npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Artificial intelligence for precision medicine in neurodevelopmental disorders

Mohammed Uddin[1,2*], Yujiang Wang[3,4] and Marc Woodbury-Smith[2,3*]

The ambition of precision medicine is to design and optimize the pathway for diagnosis, therapeutic intervention, and prognosis by using large multidimensional biological datasets that capture individual variability in genes, function and environment. This offers clinicians the opportunity to more carefully tailor early interventions— whether treatment or preventative in nature—to each individual patient. Taking advantage of high performance computer capabilities, artificial intelligence (AI) algorithms can now achieve reasonable success in predicting risk in certain cancers and cardiovascular disease from available multidimensional clinical and biological data. In contrast, less progress has been made with the neurodevelopmental disorders, which include intellectual disability (ID), autism spectrum disorder (ASD), epilepsy and broader neurodevelopmental disorders. Much hope is pinned on the opportunity to quantify risk from patterns of genomic variation, including the functional characterization of genes and variants, but this ambition is confounded by phenotypic and etiologic heterogeneity, along with the rare and variable penetrant nature of the underlying risk variants identified so far. Structural and functional brain imaging and neuropsychological and neurophysiological markers may provide further dimensionality, but often require more development to achieve sensitivity for diagnosis. Herein, therefore, lies a precision medicine conundrum: can artificial intelligence offer a breakthrough in predicting risks and prognosis for neurodevelopmental disorders? In this review we will examine these complexities, and consider some of the strategies whereby artificial intelligence may overcome them.

npj Digital Medicine (2019)2:112 ; https://doi.org/10.1038/s41746-019-0191-0

INTRODUCTION

A principle tenet of precision medicine is that subpopulations may be reasonably identified who differ in their disease risk, prognosis and response to treatment due to differences in underlying biology and other characteristics. The availability of multidimensional datasets that capture such variation can be 'trained' using artificial learning algorithms to find the cryptic phenotypic or genotypic structures, discussed subsequently, to then predict risk of disease, treatment response, prognosis and other outcomes in individual patients based on their own characteristics. The realization of this will offer clinicians the opportunity to more carefully tailor interventions—whether disease modifying or preventative in nature—to individual patients, contrasting with the current inductive process of symptom classification, and sometimes vague and inexact process of treatment decisions. One challenge of precision medicine is the high-performance computing requirements to process multidimensional datasets. However, computer capabilities have grown exponentially in recent years, and the integrated efforts of the international scientific community have made available large multidimensional biological and clinical datasets.[1–5] Recently, prediction algorithms utilizing artificial intelligence approaches for cancer[6–9] and cardiovascular disease[10,11] have shown promising results, predicting disease risk with a higher degree of precision.

In part, of course, success is predicated on the availability of accurate biological measurements, adequate quantification of relevant environmental factors and, from a genomic perspective, the identification of variants of known penetrance. Realizing a similar approach to the group of disorders of brain development termed 'neurodevelopmental disorders' (NDD) has a number of obstacles.[12] The NDDs are a group of early childhood onset disorders that impact different domains of cognitive development, motor function and other higher brain functions, and are lifelong in nature. Among the NDDs are severe disorders that impact multiple domains of cognitive functioning, such as intellectual disability (ID), as well as severe and pervasive disorders of social communication (autism spectrum disorder, ASD), motor function and cognition (epilepsy encephalopathies), and behavioral regulation (attention deficit hyperactivity disorder, ADHD). Some NDDs, particularly single gene disorders with more severe cognitive and medical consequences, are very rare. ASD and ADHD in particular are now relatively common, and result in major functional impairment, in part related to the high rates of co-morbidity. Epilepsy is one such comorbidity, with 20% of people with ASD also receiving this diagnosis. Moreover, epilepsy itself is often neurodevelopmental, although can sometimes occur de novo in adulthood or later in life. NDD co-morbidities are common and can make diagnosis challenging. Moreover, there is a degree of overlap in phenotype between different disorders, and phenotypic variability between individuals with the same diagnosis.[13–15] These complexities, often resulting in misdiagnosis or even missed diagnosis, are a major catalyst for the implementation of precision medicine. This is particularly so because, as a group, such disorders place a significant burden on healthcare. As such, early diagnosis and targeted therapeutic interventions to those who are most likely to benefit are universally agreed public health priorities.[16,17]

In this review, the ambition of precision medicine will be described, and success and implementation in medical practice so far will be briefly presented, with certain cancers and

[1]Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE. [2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada. [3]Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK. [4]School of Computing, Newcastle University, Newcastle upon Tyne, UK. *email: mohammed.uddin@gmail.com; marc.woodbury-smith@newcastle.ac.uk

cardiovascular disease as examples of success. The neurodevelopmental disorders will then be introduced and their inherent etiological and clinical complexities. Importantly, whilst large, principally genomic and clinical, datasets are available pertaining to individuals with NDD, using these data to facilitate improved diagnosis, therapeutic intervention and clinical outcomes is not straightforward. We will discuss the issues of clinical heterogeneity, lack of diagnostic clarity and biological overlap that characterize the NDDs. We will consider in detail potential approaches to address this complexity using epilepsy as exemplification, and then describe the outlook for artificial intelligence as applied to NDDs.

Precision medicine and artificial intelligence

Precision medicine is a healthcare pathway that employs numerous technologies to guide individually tailored diagnosis and treatment for patients. The availability of technologies, including high performance computing (HPC), as well as large biological datasets, are critical for the implementation of a precision medicine pathway that has the power to impact on healthcare. At the center of this strategy is a set of computer algorithms that identify patterns in multidimensional datasets that are then used to predict or optimize based on the availability of similar data on individual patients. Artificial intelligence algorithms apply learning strategies based on classification or pattern recognition to (multi-dimensional) input data in order to be able to predict from future datasets. In clinical medicine, for example, this may involve using results of pathological specimens to predict diagnosis and staging for the pathological specimen received on a new patient. There are many AI algorithms available, broadly defined according to whether they are supervised or unsupervised. Methods include the support vector, random forest, neural network and an evolutionary algorithm (EA). A brief overview of these methods is provided in Box 1. In recent years, both neural network driven machine learning and evolutionary algorithm (Fig. 1) have shown promising predictive potential for problems that are not solvable in polynomial algorithms (known as NP-hard problems).[18–20] These two models can be adapted by providing input data in supervised, unsupervised or semi-supervised models (see definition in Box 2).

In the last few decades, digitization of medical health record added a massive amount of data related to healthcare. Large digitization initiatives like EMERGE network, and 'All of Us' by NIH, USA[25]; Electronic Health Record (EHR) initiatives by Canadian Institutes of Health Research,[26] National Health Service, UK[27] are some of the world's largest electronic health record databases. The application of AI algorithms will be greatly benefitted from these large digitization efforts that can help establish genotype-phenotype relationship for genetic diseases and have the potential to infer numerous phenotypic correlations and associations. Of course, collecting large scale digital data will only be helpful if the data comprise relevant clinical information to model AI algorithms.

