbilt University Medical Center, St. Jude Children's Research Hospital).[7] An example of such CDS guidance is given in Figure 1.

However, the rapid accumulation of genomic data from DNA sequencing and clinical data from EHR systems poses significant challenges and opportunities for data scientists to extract biologically or clinically relevant information from the vast amount of genotype and phenotype data. In the past two decades, various data science techniques and artificial intelligence (AI)-based technologies have been successfully applied in genomics.

Machine learning methods, which are part of AI technology, are tools that use algorithms to automatically learn from sample data (training data) to make predictions or decisions. Both supervised and unsupervised learning methods are commonly applied to genomic data. Supervised learning provides machine learning algorithms with labelled training data, from which the model can learn to make correct predictions on the labels of new testing data. Unsupervised learning, by contrast, works on unlabelled data and uses a range of algorithms to find patterns in the data, such as clusters or outliers.

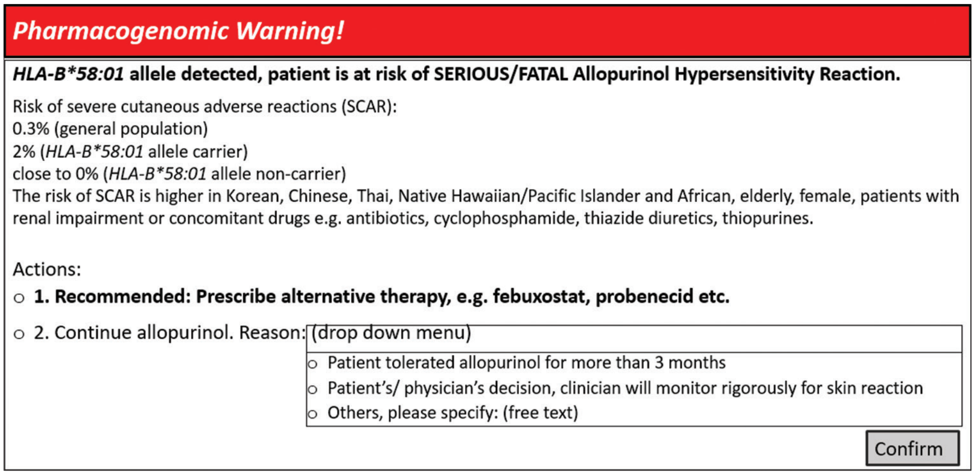
Deep learning has emerged in recent years as the leading class of machine learning algorithms.[8] It uses neural networks that are composed of hidden layers performing different operations to find complex representations of data. It is particularly powerful in handling large-scale datasets with high dimensionality and has contributed to enormous progress in numerous fields, including image classification, natural language processing (NLP), speech recognition and so on.[9]

In this review, we discuss the challenges in handling GWAS/PheWAS data and how these challenges may be overcome using data science and AI-based technologies. We further highlight a few emerging new AI technologies, which will or have started to benefit genomic studies.

## CHALLENGES AND SOLUTIONS

### Risk prediction and identification of causal SNPs

The majority of previous GWAS/PheWAS studies have relied on traditional statistical methods, such as $\chi^2$ and *t*-tests, which assign a *P* value to each SNP under investigation and subsequently assess its statistical significance by comparison to a predefined *P* value threshold (adjusted for multiple comparisons using methods such as Bonferroni, false discovery rate and permutation approaches). Linear or logistic regression methods, which have the advantage of being able to adjust for multiple covariates such as age and sex, have also been applied in many GWAS/PheWAS studies. However, very few diseases are caused by single genetic defects with large effects;



**Pharmacogenomic Warning!**

**HLA-B*58:01 allele detected, patient is at risk of SERIOUS/FATAL Allopurinol Hypersensitivity Reaction.**

Risk of severe cutaneous adverse reactions (SCAR):
0.3% (general population)
2% (HLA-B*58:01 allele carrier)
close to 0% (HLA-B*58:01 allele non-carrier)
The risk of SCAR is higher in Korean, Chinese, Thai, Native Hawaiian/Pacific Islander and African, elderly, female, patients with renal impairment or concomitant drugs e.g. antibiotics, cyclophosphamide, thiazide diuretics, thiopurines.

