Significance of Artificial Intelligence in Clinical and Genomic Diagnostics

Md Habibur Rahman[1], Md Abdur Rahman Biswash[2], Md Abu Bakar Siddique[3], Md Mostafizur Rahman[3], Moushumi Afroza Mou[3], Asim Debnath[4*], Md Fatin[2]

Abstract

AI stands for artificial intelligence, the computer systems emulating human intelligence to complete difficult tasks like interpreting data. The recent advancements in the field of AI in general, particularly the development of deep learning algorithms and hardware development in the form of GPU, have now made it possible to apply it in medical diagnostics. Artificial Intelligence frameworks are adept in treatment of vast, complex data and thus are an efficient tool for clinical assessments. AI is already transforming image-based diagnostics, electronic health records (EHRs) and clinical genomics, as we review here. We summarize AI's ability to work with problem classes like computer vision, time series analysis, and natural language processing, each of which corresponds to specific diagnostic tasks. Some novel approaches are presented in clinical genomics such as in the areas of variant calling, genome annotation and phenotype to genotype mapping. Deep learning's capacity to extract useful signals from genomic and phenotypic data with minimal human guidance is accelerating precision medicine. Convolutional and recurrent neural networks have been shown to outperform all other methods for genomic data interpretation. These tools do have limitations, including dependence on large, high quality training datasets as well as robust phenotype data. Here we discuss how advanced biobank projects are a path towards that future even if AI has not yet fully delivered on its promise to enable complex human phenotype prediction. Interpretability, bias mitigation and solving barriers to data collection are crucial elements for AI to thrive within the context of personalized medicine. The constant growth of AI has the potential to completely change genetic studies as well as clinical diagnostics.

Keywords: Artificial Intelligence (AI), Clinical Diagnostics, Genomics, Clinical Applications, Data Interpretation, Deep Learning.

Significance | Artificial intelligence revolutionizes clinical diagnostics by enhancing data interpretation, enabling precision medicine, and accelerating advancements in genomics and personalized healthcare.

*Correspondence. Asim Debnath, Burnett School of Biomedical Sciences, College of Medicine, University of Central Florida, Orlando, FL 32827, USA
Email: asim.debnath@ucf.edu

Editor Aman Shah Abdul Majid, Ph.D., And accepted by the Editorial Board Jan 10, 2025 (received for review Oct 20, 2024)

Introduction

Artificial intelligence (AI) is where a human's real intelligence is simulated on a non-living thing. We define AI in the field of clinical diagnostics as any computer system that can properly and accurately interpret health data, especially in its native form as seen by humans. These clinical applications often leverage AI frameworks to facilitate timely interpretation of large heterogeneous data sets. These AI systems are trained on external health data that have generally also been interpreted by humans and that have undergone relatively little processing before they are presented to the AI system, such as clinical images that have been annotated and interpreted by a human expert. Then the AI system learns how to perform that interpretation task on new health-associated data of the same type, which in clinical diagnostics is often a question of identifying or predicting status with a given

Author Affiliation.
[1] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh
[2] Department of Biochemistry and Molecular Biology, Tejgaon College, Dhaka, Bangladesh
[3] Department of Biological Sciences, St. John's University, New York, USA
[4] Burnett School of Biomedical Sciences, College of Medicine, University of Central Florida, Orlando, FL 32827, USA

Please cite this article.
Rahman, M. H., Biswash, M. A. R., Siddique, M. A. B., Rahman, M. M., Mou, M. A., Debnath, A., Fatin, M. (2025). "Significance of Artificial intelligence in clinical and genomic diagnostics", Journal of Precision Biosciences, 7(1),1-8,10149

disease. AI interpretation tasks may be clustered together in problem classes e.g. computer vision, time series analysis speech recognition, and natural language processing. Each of these problems is well suited to address specific types of clinical diagnostic tasks (Torkamani et al., 2017). For example, computer vision is useful for the interpretation of radiological images, time series analysis is useful for the analysis of continuously streaming health data such as those provided by an electrocardiogram (Esteva et al., 2018), speech recognition techniques can be used for detection of neurological disorders (Fraser et al., 2015), and AI-based natural language processing can be helpful in the extraction of meaningful information from electronic health record (EHR) data (Rajkomar et al., 2018). In some areas, the association between problem classes and diagnostic tasks may not be as obvious; for example, techniques from computer vision are also useful for the identification of functional regulatory elements in the human genome, where they can be used to identify recurrent motifs in DNA sequences in a manner analogous to that in which pixel patterns are detected in images by convolutional neural networks (CNNs; described in the next section) (Zou et al., 2018). Many of these problems have been addressed by a specific group of AI algorithms known as deep learning, which can learn interpretable features from large and complex datasets by using deep neural network architectures. Neural networks are computational systems of artificial neurons (also called 'nodes') that transmit signals to one another, often in interconnected layers.

The layers that are not the input or output layer are termed the 'hidden' layers. A deep neural network consists of many hidden layers of artificial neurons. Neural networks often take as input the fundamental unit of data that it is trained to interpret: for example, pixel intensity in images; diagnostic, prescription, and procedure codes in EHR data; or nucleotide sequence data in genomic applications (Eraslan et al., 2019). In other words, unlike most machine-learning approaches, minimal or no human extraction and definition of predictive features are required. A multitude of these simple features are combined in successive layers of the neural network in a variety of ways, as designed by the human neural network architect, in order to represent more sophisticated concepts or features of the input health data. Ultimately, the output of the neural network is the interpretation task that the network has been trained to execute. For example, successive layers of a computer vision algorithm might learn to detect edges in an image, then patterns of edges that represent shapes, then collections of shapes that represent certain objects, and so on. Thus, AI systems synthesize simple features into more complex concepts to derive conclusions about health data in a manner that is analogous to human interpretation, although the complex concepts used by the AI systems are not necessarily recognizable or obvious concepts to humans.

In this review, we describe the recent successes and potential future applications of AI, especially deep learning, in clinical diagnostics, with a focus on clinical genomics. We provide a brief overview of AI algorithms and the classes of problems that they are well suited to address. Next, we provide a more detailed review of how AI has been used to accomplish a variety of clinical genomics tasks, including variant calling and annotation, variant impact prediction, and phenotype-to-genotype mapping. Finally, we end by discussing the potential future applications and challenges of AI in genotype to phenotype prediction, especially as it relates to common complex diseases and individualized medicine.

Artificial intelligence and its applications

The present-day clinical diagnostic AI algorithms are referred to be "weak" or "narrow" AI. These artificial intelligence algorithms are trained to do a specific task, such as classifying skin lesion photos into diagnostic categories or generating a molecular diagnosis using phenotypic and genomic information. These algorithms do not display general intelligence and are not flexible enough to address other clinical diagnostic tasks. However, transfer learning approaches can be used to adapt a fully trained AI algorithm to accomplish closely related tasks. This is best exemplified by image-based diagnostic AI algorithms that benefit from advances in computer vision and neural networks trained for general image recognition tasks. Thus, the first step in the design of clinical diagnostic AI algorithms usually involves mapping the specific diagnostic task to a more general problem class. Here, we review these problem classes and briefly highlight the intersection of these techniques with genomics.

