UNIVERSITY OF CAMBRIDGE

phg foundation
making science work for health

Artificial intelligence for genomic medicine

Author

Sobia Raza

Acknowledgements

We are grateful to Dr Shehla Mohammed and Dr Joo Wook Ahn for reviewing this report.

Dr Laura Blackburn, Dr Emma Johnson, Alison Hall and Dr Mark Kroese at the PHG Foundation guided, reviewed and provided constructive input into this report. Nana Mensah (NHS Trainee Clinical Bioinformatician at Guy's Hospital, London) performed background research into this topic between June – July 2018 whilst on placement at the PHG Foundation.

We acknowledge with thanks the individuals who shared their expertise and insight into the subjects covered in this report. Full acknowledgments are listed on page 50.

May 2020

This report was researched and written during 2019

URLs were correct as of March 2020

This report can be downloaded from:
www.phgfoundation.org

Contact us about this report
intelligence@phgfoundation.org

The PHG Foundation is a health policy think-tank and linked exempt charity of the University of Cambridge. We work to achieve better health through the responsible and evidence based application of biomedical science. We are a registered company, no. 5823194

Contents

Executive summary 4

Priorities for policy 7

Defining AI and genomics 8
Genomics 9
Artificial intelligence 10
The importance of definitions 10
How machine learning and deep learning work 11

The growth of AI in genomics 17
Stages in genomic data analysis 19
Machine learning and deep learning in genomics 21
Rapid growth in machine learning activity 21

Existing and emerging applications 23
Phenotyping 24
Rare diseases 27
Cancers 27
Applications of machine learning and deep learning 27
Genomic sequence data processing 28
Analysis and interpretation 30
Integrated multi-omics and multi-modal data analysis 32
Other complex genomic data analyses 32
Clinical decision support 33

Considerations for policy 37
Data - improving quality, accessibility, and representation 38
Infrastructure, training and constructive collaborations 40
Regulation, explainability and interpretation 41
Privacy, security, and public perception 43

Unpicking the AI web of issues 45
Priorities for policy 47
The way ahead 48

Executive summary

Artificial intelligence (AI) techniques offer great potential for advancing genomic medicine.

This report examines the intersection between these two technologies, including the drivers behind the recent rise of AI techniques for genomics, existing and emerging applications, the limitations of AI for genomic medicine, and the challenges to realising its full potential for health. Achieving this potential will necessitate meeting the level of current enthusiasm for these technologies with the impetus, resources and collective commitment to tackle the serious issues ahead.

We offer a set of practical recommendations for policy makers to make the most of the opportunities AI presents for genomic medicine, minimise harms and speed up its effective delivery into healthcare.

The data challenge for genomic medicine

Genomic medicine has made significant strides in recent years, but the clinical application of genomics continues to evolve as new knowledge and technologies emerge. One major challenge is the ability to make sense of extremely large volumes of genomic sequence data, and effectively integrate and examine it with other relevant information, for example other molecular or clinical data.

The rise of AI

The AI techniques machine learning and deep learning (a type of machine learning) offer new computational approaches to streamlining key analytical problems in genomic medicine. Although some machine learning methods have been applied to key problems in genomic analysis for many years, activity of this kind has been increasing recently, driven by:

- Advances in high-performance computing
- Resurgence of deep learning
- Growing availability of resources for building machine learning models
- Growth of large genomic and biomedical datasets

Applications of AI in genomic medicine

Most aspects of genomic analysis have been touched in some way by machine learning and deep learning. These methods are being developed and applied across different elements of the genomic data pipeline, and to a whole spectrum of analyses, from single cell resolution to studies in large populations.

These efforts offer a significant range of potential benefits that could help advance the clinical application of genomics by:

- **Directly facilitating the steps involved in clinical genome analysis**

Examples of current activity include:

- Algorithms for better identification of genetic variants, including those that are currently difficult to accurately detect, e.g. somatic and copy-number variants
- Tools for extracting phenotype data (patient characteristics) from electronic health records, or analysing it e.g. deep-learning driven facial analysis to help inform the diagnosis of congenital conditions
- Tools for predicting the effect of genetic variants, such as their downstream impact on proteins or important molecular processes, e.g. gene expression

- **Improving understanding of genomic variation in relation to health and disease and accelerating discovery in genomic medicine**

We are still far from a complete understanding of the relationship between genomic variation and many known diseases; AI techniques applied to complex or very large datasets can provide valuable insight, and improve the underlying knowledge base upon which clinical genomic analysis relies. Research underway includes:

- Studies to examine how cancers evolve and determine which genetic changes could be drivers for tumour growth
- Algorithms to improve the efficiency and accuracy of CRISPR, a genome editing technique widely used to investigate the role of genes and other DNA sequences
- Methods to integrate and analyse genomic data together with other types of data

Current limitations

The application of AI has yet to generate clearly improved outcomes in genomic medicine, and the discovery potential within genomic datasets remains largely untapped. To make progress, multiple interconnected issues must be addressed:

Data quality and accessibility

The performance of AI algorithms is affected by the volume and quality of data used to initially 'train' (i.e. develop) them, so streamlined access to high quality genomic and healthcare data is essential

Bias

Some populations are under-represented in the databases and datasets used for training AI algorithms. This has the potential to exacerbate existing health disparities for groups that are already underserved. Algorithmic bias can also arise as a result of the availability of data, how those data are prepared and combined, how questions are framed, and because of prexisting prejudices within society

Expectations

Replicating the methods and results of AI studies and tools can be difficult. The increasing number of AI based tools for various steps of genomic data analysis will only make this more challenging

Skills and infrastructure

AI in genomics is a multidisciplinary endeavour - no single sector has a monopoly on all the necessary skills, expertise, data and resources needed to deliver all the potential benefits of AI used at scale in genomic medicine, so a focus across multiple sectors is needed

Privacy and security

Concerns around security, confidentiality, and the ethical use of data must be navigated and addressed effectively, or there is a serious risk of impeding the use and implementation of these technologies

Regulation and clinical governance

The regulatory status of many AI algorithms used within clinical genomics remains unclear. This is influenced by whether or not the algorithm qualifies as a medical device or meets an unmet need. For adaptive algorithms, questions arise about the nature of the regulatory pathway, how they should be certified and who should be liable if their use results in harm

Uncertainty

Another area of uncertainty is how algorithms used for healthcare should meet the regulatory requirements for transparency and explanation within the EU General Data Protection Regulation. These requirements could impact on how algorithms are used for clinical decision making and patient management, particularly when using black box algorithms

Considering the significant financial investment and policy work already underway to deliver AI in health and care, it is vital to address the above priorities early as part of wider efforts to accelerate the adoption of proven AI technologies. In doing so the application of AI, when experts in health, genomics, regulation and ethics are working in concert, presents a significant opportunity to unravel the complexity encoded in our genomes for health benefit.

Priorities for policy

The initial priorities for creating an environment that facilitates the application of AI in genomic medicine and realises its near-term value are to:

Establish the right conditions for facilitating AI in genomic medicine which includes improved digital infrastructure, data acquisition and management, access to specific technical skills, and cross-disciplinary collaborations

Prioritise the development of constructive AI tools that address well-defined, focused, and clinically relevant problems in genomics analysis and clinical genomics service delivery

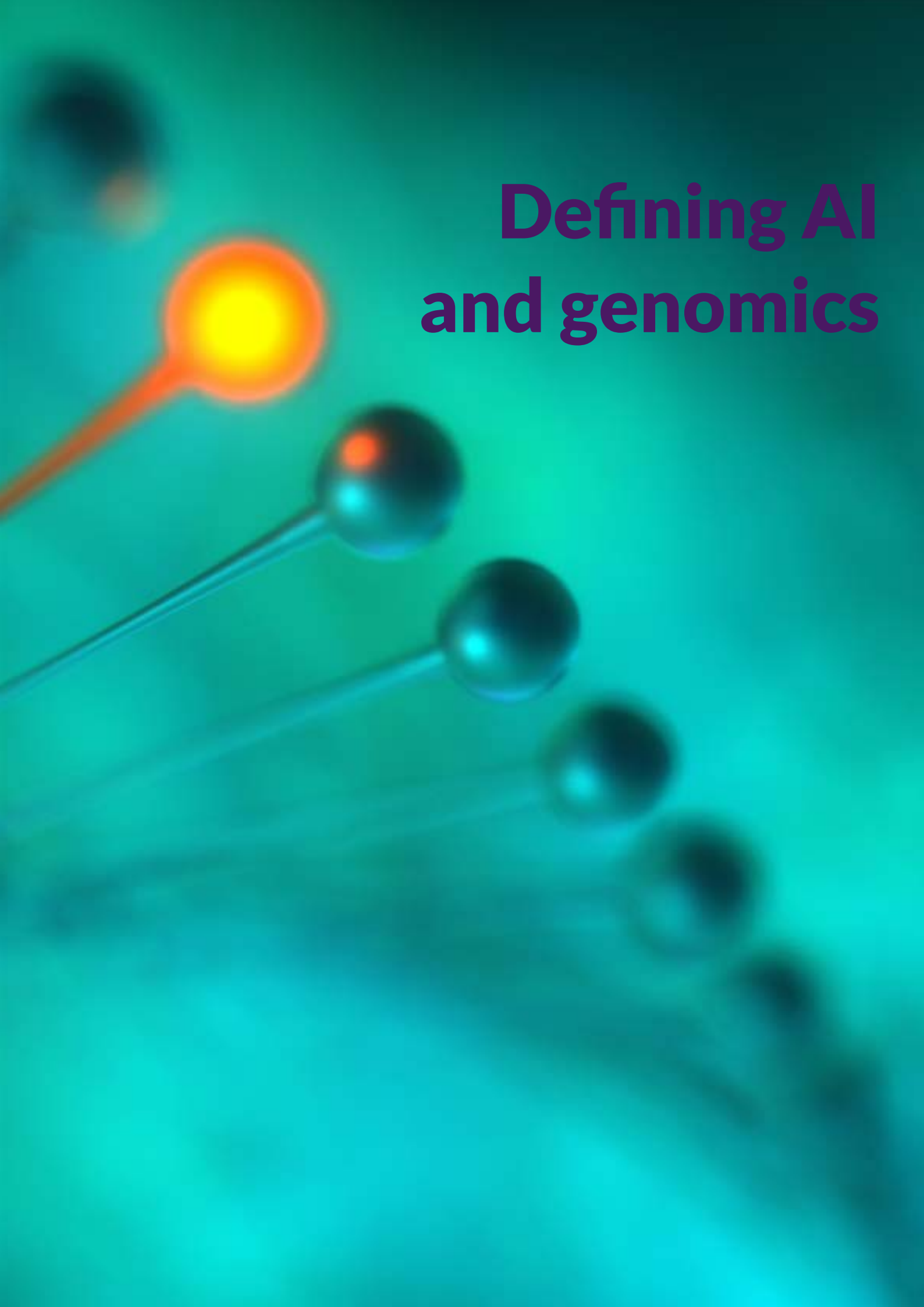
Mitigate against AI bias in genomics by promoting a workforce and research environment that is representative of societal diversity, as well as monitoring and addressing sources of bias within training datasets

Facilitate research efforts to apply machine learning (including deep learning) to well-curated, high-quality genomics and biomedical datasets, and bridging the gap between knowledge discovery and clinical practice

Support efforts driven by the clinical genomics community to benchmark, review, and determine the most effective use and integration of emerging new algorithms for clinical genome analysis

Establish sector-specific strategies to address the complex challenges and limitations of AI in genomic medicine and research

Establish the clinical governance arrangements for the use of specific AI applications in the practice of clinical genomics



Defining AI and genomics

Defining AI and genomics

Overview

- **AI, specifically machine learning and more recently deep learning, offers new computational approaches to streamline key problems in genomic medicine. Machine learning algorithms 'learn' from data to discover their own rules and can improve with experience**
- **Deep learning is a more flexible subset of machine learning, with a higher capacity for modelling complex relationships in datasets and is less dependent on prior domain knowledge**
- **While deep learning has valuable advantages over traditional machine learning methods, it has its own challenges**

The health applications of genomics and those of artificial intelligence (AI) have both been the subject of intense attention. Activity across these domains is thriving internationally, with multi-billion dollar projections of global market growth by 2026[1, 2, 3]. Independently, genomics and AI have featured extensively within recent policy discourse and strategies. Both topics have been the subject of parliamentary inquiries in the UK during the past two years[4, 5], and along with digital medicine, the transformational potential of these two technologies forms the backdrop for planning the future NHS healthcare workforce[6].

Genomics

In recent years there has been an ongoing focus on building the UK's genomics industry and genomic healthcare. This has included the 100,000 Genomes Project, the establishment of the NHS Genomic Medicine Service in England, ambitions to sequence five million genomes in five years, and plans for a new National Genomic Healthcare Strategy[7].