Box 1

Neural network is a model comprised of multiple layers of artificial neuron-based structures that are equipped with multilayer logistic regressions.[21] The model consists of an input and output layer and the artificial neurons in between are known as hidden layers (Fig. 1a). Neural network is widely used for machine learning (ML) related problems (i.e., pattern recognition or classification). Evolutionary algorithm (EA) is another model that was also adapted from nature.[22] EA is an effective optimization model that usually starts from the random assignment of an initial possible solution (known as a population) and progressively applies artificial genetic operators (mutation crossover etc.) to produce a new set of possible solutions in the subsequent generation (Fig. 1b). EA is well known for its capability in optimizing multiple objectives.[23] Although deep learning is becoming more popular, in a recent paper a type of EA algorithm was shown to outperform a deep learning algorithm in classical gaming theory.[24]

The application of AI in medicine is a burgeoning area of development in light of the major impact it could potentially have on healthcare provision. The application of machine learning in medical imaging on skin lesions[6] and treatable retinal diseases[1] has been the most impactful, and demonstrates the potential for this technology in medical practice. Deep learning algorithm to diagnose heart attack using 549 ECG records shows a sensitivity of 93.3% and specificity of 89.7%, outperforming cardiologists.[28]

Recently, DNA sequencing technology adopted machine learning to read out long stretches of DNA fragments from digital electronic signaling data. Long read technologies are important to resolve repetitive regions in the genome and detect complex structural variants. The current short reas technology can not resolve these issues and it is still unknown the disease risk contribution from repetitive region and structural variation of the genome. Nanopore sequencing technology in particular uses a neural network based deep learning method to 'call' DNA bases from the electronic signal produced by the nanopore flow cells. This method has accuracy over 98% and can produce mega base long DNA reads.[29] There has been an attempt to use AI in the clinical classification of genomic variation, based on the characterization of non-coding variants[30] splicing code,[5] DNA/RNA binding proteins[2] and non-coding RNA (ncRNA)[31] using large-scale molecular datasets. Classifying mutations according to their clinical relevance is very complex due to the largely unknown penetrance of individual variants, (i.e., the probability of diagnosis given a particular variant is identified, or mathematically, P(disease+|variant+)) Moreover, high penetrance variants are largely infrequent, with those of low penetrance much more common. Although most of the variants are non-coding in our genomes, determining pathogenicity of rare or common non-coding variants still requires major advancement in genomics. It will require multidimensional biological data and the use of artificial intelligence approaches to decipher the pathogenicity. Furthermore, many penetrant variants are also known to have more than one clinical manifestation, known as pleiotropy, and many diagnoses are characterized by variable presentation (phenotypic heterogeneity). Despite this, however, recent deep learning methods have had some degree of success in the correct

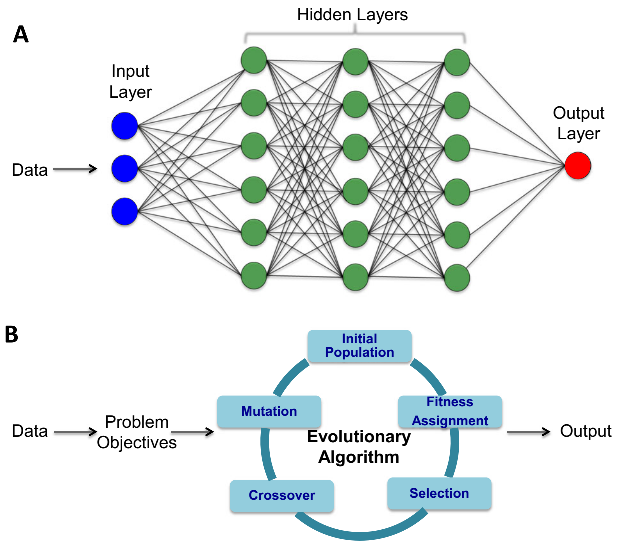

Fig. 1 Most promising artificial intelligence algorithms. **a** Simplified illustration of a basic model of neural network that is widely used in deep learning algorithms and (**b**) the components of evolutionary algorithm framework for multi objectives optimization related problem.

Table 1. Major neurodevelopmental disorders, prevalence, genetic inheritance, sex ratio, and genetic diagnostic yield.

Major neurodevelopmental disorders	Prevalence (approximately)	Sex ratio (male/female)	Genetic diagnostic yield (SNV, Indel and CNV)
Autism spectrum disorders	1.69%[CDC]	4:1	>40%[43,52]
Epilepsy	1.2%[119]	1:1	>45%[120,121]
Intellectual disabilities	1.7%[122]	2:1	>50%[123–125]
Single gene disorders	<1%	1:1, except for X linked mental retardation syndromes	100% (complete diagnosis)

CDC Centers for Disease Control and Prevention, USA

interpretation of phenotype and genomic data for disease risk in numerous types of cancer,[6,7,9,32,33] diabetic retinopathy[34,35] and pharmacogenomics.[36–38] For example, in discriminating lymph node metastases, 7 independent deep learning implementations showed greater discrimination power (i.e., in relation to pathological versus non-pathological) compared to 11 pathologists.[7] The best deep learning algorithm performed with an area under the curve (AUC) of 0.99, compared to 0.88 for 'best' clinician-derived score. The specificity found to be similar between AI and the diabetic retinopathy expert, AUC 0.96 and 0.98, respectively.

Neurodevelopmental disorders (NDDs)

Neurodevelopmental disorders have their onset early in childhood and impact on a variety of functional domains, including cognition and executive function, language and social function, and motor function and behavior control.[39–41] A number of different diagnoses are subsumed within this category, including intellectual disability (ID),[42] autism spectrum disorder (ASD),[4,43] attention deficit hyperactivity disorder (ADHD),[44] tic disorders, and other movement disorders.[45,46] Epilepsy and other early onset brain disorders, with or without associated ID, are also classified as NDDs[47,48] (Table 1). NDDs such as ASD and ADHD are common, lifelong disorders that affect males more commonly than females. In contrast, some syndromal NDDs, particularly single gene disorders, are individually very rare. As such disorders are often defined according to known biological abnormality (e.g., Down Syndrome, Fragile X syndrome, Tuberous Sclerosis) much of what we discuss in this paper is specific to common NDDs that are not defined according to known biology. Patients present with varying degree of severity, and comorbidity for two or more NDD diagnoses is common. With the exception of certain epilepsy syndromes (e.g., West Syndrome), which can be diagnosed more definitively based on the results of electroencephalography (EEG), most NDDs, including epilepsy itself, are diagnosed according to the presence of a threshold number of symptoms identified by direct observation or informant history. This is particularly problematic, as the availability of reliable information will vary from individual to individual, and even expert opinion can vary from clinician to clinician, such that diagnostic endorsement is often not definitive. Moreover, due to the developmental nature of this category of disorders, the clinical picture can vary over time,[49] with symptoms becoming more or less severe as the child grows. The availability of a more stable and objective way to classify individuals with NDDs is clearly needed, but currently this fuzziness within the diagnostic pathway is a significant barrier for the implementation of precision medicine in neurodevelopmental disorders.

All the NDDs considered in this current discussion are principally genetic in etiology.[50] For example, the early twin and family studies in ASD all supported a strong, heritable genetic component, and ASD and a lesser phenotype termed the Broader Autism Phenotype (BAP) do tend to run in families.[51] Some individual cases may result from rare, highly penetrant mutations,[4,43,52] some of which segregate in a Mendelian fashion. Some rare genetic syndromes, such as Fragile X[53] and Tuberous Sclerosis,[54] are associated with a number of NDDs. In contrast, however, most appear to result from a more complex genetic architecture that involves one or more genetic variants of variable penetrance interacting with other epigenetic mechanisms and environmental factors.[13,55,56] Understanding this genetic complexity is important, and it is anticipated that technological developments, both in silico but also laboratory based, will help unravel this. What is also striking is their degree of overlap in common-SNP based genetic etiology,[57] and pattern of differentially expressed genes. To date, over 250 genes have been reported to have strong association with NDD.[58] A very small number of genes (*SCN2A*, *CHD8*, *STXBP1*) and loci (16p11.2 microdeletion, 15q13.3 microdeletion etc.) that are found to be enriched within NDD are still below the level of 1% frequency threshold.[48,59–61] The current clinical genetic diagnostic yield for severe, syndromic NDDs associated with ID is approximately 40% and it is higher if genome sequencing data are available for other members (parents, siblings) of the family.[62]

In imaging studies, similarities in brain function evident from fMRI and diffusion tensor imaging also point to overlap at the level of intermediate phenotype between a number of NDDs such as ASD and ADHD.[63,64] Studies have examined diagnosed individuals while performing different neuropsychological tasks in the scanner, and the regions and structures in the brain that are active have been elucidated. This overlap in intermediate phenotype also extends to other mental disorders that are of later onset but are also increasingly being seen from a developmental perspective, such as schizophrenia and bipolar affective disorder.[65] However, at a clinical level, the phenotypes differ quite markedly. As we discuss subsequently, machine learning offers the opportunity to examine biological datasets in both a supervised and unsupervised manner, thereby providing both predictive models for diagnosis and treatment, as well as, theoretically, examining how multidimensional datasets may inform new models of classification. Specifically regarding classification, AI may offer new insight into how overlap at the biological level maps into disorders that are different clinically.