Computer vision

The multidisciplinary field of computer vision is concerned with the collection, processing, and analysis of pictures and/or videos. In order to generate numerical or symbolic representations of concepts encoded in the picture, computer vision algorithms synthesize (or "convolute") high dimensional image data. This process is thought to mimic the way humans identify patterns and extract meaningful features from images. The main steps in computer vision consist of image acquisition, pre-processing, feature extraction, image pattern detection or segmentation, and classification. Deep-learning algorithms such as CNNs have been designed to perform computer vision tasks. In simplified terms, a typical CNN tiles an input image with small matrices known as kernel nodes or filters. Each filter encodes a pixel intensity pattern that it 'detects' as it convolves across the input image. A multitude of filters encoding different pixel intensity patterns convolve across the image to produce two-dimensional activation maps of each filter. The pattern of features detected across the image by these filters may then be used to successively detect the presence of more complex features (Figure 1).

Surveillance, image recognition, and autonomous vehicles are some of the major applications of computer vision. In clinical diagnostics, the first applications of AI in healthcare to be cleared by the US Food and Drug Administration (FDA) have been dominated by applications of computer vision to medical scans (for example, magnetic resonance imaging (MRI) or positron emission tomography images), and pathology images (for example, histopathological slides). The first medical imaging applications include the automated quantification of blood flow through the heart via cardiac MRI (Retson et al., 2019), the determination of ejection fraction from echocardiograms (Asch et al., 2019), the detection and volumetric quantification of lung nodules from radiographs (Retson et al., 2019), the detection and quantification of breast densities via mammography (Le et al., 2019), the detection of stroke, brain bleeds, and other conditions from computerized axial tomography (Detecting Intracranial Hemorrhage With Deep Learning, 2018; FDA Approves Stroke-detecting AI Software, 2018), and automated screening for diabetic retinopathy from comprehensive dilated eye examination (Gulshan et al., 2016; Van Der Heijden et al., 2017). Imaging applications in pathology include an FDA-cleared system for whole-slide imaging (Evans et.al.,2018) and promising approaches to the automated classification of dermatological conditions (Esteva et al.,2017) as well as numerous other whole-slide imaging and AI systems in development that are expected to dramatically enhance the efficiency of pathologists (Niazi et al.,2019).

Computer vision can also inform clinical genomic testing. For example, deep learning of lung cancer histopathological images is able to identify cancer cells, determine their type, and predict what somatic mutations are present in the tumor (Velazquez et al., 2017; Coudray et al., 2018). Similarly, facial image recognition can be used to identify rare genetic disorders and to guide molecular diagnoses (Gurovich et al., 2018; Dolgin, 2019). Thus, computer vision can extract phenotypic features from medical images in order to provide recommendations for molecular testing in a manner similar to that performed by a skilled pathologist or dysmorphologist. In some cases, AI-based systems have exceeded the capabilities of human experts, for example, by accurately predicting gender from retinal fundus images, a task that human experts would perform no better than random guessing (Poplin et al., 2018).

Time series analysis

The investigation of temporal data to predict future observations, to predict the discrete state generating a sequence of observations (e.g., normal heart rhythm versus arrhythmia), or to identify anomalies within a sequence of observations is known as time series analysis. In a broader sense, time series analysis can be applied to any ordered data, such as ordered but temporally disconnected DNA sequences. Time series analysis algorithms ingest data sequences and are generally tasked to learn sequential dependencies. The primary advantage of AI algorithms in time series analysis is an improved ability to detect non-linear and/or multi-step relationships that are not efficiently interrogated by traditional approaches such as hidden Markov models. Deep-learning algorithms, especially recurrent neural networks (RNNs), have been designed for sequence analysis tasks. A typical RNN includes some form of 'memory', in which prior inputs in a sequence influence future output. This is achieved by linking the hidden state of an input to the hidden state of the next input (Fig. 1). Extensions of this concept, which are implemented in specialized networks such as long short-term memory networks (LSTMs), add network elements that enhance the ability of the network to 'remember' long-term dependencies in the input data. CNNs are often applied to time series data when the task is to define the discrete state, or context, that produces the sequential data pattern. Time series analysis has major applications in the forecasting of equity prices, weather conditions, geological events, and essentially any future event of interest. In clinical diagnostics, time series AI algorithms can be applied to medical devices producing continuous output signals, with the application of electrocardiograms being an especially active area of interest. AI applied to electrocardiograms can detect and classify arrythmias (Hannun et al., 2018), especially atrial fibrillation (Tison et al., 2018), as well as cardiac contractile dysfunction (Attia et al., 2018), and blood chemistries linked to cardiac rhythm abnormalities (Galloway et al., 2019). When applied to genomic sequence data, AI time series algorithms appear to be especially effective at detecting functional DNA sequence elements that are indicative of gene splicing (Leung et al., 2014; Jaganathan et al., 2019), large-scale regulatory elements (Quang & Xie, 2016), and gene function (Wang et al., 2018).

Automatic speech recognition

Automatic speech recognition includes a group of methodologies that enable the interpretation of spoken language. Speech-recognition algorithms ingest raw sound waves from human speech and process them to allow the recognition of basic elements of speech including tempo, pitch, timbre, and volume, as well as more complex features of speech including the spoken language, words, and sentences (Li et al.,2015). More advanced speech recognition algorithms can identify sophisticated features from audiological data, such as mood changes or emotional states (Parthasarathy et al.,2019; Trigeorgis et al., 2016). Because of the temporal complexity of speech, traditional speech-recognition algorithms have typically relied on separate models to reassemble meaning from spoken language. These steps include segmenting audio into distinct units of sound (for example, phonemes), connecting those sound units into language units (for example, words), and assembling those language units into more complex language elements (for example,

phrases) to extract meaning. Recent advances in AI algorithms that address temporal sequences through sequence-to-sequence attention-based and recurrent neural network transducer-based approaches now allow for these tasks to be executed in a single model with streaming output (Hinton et al.,2012; Prabhavalkar et al.,2017) In sequence-to-sequence models, for example, a neural network can map the sequences of phonemes produced by an acoustic model into sequences of words, or a sequence of words can be translated into another language. Thus, sequence-to-sequence and other speech recognition models can also act as powerful tools for the communication of medical and health information across language barriers.

Voice command and virtual assistant systems are the major applications of speech recognition. Speech recognition algorithms have not yet found widespread use in clinical diagnostics but they have shown great promise in the detection of neurological conditions that are often challenging to diagnose with traditional clinical tools. In these clinical applications, the same genera speech-recognition strategies are used, but the outcome targeted by the final classification step is a disease phenotype that is typically associated with characteristics of speech (tone, tempo, pitch, and so on) and not necessarily the content of the language. Speech recognition has been successfully applied to the detection of diseases with an obvious influence on speech, notably chronic pharyngitis (Li et al., 2019), and of diseases with a less obvious influence on speech, including Alzheimer's disease (Fraser et al., 2015), Parkinson's disease (Torkamani et al., 2017), major depressive disorder (Ringeval et al.,2019) posttraumatic stress disorder (Marmar et al., 2019), and even coronary artery disease (Maor et al., 2018). Like imaging, speech recognition can detect potential genetic disorders and inform downstream clinical testing. In addition, speech recognition can be used as a tool to streamline the use of EHRs through automatic transcription, benefitting clinicians and patients and enabling natural language processing (NLP) analysis (Mohr et al., 2003; Edwards et al., 2017), as described in the next section.