Beyond the United Kingdom, large-scale human genome sequencing initiatives are underway in several countries including the United States where one million genomes will be sequenced as part of the National Institutes of Health (NIH) 'All of Us' research program and the European 1+ Million Genomes Initiative. Globally, the number of medical genetic tests is growing as is the availability of direct-to-consumer genetic testing, leading to increasing scrutiny of its benefits, risks and limitations[8, 9, 10].

Deployment of genomic technologies is far more established within healthcare than the use of AI. The applications of genomic medicine span the human life cycle from conception to death[11] and cut across many different clinical fields, including prenatal and reproductive health, rare diseases, cancer, infectious diseases, regenerative medicine (gene editing and gene therapies) and pharmacotherapy.

There is still a tremendous amount more that genomic medicine has to offer, a point stressed in the 2016 annual report of the Chief Medical Officer for England, *Generation Genome*[12].

A comprehensive understanding of the genome and its functions is far from complete, and even where knowledge exists its clinical application is often lagging. A critical bottleneck is in the ability to make sense of the sheer volume and complexity of information contained within a genome. This is where AI, specifically machine learning and, more recently, deep learning, holds great promise – by offering new computational approaches to streamline key problems in genomic medicine.

Artificial intelligence

AI was addressed extensively in the 2018 annual report of the Chief Medical Officer for England, *Health 2040*[13]. In the same year a new government Office for Artificial Intelligence was established to oversee implementation of the UK's AI and Data Grand Challenge concerning early diagnosis, innovation, prevention and treatment. Following a £250 million funding pledge in 2019, NHSX (the UK government unit leading on digital transformation of the National Health Service in England) is establishing a national AI lab which will facilitate cross-government, industry and academic collaborations[14].

In the past few years numerous reports and commentaries have been published on the potential of AI for health, and its ethical, legal, and social implications[15-21]. Significant effort is being spent on realising the benefits of AI in healthcare, especially within medical imaging where some of the most promising potential examples are emerging[22, 23].

More widely, there are calls to bring together medical data repositories including electronic health records (EHRs) and clinical expertise to realise a 'deep learning' healthcare system whereby best treatment decisions can be computationally learnt with the aid of AI analyses[16, 24]. Notably, many of the global tech giants – including Google, Facebook, Amazon, Microsoft and Apple – are major investors in AI and AI expertise and have expanded into healthcare-related innovations. These include digital devices and apps, health monitoring, disease diagnosis, virtual health assistants, as well as resources for genomic analysis.

As a result, the debate about how to appraise and approve AI-based medical products is intensifying[25]. The US Food and Drug Administration (FDA) is one agency in the process of reimagining its regulatory approach around medical devices driven by advanced AI algorithms that continually adapt based on new data[26]. The UK Government has recently published a 'Code of Conduct' which sets out guidance for those developing, deploying and using intelligent algorithms and data-driven innovations in healthcare[27].

The progression of AI in healthcare is a matter of when, how, and to what extent, rather than if. However, the considerable hype surrounding AI, must not distract from its complex challenges and significant limitations.

The importance of definitions

It is important to be clear about what is meant by AI. Most references to AI within genomics relate to machine learning or deep learning. Although the terms are often used interchangeably the differences are relevant to the evolving applications within genomics and more widely across health and care. Explanations of the various terminologies surrounding AI have been provided previously[15, 28, 29]. The main terms used within this report and concepts central to the field are summarised in Table 1 (see appendices).

In brief AI is the development and use of computing systems concerned with making machines work in an intelligent way

- Machine learning is an approach for achieving artificial intelligence
- Deep learning is a branch of machine learning (figure 1)

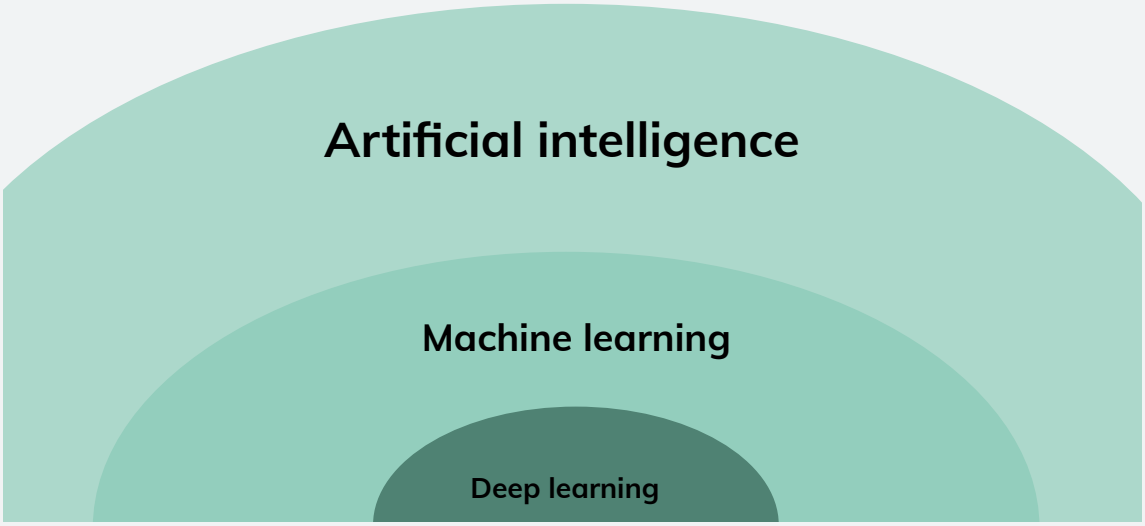
AI can be categorised as 'narrow' or 'general' whereby

- **Narrow AI** focuses on performing one specific task e.g. playing chess or filtering spam emails.Whilst 'narrow' by name, the individual uses of the technology can be broad ranging and sophisticated. Narrow AI is the focus of most current AI developments within healthcare and other industries.
- **General AI** refers to the concept of a sentient machine and one which can perform different types of intelligent tasks and 'human' reasoning. Many consider the possibility of 'general' AI to be decades away, and others consider it unfeasible altogether.

How machine learning and deep learning work

The subsequent sections of this report will refer predominantly to machine learning, since these techniques are the main subset of AI approaches that are of increasing interest in genomics, as well as deep learning – a class of machine learning methods whose impact on biomedical disciplines has been on the rise in recent years.

Figure 1: Relationship between AI, machine learning and deep learning. Machine learning is one approach for achieving artificial intelligence; deep learning is a branch of machine learning. Figure does not reflect the relative sizes of each field.



Rule-based vs. learning systems

While there is no widely accepted definition of AI, it is generally viewed as a technology that enables machines to make an intelligent decision or action. The machines or 'intelligent agents' might correspond to computing hardware, software, an application, or robotic tool.

Some approaches to achieving AI use 'rule-based' programming, which combines human crafted rules with data to deliver an answer or output. Rule-based systems are a way of encoding a human expert's knowledge into an automated system. They were part of the first generation of AI in medicine, with applications in clinical decision support systems. However the development and use of rules-based systems is constrained by a number of factors. They can be:

- **Costly and time consuming to build** – since they demand deep domain knowledge and rely on prior knowledge
- **Challenging to develop** – encoding rules for complex systems or decision processes is difficult
- **Inflexible** – the rules are hard-coded so there is little or no capacity to 'learn' new functionality, this instead has to be delivered via manual human updates.

Machine learning

Machine learning is a collection of techniques based on algorithms that use mathematical procedures to analyse patterns in data and the relationships between them. In contrast to rule-based programming, machine learning algorithms 'learn' from data to discover their own rules and can improve with experience. Relative to a rules-based approach, machine learning can in principle be:

- **Less demanding to build** – some machine learning techniques have a lower dependency on prior knowledge
- **Less challenging to encode** – since the rules for generating an output are established by the algorithm's learning process
- **More flexible** – and easier to update as they are not centred on explicitly defined human coded rules and can instead learn through new experience (e.g. new data)
- **More difficult to interpret** – some (but not all) types of machine learning techniques, particularly those based on deep learning, are not easy for humans to interpret, making it difficult to explain the logic underpinning their outputs. By contrast rule-based systems are highly interpretable since their logic is explicitly defined by a human.

What does machine learning do?

Broadly speaking, machine learning techniques can:

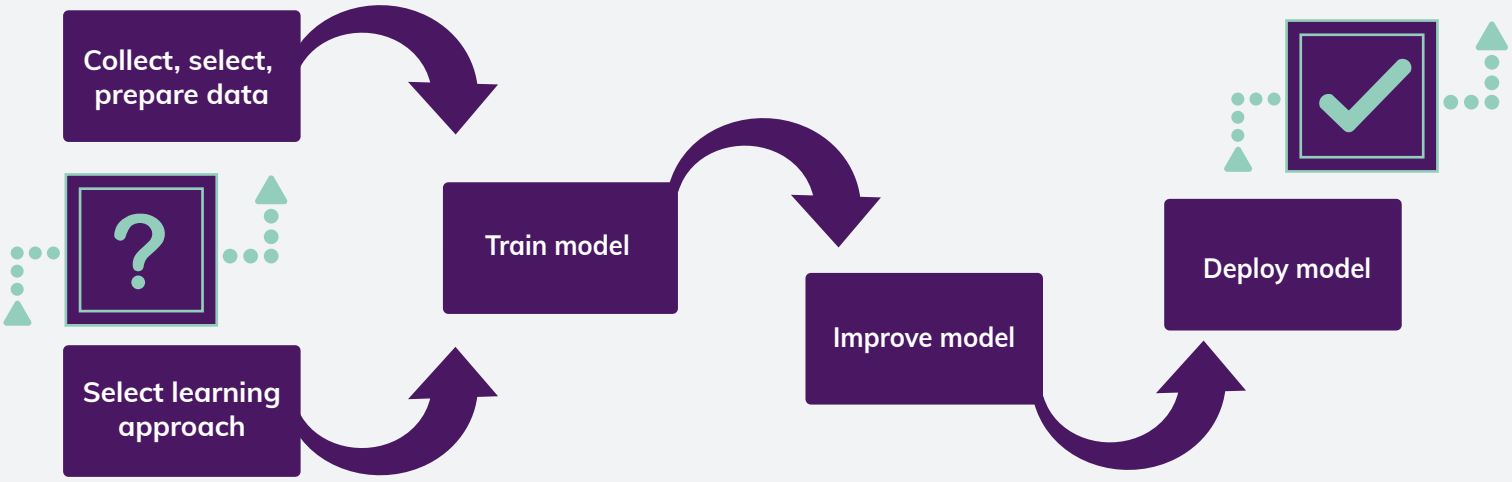
- Make predictions or classifications by learning from a provided set of labelled training data. These data are a range of inputs or features (e.g. age, blood pressure, height / weight etc.) and outputs or outcome (e.g. onset of disease within n-years). This approach is known as **supervised learning**.

- Infer underlying patterns within data without any input from the user (**unsupervised learning**). For example a set of images can be clustered into smaller sub-groups of related images based on patterns or features uncovered by the learning algorithm.

Whatever the approach, there are some steps common to building any machine learning model (figure 2). They include:

1. **Collating, selecting, and preparing relevant data** that will be used for 'training' (i.e. developing) the machine learning mathematical model. Data preparation might include cleaning and organising the data, or adding labels if using a supervised approach
2. **Selecting an appropriate learning technique** according to the problem and the dataset
3. **Developing the machine learning model** using the training dataset and chosen learning algorithm
4. **Iteratively improving the model** by assessing its performance on independent datasets and further customising to optimise
5. **Deploying the optimised learning model** if it satisfies a desired performance threshold, and if it meets other necessary ethical, technical, safety, operational, and regulatory requirements relevant to the real-world use context

Figure 2: Overview of general steps involved in building machine learning models.



The goal of many machine learning tasks is to achieve a model that can make outcome predictions when applied to new data. To do this, when developing some machine learning models, selected data are randomly split into three groups for training, validation and testing[19].

The training data are used to build the machine learning model and the validation data are used to assess the model which can then be 'tuned' to improve performance, e.g. by using more training data.

The performance of the validated model is checked using a test dataset, typically by comparing the model's predictions with observed outcomes in the test dataset. This is a necessary step to determine if the model can generalise to predictions beyond the training data. To achieve good predictive accuracy, the training data must be sufficiently representative of the test data, and ultimately be representative of 'real-world' data in the setting where the model is intended to be deployed.

Validating and testing the performance of a machine learning model in this context does not on its own mean it is ready for deployment in a medical setting. Depending on the intended application, other forms of evaluation are often necessary. These can include undertaking 'real-world' prospective validation studies, demonstrating reproducibility in different datasets, determining clinical utility (i.e. how does the machine learning tool impact on clinical outcomes), and examining how its use compares to existing practice.