Actions:

○ **1. Recommended: Prescribe alternative therapy, e.g. febuxostat, probenecid etc.**

○ 2. Continue allopurinol. Reason: (drop down menu)

○ Patient tolerated allopurinol for more than 3 months
○ Patient's/ physician's decision, clinician will monitor rigorously for skin reaction
○ Others, please specify: (free text)

**Figure 1:** Screenshot shows an example of presenting genetic test results in an electronic medical record and clinical decision-support guidance and suggestion. [Provided by Dr Elaine Lo, National University Health System, Singapore].

most complex diseases are caused by epistatic interactions of multiple genetic factors, with small effect sizes. These traditional methods are based on testing SNPs individually and in parallel, which ignore the potential interactions and correlation between different SNPs.[10]

Machine learning methods have been introduced in a number of GWAS studies to overcome the above limitation. Machine learning models built into different GWAS studies varied greatly in their complexity, ranging from relatively simple regression approaches to more complex ensemble models, as well as deep learning models.[11] For example, Deo *et al*. developed a gradient-boosting model, which is an ensemble of multiple decision-tree-based models, and successfully identified several candidate causal genes affecting cardiovascular disease-related traits, including cholesterol levels, blood pressure, and conduction system and cardiomyopathy phenotypes, and further validated three of the identified causal genes experimentally in a zebrafish model of cardiac function.[12]

When applied to the analysis of GWAS datasets, deep learning methods also achieved high accuracy in many studies. For example, deep learning models incorporating SNPs associated with obesity delivered a significant predictive performance in correctly classifying obese and non-obese patients with an area under the curve (AUC) >0.99, demonstrating the ability of deep learning algorithms to capture the combined effect of SNPs and predict complex diseases.[13]

So far, a large percentage of the published deep learning methods have focused on risk prediction rather than the identification of disease and SNP associations. This may be partially a result of the interpretability problem of deep learning models — that is, many of the deep learning models behave like black boxes and fail to provide explanations of their predictions. There is a growing need to make the models more interpretable, especially in the healthcare domain where it is crucial to build trust in a model and be able to understand its behaviour. Explainable AI is an emerging field in machine learning that aims to open the black box of machine learning methods and review the processes underlying their decisions to make the results more interpretable.

One example of applying explainable AI in the context of GWAS was by Romagnoni *et al*. on a GWAS on Crohn's disease.[14] The authors calculated the permutation feature importance (PFI) for each feature, based on the decrease in a model performance score (such as AUC) when a single feature value is randomly shuffled, and identified a few novel SNPs associated with Crohn's disease based on PFI. Another example of explainable AI is the layer-wise relevance propagation, which has been used to generate SNP relevance scores based on their contribution to the fully trained model and successfully discovered two very promising, novel SNP disease associations.[9]

An obstacle to applying machine learning methods, especially deep learning methods in GWAS/PheWAS studies, is the so-called curse of dimensionality of the genomics data,[15] which usually represent a very large number of variables (e.g., genetic variants) and a relatively small number of patient samples. With the growing amount and availability of EHR-linked genomic data and advancement of dimensionality reduction techniques, this may not be a problem in the future.

## EHR-based phenotyping

One fundamental step of using the EHR data to perform GWAS/PheWAS is EHR-based phenotyping, which is the process of identifying patients with certain characteristics of interest (e.g., exposures or outcomes). The descriptions of phenotypes may be as simple as patients with rheumatoid arthritis or far more specific and complex, such as patients with stage II thyroid cancer who are younger than 55 years of age and have bony metastases.