Natural language processing NLP is the computational extraction of meaning from natural human language. These algorithms take as input a document, or potentially the output from automatic speech recognition, and output a useful transformation of the document. This transformation could be language translation, document classification, summarization, or extraction of higher-level concepts described by the text. Typical NLP algorithms involve syntactic analysis, which involves parsing the written text in a variety of ways to extract useful computational representations of language (by sentence breaking, tagging parts of speech, and standardizing inflected word forms, for example), followed by semantic analysis to extract meaning and/or the identification of named entities from the text. A wide variety of neural network architectures have been developed for NLP depending upon the target outcome, from sequence-to-sequence networks and other RNN variants for language translation (Wu, 2016), to CNNs to extract higher-level interpretations of the text (Collobert et al.,2008).

A major challenge that is addressed by NLP is the variety of synonyms, phrases, and interrelated concepts that can be used to express a singular meaning. This problem is especially pronounced in clinical applications where controlled vocabularies are numerous and in constant flux. Thus, NLP has been effectively used to automatically standardize and synthesize these terms to produce predictions of current and future diagnoses and medical events, (Rajkomar et al., 2018; Miotto et al.,2016) Similarly, NLP can be used to make health information more accessible by translating educational materials into other languages or by converting medical terms to their lay definitions (Chen et al., 2018). AI-based chatbots have already been deployed to augment the capabilities of genetic counselors to meet rising demands on their time generated by the rapidly expanding volume of clinical and direct-to-consumer genetic testing (Kohut et al., 2019). In addition, NLP approaches to EHR analysis can overcome the high dimensionality, sparseness, incompleteness, biases, and other confounding factors present in EHR data. For example, NLP has been applied to EHRs to predict patient mortality after hospitalization. In this application, EHR data are converted to a series of patient events streamed into an RNN, which was trained to identify patterns of patient characteristics, diagnoses, demography, medications, and other events that are predictive of near-term patient mortality or hospital readmission (Rajkomar et al., 2018). Similarly, when combined with other medical data, predictions of disease severity and therapy efficacy can be made (Diller et al., 2018). When combined with genomic data, NLP-based methods have been used to predict rare disease diagnoses and to drive phenotype informed genetic analysis, resulting in automated genetic diagnoses with accuracy similar to that of human experts (Liang et al., 2019; Clark et al., 2019).

Artificial intelligence in clinical genomics Mimicking human intelligence is the inspiration for AI algorithms, but AI applications in clinical genomics tend to target tasks that are impractical to perform using human intelligence and error prone when addressed with standard statistical approaches. Many of the techniques described above have been adapted to address the various steps involved in clinical genomic analysis including variant calling, genome annotation, variant classification, and phenotype-to-genotype correspondence and perhaps eventually they can also be applied for genotype-to-phenotype predictions. Here, we describe the major classes of problems that have been addressed by AI in clinical genomics.

Variant calling the clinical interpretation of genomes is sensitive to the identification of individual genetic variants among the millions

populating each genome, necessitating extreme accuracy. Standard variant-calling tools are prone to systematic errors that are associated with the subtleties of sample preparation, sequencing technology, sequence context, and the sometimes-unpredictable influence of biology such as somatic mosaicism (Kohut et al., 2019). A mixture of statistical techniques including hand-crafted features such as strand-bias (DePristo et al., 2011) or population-level dependencies (Garrison & Marth, 2012) are used to address these issues, resulting in high accuracy but biased errors (Hwang et al., 2015). AI algorithms can learn these biases from a single genome with a known gold standard of reference variant calls and produce superior variant calls. Deep Variant, a CNN-based variant caller trained directly on read alignments without any specialized knowledge about genomics or sequencing platforms, was recently shown to outperform standard tools on some variant-calling tasks (Popli et al., 2018). The improved accuracy is thought to be due to the ability of CNNs to identify complex dependencies in sequencing data. In addition, recent results suggest that deep learning is poised to revolutionize base calling (and as a result, variant identification) for nanopore-based sequencing technologies, which have historically struggled to compete with established sequencing technology because of the error-prone nature of prior base-calling algorithms (Wick et al., 2019).

Genome annotation and variant classification After variant calling, the interpretation of human genome data relies on the identification of relevant genetic variants through prior knowledge and inference of the impact of genetic variants on functional genomic elements. AI algorithms can improve the use of prior knowledge by informing phenotype-to-genotype mapping (described in the next section). Here, we describe both genome annotation and variant classification because many of the AI algorithms that are used to predict the presence of a functional element from primary DNA sequence data are also used to predict the impact of a genetic variation on those functional elements.

Classification of coding variants

Many methods have been developed for the classification of nonsynonymous variants (Tang & Thomas, 2016). Some of these methods have been integrated into deep-learning-based meta predictors (models that process and merge the predictions produced by several other predictors) that outperform both their individual predictive components and the combination of those predictive components when integrated using regression or other machine-learning approaches (Quang et al., 2014). For example, the combined annotation dependent depletion approach (CADD) (Kircher et al., 2014) combines a variety of predictive features in a machine-learning algorithm to predict the deleteriousness of genetic variants. A deep-learning-based extension of CADD, named DANN, demonstrated improved performance using the same set of input features as CADD but combined in a deep neural network (Quang et al., 2014). This technical extension of CADD suggests that deep learning may be a superior approach for integrating known features that are predictive of deleteriousness. However, the classification accuracies of these tools are not sufficient to drive clinical reporting, although they can be useful for guiding the interpretation of clinical genomic data by prioritizing potential candidate variants for further consideration. More interesting are AI-based methods that make predictions directly from DNA or protein sequence data with minimal hand-crafting of features. One approach, Primate AI, which used CNNs trained on variants of known pathogenicity with data augmentation using cross-species information, was shown to outperform prior methods when trained directly upon sequence alignments (Sundaram et al., 2018). The network was able to learn important protein domains, conserved amino acid positions, and sequence dependencies directly from the training data consisting of about 120,000 human samples. Primate AI substantially exceeded the performance of other variant pathogenicity prediction tools in differentiating benign and pathogenic de-novo mutations in candidate developmental disorder genes, and in reproducing prior knowledge in Clinvar (Landrum et al., 2017). These results suggest that Primate AI is an important step forward for variant classification tools that may lessen the reliance of clinical reporting on prior knowledge. In addition, deep generative models have shown promise for predicting the effects of genetic variants (Riesselman et al., 2018), and are especially intriguing given their ability to evaluate the joint influence of multiple genetic variants and/or complex indels on protein function, a capability that is largely absent from most pathogenicity prediction tools. Deep generative models are a type of deep neural network that can learn to replicate data distributions and produce examples not previously observed by the model. For example, a deep generative model trained on images of birds could learn to generate novel bird images.

Classification of non-coding variants

The computational identification and prediction of noncoding pathogenic variation is an open challenge in human genomics (Chatterjee & Ahituv, 2017). Recent findings suggest that AI algorithms will substantially improve our ability to understand non-coding genetic variation. Splicing defects in genes are responsible for at least 10% of rare pathogenic genetic variation (Soemedi et al., 2017), but they can be difficult to identify because of the complexity of intronic and exonic splicing enhancers, silencers, insulators, and other long range and combinatorial DNA interactions that influence gene splicing (Baeza-Centurion et al.,2019). Splice AI, a 32-layer deep neural network, is able to predict both canonical and non-canonical splicing directly from exon–intron junction sequence data (Jaganathan et al., 2019). Remarkably, Splice AI was able to use long-range sequence information to boost prediction accuracy from 57%, using a short

window size (80 nucleotides) typical for many prior splicing prediction tools, to 95% when a 10 kb window size was ingested by the AI algorithm, and was able to identify candidate cryptic splicing variants underlying neurodevelopmental disorders.