Deep learning

Deep learning is a subset of machine learning based on large artificial neural networks, also called deep neural networks. **Neural networks are an approach to machine learning in which small computational units are connected in a way that is loosely inspired by connections in the brain**[28]. They consist of multiple internal layers of connected 'neurons', or nodes, where computation takes place. These nodes progressively detect features, e.g. pixels, from an initial input, e.g. an image. The deep learning process is commonly depicted as a network of nodes (figure 3), starting with a number of input neurons, which feed into any number of 'hidden' layers of nodes before passing to an output layer in which the final decision is presented.

Each hidden layer of the deep learning network detects and integrates information from the patterns in the neurons in the layer beneath. Many deep neural networks can have more than 100 layers, allowing them to model highly complex relationships between the input and output.

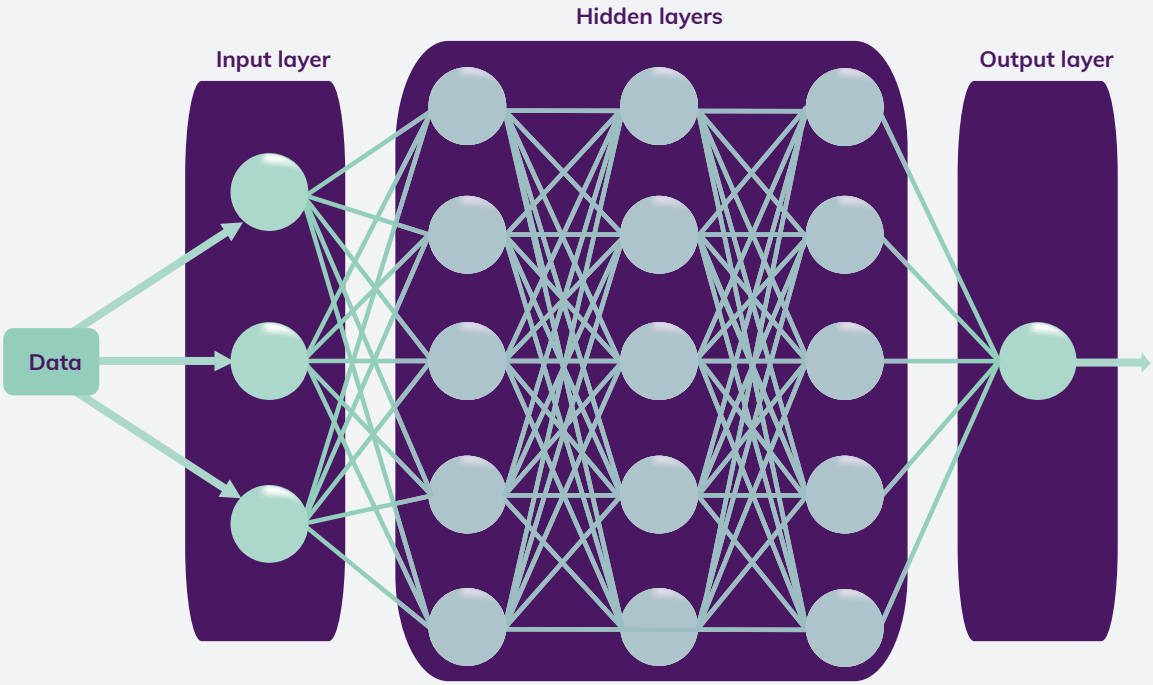
This capacity to learn and process vast quantities of data is an important advantage of deep learning over traditional machine learning methods. Accordingly, the growth of deep learning creates opportunities to leverage very large datasets for novel discoveries and better predictions in the data-rich fields of genomics and healthcare[30].

The main ways in which deep learning differs from machine learning are that deep learning is:

- **More flexible** – with a higher capacity for modelling complex relationships in datasets
- **Less dependent on prior domain knowledge** – under the right conditions, features and patterns in datasets can be learnt with less expert handcrafting compared to other types of machine learning

- **Data hungry** – usually requiring massive amounts of training data to generate accurate predictions
- **Requires greater care to train** – the learning models can be prone to 'overfitting' to spurious patterns or nuances of the specific training datasets, meaning they may not generalise well to independent datasets
- **Computationally expensive to train** – due to the large number of mathematical operations that must be conducted across a multilayer network with many connections
- **Can be more difficult to interpret** – it may not be possible to understand the logic used across the hidden layers of the deep neural network to reach a decision or output. This makes it difficult to extract biological insight or to identify weaknesses in the model

Figure 3: Representation of neural network. Deep neural networks have many hidden layers. The input data could be an image and the output could be a prediction of what the image is (e.g. cat or dog). Alternatively, the input could be a DNA sequence and the output could be a prediction of whether the sequence corresponds to a binding site for proteins. Adapted from Topol (2019)[16].



Examples of common machine learning and deep learning approaches are:

Machine learning

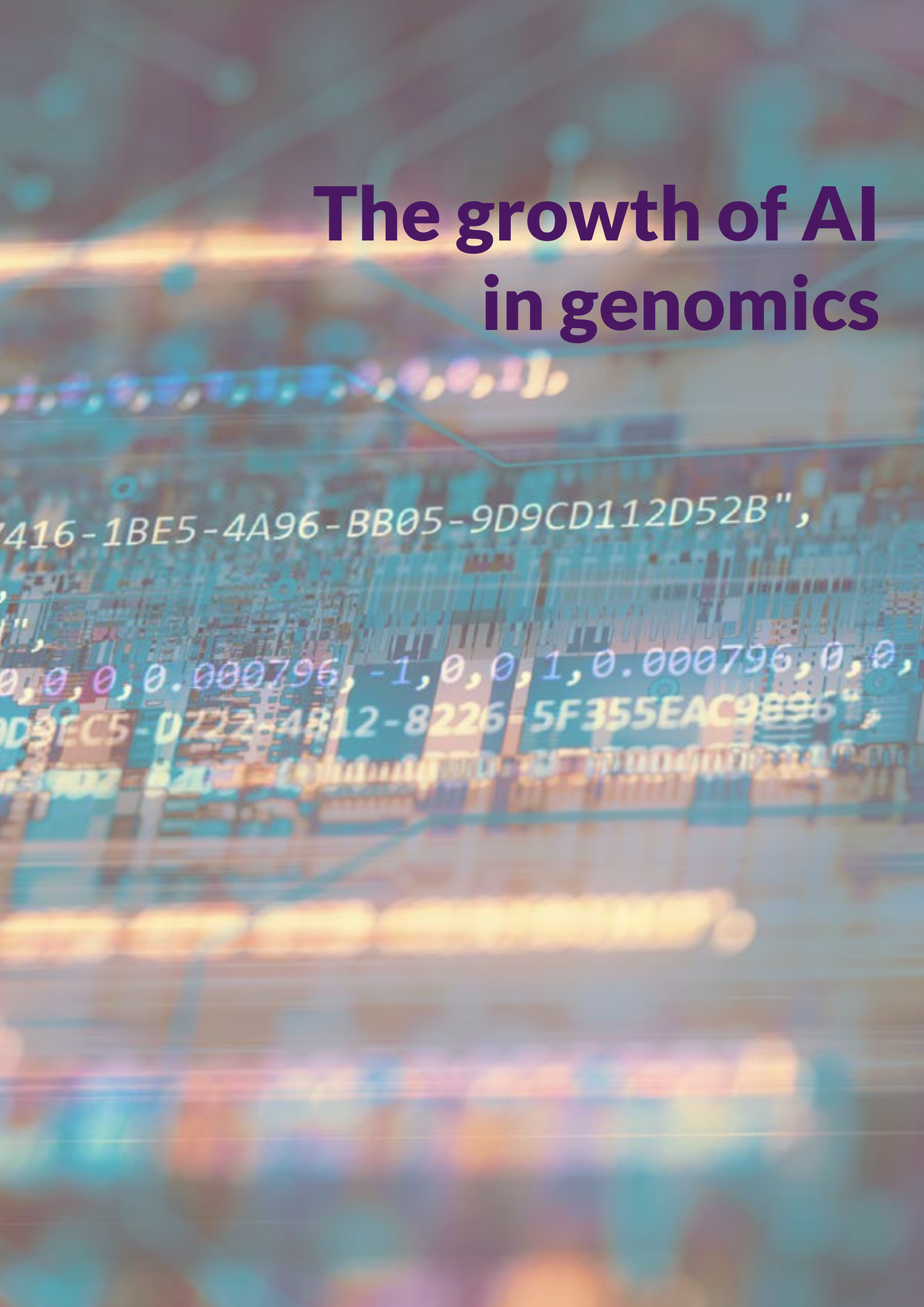
- Logistic regression
- Principal component analysis
- Support vector machine
- Random forest

Deep learning

- Feed-forward neural network
- Convolutional neural network
- Recurrent neural network
- Long short-term memory (LSTM) neural network

Broadly speaking the choice of approach – 'standard' machine learning vs. deep learning – and the specific models used for training are influenced by the datasets in question (e.g. type, size, variety, and distribution of data), the problem at hand, and the ultimate objective.

Deep learning using convolutional neural networks (CNNs) has been especially successful at image processing and classification tasks, where an image is defined according to its visual content. Prior to the advent of CNNs relevant properties or characteristics of the training data known as 'features', had to be defined and extracted. CNNs can instead automatically extract and learn hierarchies of relevant features (e.g. pixels) from images.



The growth of AI in genomics

The growth of AI in genomics

Overview

- **There is growing need for computational approaches that can tackle the analysis of large heterogeneous and high-dimensional datasets. Machine learning can facilitate new discoveries in these datasets without the need to specify explicit rules to undertake these tasks**

- **Machine learning could have utility across different stages of the genomic data pipeline. Most emerging activity in machine learning and deep learning within genomics is taking place within the analysis and interpretation phase**

Genomics is the study of the entire genetic material of an organism, in humans the genome equates to approximately 3 billion DNA base pairs. Genomic medicine makes use of an individual's genomic information to guide their clinical care and deliver more personalised strategies for diagnostic or therapeutic decision making.

One of the major goals of large-scale sequencing initiatives is to advance genomic medicine by accelerating the identification and understanding of disease and/or therapeutic associated genetic variants. Contributing to this endeavour are growing efforts to pool population-level sequence data and link genomic data with phenotypic information, clinical records, and other types of multi-omics datasets (e.g. proteomics, transcriptomics and metabolomics). This integrative analysis of 'omics and clinical data, while essential to facilitating new biomedical discoveries for personalised medicine, poses complex analytical and computational challenges for researchers and clinical services.

As a result there is a growing need for computational approaches that can tackle the analysis of large heterogeneous and high-dimensional datasets (i.e. those containing many attributes and measurements), and methods that can more generally provide faster, cheaper, more scalable, and accurate analytical solutions. Such datasets may for example bring together:

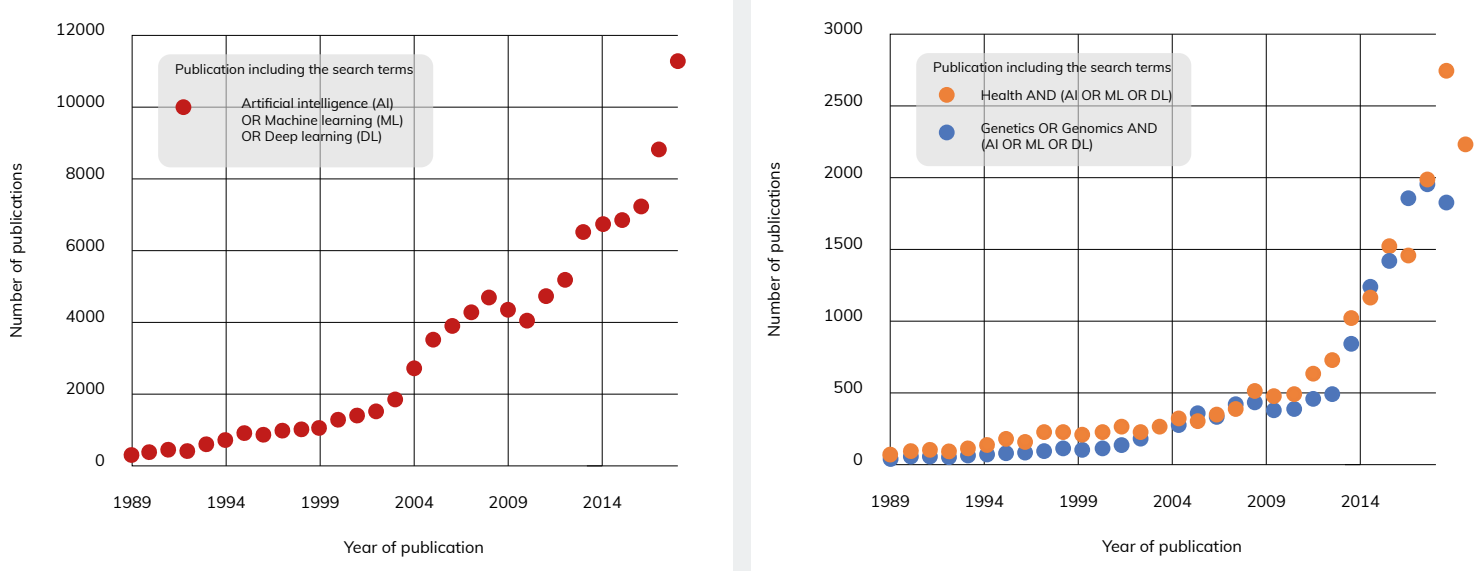
- **Molecular data** e.g. genes, proteins
- **Physiological measurements** e.g. assessments of major organ systems
- **Medical imaging data** e.g. CT, MRI scans
- **Other clinically relevant information** e.g. family histories, histopathology

Some types of data may also comprise many dimensions, where even one sample is defined by hundreds or thousands of measurements, e.g. thousands of genes in one cell. The collective analysis of these enormous datasets is inherently complex, especially when the rules for discovering new insights have to be explicitly predefined, step by step, within the computer code.