EHR-based phenotyping is challenging because of the heterogeneity, incompleteness and complexity of EHR data, which contain large repositories of both structured data (demographics, diagnosis codes, procedure codes, laboratory values, medication exposures and so on) and unstructured data (progress notes, discharge summaries, imaging and pathology reports and so on).[16] Traditionally, rule-based methods, which applies inclusion and exclusion criteria defined by clinicians based on consensus guidelines, have been adopted for EHR-based phenotyping. One simple phenotyping example is the translation table (ICD codes to phecodes) to define 776 phenotype cases and controls in the first PheWAS described earlier in this review. Combining diagnosis codes with other structured data fields, such as medications, procedures and laboratory values, tends to show better performance than a single code search, as demonstrated on phenotyping a number of diseases, including type 2 diabetes mellitus,[17] rheumatoid arthritis[18] and coronary artery disease.[19]

### *Applying AI technologies to improve phenotyping accuracy*

Unstructured data, which account for around 80% of the data in EHRs, including text from clinical notes, discharge summaries and radiology and pathology reports, contain a large chunk of phenotypic information.[20] Clinical NLP techniques (e.g., cTAKES) with the ability to parse the semantic relationships and extract structured concepts from free text have been applied in phenotyping, and when added to structured data, they showed significant improvement in phenotyping accuracy. Liao *et al*. demonstrated across several disease categories, from inflammatory bowel disease to multiple sclerosis, that the addition of NLP to structured data improved the sensitivity of phenotypes while preserving high positive predictive value.[21]

Machine learning methods have been applied to build phenotyping models. As an example, using both structured data (ICD codes and medication data) and NLP-derived clinical concepts, support vector machine (SVM) models were built for rheumatoid arthritis phenotyping and achieved higher accuracy (>93% precision and ~80% recall) compared with the rule-based method (75% precision and 52% recall).[22] This study showed that the performance of a SVM trained on their naïve dataset (with all features together and no feature engineering) was almost as good as the SVM built on a refined dataset, demonstrating the possibility of constructing a high-performing classifier without any feature engineering. Compared with the rule-based approaches, machine learning methods are more capable of capturing more complex phenotypes or working with a less-standardised dataset, and can be easily scaled.

However, building and validation of supervised machine learning models require manually labelled gold standard training and test datasets, which take significant time and expert knowledge to create. Unsupervised learning methods to automatically derive phenotype candidates (patient clusters on specific medical conditions) with no or minimum human supervision have been proposed by a number of studies. Ho *et al*. used a non-negative tensor factorisation technique called ‘Limestone’ to generate multiple phenotype candidates with no predefined phenotype definitions.[23] As reviewed by a medical expert, 82% of the top 50 automatically generated phenotype candidates are clinically meaningful, and only 40 phenotype candidates are needed as features to obtain better predictive accuracy of patients at risk of heart failure than the original set of >600 medical features. Two upgraded versions of Limestone, named Marble and Granite, respectively, have been proposed and showed improved performance.[24,25]

Following the trend towards using deep learning approaches, phenotyping methods using various techniques have been reported, such as the de-noising auto-encoders by Miotto *et al*.,[26] as well as NLP methods based on the convolutional neural network algorithm to utilise clinicians' notes or discharge summaries by Gehrmann *et al*.[27] and Yang *et al*.,[28] respectively.

### *Efforts to improve portability of EHR-based phenotyping*

Portability of the defined phenotype across healthcare systems is a challenging problem for EHR-based phenotyping. The performance of the phenotyping tools has varied across test sites,[16] which may be partially a result of the various clinical data models used.

To standardise the format and content of observational data, several common data models (CDMs) have been proposed by collaborative research networks such as Observational Health Data Sciences and Informatics (OHDSI), Informatics for Integrating Biology and the Bedside (i2b2) and Patient-Centered Clinical Research Network (PCORnet).[29,30] Among them, the most well-known is the Observational Medical Outcomes Partnership (OMOP) CDM that was offered by the OHDSI programme. It enables the capture of information (e.g., encounters, patients, providers, diagnoses, drugs, measurements and procedures) in the same way across institutions, and a large number of medical centres worldwide have transformed their EHR data into the OMOP CDM. Also, the ‘All of Us’ (AOU) programme in the United States, which is an effort to build one million patients' EHR and genomic data, is standardising its EHR data around the OMOP data model.[31] This shared data model allows standardised applications, tools and methods to be applied on data across sites and offers a foundation for creating a broad phenotyping community for collaborative testing and refining of phenotype definition.[16]