Deep-learning-based approaches have also substantially improved our ability to detect regulatory elements (Kelley et al., 2018; Alipanahi et al.,2015) and to predict the influence of genetic variation on those elements. DeepSEA, a multitask hierarchically structured CNN trained on large-scale functional genomics data (Bernstein et al., 2010), was able to learn sequence dependencies at multiple scales and simultaneously produce predictions of DNase hypersensitive sites, transcription factor binding sites, histone marks, and the influence of genetic variation on those regulatory elements, with a level of accuracy superior to those of other tools for prioritizing non-coding functional variants (Zhou & Troyanskaya, 2015). As seen for SpliceAI, the ability of DeepSEA to ingest DNA sequences of 1 kb, which is substantially larger than the input to typical motif-based search tools, was critical to this improved performance. Extensions of DeepSEA have been applied to whole-genome sequencing data from families with autism spectrum disorder to reveal several candidate non-coding mutations (Zhou et al., 2019). Further extension to the ExPecto algorithm has demonstrated its ability to predict gene expression levels directly from DNA sequence information (Zhou et al., 2018). Further investigation of these new deep-learning based frameworks for the analysis of non-coding sequence data is likely to provide new insights into the regulatory code of the human genome.

Phenotype-to-genotype mapping

Numerous genetic variations found in human genomes have either been previously identified as harmful or are anticipated to be harmful (Telenti et al., 2016), regardless of the individual health status (Erikson et al.,2016). Therefore, the molecular diagnosis of disease often requires both the identification of candidate pathogenic variants and a determination of the correspondence between the diseased individual's phenotype and those expected to result from each candidate pathogenic variant. AI algorithms can significantly enhance the mapping of phenotype to genotype, especially through the extraction of higher-level diagnostic concepts that are embedded in medical images and EHRs.

Image to genetic diagnosis

The human phenotype ontology lists 1007 distinct terms defining different abnormalities of the face (Köhler et al., 2018). These abnormalities are associated with 4526 diseases and 2142 genes. A dysmorphologist will often identify these abnormalities individually and synthesize them into a clinical diagnosis. The clinical diagnosis may then inform targeted gene sequencing or phenotype-informed analysis of more comprehensive genetic data. Often the human-provided clinical diagnosis and molecular diagnoses overlap but do not match precisely because of the phenotypic similarity of genetically distinct syndromes. DeepGestalt, a CNN-based facial image analysis algorithm, dramatically outperforms human dysmorphologists in this task and is precise enough to distinguish between molecular diagnoses that are mapped to the same clinical diagnosis (that is, distinct molecular forms of Noonan syndrome) (Gurovich et al., 2018). When combined with genomic data, PEDIA, a genome interpretation system incorporating DeepGestalt, was able to use phenotypic features extracted from facial photographs to accurately prioritize candidate pathogenic variants for 105 different monogenic disorders across 679 individuals (Hsieh et al., 2019). Deployment of DeepGestalt as a face-scanning app has the potential to both democratize and revolutionize the identification of genetic syndromes (Dolgin, 2019).

Genetic syndromes that are identified through facial analysis can be readily confirmed with DNA testing, but adequate material for somatic mutation testing is not always available in some instances of cancer. Nevertheless, knowledge of the genomic underpinnings of a tumor are critical to treatment planning. Here again, AI can bridge the gap between image-derived phenotypes and their probable genetic source. A 'survival CNN', which is a combination of a CNN with Cox proportional hazards-based outcomes (a type of statistical survival analysis), was able to learn the histological features of brain tumors that are associated with survival and correlated with somatic mutation status (Mobadersany et al., 2018). Importantly, this algorithm was not trained to predict genomic aberrations directly. Inspection of the CNN concepts used to make the survival predictions identified novel histological features that are important for prognosis determination. Like the faces of individuals with phenotypically overlapping genetic syndromes, these results suggest that the genomic aberrations underpinning an individual's tumor could potentially be predicted directly from tumor histology images. More generally, AI-based computer vision systems appear to be capable of predicting the genomic aberrations that are likely to be present in an individual's genome on the basis of the complex phenotypes embedded in relevant clinical images (Dolgin, 2019; Mobadersany et al., 2018).

EHR to genetic diagnosis Disease phenotypes can be complex and multimodal; captured not only by medical imaging, but also by biochemical and other tests that may be ordered at different times and perhaps by different physicians during the course of a differential diagnosis. These results are documented in an EHR where physicians synthesize these findings to provide diagnoses and inform clinical decision-making. Although human specialists can accomplish this task accurately within their area of expertise, AI-based algorithms can be general EHR pattern recognition experts. In a recent study involving more than 500,000 patients, an AI-based NLP approach was used to extract clinically relevant features from EHR data. A hierarchical statistical model, tiered on