This is why machine learning and deep learning are fast gaining attention in the world of healthcare and genomic medicine (figure 4). These techniques can generate predictions, or facilitate the discovery of previously unrecognised patterns and relationships within complex datasets, without the need to specify explicit rules to undertake these tasks.

Stages in genomic data analysis

Figure 4: Growth in the number of research publications archived in PubMed with search terms linked to artificial intelligence, and artificial intelligence and genomics or health.



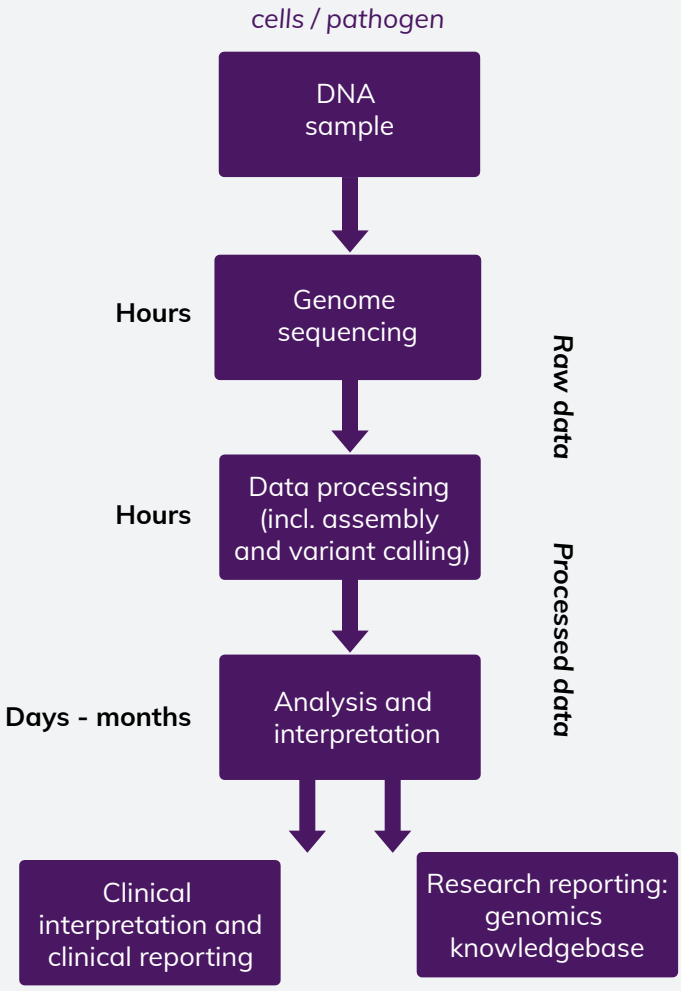
Machine learning could have utility across different stages of the genomic data pipeline:

- **Pre-sequencing** - the samples to be analysed are collected and prepared for sequencing. Depending on the purpose of the analysis the samples might include a person's blood or a tissue of interest (e.g. tumour), or can be samples of cells or pathogens
- **Sequencing** - the DNA sequencer transforms the sample into raw sequence data. Typically this is millions of fragments of sequenced DNA known as sequence 'reads'
- **Data processing** - bioinformatics pipelines are used to reconstruct a genome sequence from the sequenced reads and to then identify 'variants' in the genome – points or regions which vary when compared to a reference genome

- **Analysis and interpretation** - involves the investigation of the identified variants to determine their association with and/or functional implications in disease and health. This is usually the most intensive and time consuming phase of the entire process and can vary enormously depending on the purpose of the investigation and extent of pre-existing data and knowledge. Clinical variant analysis may be used to inform diagnostic, therapeutic, and reproductive decision making. Research analysis can range from seeking to understand genomic variation in the context of populations, to the functional molecular consequences at the cellular level

Most emerging activity in machine learning and deep learning within genomics is taking place within the analysis and interpretation phase (figure 5). The overall aim being to train algorithms to discover patterns in large datasets (e.g. associations between variants and molecular or physiological measures of health/disease), which could then provide new parameters for healthcare personalisation, identify new disease biomarkers, and refine our understanding of disease.

Figure 5: The genomic data pipeline



Machine learning and deep learning in genomics

The application of machine learning to genetics and genomics problems is not new - it would be difficult to handle complex genomic data without such algorithms. In fact, the field of AI has existed for decades, even if there has been an AI resurgence in recent years.

The more novel aspects of machine learning within genetics and genomics are:

- **Emergence of deep learning and its application** to genomics problems
- **Growing capacity to generate and to analyse large volumes** of genomics and associated biomedical data at scale

Rapid growth in machine learning activity

Several related factors are likely to have contributed to the recent expansion of machine learning activity in genomics and in healthcare more broadly:

- **Improving computing power and the declining cost of hardware,** making the use of machine learning at scale more attainable. In particular the evolution of computer graphics cards known as Graphical Processing Units (GPUs) has been game-changing for training deep learning models. GPUs are designed to perform the complex mathematical and geometric calculations needed for 3D computer graphics and advanced image processing. Since GPUs are designed to perform millions of mathematical operations in parallel they are also effective in non-graphics applications that require repetitive computation and which would otherwise be prohibitively expensive using the central processors of computers. In essence, GPUs are to deep learning research what next generation sequencing has been to genomics research.
- **The resurgence of deep learning in the past six years** – facilitated by increases in computational power - has opened up new opportunities to analyse massive health datasets such as medical imaging data, genomic sequence data and electronic health records. The volume and complexity of routinely collated healthcare data is increasing. Globally an estimated 2,314 exabytes of healthcare data will be produced in 2020 (one exabyte = one billion gigabytes)[31]. The ability of deep learning models to handle very large datasets and multiple data types as inputs makes these models attractive for healthcare and genomic medicine[29, 32].
- **Availability of computational frameworks and libraries** for building machine learning models. These are essentially interfaces, tools, and sets of routines and functions written in a given programming language that help facilitate the process of training and implementing machine learning models. As many of these frameworks and libraries are 'open source' (i.e. freely available), they are increasing the accessibility of machine learning to the growing body of researchers seeking to apply these techniques.

- **Explosion in the volume of genomics and biomedical data**, which is projected to exceed other major sources of big data within the next few years[33]. The growth and improving accessibility of these datasets is enabling the training of machine learning models. In addition, advances in biotechnology are making it possible to rapidly generate highly tailored datasets to test new hypotheses. This includes sequencing at the single cell resolution, and targeted gene editing tools to examine specific gene perturbations.

Finally, alongside the above factors, the expanding activity of machine learning can be attributed to the growing investment in the field driven by need and perceived potential.

Existing and emerging applications

Existing and emerging applications

Overview

- **AI is contributing important incremental improvements in clinical genomics analysis including phenotyping in rare diseases and cancer, and variant analysis and interpretation. However, the vast majority of AI activity in genomics is within the research phase**
- **The popularity of deep learning methods for functional genomics analysis is rising**
- **Given the potentially significant impact of AI tools in healthcare, questions are being raised around the regulatory requirements and thresholds for evidence and validation for AI algorithms in medicine**

Most aspects of genomics analysis have been touched in some way by machine learning and deep learning, from sequencing, phenotyping and variant identification, to downstream interpretation (figure 6). In fact, machine learning algorithms have been incorporated into bioinformatics tasks for many years, e.g. genome annotation and variant effect prediction. Now advances in computing, deep learning, and the growth in biomedical datasets are enabling improvements to existing areas of utility.

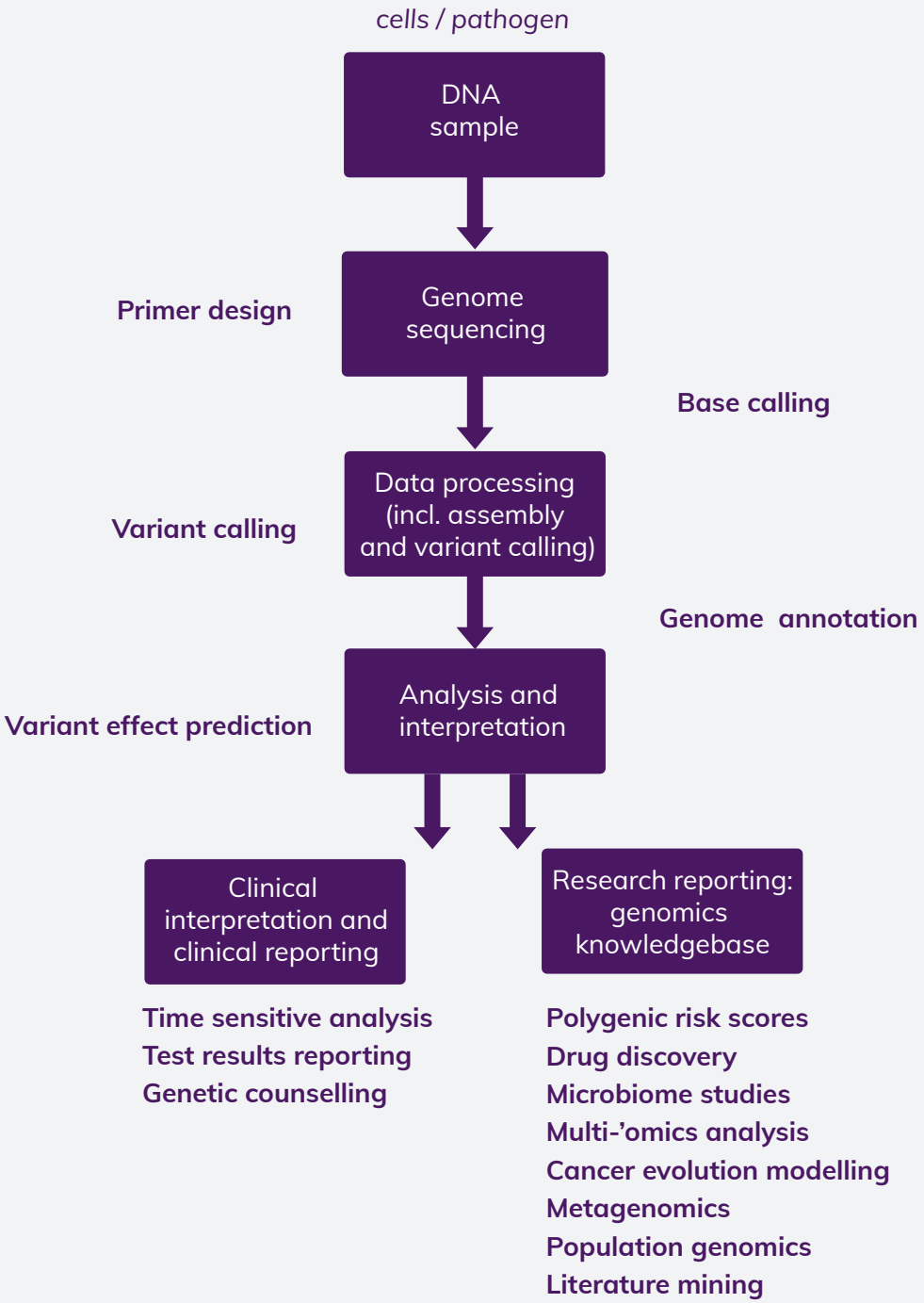
These developments, together with the increase in open-source tools and open access research, are driving the expansion and growth of AI use across different types of genomics analyses. In addition to open-source resources, proprietary software providers are incorporating machine learning algorithms within their genomics analysis tools and services. Table 2 (see appendices) provides a (non-exhaustive) list of companies engaged in AI and genomics activity.