The Electronic Medical Records and Genomics (eMERGE) network is a consortium of collaborating academic medical centres that works to develop generalisable EHR phenotype definitions in order to conduct GWAS/PheWas studies across shared clinical datasets.[32] Because of the importance of sharing and validating phenotypes in different healthcare settings, the phenotype knowledge base website, PheKB (http://phekb.org), was created as a repository of phenotypes that offers a collaborative environment to building and validating EHR-based phenotypes.[33] As of May 2021, there are approximately 80 publicly available well-defined phenotypes on PheKB, including both rule-based and machine learning methods using structured and unstructured data.

## Guide RNA design in CRISPR genetic editing

Genetic-editing techniques to modify genomic sequences have been used to validate the association between genetic variants and phenotypes identified by GWAS/PheWAS and to uncover the underlying mechanisms. The Clustered Regularly Inter-spaced Short Palindromic Repeats (CRISPR) and their associated endonuclease genes (e.g., Cas9) are a revolutionary gene-editing technology that can modify DNA with greater precision than existing technologies.[34] In this system, a single guide RNA (gRNA) guides Cas proteins to specific genomic targets. Recognition and cleavage occur via complementarity of a 20-nucleotide sequence within the gRNA to the genomic target (on-target site). By simply altering the sequence of the gRNA, Cas can be easily re-programmed to target different sites in the genome with relative ease. As a powerful gene-editing tool, CRISPR are not only used to validate the GWAS/PheWAS findings, but also offer therapeutic potential in treating diseases associated with a genetic basis.[35]

### *Clinical applications of CRISPR*

Currently, the clinical use of human germline genetic editing is still not allowed. Germline editing means the genes that are edited are heritable (in sperm, eggs or embryos) and can be passed on to the next generations.[36] The ban, however, does not apply to changes in non-germline human cells (called somatic gene editing). Since 2016, increasing numbers of CRISPR therapeutics studies have entered clinical trials, most of which focus on *ex vivo* genome editing.[37] *Ex vivo* genome editing is a therapeutic strategy in which the genome of particular cells is edited outside the patient's body, and then the modified cells are transplanted back into the patient to exert a therapeutic effect. *Ex vivo* editing guarantees that genome-editing tools only come in contact with the right target cells. This technique has been used in patients with beta thalassaemia and sickle cell disease since 2019, with promising early results. Based on analysis of the bone marrow cells from the treated patients, it was seen that the edited cells have successfully taken up residence in the bone marrow.[38]

In contrast to *ex vivo* genome editing, *in vivo* genome-editing approaches directly introduce the genome-editing components into the patient via local or systemic delivery, and so the genome editing occurs inside the patient's body. For example, an ongoing clinical trial in the United States is directly introducing CRISPR gene-editing components into the eyes of patients born with Leber's congenital amaurosis 10, a congenital vision disease, to fix the faulty photoreceptor gene in patients' vision cells.[39] Compared with the *ex vivo* strategy, a substantial problem for the *in vivo* treatment strategy is the safe and effective delivery of genome-editing components to target cells without provoking dangerous immune responses in patients or causing off-target effects.

CRISPR technology has also been used in fighting COVID-19.[40] It has been successfully used to develop rapid diagnostic tests for COVID-19 with emergency authorisation by the US Food and Drug Administration (FDA).[41] One advantage of using CRISPR over the antigen rapid test (ART) method is that the former can easily adapt to new mutations by modifying the guide RNA to detect different variants of the SARS-CoV-2 virus, whereas ART methods rely on antibodies, which require more time for redesigning. Meanwhile, scientists have examined the CRISPR-based system as a potential therapeutic strategy, using its targeted enzymatic activity to degrade SARS-CoV-2 RNA and prevent viral replication. Through the use of multiple gRNAs targeting multiple regions of the same virus or multiple strains of coronavirus, this system could possibly buffer against viral evolution and protect against future related pathogenic viruses.[42]