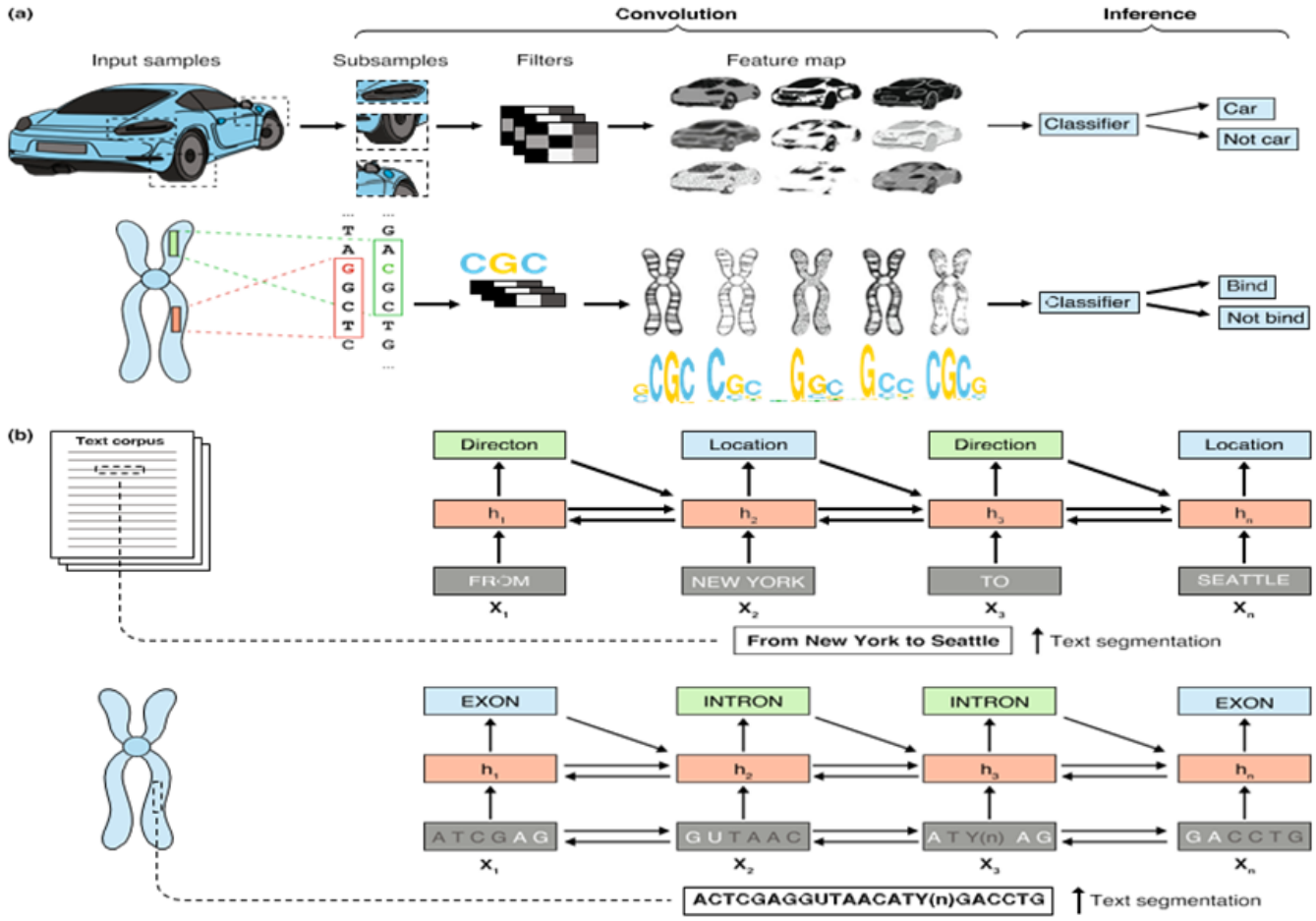

Figure 1: Examples of Different Neural Network Architectures and Their Applications in Genomics

a) Convolutional networks process input data by dividing it into smaller segments, applying specific filters to each segment, and multiplying feature values by weights. The result reveals patterns (such as conserved motifs) that are then mapped back to the original data. These patterns can be used to train models (like feedforward networks or logistic regression) to classify the data, such as determining whether a particular motif is a binding target. By selectively masking or filtering out certain base pairs, these models highlight the most important features for accurate classification.

b) Recurrent networks in natural language processing take segmented input (such as text or DNA sequences) and identify connections between elements through a network of hidden states. Typically, the hidden states are represented by unidirectional nodes that process the input in one direction. This figure shows a bidirectional recurrent network that processes input in both directions, using hidden states from neighboring elements to predict context (e.g., identifying whether a sequence is part of an intron or exon).

the basis of anatomic divisions in a manner meant to mimic the clinical reasoning of a composite of experienced physicians, was trained on the NLP output to generate a diagnostic system (Liang et al., 2019). Overall, this system was able to differentiate between 55 common pediatric diagnoses with 92% accuracy.

When linked with genomic data, an AI-based diagnostic agent coupled with a genome interpretation system can rapidly produce genetic diagnoses. For example, an NLP system was designed to extract phenotypic descriptions automatically from EHR data of pediatric patients with rare diseases, and to rank matches to the expected phenotypic features of candidate pathogenic variants in the patients' genomes (Clark et al., 2019). In 101 children with 105 genetic diseases, automated retrospective genomic diagnoses agreed with expert human interpretation at 97% recall and 99% precision. The system was also able to provide automated genomic diagnoses prospectively for three of seven seriously ill ICU infants. Intriguingly, a simpler phenotypic risk score approach, applied to an adult population with EHR and genomic data, was able to identify previously unrecognized monogenic conditions in 18 individuals from a population of 21,701 (Bastarache et al., 2018). These results suggest that AI-based phenotype to genotype mapping approaches could significantly improve the diagnostic yield of genetic testing and the identification of individuals with unrecognized genetic disorders.

Genotype-to-phenotype prediction

The ultimate clinical goal of genetics is to detect and predict future illness risk. For several frequent complicated disorders, risk stratification is both personally and therapeutically valuable thanks to relatively straightforward statistical methods for polygenic risk prediction. (Torkamani et al., 2018). A few studies have attempted genomic prediction of complex human traits using AI algorithms, but most of those reported in the literature to date are probably overfit as they purportedly explain substantially more trait variance than should be possible on the basis of heritability estimates. One application of machine learning to genomic prediction of height was able to provide relatively accurate predictions within expected bounds (Lello et al., 2018), suggesting that AI-based methods can be used to improve upon statistical techniques. However, the true utility of AI-based approaches in genotype-to-phenotype prediction will probably come from the integration of a variety of health data types and risk factors into comprehensive predictors of disease risk.

Common diseases are a result of a complex interplay between inherited genetic risk factors, environmental exposures, and behaviors. Genetic risk alone provides a baseline estimate of lifetime risk for disease, but genetic risk combined with other risk factors allows for a narrowing of that probability space into a short-term projection of disease risk. For example, several non-genetic risk factors are associated with breast cancer risk, including mammographic density, age at first birth, age at menarche, and age at menopause. Combining these nongenetic risk factors with genetic data significantly improves the accuracy of breast cancer risk models and can inform risk-based mammographic screening strategies (Lee et al., 2019). Similarly, significant improvement in risk stratification can be achieved by integrating conventional and genetic risk factors for coronary artery disease (Inouye et al., 2018). Genetic risk score models are more useful than simple pathogenicity assertions in cases where a common disease is the result of a combination of weak effects from multiple loci. However, current models integrate genetic and non-genetic risk factors in simple additive models that probably do not capture the complex causal relationships between these heterogenous risk factors. AI algorithms, given an appropriate volume of data, excel at dissecting this complexity. Unraveling the complex interplay between genetic data, EHR data, digital health monitoring devices, and other sources of health information with AI-based algorithms is a compelling prospect for the future.

Challenges and limitations

The capacity of AI-based systems to comprehend complicated data can be superhuman. However, when used to data on human health, their strength and complexity can also lead to erroneous, immoral, and prejudiced findings. Without careful consideration of the methods and biases embedded in a trained AI system, the practical utility of these systems in clinical diagnostics is limited. Thus, we end with a discussion on the challenges and limitations of AI in clinical diagnostics.

Regulatory issues

The FDA has authorized an increasing number of AI algorithms (Topol., 2018). These algorithms raise a number of regulatory and ethical challenges around the sourcing and privacy of the data used to train the algorithms (Dias & Torkamani, 2019), the transparency and generalizability of the underlying algorithms themselves, the regulatory process for refreshing these algorithms as further data become available, and the liability associated with prediction errors (Vayena et al., 2018). Some of these issues can and should be addressed by open sharing of AI models in detail (including source codes, model weights, meta graphs, and so on) with the scientific and medical community to improve transparency. Other issues will need to be addressed by the development of: (i) best practices for the interpretability of predictions to protect patient autonomy and shared decision-making; (ii) fairness standards to minimize disparities induced by machine bias; and (iii) ad hoc guidance to allow for continuous improvement of the algorithms (Vayena et al., 2018). As with most biomedical advances, the cost and expertise necessary to deploy AI algorithms is another concern, although these concerns diminish as interpretability and fairness issues are addressed. We explore these issues in further detail below. AI interpretability AI is often criticized for being a 'black box': a system

that produces an output without any explanation or justification. While this is perfectly acceptable in low-risk situations, clinical decision-making is not a low-risk situation. 'What?' may sufficiently encompass the question of interest in a general object-detection task, but 'why?' is an inherent part of the question in most clinical diagnostic tasks, because it is often crucial to subsequent clinical decision-making or at least necessary for acceptance of the prediction by both physicians and patients. An ideal AI-based clinical diagnostic system should produce accurate predictions and provide human interpretable explanations of those predictions. A common approach to answering 'why?' in computer vision applications is to generate a visual overlay of the portions of an image that contribute most strongly to an output prediction (Selvaraju et al., 2017; Olah et al., 2017). This strategy works well for image-based and other CNN-based clinical diagnostic tasks. In fact, many of the AI-based clinical diagnostic methods described in this review include some form of interpretive analysis. Thus, although AI interpretability is an important problem in general, the criticism of 'black box' systems in current AI-based clinical diagnostics may be overstated.

