COMMENTARY

MAYO CLINIC PROCEEDINGS:
DIGITAL HEALTH

MAYO CLINIC

The Potential for Artificial Intelligence Applied to Epigenetics

Manlio Vinciguerra, MSc, PhD

From the Department of Translational Stem Cell Biology, Research Institute of the Medical University, Varna, Bulgaria; and Liverpool Centre for Cardiovascular Science, Liverpool John Moores University, Liverpool, United Kingdom.

Artificial intelligence (AI) has been increasingly applied to various fields in medicine, including diagnosis, treatment, drug discovery, and health management. Although AI is a broad concept, machine learning (ML) is the application of AI into a system or machine, which helps it to self-learn and improve continually. Machine learning approaches can be supervised, semi-supervised, or unsupervised. Supervised learning is the type of ML in which computers are trained using well "labeled" training data (tagged with the correct output), and on the basis of that data, they predict the output. Deep learning (DL) is a subset of the broader family of ML methods. Deep learning uses complex algorithms and deep neural networks to repetitively train a specific model or pattern. The term "deep" in DL reflects the use of multiple layers in the network. Deep learning approaches can employ several types of architectures or networks, such as convolutional neural networks (CNNs) in the medical field. Artificial intelligence has been found to improve the accuracy and efficiency of medical tasks and has the potential to revolutionize health care in all fields of medicine.[1] Artificial intelligence has been used to improve the accuracy of medical imaging. For example, DL can analyze computed tomography scans and magnetic resonance imaging scans using CNNs to identify patterns that may indicate a specific disease. Convolutional neural networks in turn use a mathematical operation called convolution in place of general matrix multiplication, which is specifically designed to process pixel data for image recognition and processing. This can help make faster and more accurate diagnoses. Artificial intelligence has also been used to develop predictive models for diseases such as cancer. By analyzing patient data, AI can identify risk factors, predict the likelihood of developing a disease, and develop personalized treatment plans for patients. By analyzing patient data, such as medical history and genetics, AI can identify the most effective treatment options for each patient. This can improve treatment outcomes and reduce the risk of adverse effects. Artificial intelligence has also been used to develop predictive models for treatment response. By analyzing patient data, AI can predict how a patient will respond to a specific treatment, allowing doctors to tailor treatment plans accordingly. Artificial intelligence has been used to accelerate the drug discovery process.[2] By analyzing large datasets of chemical compounds, AI can identify potential drug candidates that are more likely to be effective. This can reduce the time and cost of developing new drugs. Artificial intelligence has also been used to develop predictive models for drug toxicity. By analyzing the chemical properties of drugs, AI can predict the likelihood of adverse effects, allowing drug developers to prioritize safer compounds.[2] Artificial intelligence has also been used to develop predictive models for disease outbreaks and personalized health monitoring systems. By analyzing patient data, such as activity levels and vital signs, AI can provide personalized recommendations for improving health and preventing disease.[3,4] Overall, AI has the potential to revolutionize health care by improving the accuracy and efficiency of medical tasks, developing personalized treatment plans, accelerating the drug discovery process, and predicting disease outbreaks. However, there are also challenges associated with the use of AI in health care, such as the need for large datasets, the potential for bias, and the ethical implications of using AI to make medical decisions.

Epigenetics

Epigenetics includes 3 main heritable regulatory systems that determine chromatin remodeling and gene transcription regulation: DNA

Mayo Clin Proc Digital Health ■ December 2023;1(4):476-479 ■ https://doi.org/10.1016/j.mcpdig.2023.07.005
www.mcpdigitalhealth.org ■