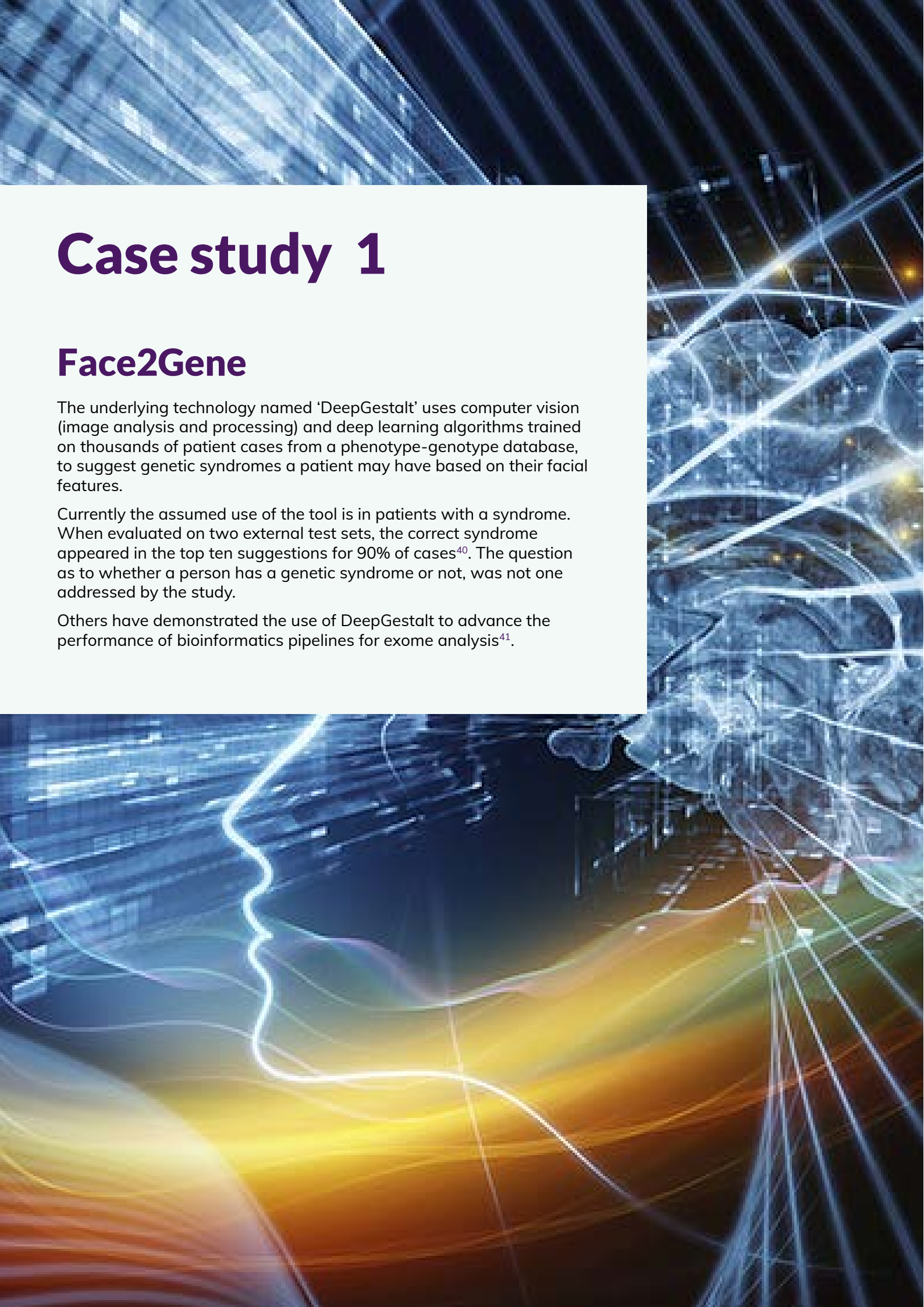
The vast majority of AI activity in genomics is within the research phase. A great deal of excitement and anticipation surrounds deep learning in particular, with a large body of research underway deploying these methods to study the fundamental biological processes underlying disease[34].

Phenotyping

In a clinical context, phenotyping is the process of observing and reporting a patient's features. Phenotype information can be used at several stages of a diagnostic pathway from informing the choice of genetic test, to supporting the interpretation of genetic results. Machine learning methods are being developed to extract phenotype data from electronic health records[35], to refine phenotype classification[36], and facilitate the analysis of phenotype data. In particular, deep learning methods to support image interpretation for rare disease and cancer phenotyping are showing great promise.

Figure 6: AI applications and developments underway along the genomic data pipeline





Case study 1

Face2Gene

The underlying technology named ‘DeepGestalt’ uses computer vision (image analysis and processing) and deep learning algorithms trained on thousands of patient cases from a phenotype-genotype database, to suggest genetic syndromes a patient may have based on their facial features.

Currently the assumed use of the tool is in patients with a syndrome. When evaluated on two external test sets, the correct syndrome appeared in the top ten suggestions for 90% of cases[40]. The question as to whether a person has a genetic syndrome or not, was not one addressed by the study.

Others have demonstrated the use of DeepGestalt to advance the performance of bioinformatics pipelines for exome analysis[41].

Rare diseases

For rare genetic diseases, phenotype data can include specific facial patterns since many syndromic genetic conditions are associated with cranio-facial dysmorphologies. Research to develop computer-aided facial analysis for rare diseases has been underway for several years[37].

Boston based company FDNA have created a smartphone-based facial image analysis framework, Face2Gene, to classify distinctive facial features in photos of people with congenital and neurodevelopmental disorders. The underpinning technology uses deep learning algorithms trained on tens of thousands of patient images to distinguish subtle facial patterns. Trained algorithms then predict the most likely genetic syndrome in patients through the detection of distinctive facial features in photos, and also suggest genes for prioritisation at the analysis stage based on the association of certain genotypes with specific syndromes. The aim of the tool is to support medical genetics in clinical and laboratory practices[38].

Challenges to improving the tool include: ethnic bias in the training data sets (based largely on data from patients of European descent), the fragmentation of databases, and privacy implications surrounding the use of facial images – a sensitive yet easily accessible source of information. Efforts to examine these issues continue. The results of one study suggest that the tool could already be useful in its current form in patients with congenital dysmorphic syndromes in Japan[39] .

Cancers

The intersection of machine learning with genomics and imaging technologies is not limited to rare diseases. Histology (tissue-level) image analysis has been an important tool in cancer diagnosis and prognostication for over a century and is another area where machine learning could support phenotypic analyses. There is a growing body of work demonstrating the potential of deep learning, in particular to facilitate digital pathology workflows[42]. Some of this research has sought to combine pathology image analysis with genomics or other 'omics based measurements for better predictive modelling[43].

Many of the datasets in question are high-dimensional in nature, meaning the number of dimensions or features being measured are very high e.g. gene expression data assaying thousands of genes. Examining these large high-dimensional and multi-attribute datasets in tandem is computationally demanding so has been challenging to date. However improved computing power (e.g. GPUs) in combination with deep learning neural networks that can process large datasets and model complex relations is opening opportunities to analyse and gain new insights from these combined datasets.

In one study a deep learning approach integrating histology images and genomic data predicted the overall survival of patients diagnosed with brain tumours with the same or greater accuracy than human experts[44]. The approach used deep learning to learn visual patterns (from the images) and molecular biomarkers associated with patient outcomes. In another study an integrative model combining 'omics data with histopathology images generated better prognostic predictions in stage 1 lung adenocarcinoma patients compared to predictions with image or 'omics analysis alone[45].

There is also a growing interest in leveraging large multi-attribute and high-dimensional datasets that fuse non-molecular measurements (e.g. radiomics – quantitative features extracted from digital medical images - and digital pathology imaging) with 'omics data to train algorithms for prognostics and diagnostics[43]. However several factors are constraining the ability to train, validate, and deploy these algorithms in a clinical setting. One issue is the slow adoption of whole slide imaging technology (to convert glass slides into digital images of the tissue sample) – a crucial prerequisite where machine learning for pathology imaging is concerned. Another is the standardisation and aggregation of data (imaging and 'omics) across different research or medical centres.

DNA sequencing

Data obtained from any sequencing technology can contain errors and noise – and the types of errors can vary depending on the sequencing method and platform. Machine learning can help improve accuracy in the sequencing process.

Certain sequencing technologies rely on the capture of DNA target regions by complementary DNA 'probes', which can vary in their binding efficiency by a factor of 10,000. To help inform the design of efficient probes, scientists have trained a machine learning algorithm to predict DNA binding rates from sequence data[46].

Nucleotide base-calling from raw DNA-sequencing data can be another source of error. Various deep learning tools have been developed to better predict base identity from changes in electric current measured by Oxford Nanopore long read sequencers[47-49]. Methods for improved base-calling are one approach to improving the accuracy of long-read sequencing which can be lower compared to certain short-read sequencing platforms. Deep learning may offer computational methods for addressing the accuracy, and by extension the clinical usability of long-read sequencing data.

Genomic sequence data processing

Variant identification

Variant identification, also known as variant calling, is the bioinformatics analysis concerned with determining which points in an individual genome differ relative to a reference sequence. The accurate identification of variants is essential for correctly detecting variants that may underlie disease.

While there are well-established methods for variant calling, a number of deep learning tools are being developed with the aim of further improving accuracy of the variant calls.

One of these tools, Google's 'DeepVariant', has outperformed existing state-of-the-art methods on certain datasets despite the model being trained without specialised knowledge about genomics or next-generation sequencing[50]. The tool approaches variant calling as an image classification problem, something which deep learning excels at (see p.15-16), whereby genomic data is converted into images and image analysis is performed to classify points in the genome as a variant or non-variant. However this also adds to the computing overhead – the required power and resources – a cost which may not be justified for some purposes and settings. Collaborative projects are examining its utility in different use cases, including variant calling from long-read sequence data[51], and in non-human genomes[52].

Other groups are developing deep learning based variant callers to better address accuracy issues related to platform e.g. single molecule long-read sequencing technologies, or variants e.g. somatic cancer variants.

Somatic genetic variants are genetic changes that are not inherited or passed onto progeny, but arise in particular cell subsets over time. Although most are harmless, some can lead to local changes in the surrounding tissue, so are of interest in research and for informing patient treatment in certain cancers.

Accurately identifying somatic variants is inherently challenging given the complex nature of tumour biology, tumour-normal cross contamination, sequencing artefacts, and because these variants often occur at low frequencies. A number of machine learning methods[53] have been applied to optimise sensitivity and specificity for detecting true somatic variants, and deep learning approaches are emerging too[54, 55]. The deep neural networks can learn features from observed training data, to better differentiate true variant calls from artefacts introduced by sequencing errors, cross contamination, or coverage biases[55].

Copy number variants (CNVs) – a type of change where regions of DNA are deleted or duplicated – are another challenging-to-detect group of variants to which machine learning methods are being applied[56]. By learning genomic features from a small subset of validated CNVs and using data (CNV calls) from multiple existing CNV detection algorithms, a machine learning approach was trained to accurately identify true CNVs at a higher precision compared to individual CNV callers[56]. Improvements in accurately identifying this class of variants is crucial for medical genomics and research. An estimated 4.8–9.5% of the genome contributes to CNVs[57], with some having no effect on health and others being implicated in a range of inherited and sporadic genetic disorders.

While machine learning and deep learning are assisting in better identifying variants in the genome, understanding the significance of these changes remains challenging. Indeed variant interpretation is another area where machine learning and deep learning techniques are being utilised (see p.30).

Genome annotation

Machine learning methods are used extensively to identify and classify specific sequences and elements within the genome – examples include features known as splice sites, transcription start sites, promoters, and enhancers[58]. These genomic features relate to important functional, structural, and regulatory mechanisms, so their accurate identification is fundamental to clinical genome analysis.

Typically the machine learning methods learn and then detect specific patterns (e.g. DNA sequences) that relate to these DNA elements. Developments in high-throughput sequencing and functional genomics techniques (e.g. methods to analyse protein interaction with DNA) are generating larger and more detailed datasets that can aid in the discovery and prediction of these genomic features. Deep learning approaches are being leveraged to analyse these datasets given their higher capacity for modelling complexity and discovering patterns buried within the detail of these data[30].

Some methods for modelling genomic features have been extended to predict how they are affected by genomic variants and whether this might also impact disease risk[59], i.e. variant effect prediction (see p.30).

Analysis and interpretation

Variant filtering and effect prediction

Prioritising and classifying variants on the basis of their likelihood to cause disease is fundamental to clinical genomics and has implications for patient care and treatment. Together with other sources of evidence, software programs to predict the molecular impact of genomic variants are widely used by clinical laboratories when assessing variants (functional analysis). One major consideration is the effect of genetic variants on proteins, since a protein's shape determines its function and dysfunction in disease.