Artificial intelligence in NDDs

The availability of fMRI that enabled the high-resolution capture of brain activity was a major milestone[66] in NDD diagnosis and therapeutics in the 90s (Fig. 2). Since then, the human genome has been mapped[67,68] and exome and whole-genome sequencing technologies have led to the detection of hundreds of disease causal genes and loci for ASD and other NDDs.[43,69–71] Indeed, conducting exome or genome sequencing for newborn babies at high risk of genetic abnormalities is now becoming more frequent and cost effective.[72] Subsequently, the advent of transcriptome sequencing dependent technologies led to the establishment of

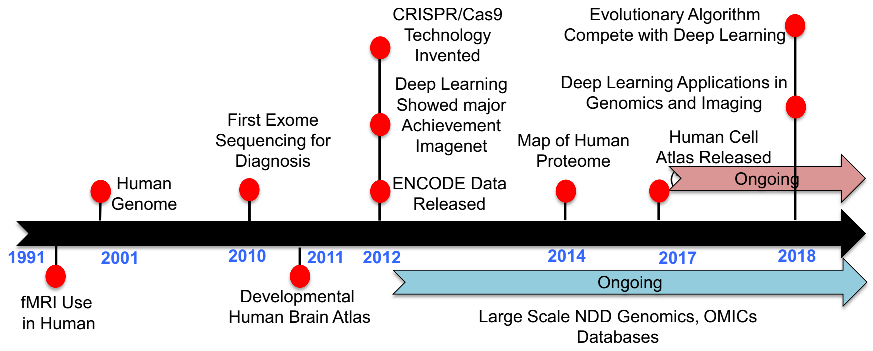
CRISPR/Cas9 Technology Invented
Evolutionary Algorithm Compete with Deep Learning
Deep Learning Showed major Achievement Imagenet
Deep Learning Applications in Genomics and Imaging
First Exome Sequencing for Diagnosis
Map of Human Proteome
Human Cell Atlas Released
Ongoing
Human Genome
ENCODE Data Released
1991
2001
2010
2011
2012
2014
2017
2018
Ongoing
fMRI Use in Human
Developmental Human Brain Atlas
Large Scale NDD Genomics, OMICs Databases

Fig. 2 Historical milestones related to precision medicine and artificial intelligence.

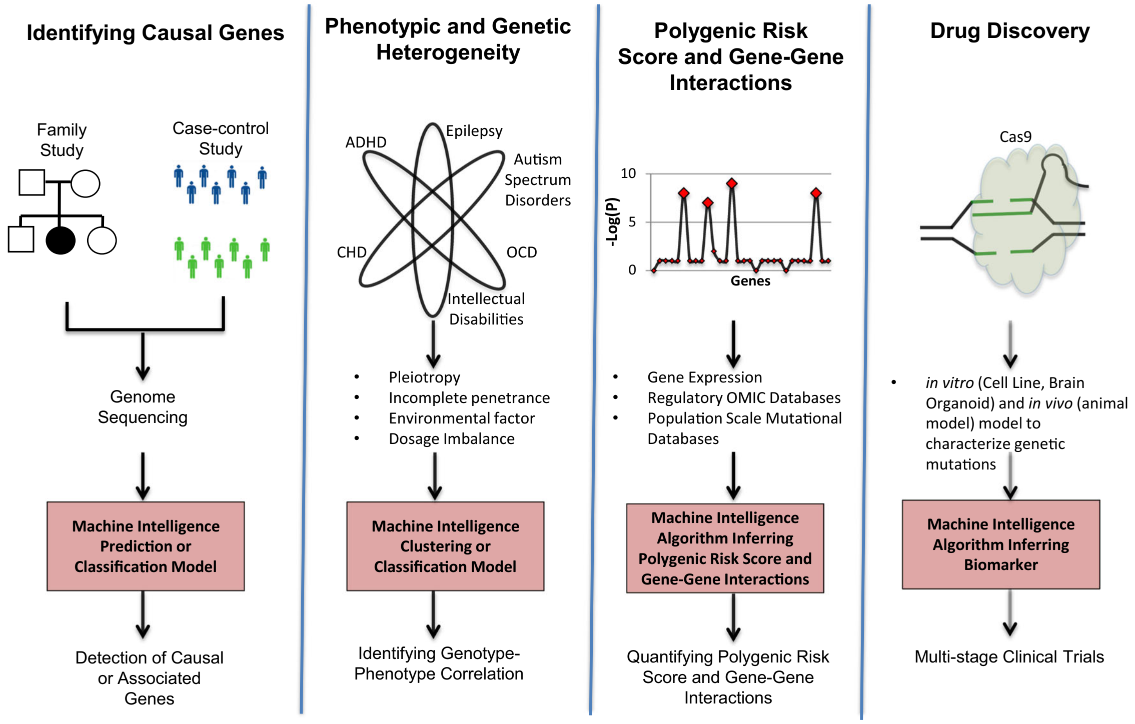
Identifying Causal Genes
Family Study
Case-control Study
Genome Sequencing
Machine Intelligence Prediction or Classification Model
Detection of Causal or Associated Genes
Phenotypic and Genetic Heterogeneity
ADHD
Epilepsy
Autism Spectrum Disorders
CHD
OCD
Intellectual Disabilities
Pleiotropy
Incomplete penetrance
Environmental factor
Dosage Imbalance
Machine Intelligence Clustering or Classification Model
Identifying Genotype-Phenotype Correlation
Polygenic Risk Score and Gene-Gene Interactions
10
5
0
-Log(P)
Genes
Gene Expression
Regulatory OMIC Databases
Population Scale Mutational Databases
Machine Intelligence Algorithm Inferring Polygenic Risk Score and Gene-Gene Interactions
Quantifying Polygenic Risk Score and Gene-Gene Interactions
Drug Discovery
Cas9
in vitro (Cell Line, Brain Organoid) and in vivo (animal model) model to characterize genetic mutations
Machine Intelligence Algorithm Inferring Biomarker
Multi-stage Clinical Trials

Fig. 3 Complex unresolved problems in neurodevelopmental disorders that artificial intelligence algorithms can create an impact.

the Allen developmental human brain atlas[73] in 2011, ENCODE database profiling the non-coding elements in the human genome[74] in 2012, and the Human Cell Atlas[75] in 2017. Multiple sequencing consortiums focussed on the NDDs were also started during the period of 2012 and 2014 with the aim of identifying disease-implicated variants, and making exome and WGS data available to the scientific community for further study.[52,70,76,77] Bearing in mind that most identified genetic variation is of unknown pathogenicity, and little is known about functional consequences, the discovery of CRISPR/Cas as a gene editing tool in 2012 has allowed scientists to better characterize identified genetic variants.[78,79]

In recent years, artificial intelligence approaches has been used in autism spectrum disorder,[5,12,14,15] epileptic encephalopathy,[80–82] intellectual disability,[83–85] attention deficit hyperactivity disorder (ADHD),[86] and rare genetic disorders.[2] In our discussion of AI in NDDs, three layers of analyses will be considered. The overriding theme will be the application of these methods to multidimensional NDD biological datasets, and the complexities therein (Fig. 3).