methylation, noncoding RNAs, and chromatin remodeling (histone modifications and histone variants), and the functional crosstalk among these epigenetic processes determines cell phenotype. Among epigenetic mechanisms, DNA methylation is the most commonly studied and involved in various biological processes, such as organism development, cancer, cardiovascular, and neurological disorders.[5] DNA methylation involves the transfer of a methyl group onto the C5 position of the cytosine to form 5-methylcytosine. It regulates gene expression by recruiting proteins involved in gene repression or by inhibiting the binding of transcription factors to DNA. It is controlled by DNA methyltransferases, methyl-CpG binding proteins, and other chromatin-remodeling factors. The field of genomics has been revolutionized by genome-wide association studies (GWASs), which are observational studies of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait or a phenotype. Genome-wide association studies typically focus on associations between single-nucleotide polymorphisms and human diseases but can equally be applied to any other genetic variant and any other organism. Similarly, in the field of epigenetics, epigenome-wide association studies (EWASs) investigate the association between a phenotype and epigenetic variants, most commonly the above mentioned DNA methylation. The concept of EWASs was first introduced in 2011.[6] Since then, the decreasing cost of measuring DNA methylation sites in EWAS and the increasing availability of bioinformatic analytical tools have contributed to the exponential rise in published EWAS.[7] Accordingly, EWASs have shed light on the pathogenesis of cardiovascular,[8] neurological,[9] psychiatric[10] and many other disorders.

Artificial Intelligence Applied to Epigenetics

In the AI field, DL and, in particular, CNN analysis of large-scale genetic data, such as GWASs and/or EWASs, is expected to be faster, more efficient and accurate, both for rare diseases[11] and common noncommunicable diseases.

In this commentary, I will bring selected examples of the use of AI, ML, and DL approaches to identify and study the molecular pathogenesis of 3 common noncommunicable diseases with a polygenic component: schizophrenia (SZ), Alzheimer disease (AD), and atrial fibrillation (AF).

Alzheimer disease is an age-associated, neurodegenerative disorder. It is the leading cause of dementia and a major public health problem worldwide. It is a complex illness due to not fully elucidated environmental and genetic factors with a heritability of approximately 75%.[12] Multiple AD EWAS have identified differential methylated DNA and new genes associated with AD in different regions of the human brain[13-15]; however with limited depth and technical challenges. Huang et al[16] recently developed EWASplus, a computational method that uses a supervised ML strategy to extend EWAS coverage to the entire genome. Application of EWASplus to 6 AD-related traits predicted hundreds of new significant brain CpGs associated with AD. Genes found near top ranked EWASplus loci are enriched for kinases and for genes previously shown to be physically interacting with known AD genes.[16] Similar approaches have been employed to predict AD progression. Chen et al[17] benchmarked on longitudinal DNA methylation data collected from the peripheral blood in the Alzheimer's Disease Neuroimaging Initiative, and proposed new multi-task deep autoencoders, belonging to multi-task learning (a sub-field of ML) that outperform other ML approaches for predicting AD progression. Chen et al[17] recently showed that it is possible to predict AD on the basis of a deep neural network by integrating gene expression and DNA methylation datasets. The most challenging problem in constructing a model to diagnose AD based on a multiomics dataset (but this is valid for any disease) is how to integrate different omics data and how to deal with high-dimensional and low-sample-size data. In their study, Chen et al[17] proposed to reduce the number of features based on a differentially expressed gene and a differentially methylated position in the multiomics dataset. They started using 2 previously published large-scale gene expression profiles and 1 large DNA methylation profile from the postmortem prefrontal cortex of hundreds of individuals. They thus developed a deep neural network-based prediction model that improves performance

compared to that of conventional ML algorithms.[17] The model finally included 35 genes selected that had discriminative ability to classify the AD status compared to the health status.[17] Integrating gene expression and DNA methylation data could improve prediction accuracy.

Schizophrenia is a mental disorder characterized by continuous or relapsing episodes of psychosis. Schizophrenia affects approximately 24 million people, or 1 in 300 people (0.32%) worldwide.[18] Males are more often affected and, on average, have an earlier onset. The causes of SZ include genetic and environmental factors. However, the epigenetic dysregulation contributing to the etiology of SZ is unclear. In particular, cell type-specificity of DNA methylation makes population-based epigenetic studies of SZ challenging. A few studies analyzed whether a blood marker of epigenetic risk for SZ could be derived that is specific for the disease and predicts disease-associated brain function. In a case-control study conducted from 2008 to 2018 in sites in Germany, the United Kingdom, the United States, and Australia, blood DNA methylation data (quantified as poly-methylation score) from whole-blood samples of 7488 participants, of whom 3158 received a diagnosis of SZ, together with GWAS and neuroimaging, were analyzed using ML approaches. The poly-methylation score signature was significantly associated with SZ across independent data sets (area under the curve [AUC]=0.69-0.78; P=.049-1.24 × 10^{-7}) but not with other psychiatric ailments such as major depressive disorder (AUC=0.51; P=.16), autism (AUC=0.53; P=.66), or bipolar disorder (AUC=0.58; P=.21).[19] In the latter work, an updated biologically informed ML approach for epigenetic fingerprints was used. Biologically informed ML is a 2-stage ML approach that first compresses data from individual DNA methylation sites into a pathway-level feature. Then, a second-stage algorithm integrates these pathway-level features into a system-level classifier. In another recent report, DNA methylation data on whole blood from 414 SZ cases and 433 nonpsychiatric controls were used as training data for a ML-classification algorithm with built-in feature selection, sparse partial least squares discriminate analysis, to calculate a "risk distance" to identify individuals with the highest probability of SZ. The model was then evaluated on an independent data set of 353 SZ cases and 322 nonpsychiatric controls. This model classified 303 individuals as cases with a positive predictive value of 80%, far surpassing the performance of a model based on polygenic risk score, and it was not associated with medication use.[20] These AI/ML-based results indicate that systemic epigenetic variants may classify patients with SZ accurately.