### *Prediction of gRNA on-target efficacy and off-target effect with machine learning methods*

One major challenge for effective application of CRISPR systems is the optimal design of gRNA with high sensitivity and specificity. Previous studies have demonstrated that multiple mismatches as well as DNA or RNA bulges can be tolerated in gRNA target sequences, resulting in cleavage of unintended genomic sites (off-targets).[43] As confirmed by Fu *et al*., one to five base mismatches could be tolerated by gRNA (guide RNA) during the guiding process, which in turn causes unintended sequences to be erroneously edited.[44] The on-target activity and off-target effects of individual gRNAs can vary widely, and accurate prediction of these effects would facilitate the optimal design of gRNAs by minimising their off-target effects (high specificity) and maximising their on-target efficacy (high sensitivity).

Machine learning has been gradually applied to gRNA design. Various machine learning-based gRNA design models have been developed and applied.[45] Most of these models take into account the sequence features and/or secondary structure features to guide gRNA design to improve gene-editing results. For example, a collaborative project involving Microsoft, the Broad Institute of Massachusetts Institute of Technology and University of California Berkeley and others has developed two predictive modelling approaches, named Azimuth and Elevation, for on-target and off-target activity prediction, respectively, and these tools are currently provided on Microsoft Azure as a cloud-based, end-to-end guide-design service.[46,47]

One obstacle to building a good learning model to predict gRNA activity is the data heterogeneity and inconsistency issue. Current data are mostly collected from experiments using different cell types, different experimental platforms or even different types of Cas enzymes. gRNA activity depends not only on its own sequence, but also on the experimental conditions (e.g., *in vitro* vs *in vivo*), and effective integration of data is therefore required. A few studies have applied deep learning methods to predict gRNA effect.[48] DeepCRISPR, as an example, is a comprehensive deep learning framework to simultaneously predict gRNA on-target efficacy and a whole-genome off-target cleavage profile.[49] In addition to the sequence feature, epigenetic features that may affect gRNA knockout efficacy were also used as input data to train the models. Although trained on limited cell type data, by taking advantage of the powerful deep learning framework, DeepCRISPR showed a generally good prediction ability when adapting to new cell types.

## EMERGING MACHINE LEARNING AND AI TRENDS RELATED TO THE GENOMICS FIELD

### Transfer learning in NLP

Transfer learning is a technique of training a deep learning model on a large dataset and then using the pre-trained model to perform similar tasks that may be in a different domain on another dataset. This breakthrough technology has been applied in the field of computer vision (e.g., image classification) and has gained much success since 2012.[50] One of the main advantages of using a pre-trained model as a starting point is the relatively small sample size required. As the model has been pre-trained on a large-scale dataset, the pre-trained model just needs to be trained or fine-tuned when it is applied to a similar task using a relatively smaller dataset. The knowledge that the pre-trained model has learned from the large dataset during pre-training can be transferred to a new task. This is especially useful for deep learning approaches because they usually have a huge number (millions) of parameters requiring a large dataset to train. The lack of training data may lead to overfitting and decrease in accuracy.

Transfer learning was introduced into the field of NLP in 2018 with Bidirectional Encoder Representations from Transformers (BERT), which is a highly successful model developed by the Google AI team.[51] BERT is a pre-trained model trained on Wikipedia (2.5 billion words) and BooksCorpus (0.8 billion words) and has dramatically improved performance for many NLP tasks, such as text classification, sentence classification, semantic similarity between pairs of sentences, question answering task with paragraph, text summarisation and so on. Google has been leveraging BERT to better understand user searches. Researchers have started to apply BERT for improving the performance of biomedical and clinical text mining models in recent years, and several pre-trained models have been built on top of BERT.

Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) is a domain-specific language representation model that has been pre-trained on large-scale biomedical corpora, including PubMed Abstracts (4.5 billion words) and PMC Full-text articles (13.5 billion words) to understand complex biomedical texts.[52] When applied on three representative biomedical text mining tasks, including biomedical named entity recognition, biomedical relation extraction and biomedical question answering, BioBERT largely outperforms BERT and previous state-of-the-art models. Our group in the National University Health System is currently leveraging it in several clinical projects, such as high-risk pregnancy prediction, which involves text mining from clinician's notes.