Algorithms incorporated by tools such as Polyphen, Mutation Taster and CADD determine the degree of protein disruption caused by a given variant based on probabilities learned from labelled genomic data[60-62]. Other tools, such as Exomiser and eXtasy, incorporate phenotype as well as genotype information to score and rank disease causing variants. A challenge for clinical genomics laboratories is that different *in silico* tools can generate different predictions[63]. Discordant results might arise because of differences in the datasets underpinning the tools, user-defined variables, or varying performance characteristics of the algorithms. Studies have sought to compare the performance of various tools[64] and to determine combinations of algorithms with increased concordance[63]. Updated versions of these prediction programs are often released over time as training data sets improve and machine learning technology develops.

As with other elements of the genomics pipeline, the popularity of deep learning methods for functional analysis is rising. Harvard researchers have published an open-source software to predict how proteins fold based on their amino acid sequence[65]. Similarly DeepMind's 'AlphaFold' tool models properties of a protein from its genetic sequence[66].

Machine and deep learning algorithms also exist for predicting the effect of variants that lie outside of protein-coding regions, in non-coding regulatory DNA[67], and for analysis relating to other important molecular processes including gene–gene/gene–protein/protein–protein interactions[68-70], gene expression[71, 72], and methylation – a type of chemical modification to DNA that influences gene expression[73].

Together with deep learning the growing availability of 'omics datasets is bolstering research efforts to improve variant effect prediction. Strategies have included:

- The use of genomic data from non-human primates to predict the clinical impact of human variants[74]
- RNA sequence data to predict variants which affect RNA splicing (an important process for protein diversity)[75]
- DNA sequence data to predict the tissue specific gene expression effects of variants[76].

Over the next few years it is likely that significant in-roads will be made to improve *in silico* functional analysis and prediction of variant pathogenicity. Arguably, this progress will place even greater emphasis on discerning the optimal combination of algorithms for clinical applications.

Case study 2

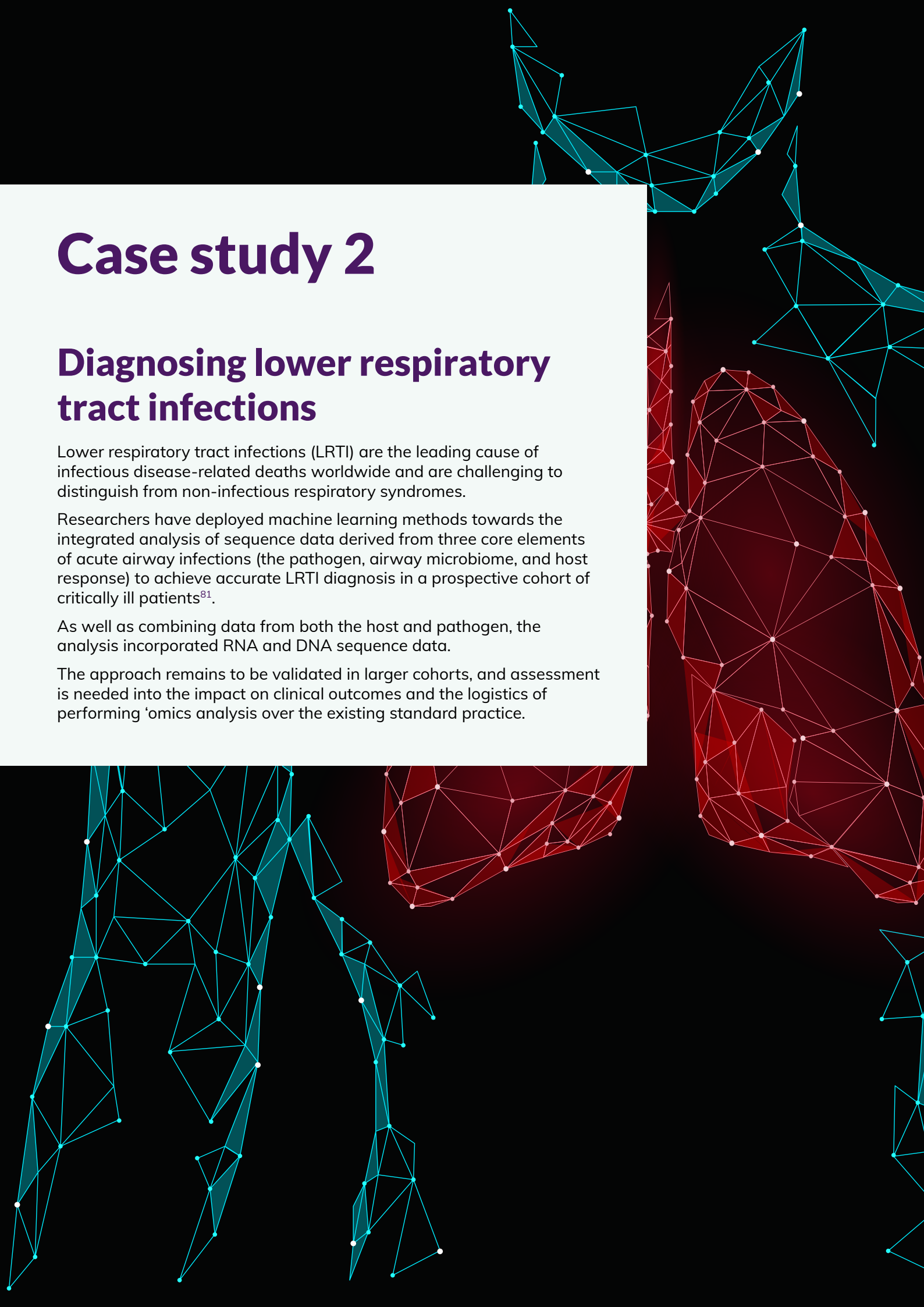
Diagnosing lower respiratory tract infections

Lower respiratory tract infections (LRTI) are the leading cause of infectious disease-related deaths worldwide and are challenging to distinguish from non-infectious respiratory syndromes.

Researchers have deployed machine learning methods towards the integrated analysis of sequence data derived from three core elements of acute airway infections (the pathogen, airway microbiome, and host response) to achieve accurate LRTI diagnosis in a prospective cohort of critically ill patients[81].

As well as combining data from both the host and pathogen, the analysis incorporated RNA and DNA sequence data.

The approach remains to be validated in larger cohorts, and assessment is needed into the impact on clinical outcomes and the logistics of performing 'omics analysis over the existing standard practice.



Integrated multi-omics and multi-modal data analysis

DNA sequencing is just one of many tools for examining the functional genomics pathway. The ability to measure and analyse other molecular constituents of cells can be just as important for understanding the significance of genomic variation. In fact many patients referred for genetic testing will be faced with an inconclusive result due to the limits of our scientific knowledge, variations in cellular processes, environmental dynamics, and non-inherited variants. Various factors can affect the molecular processes between genotype and phenotype.

To better understand the biological complexity of diseases, numerous research efforts are attempting to apply an integrative 'omics approach[77]. Typically, studies combine data from multiple 'omics technologies together with health records and, in some instances, environmental monitors[78]. Machine learning techniques are gaining traction for the analysis of these high-dimensional datasets, offering the advantage of being able to sift through large volumes of disparate data types to discover patterns[79, 80].

Clinical deployment is currently constrained by costs associated with data collection, storage, and analysis, as well as issues of standardisation, reproducibility and utility. Although medicine is a long way from routine multi-omics diagnostics, the growing trend for combining advanced analytics with detailed biomedical datasets will be key for advancing personalised medicine and addressing multifactorial diseases. One example is the integration of host and pathogen data to examine the manifestation of infectious diseases (case study 2).

Other complex genomic data analyses

Machine learning methods are shifting analytical capabilities and expanding the options available to address complex problems in other areas of genomics research involving large, heterogeneous, or multi-model data types. Examples include:

Population genetics

Population genetics is the study of genetic variation within populations and the factors which shape this variation over space and time. It is argued that supervised machine learning could attain greater predictive power than classical statistical estimation approaches widely used in population genetics. Compared to competing methods machine learning makes fewer assumptions and is agnostic about the processes used to create datasets, so could be better at recognising phenomena as they are in nature, rather than how scientists choose to represent them in a model[82].

Polygenic scores

Polygenic scores (PGS) discern the cumulative effect of common single nucleotide polymorphisms (SNPs), a type of genomic variant, which individually have a small effect on a trait. They have been developed as a means of investigating the genetic basis of complex traits, which are influenced by multiple SNPs. Although PGS are not yet widely used in the clinic, there is interest in utilising them for prediction of common diseases, with the assertion that polygenic risk prediction could potentially lead to actionable outcomes[83].

Polygenic scores are calculated using polygenic models, and a limitation of these models is that they incorporate strict assumptions about the underlying data which do not necessarily reflect the complex biology of polygenic traits, and so result in reduced predictive efficacy. For example one assumption is that different data observations e.g. SNPs within genes are non-correlated, whereas in reality many complex traits are underpinned by different gene-gene interactions as well as the interplay between genetic and non-genetic (lifestyle/environmental) factors.

It is suggested that certain machine learning methods could improve the predictive power of these models as they make fewer assumptions and have greater capacity to recognise patterns in strongly correlated data. They could also be used to develop more dynamic methods that better account for complex interactions e.g. between genes and other factors e.g. the shifting influence of genetic factors over the human lifespan[83, 84].

Microbiome studies

Microbiome studies examine all genetic material within a microbiota – the collection of microorganisms present at particular body sites e.g. the gut, skin. Machine learning methods are being applied in microbiome research to classify specific microbial sequences in a sample, and to investigate the link between dynamic changes in the microbiome and host phenotype and disease[85, 86].

Single cell analysis

Single cell analysis is the application of 'omics technologies to individual cells. Advances in single cell sequencing techniques are helping to capture the complexity and diversity of cell populations, and provide greater detail into the molecular characteristics of disease. Machine learning algorithms are being trained to analyse the growing volume of single cell data and address interpretation issues linked to data quality, noise, and heterogeneity[87].

Cancer evolution modelling

Cancer evolution modelling aims to determine the temporal order of genetic changes that occur in different cancers as they evolve and change. This information could inform strategies for early detection and for anticipating disease progression. Several research groups are developing machine learning methods to track cancer evolution and to determine which genetic changes are drivers for cancer growth[88-90].

Clinical decision support

Time sensitive analyses and periodic reanalyses

Broadly speaking the value of machine learning comes from the opportunity to accelerate discoveries of significance to genomic medicine. A group at the Rady Children's Hospital, San Diego, have applied this notion in a very literal sense by using machine learning to support the rapid analysis of whole genome sequence data for the diagnosis of critically-ill newborns in less than 24 hours[91].