Atrial fibrillation is an abnormal heart rhythm (arrhythmia) characterized by the rapid and irregular beating of the atrial chambers of the heart. It is a type of supraventricular tachycardia. In Europe and North America, as of 2014, it affects approximately 2%-3% of the population. Atrial fibrillation is associated with an increased risk of heart failure, dementia, and stroke. In fact, AF is the cause of 20%-30% of all ischemic strokes, and patients with AF have a 5-fold increased risk of developing an ischemic stroke. Application of a CNN analysis to a multiethnic AF network GWAS for early-onset AF, including 6358 subjects from 4 independent cohorts (Korean,

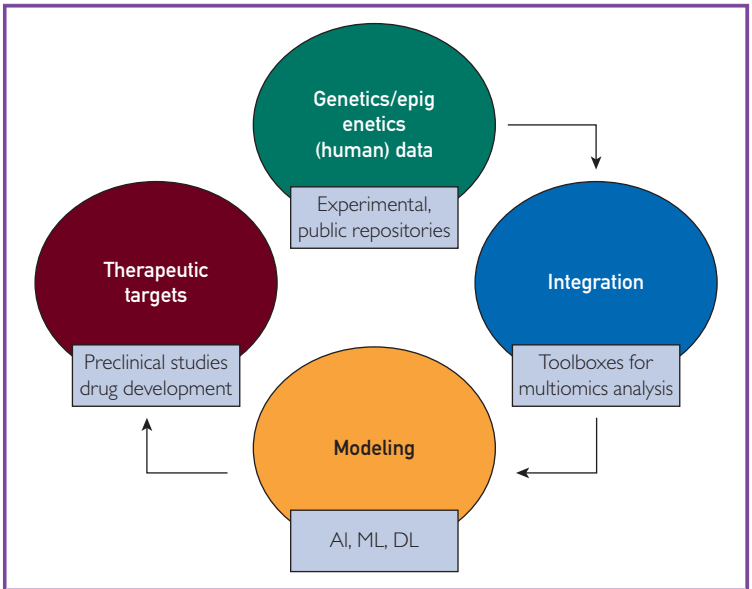

FIGURE. Integration of genomic and epigenomics data by multiomic analyses bioinformatics tools and modeling by artificial intelligence (AI)-dependent approaches is a promising avenue to identify new therapeutic targets, to be tested at the preclinical and clinical level. DL, deep learning; ML, machine learning.

Japanese, European, and multiethnic), led to moderate-to-high predictive power and assigned a high saliency score for PITX2 among the AF associated single-nucleotide polymorphisms.[21] Libiseller-Egger et al[23] used a previously published CNN[22] to predict the cardiovascular age of 36,349 participants of the UK Biobank from their electrocardiograms, and performed a GWAS on the difference between predicted and chronological age (delta age). The analysis identified 8 loci associated with delta age, including SCN5A and TTN.[23] Interestingly, PITX2 (a transcription factor that plays a critical role in early development), SCN5A (an integral membrane protein and a voltage-gated sodium channel subunit), and TTN (a giant protein that functions as a molecular spring that is responsible for the passive elasticity of muscle) have been previously and consistently associated to AF by conventional non AI-based genetic linkage studies.[24]

In summary, AI/ML/DL-driven approaches described in this commentary are increasingly robust and useful tools for detecting common polygenic diseases that affect our communities. By capturing the cumulative effects and interactions with the epigenome, AI finds patterns in constantly growing genetic and epigenetic data sets that relate to the development of diseases (Figure). In particular, DL methods are accurate but will need to undergo further refinement before clinicians can confidently use such tools.