AI approaches are critical for identifying causal genes and loci. Although the current genetic diagnostic yield (including copy number variation (CNV), single nucleotide variants (SNV), and indel) for severe, syndromal ID is around 50% (Table 1) we still do not know genes or loci for NDDs more generally, which includes many of the cases whether there is no ID and/or evidence of craino-facial dysmorphology.[58,62] In addition, many identified loci are confounded by unknown penetrance, and, beyond bioinformatic prediction, do not have a strong evidential basis of support. De novo CNVs and SNVs and loss of function (LOF) mutations are certainly significantly enriched in individuals with NDD compared to typically developing controls.[4,71,76] Unfortunately, bioinformatic prediction is still unable to confidently classify the more common missense mutations according to pathogenicity. Indeed, identifying causal genes from these 'variants of uncertain significance' (VUS) remains a major unresolved problem that does lend itself to a AI solution. Recently, the identification of *OTUD7A* as a pathological gene in the 15q13.3 microdeletion syndrome locus illustrated the power of integrating computational and molecular approaches to resolving causality in CNVs.[87] Whilst this approach may certainly provide one solution, it is costly and, importantly, time intensive. Recently, post-zygotic mutations from blood and brain have been shown to be associated with ASD, epilepsy and ID.[48,60,88] The abundance of neuron specific mutations has also been reported in the literature.[89] What is unclear is the proportion of cells with potentially pathogenic postzygotic mutation. To comprehensively resolve genetic risk in relation to NDDs, there are other genomic regions that still need careful evaluation, such as non-coding variants, common variants and repeated sequences (over 40% of the entire genome). Moreover, an unsupervised learning approach, discussed below, may offer the opportunity to identify new patterns to data independent of these diagnostic categories.

Despite these limitations, AI approaches have recently shown reasonable success for improving genetic diagnostics in NDDs. As indicated above, one of challenging task is the correct classification of missense variants and Human Splicing Code,[5] and DeepSEA[30] showed very promising results in missense variant interpretations. The application of Human Splicing Code is one of the first machine-learning algorithms that shows convincing evidence of accurately classifying disease-causing variants, including those that are intronic. This method applies a Bayesian machine learning algorithm to model splicing dysregulation from a set of three or triplet exons. The method demonstrated pathogenic missense variants in ASD and in spinal muscular atrophy,[5] including variants that had not previously been classifiable in this way. In contrast, DeepSEA is a deep learning based algorithm that predicts the noncoding variant effects de novo from sequence data. The model uses large-scale chromatin-profiling data, including transcription factor binding, DNase I sensitivity and histone-mark profiles to predict the functional consequences of a non-coding variant. These and other algorithms are performed independent of established diagnostic categories, and serve to enrich the information for each genomic element for incorporation into downstream analyses discussed subsequently. Such holistic approaches, therefore, resolve variant pathogenicity through the interpretation of multidimensional omics data in the context of different NDD diagnoses. The recent advent of long range sequencing technologies (e.g., Pacific Biosciences, Oxford Nanopore Technology and others) are producing high quality DNA sequencing data that allow repeated sequences to be resolved.

AI approaches are critical to elucidate hidden structure in phenotype and genetic heterogeneity. As indicated above, both phenotypic and genetic heterogeneity characterize NDDs. For example, 15q13.3 microdeletion syndrome impacts multiple domains of cognitive function and is associated with heterogenous phenotypes, including epilepsy/seizure (57%), speech delay (16%), and ASD (11%).[90] There are hundreds of such CNVs with no straightforward mapping between manifested phenotypes and the variants/genes.[40,58] Despite the possibilities, there remains the problem of phenotype, and in particular, the oversimplification of dichotomizing phenotypes such as ASD and ADHD into 'caseness'. Variant information classified according to algorithms such as those defined above, as well as incorporation of other layers of biological data (neuroimaging, neurophysiology, neuropsychology), can be used to identify hidden structure in data, particularly if a unsupervised approach is used. These hidden structures may or may not map onto existing diagnostic categories, but, crucially, may be more closely aligned with endophenotypes, treatment response, prognosis and other clinical and outcome parameters. This discovery-driven approach may validate existing clinical diagnostic models of disorder classification, as well as potentially identify new models of classification that are driven principally (or, indeed, entirely) by the clustering of biological data. In NDDs in particular, diagnostic criteria have evolved significantly over time, principally due to a lack of clear, objective *sine qua non* for each disorder. Unfortunately, this evolution has seen the boundaries between disorders dissolving, milder forms being pathologized and discrete diagnostic categories morphing into spectra. Whilst biology does to some extent inform this nosological evolution, greater emphasis needs to be placed on using AI approaches on large-scale datasets to validate or challenge existing classification paradigms. Moreover, even if syndromes such as ADHD and ASD do truly exist as spectra, AI may be useful in identifying boundaries, perhaps informed by outcome and prognosis.

Although attempts with neural network deep learning approach showed that by combining fMRI with phenotypic data ASD classification can be improved,[91] this is still predicated on the fundamental existence of a categorical diagnostic label, *viz*. ASD, that may not correctly capture the structure in the underlying data.[92] Similarly, in epilepsy, EEG endophenotypes have been proposed[93] and purely EEG-based classification of seizures have been investigated theoretically and clinically.[94] However, none of these methods have been applied in a quantitative context, perhaps as the diagnosis of subtypes of epilepsy often rests heavily on qualitative EEG observations.

Methods are also needed that allow individuals to be assigned to more than one category in a probabilistic manner. For example, an individual may fall into diagnostic or endophenotypic category A with a probability of 0.9, and category B with a probability of 0.6 (we will highlight some examples of this in psychiatric conditions, and suggest that similar approaches can be taken in NDD). This closely reflects the reality of NDD symptom manifestation, whereby an individual with, say, ASD is also very likely to manifest ADHD or one of the other NDD diagnosis. In other words, co-morbid conditions may share an endophenotype that has clear diagnostic biomarkers. We need methods to both identify such diagnostic biomarkers, and to evaluate risk of different diagnostic categories for an given individual.

The availability of data-driven clinical diagnostic entities may also facilitate the triaging of patients in clinical practice. There is currently little opportunity to do this, as even well-designed screening instruments have limited reliability between sexes and across different ethnic groups. Although diagnostic criteria exist, there is much variation between clinicians on diagnostic thresholds used, which beyond the need for symptoms to impact on functioning are otherwise not explicitly written into these criteria. With the availability of data-driven categorization, there may be an opportunity, therefore, for the results of biological tests to inform who should be referred for further evaluation and or monitoring in a similar way to other medical tests. In addition, current diagnostic assessments for NDDs can be lengthy, and their multidisciplinary nature costly, leading to long waiting lists for children to receive diagnostic assessments. There is, therefore, a real opportunity for AI to automate some of the tasks in the diagnostic pathway and thereby have far reaching implications for clinical care and healthcare economics.

AI algorithms require major push to determine polygenic risk scores and gene-gene interactions. As we have stated before, hundreds of genes are involved in NDDs and the variability of gene variants (both common and rare) contributes to the overall pattern of brain function, as evidenced by fMRI and EEG, and phenotype at the clinical level. The risk prediction for each individual mutation is a complex process, as it will require a large number cases and controls to quantify significance. Similarly, predicting disease phenotypes from MRI and EEG data has similar problems, and incorporating data across these biological levels has challenges. For example, in epilepsy, siblings can share a very similar genetic and environmental background, but some develop epileptic seizures, while their siblings do not. In such cases, even the background EEG can appear very similar in siblings, but the exact factors causing one to have seizures is not well understood.

Even after quantifying risk factors, genetics lack well-established statistical or computational model that can utilize multiple variant or gene risk factors and combine them into a unified polygenic risk score. A neural network based approach on quantifying gene score for polygenic trait (i.e height) using single nucleotide polymorphism data showed promising improvements out performing previous methods.[95]

Polygenic risk prediction in NDDs remains problematic in light of the largely negative findings from underpowered genome-wide association studies (GWAS). Recently, a genome wide association study on large autism spectrum disorder and control cohort identified five common variants that confers very small risk factor.[57] When the proportion from de novo risk factor is substantially large, it is still not clear to what extent common variants contributes into the genetic risk factor of neurodevelopmental disorders. Therefore, although common variants are likely to play an important role, particularly in relation to quantitative traits that merge with those in the population-at-large.