Case study 3

Automating analysis of genetic data

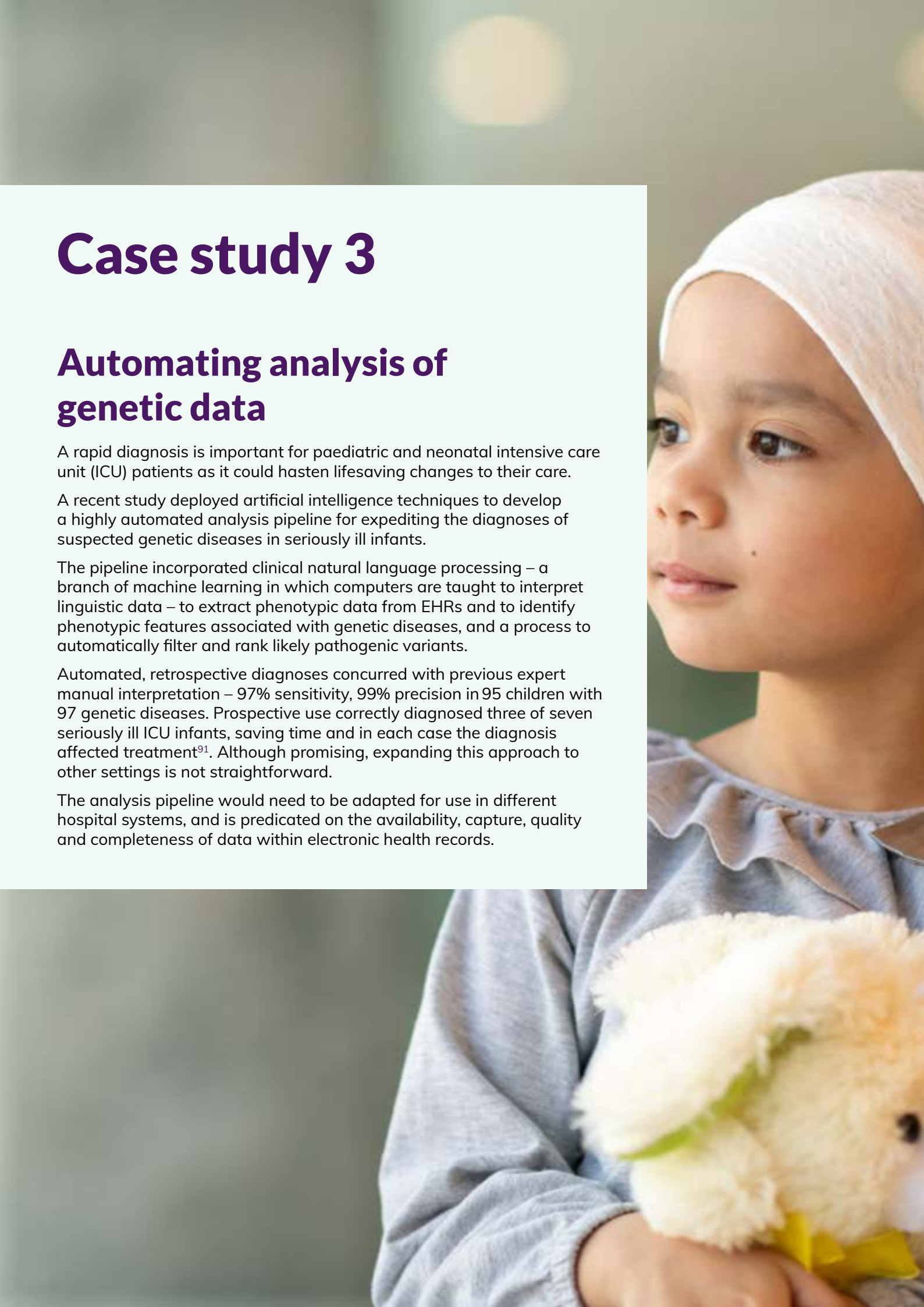
A rapid diagnosis is important for paediatric and neonatal intensive care unit (ICU) patients as it could hasten lifesaving changes to their care.

A recent study deployed artificial intelligence techniques to develop a highly automated analysis pipeline for expediting the diagnoses of suspected genetic diseases in seriously ill infants.

The pipeline incorporated clinical natural language processing – a branch of machine learning in which computers are taught to interpret linguistic data – to extract phenotypic data from EHRs and to identify phenotypic features associated with genetic diseases, and a process to automatically filter and rank likely pathogenic variants.

Automated, retrospective diagnoses concurred with previous expert manual interpretation – 97% sensitivity, 99% precision in 95 children with 97 genetic diseases. Prospective use correctly diagnosed three of seven seriously ill ICU infants, saving time and in each case the diagnosis affected treatment[91]. Although promising, expanding this approach to other settings is not straightforward.

The analysis pipeline would need to be adapted for use in different hospital systems, and is predicated on the availability, capture, quality and completeness of data within electronic health records.



The study authors suggest that such a rapid automated analysis pipeline could have a number of uses, including immediate provisional diagnosis, independent re-evaluation in cases where manual interpretation fails to provide a diagnosis, and the periodic reanalysis of unsolved cases.

The subject of re-evaluation and reanalysis of test results is coming into sharp focus as the use of genetic and genomic testing expands[92, 93] and as associated databases and knowledge improve over time. As more patients are tested and larger regions of the genome analysed, the potential for uncertain findings increases. It is possible that in the future, machine learning methods could support the semi-automated and systematic reanalysis of variants to determine potential changes to the interpretation of existing test results.

Clinical genomics result feedback

An inherent challenge in clinical genomics is the communication of complex information and test results to patients and mainstream physicians. Many health systems have a limited pool of clinical geneticists and genetic counsellors which could be overwhelmed by the increasing volume of genetic testing. As a means of supplementing and scaling genetic counselling, companies are developing AI chatbots for patients to talk about genetics.

A number of health centres are testing the 'Genetic Information Assistant' created by Clear Genetics. The tool has been used by Geisinger Health in Pennsylvania, US, for patients undergoing sequencing as part of their MyCode study. Another company, OptraHealth, has created a similar chatbot 'GeneFAX', as well as a digital genetics assistant 'OptraGuru' that can be queried via the virtual assistant tools Amazon Alexa and Microsoft Cortana.

The growing use of direct-to-consumer (DTC) genetic testing is further accentuating the need to strengthen genomic literacy among the public and healthcare professionals. In an age where many people use the internet to search for health information, carefully designed and rigorously tested chatbots could play a constructive role in disseminating genomic knowledge.

Drug discovery and therapeutics

Drug discovery

The opportunities to apply machine learning across all stages of drug discovery have motivated many pharmaceutical companies to invest resources into this area[94]. A comprehensive account of this activity is beyond the scope of this report. With respect to genomics, machine learning is being applied to these datasets for a number of purposes including defining disease sub-types, identifying disease biomarkers[95], target discovery[94], drug repurposing[96], and predicting drug responses[97].

Many major pharmaceutical companies have AI-focused R&D initiatives or collaborations underway. For example, AstraZeneca and BenevolentAI are applying AI to genomics, chemistry and clinical data to accelerate the discovery of new potential drug targets. GlaxoSmithKline (GSK) has invested in consumer genetics company 23andMe, gaining access to the datasets they hold to develop drug targets using machine learning. The drug-maker has also established collaborations with AI drug discovery companies.

To some extent the hype surrounding AI for pharma has been tempered by the high-profile failure of IBM Watson Health's Oncology AI software. The cognitive computing cloud platform, used by hundreds of hospitals globally, returned multiple examples of unsafe and incorrect treatment recommendations for cancer patients. Given the potential significant impact within healthcare, this example and others[98] are raising important questions around the regulatory requirements, and thresholds for evidence and validation for AI algorithms in medicine[25].

Genome editing

Genome editing, whereby sections of DNA are removed, added or altered, is another area of therapeutics research being facilitated by machine learning. Genome editing techniques are widely used in research to investigate the role of genes and DNA sequences, and increasingly for therapeutic ends, to replace or alter a defective gene in patients.

Machine learning and deep learning algorithms are being trained to increase the efficiency and accuracy of deploying CRISPR – currently the most versatile, cheapest, and simplest tool for genetic manipulation. Algorithmic methods have been developed to predict the activity of the editing system [99, 100], the exact changes resulting from edits[101], and off-target effects – unintended DNA changes that can complicate or hinder use of the technology[102]. Advances in *in silico* predictions will be vital for enabling better research models to study disease and for accelerating and informing the design of safer and more precise therapeutics. For these reasons CRISPR technologies are rising up the agenda of pharmaceutical companies. GSK have announced a multi-million dollar partnership with the University of California to establish a CRISPR laboratory, with data analysis to be supported by GSK's artificial intelligence group.

AI for genomic medicine - conclusions

The number and range of applications for AI in genomics is rapidly expanding. While AI has not yet brought about a watershed moment for clinical genomics analysis, it is contributing important incremental improvements in the quality and accuracy of predictions made along the genomics analysis pipeline. Collectively these changes could lead to substantial progress, particularly given the escalating scale and pace of activity.

The advantages presented by AI models for evaluating large, complex biomedical datasets, hold enormous potential for accelerating discoveries of significance to genomic medicine. As machine learning and deep learning speed up the pace of discovery, the key challenge will lie in bridging the research to clinic gap.

Despite the vast potential, major hurdles remain to be addressed if AI is to live up to the high expectations that it will transform genomic medicine.

Considerations for policy

Considerations for policy

Overview

- **The accuracy of a machine learning model is highly dependent on the quality and reliability of the training data. Healthcare datasets are noisy, complex, heterogeneous, poorly annotated and generally unstructured**
- **Better data management, and access to specific technical skill sets, are key to enabling a clinical environment that can readily engage with advances in machine learning**
- **Challenges around transparency, privacy and security could undermine important efforts to build patients' and healthcare professionals' trust in AI systems**
- **If left unresolved, the Eurocentric bias of genomic data could exacerbate existing health disparities**
- **An understanding of the limitations of AI systems and reasonable expectations for their incorporation into healthcare is key. AI is likely to augment and support, rather than replace healthcare professionals**

Data - improving quality, accessibility, and representation

The 'quick-wins' claimed for AI are often at odds with the reality of building a healthcare-ready algorithm - for instance, the importance and difficulty of the initial steps of data curation, cleaning and preparation is often underestimated.

In an ideal scenario, training data would be fully annotated, well structured, contain minimal noise (corruption), and be appropriate for the specific task at hand[103]. In reality, healthcare datasets are noisy, complex, heterogeneous, poorly annotated and generally unstructured[32]. In some cases, valuable health data may not even be captured digitally – the most fundamental prerequisite for building AI models. For these reasons, extensive effort has to be expended on collating, cleaning, standardising, and formatting datasets before they are used to develop algorithms.

Genomic data has several sources of error and biases including those stemming from differences across various laboratory sequencing kits, methods and technologies, as well as technical sequencing artefacts.

The Eurocentric bias of the data is a well-recognised issue in genomics and if left unresolved has the potential to exacerbate existing health disparities for groups that are already underserved[104-106].

Machine learning algorithms trained on datasets that are predominantly derived from individuals of European ancestry will be less effective than those trained on a fully representative dataset, and potentially even incorrect and harmful, if used to make predictions on individuals of non-European descent.

Sequencing initiatives that seek to gather datasets that are representative of the diversity within society are one important step towards mitigating negative algorithmic biases. The UK government intends to include under-represented groups as part of the plans to sequence five million genomes[107]. Similarly the NIH's 'All of Us' research program aims to sequence a diverse sampling of Americans. Greater transparency around the limitations of the data used for training, including the extent of diversity is another important area for action.

Bias can also arise as a result of the availability of data, how those data are prepared and combined, how questions are framed and because of preexisting prejudices within society.

Scientific research is not immune to these problems. How to assess, address, and mitigate against these challenges will be crucial as the deployment of machine learning methods on large biomedical datasets grows.

Managing expectations

The degree of hype surrounding AI in healthcare has yet to be equalled by the volume of solid evidence for clinical deployment. Descriptions in the popular press and by some AI companies seeking to promote their tools can project unrealistic expectations around the clinical readiness of AI in genomics. This is at a time where the AI field is grappling with what has been described as 'a reproducibility crisis'[108].

Replicating the methods and results of AI studies is a critical component of scientific research, and the development of tools for medical applications. Achieving this in practice is not trivial.

One factor impeding reproducibility is the inaccessibility of exact research methods, an algorithm's code, or the underlying training data. This can be because the developers are either unwilling or unable – for technical, commercial, or privacy reasons – to share specific details. Even when methods and code are available, there may be insufficient information within the reported documentation to precisely repeat an experiment[108,109].

Certain proprietary analysis and decision support tools are marketed for their potential to impact genomic medicine. However details about their use of AI are scarce - while these platforms may facilitate data analysis, the contribution of AI could be inflated - making it difficult to assess the validity of their claims.

Clinical genomic analysis incorporates machine learning algorithms at various stages, but this is not the same as a complete end-to-end AI-driven analysis pipeline, or one that negates the need for human input. Humans take key decisions about how to design and deploy algorithms. In fact most experts agree that AI is likely to augment and support, rather than replace healthcare professionals. This difference is critical from a regulatory perspective, because solely automated individual data processing which produces legal or significantly similar effects as might be the case in healthcare requires additional regulatory safeguards (see p.41-42).

AI activity in genomics research and the life sciences is progressing far faster than in clinical practice. This has been attributed to the different standards for validation and regulatory oversight, as well as the willingness of the scientific community to implement[16].

For healthcare, the thresholds for adopting new technologies are higher than for other sectors. Currently AI adoption in healthcare is limited by the dearth of robust, prospective validation in clinical settings, even before considering the crucial question of clinical utility.

A lack of external validation of the predictive performance of machine learning models is also a challenge within research. However, research is generally better primed to take advantage of these algorithms. Whilst not completely seamless, research data tends to be curated and organised in way that makes it more amenable to AI than healthcare data.

The UK Biobank, a collation of genotype data and medical information on 500,000 participants, is one example. Established over ten years ago, the Biobank has over >150 studies listed that purport to incorporate machine learning. The 100,000 Genomes Project has a specific consortium dedicated to 'Quantitative Methods, Machine Learning and Functional Genomics'.

As well as having better access to the infrastructure and personnel required for developing and deploying AI, the research sector tends to adapt more responsively to informatics developments. Improved infrastructure, better data management, and access to specific technical skills sets are key to enabling a clinical environment that can readily engage with advances in machine learning.

Infrastructure, training and constructive collaborations

Effective use of AI in healthcare requires robust data infrastructure, high quality datasets, interoperability and sharing standards[15]. Genomic data in particular places considerable demands on computing and digital storage. Moreover training deep learning models is computationally expensive and this constraint can restrict AI development to institutions with access to high-performance computing. However the underlying data infrastructure in the NHS is not considered fit for purpose for AI[15].