Two clinically trained BERT models, Clinical BERT and Discharge Summary BERT, which are specific for NLP tasks in the clinical domain have been built by Alsentzer *et al*.[53] These two models, ClinicalBERT and Discharge Summary BERT, were pre-trained using all clinical note types or only discharge summaries, respectively, from approximately two million notes in the MIMIC-III database and resulted in performance improvements over BERT and BioBERT on three clinical NLP tasks. Another clinical BERT specific to the clinical domain is ClinicalBERT developed by Huang *et al*.[54] ClinicalBERT has also been trained on clinical notes from the MIMIC-III database and used to predict hospital re-admission. However, as the authors from both studies pointed out, MIMIC only contains notes from the intensive care unit of a single healthcare institution and may result in limitations in the built models. Thus re-training on a larger collection of clinical notes is recommended for better performance.

## Multi-view learning in multi-omics

In addition to genomics, high-throughput data have now become widely available in other omics, such as proteomics, transcriptomics and metabolomics, which together are called 'multi-omics' data. The flow of genetic information in the central dogma is complex and involves many levels of molecules and interactions. Multi-omics data provide information on biomolecules from different layers and, when linked to clinical phenotypes, help in bridging the gap from genotype to phenotype. Studies integrating multi-omics data have great potential in the exploration of complex biological systems and would lead to a better understanding of human health and disease, eventually aiding in better treatment and prevention.

However, integrating large-scale multi-omics data to discover functional insights into biological systems is a challenging task. To address these challenges, machine learning has been applied for multi-omics data integration and analysis. Various machine learning approaches have already been explored for a wide range of applications.[8] Compared with traditional machine learning methods, which have difficulty in integrating the heterogeneous and noisy omics data, multi-view learning, an emerging machine learning method, is more effective in studying heterogeneity of data and revealing cross-talk patterns, and researchers have started to apply it to multi-omics data.

In multi-view learning, data from multiple omics sources can be encoded by multiple data views. Each view is an aspect of the whole complex biological phenomenon that is compatible and complementary to other views. Multi-view learning algorithms applied to these data aims to capture the interactions within each omic, as well as the interactions across all omics, to get a comprehensive understanding of complex biological phenomena. Multi-view learning enables data from multiple omics sources, such as gene expression, chromatin accessibility and protein expression, to be represented in a common space, so that they can be simultaneously clustered and a group of genes or a group of proteins that function together can be identified. More importantly, the functional linkage between genes, regulatory elements and proteins can be revealed (e.g., protein $\alpha$ binds to chromatin region $\beta$ to regulate the expression of gene $\gamma$), which may be further linked with a specific subtype of diseases if phenotypic data are added.[55] A few studies have applied multi-view learning methods to multi-omics data from The Cancer Genome Atlas (TCGA) — a large multi-omic repository of data on thousands of cancer patients. Their results showed that multi-view learning outperformed the state-of-the-art methods or single-view approaches, and successfully identified different disease subtypes.[56,57]

## Next-generation sequencing data pre-processing

With its ultra-high throughput, scalability and speed, NGS enables researchers to sequence whole genomes/exomes and has become one of the major sources of genomic data used in both biological studies and clinical practice. However, the large volume of raw data generated by NGS also pose significant challenges in data storage and data pre-processing. AI technology has been applied in NGS raw data pre-processing and has shown promising results.[58]

For example, DeepVariant, developed by Google AI, is an analysis pipeline that uses a deep neural network to call genetic variants from next-generation DNA sequencing data.[59] It transformed a variant calling problem into an image recognition problem by converting BAM files into images similar to genome browser snapshots and applied the TensorFlow deep learning method to call variants in sequencing data. DeepVariant was reported to be the most accurate pipeline in variant calling in the Precision FDA Truth challenge (2016) and outperformed the other variant callers, including the Genome Analysis Toolkit gold standard pipeline in a comparison conducted by Supernat *et al.*[60]

## CONCLUSION

This review provided a summary of how AI-based technology, especially deep learning methods, has been increasingly applied in the fields of genomics and pharmacogenomics and gained a lot of success. There remain significant challenges in using AI technologies in the near future because of legacy IT infrastructure and data constructs. There are also limitations in awareness of the capabilities that AI can bring to healthcare and the training needed to build AI models at scale. It is estimated that currently, only 20% of all datasets are sufficiently processed to be used for AI modelling. There are still technical and legal challenges to be overcome in order to facilitate the use of the remaining datasets. When these challenges are overcome, we may truly reap the benefits of automation afforded by the use of AI in healthcare.