Regarding the inference of gene–gene interaction, the number of permutations and combinations involving all the genes in our genome is a massive exponential search space and there is no efficient algorithm that can infer gene-gene interactions involving more than few genes. Statistical significance for the large number of interactions also suffers from the impact of multiple testing thresholds.[96,97] Although gene–gene interaction is likely a major contributor to the phenotypic variance of NDDs, there is currently no credible artificial intelligence algorithm able to cope with data on this scale. Certainly, a large number of genes can be simplified into a smaller number of protein–protein interactions or co-expression networks using traditional statistical model or algorithms, but as discussed subsequently, this is computationally NP-hard.[98] Adding to this complexity, there may be significant overlaps between gene lists and/or protein/co-expression networks for different neurodevelopmental disorders, and so discriminatory classification of, say, ASD vis-à-vis schizophrenia adds further layers of complexity. This latter problem is perhaps the most important one to consider, as this will inform the translational capacity of the algorithm for precision medicine.

Although NDDs are mostly genetic in etiology, environment will still impact on genetically driven brain patterning, and therefore have the potential to influence disease severity. Recently multiple independent reports have shown an association between postzygotic mosaic mutations and autism spectrum disorders, intellectual disability, epilepsy and other NDDs.[48,60,88] Currently the complex interactions between post zygotic mosaic mutations and environment is poorly understood.

AI approaches are at the frontier for therapeutic intervention and drug design. Currently, there are 51 food and drug administration (FDA) approved targeted gene specific drugs for neurology and psychiatric conditions. The advent of sequencing technology has principally been focussed on facilitating the implementation of early precision diagnostics. Precision therapeutics remains a major challenge for NDDs. Recently the advent of genome editing technologies (i.e., CRISPR/cas9), and antisense oligonucleotide therapy has allowed scientists to mimic cellular phenotype, and help identify precise molecular targets. For example, the application of CRISPR/cas9 helped knocked out a functional copy of *CHD8* gene in induced pluripotent stem cells (iPSCs). The knockout iPSCs showed differential expression of several thousands of genes in neural progenitors and impacts early differentiating neurons.[99] CRISPR/cas9 or other cas family proteins are still error prone, and the experimental success is not highly accurate. The future hope is that CRISPR/cas9, antisense oligonucleotide therapy and gene therapy based technologies will allow us to detect precise target molecules for most of the mutated genes in NDD. This will eventually lead to the experimental pathway to design target molecules (i.e., antisense oligonucleotide, or siRNA) to inhibit or disrupt the faulty pathway. Such drug design will require a major push on artificial intelligence algorithm implementation.

Recently the idea of repurposing drug is becoming a major area of research as well. Finding out common pathway for approved drugs can benefit multiple diseases. Finding out these shared pathway relationship is complex and do not have enough molecular and genomic data to establish a connection. For example, mTOR pathway impacts a certain group of epilepsy individuals and the same pathway found to be dysregulated in tuberous sclerosis individuals.[54,100] Hence, mTOR inhibitors have a great potential to impact treatment outcome for individuals with epilepsy carrying mTOR mutation or tuberous sclerosis related epilepsy.

Challenges for artificial intelligence in relation to NDDs

There exist major complexities involving deep phenotypic and large scale omics data. Artificial intelligence will eventually radically transform healthcare delivery for patients with NDDs, but there are major hurdles that need to be resolved. For example, the modifying effect of environment is not well understood, but may explain disease discordance in monozygotic twins and the observation of different genetic risk factors in siblings. Identical variants may have different phenotypic consequences, and even recurrent large deletions, or bioinformatically predicted damaging mutations, may result in phenotype among some but no apparent consequences among others. Except in a few specific situations (for example, fetal alcohol syndrome,[101] and microcephaly through infectious agents[102]), major environmental influences on NDDs and their contributions to phenotypic severity or heterogeneity are still unknown. Environmental impact is also highly likely to be a source for inducing post-zygotic mutations, recently shown to be associated with NDD.[48,60] Moreover, these environmental factors may differ between countries and continents. As such, artificial intelligence might capture structure in data for one geographic location that is not relevant for disease risk in another location. Unsupervised AI models (Box 2) can be utilized to identify previously unknown sub-structures within NDD cases based on environmental factors that are local population (Fig. 3).

On the technical side, similar problems also arise due to different methods, tools, and protocols being used to collect data. For example, reliability and reproducibility of neuroimaging findings depend hugely on many experimental factors.[103] Similarly, population scale omics data suffer from batch effect and technology specific biases.[104,105] Thus, although large databases may be available for machine learning approaches, great care has to be taken in the quality and comparability of datasets used. Otherwise, any structure and information extracted from the data using artificial intelligence may be completely trivially driven by the composition of the data.

Omics data are necessarily multidimensional, and characterized by a large computational burden.[106–108] Compounding this, with the advent of single cell genomics, the genomic architecture of NDDs is becoming apparent, and the formidable challenge this introduces to the understanding of disease pathophysiology. There are a large number of somatic variants that have been identified that have the potential to impact phenotypic severity.

Box 2

Supervised AI Algorithm: In supervised learning algorithm, the training data helps the algorithm learning a function that maps an input to an output based on known or labeled input-output pairs.

Unsupervised AI Algorithm: Unsupervised learning is a type of machine learning that involves unlabeled training data where the input-output relationship is not known and the algorithm infers patterns (or possible solutions) within datasets.

Semi-supervised AI Algorithm: Semi unsupervised learning is a type of machine learning that involves a mix of known and unknown training data that helps the algorithm to infer input-output relationships.

For example, there is evidence of certain somatic mutations associated with autism spectrum disorder, microcephaly, and epilepsy.[48,60,88] Recent analysis has also shown that up to 40% of neurons could have a large mega base scale copy number variation.[109] Single cell genomics has also identified private somatic mutations within each neuron.[89] Although the contribution of somatic mutation to disease risk is not well understood, this particular type of mutation will add unpredictable variance within machine learning approaches and will impact replication significantly.

Lack of proper training datasets (control and case) and complexity on interpretability are two major issues in AI. Although collaborative initiatives have resulted in the growing availability of datasets that can be used for training, there is still a need for larger, more complete biomedical datasets that are representative of different populations and tissues. Initiatives such as 100,000 Genomes Consortium, MSSNG,[76] The Simons Foundation Autism Research Initiative,[110] The Exome Aggregation Consortium,[111] ENCODE Project Consortium,[74] The Genotype-Tissue Expression Consortium,[112] Allen Brain Atlas[73] offer huge potential for training and testing AI algorithms, but only if the data are complete, and similar, robust, measures have been used across samples and tissues (Fig. 3). This is often an issue in NDDs where on the one hand it may be difficult to engage the participant in all potential assessments employed, and on the other hand different studies may have used different measures.

Clinical implementation of artificial intelligence algorithms should be informed by the needs of the healthcare practitioner and their patients. Most artificial intelligence algorithms work as a 'black box', which may raise concern among health professionals, and raise questions from clinical service users. To overcome this, gold standard AI protocols need to be established that can be understood by healthcare professionals. There is also a need to be transparent about the limitations of AI methods. For example, in genetic algorithms, it can be extremely difficult for clinicians to decipher how through random operations (i.e., mutation, crossover) and variables the model reaches fitness convergence for optimum solutions in a multidimensional search space.[113] Ultimately, however, despite the complexity of different algorithms, statistical models and tests are used to favor or refute evidence (i.e., *p*-values, false discovery rates, area under the curve), which can be understood by many professionals working in a clinical setting.[114]

The datasets themselves, multidimensional in nature, will also have been collected from multidisciplinary experts who may not necessarily 'talk the same language'. What one person may call 'case', therefore, may be 'non-case' to another, and genetic variants may similarly vary in their interpretation in relation to significance. In addition, pre-processing of data and, indeed, even the design of the original study itself from which data are being collated, may present additional confounds to data interpretation. Fortunately, scientific methods have become much more transparent in recent years, and accessing detailed information pertaining to the methods used is often readily available for datasets. Of course, it is equally important for such information to be transparent in relation to the use of AI methods themselves.