When complex interdependencies form the basis of a prediction, however, accurate interpretation of AI output becomes quite challenging (Mittelstadt et al., 2019). Interpretable machine learning methods are an active area of computer science research (Doshi-Velez & Kim, 2017), but most interpretable AI approaches involve the production of a simplified and potentially inaccurate approximation of the more complex AI system (Mittelstadt et al., 2019). Recently, a move towards more interactive models of interpretability through 'dialogue' with the AI system has been proposed (Mittelstadt et al., 2019). This approach allows the human user to ask contrastive questions of the AI system in order to explore how its output predictions would change if inputs were modified. This approach could also facilitate a dialogue between physician and patient, with the aid of the AI interpretation system, to help them to understand the clinical diagnosis and, in some instances, the risk factors that could be modified to change the predicted outcome. Thus, further improvements to interpretable AI systems could not only substantially enhance the acceptability of AI predictions but also enhance the transparency of health communication between physicians and patients.

Data and machine bias

Interpretative output is crucial for revealing the information found by AI systems and identifying biases that might lead to unwanted behavior, in addition to being required for acceptability in clinical practice. There is substructure embedded in genomic and health data. Some substructure is due to truly differing causal relationships between alleged risk factors and health outcomes, whereas other substructure can be attributed to external factors such as socioeconomic status, cultural practices, unequal representation, and other non-causal factors that relate to the delivery and accessibility of medicine and clinical tests rather than to their efficacy (Gianfrancesco et al., 2018; Sirugo et al., 2019). AI systems must be carefully applied to differentiate between these types of bias. When medical AI systems are not inspected for non-causal bias, they can act as propagators of disparity. For example, DeepGestalt, the previously described AI system for facial dysmorphology analysis, displayed poor accuracy for the identification of Down syndrome in individuals of African versus European ancestry (36.8% versus 80%, respectively) (Lumaka et al., 2016). Retraining the model with examples of Down syndrome in individuals of African ancestry improved the diagnosis of Down syndrome in individuals of African ancestry to 94.7% (Lumaka et al., 2016). Genetic risk prediction is also prone to unequal performance in different population groups because of underrepresentation in the training data (Martin et al., 2019).

However, not all machine bias can be resolved by addressing underrepresentation in training data. In some cases, the bias is embedded in ostensibly representative training data. For example, gender bias is common in written documents and can be rapidly incorporated into NLP systems (Bolukbasi et al., 2016). Extensions to these models were required to 'debias' word embeddings. In clinical applications, EHR data may be representative overall, but the contents may include biases that result from the delivery of care or physician bias. For example, recent immigrants in Canada are more likely to receive aggressive care and die in intensive care units than are other residents (Yarnell et al., 2017). Furthermore, the substructure of genomic data is correlated with population structure, which can lead to the appearance of non-causal trait associations (Sohail et al., 2019). However, tools that will help to address machine bias are being developed, and careful attention to these issues could not only help to resolve machine bias issues but could eventually lead to diagnostic systems that are free from human bias (Chen et al., 2019).

Conclusions and future directions

AI systems outperformed state-of-the-art methods and have received FDA-cleared for a range of clinical diagnostics, predominantly imaging-based diagnostics. This productivity surge is driven by the availability of large datasets for training, e.g. large collections of annotated medical images or large functional genomics data sets, and the developments of AI algorithms and the GPU systems used to train them. Currently, the most promising applications of AI in clinical genomics appear to be the AI extraction of deep phenotypic information from images, EHRs, and other medical devices to inform downstream genetic analysis. However, deep-learning algorithms have also shown tremendous promise in a variety of clinical genomics tasks such as variant calling, genome annotation, and functional impact prediction. It is

possible that more generalized AI tools will become the standard in these areas, especially for clinical genomics tasks where inference from complex data (that is, variant calling) is a frequently recurring task. These applications have benefited from advances in CNNs and RNNs which appear to be particularly well suited for the analysis of genomic data. Yet, the utility of AI algorithms as the ultimate clinical decision support tool in predicting common complex human phenotypes has not been convincingly demonstrated. The rise of biobank-scale efforts with longitudinal health data collection, such as the UK Biobank (Sudlow et al., 2015) and All of Us Research Program (Sankar & Parker, 2016), will potentially provide the training datasets necessary to make this goal a reality. Given the reliance of AI on large-scale training datasets, it is likely that the scalable collection of phenotype data, and not genomic data, will be the more difficult barrier to overcome in realizing this ambition. Modern DNA sequencing technology allows for the generation of genomic data uniformly and at scale, but the collection of phenotype data requires numerous data collection modes, and tends to be slow, expensive, and highly variable across collection sites. Finally, the interpretability and identification of machine bias are essential to broad acceptance of AI technology in any clinical diagnostic modality.

Author contributions

M.H.R. and A.D. conceptualized and developed the methodology. M.A.R.B., M.M.R., and M.A.B.S. prepared the original draft and reviewed and edited the writing. M.A.M. and M.F. analyzed the data and reviewed and edited the writing.

Acknowledgment

None declared.

Competing financial interests

The authors have no conflict of interest.

References

Torkamani, A., Andersen, K. G., Steinhubl, S. R., & Topol, E. J. (2017). High-Definition Medicine. Cell, 170(5), 828–843. https://doi.org/10.1016/j.cell.2017.08.007

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2018). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer S Disease, 49(2), 407–422. https://doi.org/10.3233/jad-150520

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., . . . Dean, J. (2018). Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 1(1). https://doi.org/10.1038/s41746-018-0029-1

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2018). A primer on deep learning in genomics. Nature Genetics, 51(1), 12–18. https://doi.org/10.1038/s41588-018-0295-5

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics, 20(7), 389–403. https://doi.org/10.1038/s41576-019-0122-6

Retson, T. A., Besser, A. H., Sall, S., Golden, D., & Hsiao, A. (2019). Machine learning and deep neural networks in thoracic and cardiovascular imaging. Journal of Thoracic Imaging, 34(3), 192–201. https://doi.org/10.1097/rti.0000000000000385

Asch, F. M., Abraham, T., Jankowski, M., Cleve, J., Adams, M., Romano, N., Polivert, N., Hong, H., & Lang, R. (2019). ACCURACY AND REPRODUCIBILITY OF a NOVEL ARTIFICIAL INTELLIGENCE DEEP LEARNING-BASED ALGORITHM FOR AUTOMATED CALCULATION OF EJECTION FRACTION IN ECHOCARDIOGRAPHY. Journal of the American College of Cardiology, 73(9), 1447. https://doi.org/10.1016/s0735-1097(19)32053-4

Le, E., Wang, Y., Huang, Y., Hickman, S., & Gilbert, F. (2019). Artificial intelligence in breast imaging. Clinical Radiology, 74(5), 357–366. https://doi.org/10.1016/j.crad.2019.02.006

Detecting Intracranial Hemorrhage with Deep Learning. (2018, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8512336

FDA approves stroke-detecting AI software. (2018). Nature Biotechnology, 36(4), 290. https://doi.org/10.1038/nbt0418-290

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama, 316(22), 2402-2410.