### Financial support and sponsorship

Nil.

### Conflicts of interest

There are no conflicts of interest.

## REFERENCES

1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
2. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, *et al*. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47:D1005-12.
3. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, *et al*. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013;31:1102-10.
4. Karczewski KJ, Daneshjou R, Altman RB. Chapter 7: Pharmacogenomics. PLoS Comput Biol 2012;8:e1002817.
5. Jorgensen AL, FitzGerald RJ, Oyee J, Pirmohamed M, Williamson PR. Influence of CYP2C9 and VKORC1 on patient response to warfarin: A systematic review and meta-analysis. PLoS One 2012;7:e44064.
6. Shuldiner AR, O'Connell JR, Bliden KP, Gandhi A, Ryan K, Horenstein RB, *et al*. Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. JAMA 2009;302:849-57.
7. Freimuth RR, Formea CM, Hoffman JM, Matey E, Peterson JF, Boyce RD. Implementing genomic clinical decision support for drug-based precision medicine. CPT Pharmacometrics Syst Pharmacol 2017;6:153-5.
8. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv 2021;49:107739.
9. Mieth B, Rozier A, Rodriguez JA, Höhne MMC, Görnitz N, Müller KR. DeepCOMBI: Explainable artificial intelligence for the analysis and discovery in genome-wide association studies. NAR Genom Bioinform 2021;3:lqab065.
10. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol 2012;8:e1002822.
11. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Front Genet 2020;11:350.
12. Deo RC, Musso G, Tasan M, Tang P, Poon A, Yuan C, *et al*. Prioritizing causal disease genes using unbiased genomic features. Genome Biol 2014;15:534.
13. Montaez CAC, Fergus P, Montaez AC, Hussain A, Al-Jumeily D, Chalmers C, editors. Deep learning classification of polygenic obesity using genome wide association study SNPs 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018.
14. Romagnoni A, Jégou S, Van Steen K, Wainrib G, Hugot JP. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci Rep 2019;9:10351.
15. Koumakis L. Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J 2020;18:1466-73.
16. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: From rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018;1:53-68.
17. 16. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19:212-8.
18. Nicholson A, Ford E, Davies KA, Smith HE, Rait G, Tate AR, *et al*. Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: A strategy for developing code lists. PLoS One 2013;8:e54878.
19. Esteban S, Rodríguez Tablado M, Ricci RI, Terrasa S, Kopitowski K. A rule-based electronic phenotyping algorithm for detecting clinically relevant cardiovascular disease cases. BMC Res Notes 2017;10:281.
20. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearb Med Inform 2014;9:14-20.
21. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, *et al*. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885.
22. Carroll RJ, Eyler AE, Denny JC. Naïve electronic health record phenotype identification for rheumatoid arthritis. AMIA Annu Symp Proc 2011;2011:189-96.
23. Ho JC, Ghosh J, Sun J, editors. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014.
24. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, *et al*. Limestone: High-throughput candidate phenotype generation via tensor factorization. J Biomed Inform 2014;52:199-211.
25. Henderson J, Ho JC, Kho AN, Denny JC, Malin BA, Sun J, *et al*., editors. Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping. IEEE International Conference on Healthcare Informatics (ICHI); 2017.
26. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:26094.
27. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, *et al*. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One 2018;13:e0192360.
28. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for patient phenotyping from electronic health records. Sci Rep 2020;10:1432.
29. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, *et al*. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-8.

30. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016;23:909-15.
31. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the all of us research program: Transforming i2b2 data into the OMOP common data model. PLoS One 2019;14:e0212463.
32. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, *et al*. The electronic medical records and genomics (eMERGE) network: Past, present, and future. Genet Med 2013;15:761-71.
33. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, *et al*. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016;23:1046-52.
34. 33. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, *et al*. Multiplex genome engineering using CRISPR/Cas systems. Science 2013;339:819-23.
35. Rao S, Yao Y, Bauer DE. Editing GWAS: Experimental approaches to dissect and exploit disease-associated genetic variation. Genome Med 2021;13:41.
36. Lander ES, Baylis F, Zhang F, Charpentier E, Berg P, Bourgain C, *et al*. Adopt a moratorium on heritable genome editing. Nature 2019;567:165-8.
37. Li Y, Glass Z, Huang M, Chen ZY, Xu Q. Ex vivo cell-based CRISPR/Cas9 genome editing for therapeutic applications. Biomaterials 2020;234:119711.
38. Frangoul H, Altshuler D, Cappellini MD, Chen YS, Domm J, Eustace BK, *et al*. CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. N Engl J Med 2021;384:252-60.
39. Maeder ML, Stefanidakis M, Wilson CJ, Baral R, Barrera LA, Bounoutas GS, *et al*. Development of a gene-editing approach to restore vision loss in Leber congenital amaurosis type 10. Nat Med 2019;25:229-33.
40. Ooi KH, Liu MM, Tay JWD, Teo SY, Kaewsapsak P, Jin S, *et al*. An engineered CRISPR-Cas12a variant and DNA-RNA hybrid guides enable robust and rapid COVID-19 testing. Nat Commun 2021;12:1739.
41. Broughton JP, Deng X, Yu G, Fasching CL, Servellita V, Singh J, *et al*. CRISPR–Cas12-based detection of SARS-CoV-2. Nature Biotechnology 2020;38:870-4.
42. Abbott TR, Dhamdhere G, Liu Y, Lin X, Goudy L, Zeng L, *et al*. Development of CRISPR as an antiviral strategy to combat SARS-CoV-2 and influenza. Cell 2020;181:865-76.e12.
43. Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, *et al*. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res 2014;42:7473-85.
44. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, *et al*. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol 2013;31:822-6.
45. Liu Q, Cheng X, Liu G, Li B, Liu X. Deep learning improves the ability of sgRNA off-target propensity prediction. BMC Bioinformatics 2020;21:51.
46. Listgarten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J, *et al*. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomed Eng 2018;2:38-47.
47. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, *et al*. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 2016;34:184-91.
48. Fu BX, St Onge RP, Fire AZ, Smith JD. Distinct patterns of Cas9 mismatch tolerance *in vitro* and in vivo. Nucleic Acids Res 2016;44:5365-77.
49. Chuai G, Ma H, Yan J, Chen M, Hong N, Xue D, *et al*. DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. Genome Biol 2018;19:80.
50. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol 2020;1213:3-21.
51. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 181004805. 2018.
52. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, *et al*. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234-40.
53. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, *et al*. Publicly available clinical BERT embeddings. arXiv Preprint arXiv: 190403323. 2019.
54. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv: 190405342. 2019.
55. Nguyen ND, Wang D. Multiview learning for understanding functional multiomics. PLoS Comput Biol 2020;16:e1007677.
56. Yu Y, Zhang L-H, Zhang S. Simultaneous clustering of multiview biomedical data using manifold optimization. Bioinformatics 2019;35:4029-37.
57. Chaudhary K, Poirion OB, Lu L, Huang S, Ching T, Garmire LX. Multimodal meta-analysis of 1,494 hepatocellular carcinoma samples reveals significant impact of consensus driver genes on phenotypes. Clin Cancer Res 2019;25:463-72.
58. Luo R, Sedlazeck FJ, Lam T-W, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun 2019;10:998.
59. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, *et al*. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 2018;36:983-7.
60. Supernat A, Vidarsson OV, Steen VM, Stokowy T. Comparison of three variant callers for human whole genome sequencing. Sci Rep 2018;8:1-6.