One of the major setbacks in NDDs is the paucity of available treatments. There has been a downshift in industry-sponsored trials of potential compounds for the treatment of different NDDs. This downshift includes issues related to drug design, the lack of positive control and replication. This is unfortunate, as we are now beginning to uncover different aspects of brain structure and function at the molecular level that are associated with phenotypic consequence. These compounds may be the focus for potential drug development, and their known pathophysiology may inform repurposing of existing compounds (Fig. 3). One of the central complicating factor in compound screening, is the three dimensional structure of proteins. The complexity of predicting the tertiary structure from polypeptide sequence is a computationally intractable problem.[98] AI based prediction algorithms can overcome such barriers through rigorous training datasets of polypeptide sequences.

Genome editing, antisense oligonucleotide therapy are two major technologies that show promise in facilitating an understanding of biology and consequently addressing the paucity of available treatments for NDDs. Recently, the clustered regularly interspaced short palindromic repeats (CRISPR) system showed the ability to correct mutations in vitro[79] and in vivo[115] in numerous diseases. Unfortunately, the CRISPR-Cas9 system currently lacks target precision. Moreover, the blood–brain barrier is a major challenge to deliver CRISPR like editing system in vivo into the brain cells. Regarding the delivery of genome editing machineries, recent efforts on vector and non-vector based CRISPR system delivery shows limited success on breaking the blood-brain barrier.[116,117] For NDD, future treatment options should implement AI based algorithms that can design genome editing or antisense oligonucleotide design tools that are compatible with the in vivo delivery mechanism. Without the integration AI based algorithms, the potential of precision medicine will not be fully realized in NDDs.

The paradigm shift promised by precision medicine will of course impact frontline healthcare staff, who will need training in its strengths and limitations, as well as the interpretation and translation of AI-driven knowledge into information that is clinically meaningful for patients. The healthcare sector will need to build its high performance computation (HPC) capacity, and innovators will need to devise sophisticated AI-platforms. Consideration will need to be given to protection of data and the legal framework by which such data are stored and shared. Indeed, such considerations need to be happening significantly before any implementation, meaning that even now such discussions should be taking place. The sensitive nature of healthcare (and social care) information demands an absolute watertight system both in terms of storage and sharing, but also algorithm performance. Having one's data breached, or being given someone else's clinical information should not happen. Moreover, being given an incorrect diagnosis as a result of algorithmic failure should also never happen, particularly as it may be more difficult to identify than clinician error.

Opportunities for ML in NDDs

The challenges described above, both in terms of the vicissitudinous nature of diagnostic labels that are poorly defined in the first place, and the high levels of variability observed across multiple levels of biological measurements, may also be reframed as real opportunities for ML. At its most radical, a completely unsupervised approach may identify new, and more biologically meaningful, diagnostic categories. One paradigm would see clinical and biological data pooled together, or alternatively biological data could 'drive' the generation of new diagnostic entities. This is very different from existing uses of ML which are predicated on the existence of diagnostic categories such as ASD, ADHD, and others. This does, however, introduce the risk of 'overfitting' data. In addition, of course, throwing out the baby with the bathwater may not be entirely appropriate, and so we also advocate ML endeavors that attempt to use existing diagnostic constructs according to available underlying biological data. Some such approaches have been reasonably successful in correct diagnostic assignment, and are more immediately implementable in clinical practice, as current treatment algorithms are very much focussed on these very diagnostic categories. Natural language processing (NLP) is another emerging field of machine intelligence that can automatically transform clinical text into structured clinical data.[118] NLP algorithms can analyze digital health records and psychiatric notes to identify relatedness among patients' phenotypes and their associated genetic markers. Although scientists have been working on NLP algorithms for the last few decades, significant improvements are still required from extracting text to understanding the clinical and biological relevance. ML approaches may be used to parse clinical text in the form of diagnostic reports. It is fairly standard practice for NDD diagnoses to be made following detailed assessment comprising clinical interview and objective testing such as IQ and other aspects of cognition. These assessments are typically summarized as a detailed report. Such reports contain rich data, that, theoretically, could be extracted using a suitable algorithm. This approach is based on the valid assumption that such data contain the unique insight of expert clinicians that might otherwise be overlooked in more formal measurements, or using pure biological data.

Analyzing data across multiple levels of biological function is also an attractive proposal in ML. By way of example, historically EEG has shown variable success in the identification of distinct neurophysiological patterns of impairment between ASD and controls. In brief, there is no consistent pattern of brain activation in response to particular stimuli that consistently differentiates the ASD from the non-ASD brain. However, there is strong reason to continue to pursue EEG-identified biomarkers for brain disorders, because as a method it represents a cost-effective, objective way of facilitating diagnosis that could easily be implemented in the clinical setting. ML algorithms with the capacity to handle longer EEG tracings (24 h for example) may be one potential avenue for exploration. Alternatively, consideration might be given to the ways in which other biological signals, from fMRI perhaps, improves the interpretability of EEG signals.

Finally, NDDs are not static disorders, but evolve over time, and one of its biggest challenges is the unpredictable nature of the progression. Thus, there is considerable within-subject variability in how the disorder changes over time, which is often neglected in the context of research studies. The challenge is to understand the manifestation of wide ranges of phenotypes during developmental stages in an individual that arises from the same genetic and neuronal substrate.[39,47,58] Application of artificial intelligence algorithm on longitudinal studies can be designed to capture the pattern of disease progression over time and the variability at the personal or sub-population level.

Bearing in mind how long it takes a compound to go from original identification to eventual therapeutic use, ML algorithms will also have a significant role to play in parsing the large volumes of data generated during drug-development, as well as prioritizing molecules based on their known biological properties. Integrating genomics within the artificial intelligence drug development algorithm will enhance the implementation of precision medicine for NDD. Genomic profile can add the sensitivity that the artificial intelligence algorithm requires to design drug at the individual or a sub population level. Treatment response in NDDs is one other area of healthcare delivery that could benefit from ML, and this extends to the management of mental health disorders more generally. Despite much research, predicting treatment responsiveness remains very poorly understood. This is particularly important as many treatments require a period of time before efficacy (or lack thereof) is established. Consequently, patients may remain essentially untreated for many weeks, months if identifying a successful drug requires several attempts. There will, of course, be many reasons why there is such large variation in treatment response, and ML offers the opportunity to identify structure in multidimensional data that captures such things as metabolism, absorption, and disorder characteristics (severity, comorbidity and so forth).

CONCLUSIONS

Artificial intelligence is already impacting healthcare, and it is hoped that some of the successes achieved so far in cancer and cardiovascular disease will also be seen in the NDDs. This will necessarily involve the integrated use of existing and new supervised and unsupervised learning approaches, as well as an HPC infrastructure that can manage the multidimensional nature of the emerging omics data. There needs to be major investment in new treatments that will map onto different disease categories and subcategories, and researchers need to step away from existing diagnostic constructs to embrace a more intermediate biologically driven level of phenotype that may map more neatly onto treatment response and clinical outcomes. The healthcare sector, which is already financially stretched, has a formidable task ahead: there needs to be a development of infrastructure, expertize in knowledge translation among healthcare professionals, and engagement with the service users themselves in developing new clinical pathways. Short term investment in ML will certainly have long term gains, both in terms of financial savings resulting from precision medicine, but also the ultimate improvement in the health of the population (Box 2).