Van Der Heijden, A. A., Abramoff, M. D., Verbraak, F., Van Hecke, M. V., Liem, A., & Nijpels, G. (2017). Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmologica, 96(1), 63–68. https://doi.org/10.1111/aos.13613

Evans, A. J., Bauer, T. W., Bui, M. M., Cornish, T. C., Duncan, H., Glassy, E. F., ... & Pantanowitz, L. (2018). US Food and Drug Administration approval of whole slide imaging for primary diagnosis: a key milestone is reached and new questions are raised. Archives of pathology & laboratory medicine, 142(11), 1383-1387.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639), 115-118.

Niazi, M. K. K., Parwani, A. V., & Gurcan, M. N. (2019). Digital pathology and artificial intelligence. The lancet oncology, 20(5), e253-e261.

Velazquez, E. R., Parmar, C., Liu, Y., Coroller, T. P., Cruz, G., Stringfield, O., Ye, Z., Makrigiorgos, M., Fennessy, F., Mak, R. H., Gillies, R., Quackenbush, J., & Aerts, H. J. (2017). Somatic mutations drive distinct imaging phenotypes in lung cancer. Cancer Research, 77(14), 3922–3930. https://doi.org/10.1158/0008-5472.can-17-0122

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction

from non–small cell lung cancer histopathology images using deep learning. Nature Medicine, 24(10), 1559–1567. https://doi.org/10.1038/s41591-018-0177-5

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P. M., Kamphausen, S. B., Zenker, M., Bird, L. M., & Gripp, K. W. (2018). Identifying facial phenotypes of genetic disorders using deep learning. Nature Medicine, 25(1), 60–64. https://doi.org/10.1038/s41591-018-0279-0

Dolgin, E. (2019). AI face-scanning app spots signs of rare genetic disorders. Nature. https://doi.org/10.1038/d41586-019-00027-x

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering, 2(3), 158–164. https://doi.org/10.1038/s41551-018-0195-0

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2018). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., Vittinghoff, E., Lee, E. S., Fan, S. M., Gladstone, R. A., Mikell, C., Sohoni, N., Hsieh, J., & Marcus, G. M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiology, 3(5), 409. https://doi.org/10.1001/jamacardio.2018.0136

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., & Friedman, P. A. (2018). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70–74. https://doi.org/10.1038/s41591-018-0240-2

Galloway, C. D., Valys, A. V., Shreibati, J. B., Treiman, D. L., Petterson, F. L., Gundotra, V. P., Albert, D. E., Attia, Z. I., Carter, R. E., Asirvatham, S. J., Ackerman, M. J., Noseworthy, P. A., Dillon, J. J., & Friedman, P. A. (2019). Development and validation of a Deep-Learning model to screen for hyperkalemia from the electrocardiogram. JAMA Cardiology, 4(5), 428. https://doi.org/10.1001/jamacardio.2019.0640

Leung, M. K. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. Bioinformatics, 30(12), i121–i129. https://doi.org/10.1093/bioinformatics/btu277

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... & Farh, K. K. H. (2019). Predicting splicing from primary sequence with deep learning. Cell, 176(3), 535-548.

Quang, D., & Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research, 44(11), e107. https://doi.org/10.1093/nar/gkw226

Wang, J., Cao, H., Zhang, J. Z. H., & Qi, Y. (2018). Computational Protein Design with Deep Learning Neural Networks. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-24760-x

Li, J., Deng, L., Haeb-Umbach, R., & Gong, Y. (2015). Robust automatic speech recognition: a bridge to practical applications.

Parthasarathy, S., Rozgic, V., Sun, M., & Wang, C. (2019, May). Improving emotion classification through variational inference of latent variables. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7410-7414). IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5200-5204). IEEE.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97.

Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L., & Jaitly, N. (2017, August). A Comparison of sequence-to-sequence models for speech recognition. In Interspeech (pp. 939-943).

Li, Z., Huang, J., & Hu, Z. (2019). Screening and diagnosis of chronic pharyngitis based on deep learning. International Journal of Environmental Research and Public Health, 16(10), 1688. https://doi.org/10.3390/ijerph16101688

Torkamani, A., Andersen, K. G., Steinhubl, S. R., & Topol, E. J. (2017). High-definition medicine. Cell, 170(5), 828-843.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., ... & Pantic, M. (2019, October). AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop (pp. 3-12). https://doi.org/10.1145/3347320.3357688

Marmar, C. R., Brown, A. D., Qian, M., Laska, E., Siegel, C., Li, M., Abu-Amara, D., Tsiartas, A., Richey, C., Smith, J., Knoth, B., & Vergyri, D. (2019). Speech-based markers for posttraumatic stress disorder in US veterans. Depression and Anxiety, 36(7), 607–616. https://doi.org/10.1002/da.22890

Maor, E., Sara, J. D., Orbelo, D. M., Lerman, L. O., Levanon, Y., & Lerman, A. (2018). Voice signal characteristics are independently associated with coronary artery disease. Mayo Clinic Proceedings, 93(7), 840–847. https://doi.org/10.1016/j.mayocp.2017.12.025

Mohr, D. N., Turner, D. W., Pond, G. R., Kamath, J. S., De Vos, C. B., & Carpenter, P. C. (2003). Speech recognition as a transcription aid: A randomized comparison with standard transcription. Journal of the American Medical Informatics Association, 10(1), 85–93. https://doi.org/10.1197/jamia.m1130

Edwards, E., Salloum, W., Finley, G. P., Fone, J., Cardiff, G., Miller, M., & Suendermann-Oeft, D. (2017). Medical Speech Recognition: Reaching Parity with Humans. In Lecture notes in computer science (pp. 512–524). https://doi.org/10.1007/978-3-319-66429-3_51

Wu, Y. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitasks learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6(1), 1-10.

Chen, J., Druhl, E., Ramesh, B. P., Houston, T. K., Brandt, C. A., Zulman, D. M., Vimalananda, V. G., Malkani, S., & Yu, H. (2018). A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. Journal of Medical Internet Research, 20(1), e26. https://doi.org/10.2196/jmir.8669

Kohut, K., Limb, S., & Crawford, G. (2019). The changing role of the genetic counsellor in the Genomics era. Current Genetic Medicine Reports, 7(2), 75–84. https://doi.org/10.1007/s40142-019-00163-w

Diller, G., Kempny, A., Babu-Narayan, S. V., Henrichs, M., Brida, M., Uebing, A., Lammers, A. E., Baumgartner, H., Li, W., Wort, S. J., Dimopoulos, K., & Gatzoulis, M. A. (2018). Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. European Heart Journal, 40(13), 1069–1077. https://doi.org/10.1093/eurheartj/ehy915

Liang, H., Tsui, B. Y., Ni, H., Valentim, C. C. S., Baxter, S. L., Liu, G., Cai, W., Kermany, D. S., Sun, X., Chen, J., He, L., Zhu, J., Tian, P., Shao, H., Zheng, L., Hou, R., Hewett, S., Li, G., Liang, P., . . . Xia, H. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature Medicine, 25(3), 433–438. https://doi.org/10.1038/s41591-018-0335-9