POTENTIAL COMPETING INTERESTS

The author reports no competing interests.

Abbreviations and Acronyms: AD, Alzheimer disease; **AF,** atrial fibrillation; **AI,** artificial intelligence; **AUC,** area under the curve; **CNN,** convolutional neural network; **DL,** deep learning; **EWAS,** epigenome-wide association study; **GWAS,** genome-wide association study; **ML,** machine learning; **SZ,** schizophrenia

Correspondence: Address to Manlio Vinciguerra, MSc, PhD, Medical University of Varna, ul. "Professor Marin Drinov" 55, 9002 Varna, Bulgaria (manlio.vinciguerra@mu-varna.bg).

REFERENCES

1. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29.
2. Ming J, Sun B, Li Z, et al. Aspirin inhibits the SHH/GLI1 signaling pathway and sensitizes malignant glioma cells to temozolomide therapy. *Aging (Albany NY)*. 2017;9(4):1233-1247.
3. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-438.
4. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol*. 2017;69(21):2657-2664.
5. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol*. 2019;20(10):590-607.
6. Wei S, Tao J, Xu J, et al. Ten years of EWAS. *Adv Sci (Weinh)*. 2021;8(20):e2100727.
7. Campagna MP, Xavier A, Lechner-Scott J, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clin Epigenetics*. 2021;13(1):214.
8. Kessler T, Vilne B, Schunkert H. The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol Med*. 2016;8(7):688-701.
9. Smith RG, Pishva E, Shireby G, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. *Nat Commun*. 2021;12(1):3517.
10. Stamawska A, Demontis D. Role of DNA methylation in mediating genetic risk of psychiatric disorders. *Front Psychiatry*. 2021;12:596821.
11. Brasil S, Neves CJ, Rijoff T, et al. Artificial intelligence in epigenetic studies: shedding light on rare diseases. *Front Mol Biosci*. 2021;8:648012.
12. Wingo TS, Lah JJ, Levey AI, Cutler DJ. Autosomal recessive causes likely in early-onset Alzheimer disease. *Arch Neurol*. 2012;69(1):59-64.
13. De Jager PL, Srivastava G, Lunnon K, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci*. 2014;17(9):1156-1163.
14. Lunnon K, Smith R, Hannon E, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nat Neurosci*. 2014;17(9):1164-1170.
15. Watson CT, Roussos P, Garg P, et al. Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease. *Genome Med*. 2016;8(1):5.
16. Huang Y, Sun X, Jiang H, et al. A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer's disease. *Nat Commun*. 2021;12(1):4472.
17. Chen L, Saykin AJ, Yao B, Zhao F; Alzheimer's Disease Neuroimaging Initiative (ADNI). Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood. *Comput Struct Biotechnol J*. 2022;20:5761-5774.
18. Velligan DI, Rao S. The epidemiology and global burden of schizophrenia. *J Clin Psychiatry*. 2023;84(1):MS21078COM5.
19. Chen J, Zang Z, Braun U, et al. Association of a reproducible epigenetic risk profile for schizophrenia with brain methylation and function. *JAMA Psychiatry*. 2020;77(6):628-636.
20. Gunasekara CJ, Hannon E, MacKay H, et al. A machine learning case-control classifier for schizophrenia based on DNA methylation in blood. *Transl Psychiatry*. 2021;11(1):412.
21. Kwon OS, Hong M, Kim TH, et al. Genome-wide association study-based prediction of atrial fibrillation using artificial intelligence. *Open Heart*. 2022;9(1):e001898.
22. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol*. 2019;12(9):e007284.
23. Libiseller-Egger J, Phelan JE, Attia ZI, et al. Deep learning-derived cardiovascular age shares a genetic basis with other cardiac phenotypes. *Sci Rep*. 2022;12(1):22625.
24. Kim JA, Chelu MG, Li N. Genetics of atrial fibrillation. *Curr Opin Cardiol*. 2021;36(3):281-287.