Received: 3 December 2018; Accepted: 29 October 2019;
Published online: 21 November 2019

REFERENCES

1. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 e1129 (2018).
2. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
3. Gurovich, Y. et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).
4. Uddin, M. et al. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nat. Genet.* **46**, 742–747 (2014).
5. Xiong, H. Y. et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
6. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
7. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
8. Ainscough, B. J. et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.* **50**, 1735–1743 (2018).
9. Chang, P. et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am. J. Neuroradiol.* **39**, 1201–1207 (2018).
10. Bello, G. A. et al. Deep learning cardiac motion analysis for human survival prediction. *Nat. Mach. Intell.* **1**, 95–104 (2019).
11. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
12. Bone, D. et al. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *J. Autism Dev. Disord.* **45**, 1121–1136 (2015).
13. Coe, B. P., Girirajan, S. & Eichler, E. E. The genetic variability and commonality of neurodevelopmental disease. *Am. J. Med. Genet. C Semin. Med. Genet.* **160C**, 118–129 (2012).
14. Kosmicki, J. A., Sochat, V., Duda, M. & Wall, D. P. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Transl. Psychiatry* **5**, e514 (2015).
15. Wall, D. P., Dally, R., Luyster, R., Jung, J. Y. & Deluca, T. F. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE* **7**, e43855 (2012).
16. Bitta, M., Kariuki, S. M., Abubakar, A. & Newton, C. Burden of neurodevelopmental disorders in low and middle-income countries: A systematic review and meta-analysis. *Wellcome Open Res.* **2**, 121 (2017).
17. Mazurek, M. O., Curran, A., Burnette, C. & Sohl, K. ECHO Autism STAT: accelerating early access to autism diagnosis. *J. Autism Dev. Disord.* **49**, 127–137 (2019).
18. Padovani de Souza, K. et al. Machine learning meets genome assembly. *Brief Bioinform.* https://doi.org/10.1093/bib/bby072 (2018).
19. Rizzi, R., Cairo, M., Makinen, V., Tomescu, A. I. & Valenzuela, D. Hardness of covering alignment: phase transition in post-sequence genomics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 23–30 (2019).

20. Kapun, E. & Tsarev, F. De Bruijn superwalk with multiplicities problem is NP-hard. *BMC Bioinforma.* **14**(Suppl 5), S7 (2013).
21. Lee, H. K. H. Model selection for neural network classification. *J. Classification* **18**, 227 (2001).
22. Pelikan, M. *Hierarchical Bayesian Optimization Algorithm: Toward A New Generation of Evolutionary Algorithms* (Springer-Verlag, 2005).
23. Gen, M., Cheng, R. & Lin, L. *Network Models and Optimization: Multiobjective Genetic Algorithm Approach* (Springer, 2008).
24. Wilson, D. G., Cussat-Blanc, S., Luga, H. & Miller, J. F. Evolving simple programs for playing atari games. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 229–236 (2018).
25. Sankar, P. L. & Parker, L. S. The precision medicine initiative's all of us research program: an agenda for research on its ethical, legal, and social issues. *Genet. Med.* **19**, 743–750 (2017).
26. Gagnon, M. P. et al. Electronic health record acceptance by physicians: testing an integrated theoretical model. *J. Biomed. Inf.* **48**, 17–27 (2014).
27. Sheikh, A. et al. Implementation and adoption of nationwide electronic health records in secondary care in England: final qualitative results from prospective national evaluation in "early adopter" hospitals. *BMJ* **343**, d6054 (2011).
28. Strodthoff, N. & Strodthoff, C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiol. Meas.* **40**, 015001 (2019).
29. Boza, V., Brejova, B. & Vinar, T. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS ONE* **12**, e0178751 (2017).
30. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
31. Yi, H. C. et al. A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. *Mol. Ther. Nucleic Acids* **11**, 337–344 (2018).
32. Shu, C., Wang, Q., Yan, X. & Wang, J. Whole-genome expression microarray combined with machine learning to identify prognostic biomarkers for high-grade glioma. *J. Mol. Neurosci.* **64**, 491–500 (2018).
33. Wood, D. E. et al. A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.* **10**, https://doi.org/10.1126/scitranslmed.aar7939 (2018).
34. Raumviboonsuk, P. K. et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Med.* **25**, 1–9 (2019).
35. Ting, D. S. W. et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study. *npj Digital Med.* **24**, 1–8 (2019).
36. Athreya, A. P. et al. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine learning approach with multi-trial replication. *Clin. Pharmacol. Ther.* https://doi.org/10.1002/cpt.1482 (2019).
37. Kalinin, A. A. et al. Deep learning in pharmacogenomics: from gene regulation to patient stratification. *Pharmacogenomics* **19**, 629–650 (2018).
38. Preuer, K. et al. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* **34**, 1538–1546 (2018).
39. Thapar, A., Cooper, M. & Rutter, M. Neurodevelopmental disorders. *Lancet Psychiatry* **4**, 339–346 (2017).
40. Uddin, M. et al. Indexing effects of copy number variation on genes involved in developmental delay. *Sci. Rep.* **6**, 28663 (2016).
41. Hu, W. F., Chahrour, M. H. & Walsh, C. A. The diverse genetic landscape of neurodevelopmental disorders. *Annu. Rev. Genomics Hum. Genet.* **15**, 195–213 (2014).
42. Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T. & Saxena, S. Prevalence of intellectual disability: a meta-analysis of population-based studies. *Res. Dev. Disabil.* **32**, 419–436 (2011).
43. Jiang, Y. H. et al. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
44. Lionel, A. C. et al. Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD. *Sci. Transl. Med.* **3**, 95ra75 (2011).
45. Huisman-van Dijk, H. M., Schoot, R., Rijkeboer, M. M., Mathews, C. A. & Cath, D. C. The relationship between tics, OC, ADHD and autism symptoms: a cross- disorder symptom analysis in Gilles de la Tourette syndrome patients and family-members. *Psychiatry Res.* **237**, 138–146 (2016).
46. Zarrei, M. et al. De novo and rare inherited copy-number variations in the hemiplegic form of cerebral palsy. *Genet. Med.* **20**, 172–180 (2018).
47. Loussouarn, A., Dozieres-Puyravel, B. & Auvin, S. Autistic spectrum disorder and epilepsy: diagnostic challenges. *Expert Rev. Neurother.* 1–7, https://doi.org/10.1080/14737175.2019.1617699 (2019).
48. Uddin, M. et al. Germline and somatic mutations in STXBP1 with diverse neurodevelopmental phenotypes. *Neurol. Genet.* **3**, e199 (2017).
49. Krol, A. & Feng, G. Windows of opportunity: timing in neurodevelopmental disorders. *Curr. Opin. Neurobiol.* **48**, 59–63 (2018).
50. Woodbury-Smith, M. & Scherer, S. W. Progress in the genetics of autism spectrum disorder. *Dev. Med. Child Neurol.* **60**, 445–451 (2018).
51. Tick, B., Bolton, P., Happe, F., Rutter, M. & Rijsdijk, F. Heritability of autism spectrum disorders: a meta-analysis of twin studies. *J. Child Psychol. Psychiatry* **57**, 585–595 (2016).
52. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
53. Berry-Kravis, E. Mechanism-based treatments in neurodevelopmental disorders: fragile X syndrome. *Pediatr. Neurol.* **50**, 297–302 (2014).
54. Curatolo, P., Moavero, R. & de Vries, P. J. Neurological and neuropsychiatric aspects of tuberous sclerosis complex. *Lancet Neurol.* **14**, 733–745 (2015).
55. van Loo, K. M. & Martens, G. J. Genetic and environmental factors in complex neurodevelopmental disorders. *Curr. Genomics* **8**, 429–444 (2007).
56. Tran, N. Q. V. & Miyake, K. Neurodevelopmental disorders and environmental toxicants: epigenetics as an underlying mechanism. *Int. J. Genomics* **2017**, 7526592 (2017).
57. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
58. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
59. Uddin, M. et al. Genomic context analysis of de novo STXBP1 mutations identifies evidence of splice site DNA-motif associated hotspots. *G3* **8**, 1115–1118 (2018).
60. Lim, E. T. et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.* **20**, 1217–1224 (2017).
61. Katayama, Y. et al. CHD8 haploinsufficiency results in autistic-like phenotypes in mice. *Nature* **537**, 675–679 (2016).
62. Wright, C. F. et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.* **20**, 1216–1223 (2018).
63. Ha, S., Sohn, I. J., Kim, N., Sim, H. J. & Cheon, K. A. Characteristics of brains in autism spectrum disorder: structure, function and connectivity across the lifespan. *Exp. Neurobiol.* **24**, 273–284 (2015).
64. Qiu, M. G. et al. Changes of brain structure and function in ADHD children. *Brain Topogr.* **24**, 243–252 (2011).
65. Ameis, S. H. et al. A diffusion tensor imaging study in children With ADHD, autism spectrum disorder, OCD, and matched controls: distinct and non-distinct white matter disruption and dimensional brain-behavior relationships. *Am. J. Psychiatry* **173**, 1213–1222 (2016).
66. Raichle, M. E. A brief history of human brain mapping. *Trends Neurosci.* **32**, 118–126 (2009).
67. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
68. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
69. Ng, S. B. et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
70. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
71. Tammimies, K. et al. Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. *JAMA* **314**, 895–903 (2015).
72. Farnaes, L. et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom. Med.* **3**, 10 (2018).
73. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
74. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
75. Regev, A. et al. The human cell atlas. *Elife* **6**, https://doi.org/10.7554/eLife.27041 (2017).
76. RK, C. Y. et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
77. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
78. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
79. Ran, F. A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
80. Munsell, B. C. et al. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *Neuroimage* **118**, 219–230 (2015).
81. Shoeb, A. H. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment* (Harvard-MIT Division of Health Sciences and Technology, 2009).
82. Yuan, Q., Zhou, W., Li, S. & Cai, D. Epileptic EEG classification based on extreme learning machine and nonlinear features. *Epilepsy Res.* **96**, 29–38 (2011).