Clark, M. M., Hildreth, A., Batalov, S., Ding, Y., Chowdhury, S., Watkins, K., Ellsworth, K., Camp, B., Kint, C. I., Yacoubian, C., Farnaes, L., Bainbridge, M. N., Beebe, C., Braun, J. J. A., Bray, M., Carroll, J., Cakici, J. A., Caylor, S. A., Clarke, C., . . . Kingsmore, S. F. (2019). Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Science Translational Medicine, 11(489). https://doi.org/10.1126/scitranslmed.aat6177

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491–498. https://doi.org/10.1038/ng.806

Garrison, E., & Marth, G. (2012, July 17). Haplotype-based variant detection from short-read sequencing. arXiv.org. https://arxiv.org/abs/1207.3907

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific Reports, 5(1). https://doi.org/10.1038/srep17875

Poplin, R., Chang, P., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 36(10), 983–987. https://doi.org/10.1038/nbt.4235

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1727-y

Tang, H., & Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics, 203(2), 635–647. https://doi.org/10.1534/genetics.116.190033

Quang, D., Chen, Y., & Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics, 31(5), 761–763. https://doi.org/10.1093/bioinformatics/btu703

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics, 46(3), 310–315. https://doi.org/10.1038/ng.2892

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., Xu, J., Batzoglou, S., Li, X., & Farh, K. K. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nature Genetics, 50(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., . . . Maglott, D. R. (2017). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Riesselman, A. J., Ingraham, J. B., & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nature Methods, 15(10), 816–822. https://doi.org/10.1038/s41592-018-0138-4

Chatterjee, S., & Ahituv, N. (2017). Gene regulatory elements, major drivers of human disease. Annual Review of Genomics and Human Genetics, 18(1), 45–63. https://doi.org/10.1146/annurev-genom-091416-035537

Soemedi, R., Cygan, K. J., Rhine, C. L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., & Fairbrother, W. G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nature Genetics, 49(6), 848–855. https://doi.org/10.1038/ng.3837

Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J., & Lehner, B. (2019). Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. Cell, 176(3), 549-563.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research, 28(5), 739–750. https://doi.org/10.1101/gr.227819.117

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831-838.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nature Biotechnology, 28(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931–934. https://doi.org/10.1038/nmeth.3547

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak, J. J., Funk, J., Yao, K., Tajima, Y., Packer, A., Darnell, R. B., & Troyanskaya, O. G. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding

mutations to autism risk. Nature Genetics, 51(6), 973–980. https://doi.org/10.1038/s41588-019-0420-0

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics, 50(8), 1171–1179. https://doi.org/10.1038/s41588-018-0160-6

Telenti, A., Pierce, L. C. T., Biggs, W. H., Di Iulio, J., Wong, E. H. M., Fabani, M. M., Kirkness, E. F., Moustafa, A., Shah, N., Xie, C., Brewerton, S. C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B. A., Och, F. J., Turpaz, Y., & Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences, 113(42), 11901–11906. https://doi.org/10.1073/pnas.1613365113

Erikson, G. A., Bodian, D. L., Rueda, M., Molparia, B., Scott, E. R., Scott-Van Zeeland, A. A., ... & Torkamani, A. (2016). Whole-genome sequencing of a healthy aging cohort. Cell, 165(4), 1002-1011.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., . . . Robinson, P. N. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research, 47(D1), D1018–D1027. https://doi.org/10.1093/nar/gky1105

Hsieh, T., Mensah, M. A., Pantel, J. T., Aguilar, D., Bar, O., Bayat, A., Becerra-Solano, L., Bentzen, H. B., Biskup, S., Borisov, O., Braaten, O., Ciaccio, C., Coutelier, M., Cremer, K., Danyel, M., Daschkey, S., Eden, H. D., Devriendt, K., Wilson, S., . . . Krawitz, P. M. (2019). PEDIA: prioritization of exome data by image analysis. Genetics in Medicine, 21(12), 2807–2814. https://doi.org/10.1038/s41436-019-0566-2

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Vega, J. E. V., Brat, D. J., & Cooper, L. a. D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences, 115(13). https://doi.org/10.1073/pnas.1717139115

Bastarache, L., Hughey, J. J., Hebbring, S., Marlo, J., Zhao, W., Ho, W. T., Van Driest, S. L., McGregor, T. L., Mosley, J. D., Wells, Q. S., Temple, M., Ramirez, A. H., Carroll, R., Osterman, T., Edwards, T., Ruderfer, D., Edwards, D. R. V., Hamid, R., Cogan, J., . . . Denny, J. C. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science, 359(6381), 1233–1239. https://doi.org/10.1126/science.aal4043

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics, 19(9), 581–590. https://doi.org/10.1038/s41576-018-0018-x

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., De Los Campos, G., & Hsu, S. D. H. (2018). Accurate genomic prediction of human height. Genetics, 210(2), 477–497. https://doi.org/10.1534/genetics.118.301267

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., De Villiers, C. B., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F., & Antoniou, A. C. (2019). BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genetics in Medicine, 21(8), 1708–1718. https://doi.org/10.1038/s41436-018-0406-9

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., Lai, F. Y., Kaptoge, S., Brozynska, M., Wang, T., Ye, S., Webb, T. R., Rutter, M. K., Tzoulaki, I., Patel, R. S., Loos, R. J., Keavney, B., Hemingway, H., Thompson, J., . . . Samani, N. J. (2018). Genomic risk prediction of coronary artery disease in 480,000 adults. Journal of the American College of Cardiology, 72(16), 1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079

Topol, E. J. (2018). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. Genome Medicine, 11(1). https://doi.org/10.1186/s13073-019-0689-8

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via Gradient-Based Localization. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. Distill, 2(11). https://doi.org/10.23915/distill.00007

Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279-288).

Doshi-Velez, F., & Kim, B. (2017, February 28). Towards a rigorous science of interpretable machine learning. arXiv.org. https://arxiv.org/abs/1702.08608

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544. https://doi.org/10.1001/jamainternmed.2018.3763

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. Cell, 177(1), 26-31.

Lumaka, A., Cosemans, N., Mampasi, A. L., Mubungu, G., Mvuama, N., Lubala, T., Mbuyi-Musanzayi, S., Breckpot, J., Holvoet, M., De Ravel, T., Van Buggenhout, G., Peeters, H., Donnai, D., Mutesa, L., Verloes, A., Tshilobo, P. L., & Devriendt, K. (2016). Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. Clinical Genetics, 92(2), 166–171. https://doi.org/10.1111/cge.12948

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics, 51(4), 584–591. https://doi.org/10.1038/s41588-019-0379-x

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Yarnell, C. J., Fu, L., Manuel, D., Tanuseputro, P., Stukel, T., Pinto, R., Scales, D. C., Laupacis, A., & Fowler, R. A. (2017). Association between Immigrant Status and End-of-Life Care in Ontario, Canada. JAMA, 318(15), 1479. https://doi.org/10.1001/jama.2017.14418

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W., Hirschhorn, J., Daly, M. J., Patterson, N., Neale, B., Mathieson, I., Reich, D.,

& Sunyaev, S. R. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife, 8. https://doi.org/10.7554/elife.39702

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? The AMA Journal of Ethic, 21(2), E167-179. https://doi.org/10.1001/amajethics.2019.167

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine, 12(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sankar, P. L., & Parker, L. S. (2016). The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. Genetics in Medicine, 19(7), 743–750. https://doi.org/10.1038/gim.2016.183