83. Quinodoz, M. et al. DOMINO: using machine learning to predict genes associated with dominant disorders. *Am. J. Hum. Genet.* **101**, 623–629 (2017).
84. Smyser, C. D. et al. Prediction of brain maturity in infants using machine-learning algorithms. *Neuroimage* **136**, 1–9 (2016).
85. Crippa, A. et al. Use of machine learning to identify children with autism and their motor abnormalities. *J. Autism Dev. Disord.* **45**, 2146–2156 (2015).
86. Duda, M., Ma, R., Haber, N. & Wall, D. P. Use of machine learning for behavioral distinction of autism and ADHD. *Transl. Psychiatry* **6**, e732 (2016).
87. Uddin, M. et al. OTUD7A regulates neurodevelopmental phenotypes in the 15q13.3 microdeletion syndrome. *Am. J. Hum. Genet.* **102**, 278–295 (2018).
88. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 1237758 (2013).
89. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
90. Lowther, C. et al. Delineating the 15q13.3 microdeletion phenotype: a case series and comprehensive review of the literature. *Genet. Med.* **17**, 149–157 (2015).
91. Dvornek, N. C., Ventola, P. & Duncan, J. S. Combining phenotypic and resting-state FMRI data for autism classification with recurrent neural networks. *Proc. IEEE Int. Symp. Biomed. Imaging* **2018**, 725–728 (2018).
92. Du, Y., Fu, Z. & Calhoun, V. D. Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Front. Neurosci.* **12**, 525 (2018).
93. Chowdhury, F. A. et al. Revealing a brain network endophenotype in families with idiopathic generalised epilepsy. *PLoS One* **9**, e110136 (2014).
94. Lagarde, S. et al. The repertoire of seizure onset patterns in human focal epilepsies: determinants and prognostic values. *Epilepsia* **60**, 85–95 (2019).
95. Pare, G., Mao, S. & Deng, W. Q. A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci. Rep.* **7**, 12665 (2017).
96. Motsinger, A. A. & Ritchie, M. D. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Hum. Genomics* **2**, 318–328 (2006).
97. Yosef, N. et al. A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data. *Bioinformatics* **23**, e91–e98 (2007).
98. Lathrop, R. H. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* **7**, 1059–1068 (1994).
99. Wang, P. et al. CRISPR/Cas9-mediated heterozygous knockout of the autism gene CHD8 and characterization of its transcriptional networks in neurodevelopment. *Mol. Autism* **6**, 55 (2015).
100. Curatolo, P., Moavero, R., van Scheppingen, J. & Aronica, E. mTOR dysregulation and tuberous sclerosis-related epilepsy. *Expert Rev. Neurother.* **18**, 185–201 (2018).
101. Mattson, S. N., Bernes, G. A. & Doyle, L. R. Fetal alcohol spectrum disorders: a review of the neurobehavioral deficits associated with prenatal alcohol exposure. *Alcohol Clin. Exp. Res.* https://doi.org/10.1111/acer.14040 (2019).
102. Mlakar, J. et al. Zika virus associated with microcephaly. *N. Engl. J. Med.* **374**, 951–958 (2016).
103. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
104. t Hoen, P. A. et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
105. Leek, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
106. Unger, R. & Moult, J. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. *Bull. Math. Biol.* **55**, 1183–1198 (1993).
107. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A. & Eskin, E. Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics* **26**, i183–i190 (2010).
108. Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med. Genomics* **8**, 33 (2015).
109. McConnell, M. J. et al. Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
110. Abrahams, B. S. et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 36 (2013).
111. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
112. Carithers, L. J. & Moore, H. M. The genotype-tissue expression (GTEx) project. *Biopreserv. Biobank* **13**, 307–308 (2015).
113. Deb, K. e. *Evolutionary Multi-criterion Optimization: 10th International Conference* In *2019 Proceeding*s. March 10–13 (EMO, East Lansing, 2019).
114. Huynh-Thu, V. A., Saeys, Y., Wehenkel, L. & Geurts, P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics* **28**, 1766–1774 (2012).
115. Maddalo, D. et al. Corrigendum: In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **524**, 502 (2015).
116. Lau, C. H. & Suh, Y. In vivo genome editing in animals using AAV-CRISPR system: applications to translational research of human disease. *F1000Res* **6**, 2153 (2017).
117. Swiech, L. et al. In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat. Biotechnol.* **33**, 102–106 (2015).
118. Sheikhalishahi, S. et al. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med. Inf.* **7**, e12239 (2019).
119. Zack, M. M. & Kobau, R. National and state estimates of the numbers of adults and children with active epilepsy-United States, 2015. *MMWR Morb. Mortal. Wkly. Rep.* **66**, 821–825 (2017).
120. Helbig, K. L. et al. Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genet. Med.* **18**, 898–905 (2016).
121. Allen, N. M. et al. Chromosomal microarray in unexplained severe early onset epilepsy-a single centre cohort. *Eur. J. Paediatr. Neurol.* **19**, 390–394 (2015).
122. Bourke, J., de Klerk, N., Smith, T. & Leonard, H. Population-based prevalence of intellectual disability and autism spectrum disorders in western australia: a comparison with previous estimates. *Medicine* **95**, e3737 (2016).
123. Monroe, G. R. et al. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet. Med.* **18**, 949–956 (2016).
124. Chong, W. W. et al. Performance of chromosomal microarray for patients with intellectual disabilities/developmental delay, autism, and multiple congenital anomalies in a Chinese cohort. *Mol. Cytogenet.* **7**, 34 (2014).
125. Chen, X. et al. Genome-wide array analysis reveals novel genomic regions and candidate gene for intellectual disability. *Mol. Diagn. Ther.* **22**, 749–757 (2018).

ACKNOWLEDGEMENTS

The authors would like to thank the grant agencies funded this work—Al Jalila Foundation, Sandooq Al Watan, and grants from College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, NeuroGen Technologies Research Grant, and Harvard Medical School Center for Global Health Delivery-Dubai.

AUTHOR CONTRIBUTIONS

M.U. and M.W.S. conceptualized the content and the outline of the paper. M.U., Y.W., and M.W.S. conducted extensive literature review and wrote the draft version of the paper.

COMPETING INTERESTS

M.W.S. has received financial reimbursement from Servier, a pharmaceutical company, in connection with his role as national co-ordinator for ASD clinical trials. M.U. is a member of the scientific advisory board of NeuroGen Technologies Ltd, a genetic diagnostic company. Y.W. declares no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to M.U. or M.W.-S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.