The evolving demands of large-scale biomedical data analyses, and the rapid progress in GPU technology, means computing infrastructure must be sufficiently scalable, secure, flexible and accommodating of advances to meet the evolving needs of AI in genomics. Inflexible computer hardware systems are at risk of becoming quickly outdated, such that they may not be able to perform the necessary analyses.

Cloud computing services are one solution to these technical challenges. Moving to cloud services reduces the overhead of managing the computing infrastructure and offers the ability to add computational capacity on demand, in real time. Specialist expertise is still required to manage the interface between the cloud service and organisation, and to configure the platform for building AI models.

Moreover, there are a number expectations that healthcare organisations must meet to ensure the safe, secure, and effective use of cloud services[110]. These include undertaking a risk assessment, monitoring implementation, and selecting a cloud service that meets the relevant security and regulatory requirements. The use of cloud computing can raise regulatory challenges if the cloud is located outside a designated geographical area[111].

The successful application of AI requires a combination of expertise, spanning statistics, machine learning and deep learning, and genomics. The need to recruit and retain staff with AI knowledge was highlighted in the NHS workforce review led by Eric Topol[6]. Some of this expertise is moving from academia to industry as commercialisation facilitates access to state-of-the-art computing infrastructures and greater remuneration.

One suggestion for attracting the necessary technical skills into the NHS is to create long-term roles that share time between the NHS and industry and/or academia[6]. This and other strategies for fostering greater cross-discipline collaborations are paramount to guiding the effective and appropriate development of AI in genomics.

The importance of domain knowledge in the specific area of genomics should not be underestimated, nor should the significance of understanding the environment in which the algorithms will be deployed.

A poor understanding of the healthcare ecosystem and stakeholder needs, including thresholds for evidence and validation, has contributed to the downfall of numerous healthcare start-ups. A centralised agency, NHSX, has been created with responsibility to improve health and care by 'giving people the technology they need'.

A code of conduct published in 2018 clearly sets out the principles expected from those developing, deploying and using data-driven technologies in the NHS[27]. Additional guidance to assist in interpreting certain key principles in the code is being developed. NICE has developed evidence standards[112] and a user guide[113] to support the development process.

Domain knowledge is required to inform the choice and design of the learning models. Depending on the problem and dataset in question, deep learning methods may not be the optimal option despite their growing popularity[30]. In fact, in some cases they may lead to poorer predictive performance than simpler machine learning models.

AI in genomics is undoubtedly a team endeavour. Several multidisciplinary and cross-organisational interactions are underway between computing, AI, and biomedical/genomics sectors.

Oxford Nanopore's MinIT, a hand-held AI supercomputer for nanopore sequencing, is powered by a GPU manufactured by NVIDIA, a major AI computing company. Separately, NVIDIA and the Scripps Research Translational Institute (San Diego) are collaborating to develop AI tools and infrastructure for the analysis of genomic and digital medical sensor data. A collaboration between Microsoft and academic research institutions led to a set of computational tools to inform more efficient and accurate use of CRISPR genome editing[102].

The trend for cross-disciplinary partnerships is likely to continue. Currently no one sector has a monopoly on all the necessary skills, expertise, data and resources to deliver the benefits of AI in genomic medicine at scale.

Regulation, explainability and interpretation

The regulation of machine learning systems is a complex and hotly debated topic in technology law and policy. In the UK, for the forseeable future, EU regulations are expected to apply in this area, either in their current form or transposed into UK regulation.

The General Data Protection Regulation (EU) 2016/678 (GDPR) grants rights to EU citizens as data subjects, including transparency over data use, a right to data access and the right to be forgotten. In addition, the EU Medical Devices Regulation (EU 2017/745) and *In vitro* diagnostic Medical Devices Regulation (EU 2017/746) determine compliance standards for AI software that qualifies as a medical device or an *in vitro* diagnostic medical device.

Other jurisdictions face similar problems in trying to provide sufficient oversight of highly adaptive and dynamic systems incorporating complex algorithms. The legislation on data acquisition, product certification and product liability for algorithms will require clarity in light of the growing use of AI in healthcare.

Two areas are intensely occupying legal scholars and those developing machine learning tools for health or research.

Does the current regulatory framework strike a fair balance between the need for medical innovation and patient safety?

Assuring compliance with the GDPR is likely to make the development process more burdensome prescribing the information to be given to data subjects, and the Regulation's data minimisation requirements are likely to reduce the amount of data available for training models. However Article 9(2)(j) and Article 89(1) of the GDPR provide an exemption for research, acknowledging the need to balance the value of data processing in the public interest, for scientific or historical purposes against the rights of data subjects. Other provisions facilitate the processing of personal data for health and for other areas of public interest.

The medical devices regulations could have implications for AI applications that are used to determine disease risk and guide diagnosis for individuals. Qualification as a medical or *in vitro* diagnostic medical device will largely depend on whether the manufacturer intends the algorithms to be used for a medical purpose. If they qualify, they will be subject to the safety, performance, and quality management criteria that these regulations impose.

Does the GDPR contain a right to explanation?

There has also been vigorous debate concerning the existence and nature of a right to explanation. A right to explanation could potentially cover both rights to transparency about the algorithm and its application, and specific provisions in the GDPR which potentially allow EU citizens to contest 'legal effects or similarly significant' decisions made about them solely using algorithms. This has aroused substantial debate as to when this right might be triggered, and how it might be mitigated as some machine learning algorithms, especially those based on deep learning, are 'black boxes', i.e. the processes between data input and decision output are opaque.

Debate includes the extent to which a potential right might cover the internal logic of the algorithm and why the application of machine learning generated the outcome it did. Explaining exactly why the algorithm came to the decision it did, can be impractical if not impossible.

Consequently, explainable AI is emerging as a field to address how black-box decisions of AI systems are made. As well as the prospect of a legal right to explanation, there is an expectation in medicine – and in other sectors such as justice systems, and recruitment – for transparency, ensuring fairness and accuracy, and engendering trust[114, 115].

On one level, model interpretability can help to discover and avoid discriminatory or flawed predictions. These might, for example, have arisen due to bias, or quality issues in the training data. It might also make it easier to detect catastrophic failure of a model[116, 117].

At another level, interpretable models could identify relationships in data that are potentially important for health. Some experts have argued that 'the very high complexity of biological systems will intrinsically limit applications of current 'black box' machine learning in patient data'[118]. In other words, the inscrutable logic of black box algorithms could result in a missed opportunity to gain the causal mechanistic insights that are essential to understanding disease, basic biology, and identifying drug targets[118]. Indeed, commentators have suggested that for some 'high stakes decisions', interpretable models are always preferable[119].

Privacy, security, and public perception

Various data sharing scandals have heightened public concerns around confidentiality, privacy and the ethical use of data. These must be recognised and addressed, or they risk seriously impeding the effective use of these technologies.

Among the concerns is the sharing of healthcare data with commercial organisations[120]. Past scandals have underscored the need for greater transparency about business models, better engagement with the wider public, as well as assurances around how data are used by the commercial companies.

Collaborations between the computing/AI industry and healthcare institutions involves a 'value exchange'[121], for example high-end infrastructure and machine learning expertise in exchange for access to health datasets and medical expertise. This has raised important questions about how the benefits of these public-private partnerships are shared fairly.

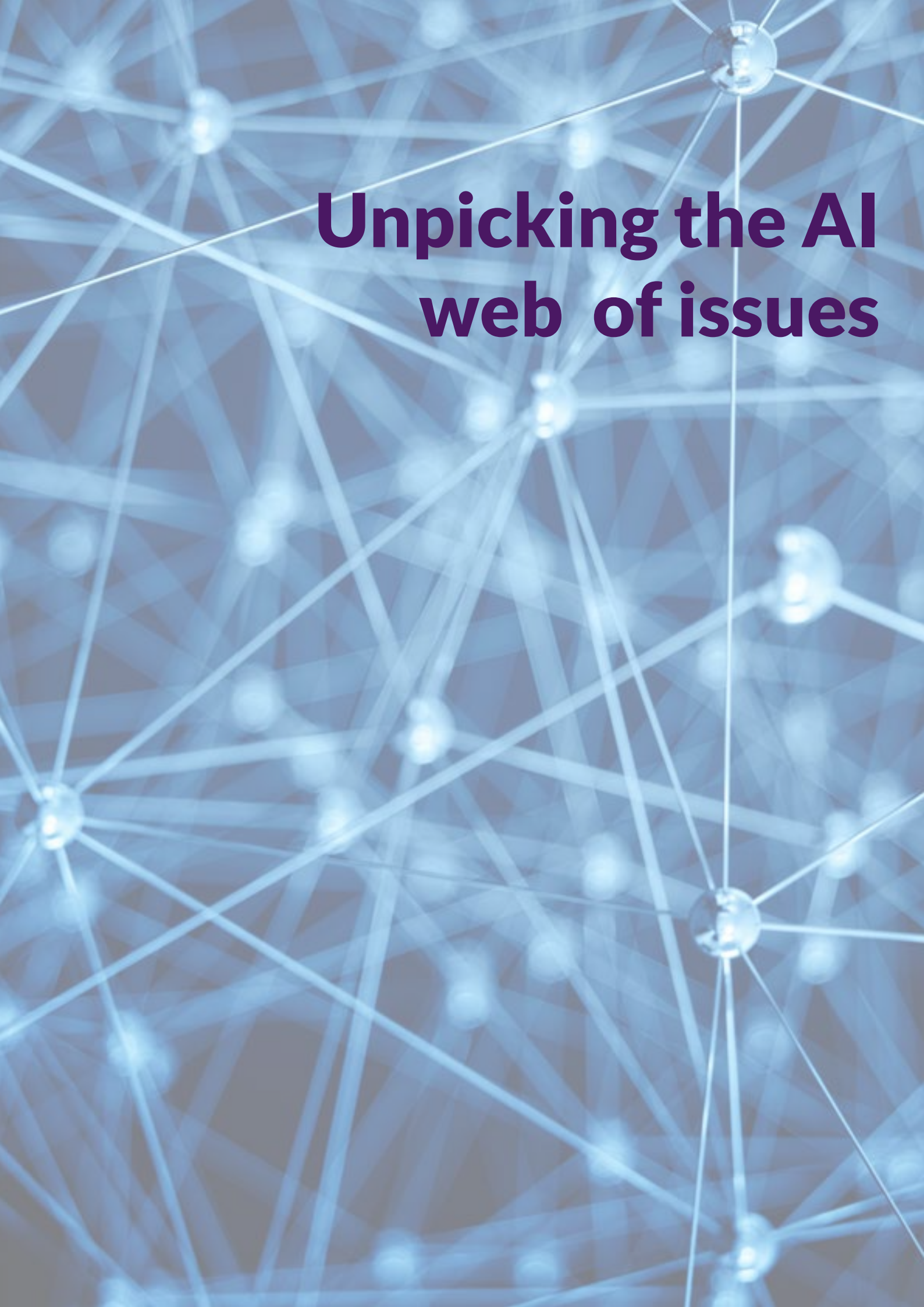
Another challenge is preserving people's privacy when using amassed data about them. Data de-identification is widely regarded as the solution, but the robustness of this approach is being questioned in light of a growing number of cases where data have been re-identified. One study found that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes[122]. This issue is a major one for AI in genomic medicine as one of the major objectives is to harness the untapped potential of large detailed health and biomedical datasets.

An added difficulty is the vulnerability of machine learning models to a range of cybersecurity attacks that are increasing in sophistication. Even without sharing data, in certain cases, information about the training data used for a machine learned system can be reconstructed from the model.

Security challenges have stimulated the development of privacy enhancing technologies and privacy preserving machine learning techniques[123]. Measures to prevent data breaches and actively mitigate against security risks must be a core – rather than an adjunct – component of genomics – AI activity.

Another concern is the potential for the malicious or discriminatory use of AI tools. For example the deliberate hacking of health algorithms to cause harm, or the use of open-source predictive tools to reveal sensitive information about someone. Even the misuse of non-clinical AI tools in society could incite a negative perception of their use in medicine.

The controversy surrounding facial recognition and analysis technologies in the public and private sectors is one example. How stakeholders and policy makers respond to and curb the malevolent use of AI driven tools that undertake highly sensitive analyses will be critical, especially as they become more easily accessible and simpler to deploy.



Unpicking the AI web of issues

Unpicking the AI web of issues

Overview

- **The application of AI for genomics poses multiple interconnected issues requiring resolution**
- **These issues cannot be effectively and sustainably addressed in isolation**
- **Seven priority actions can go a long way towards meeting these policy challenges**

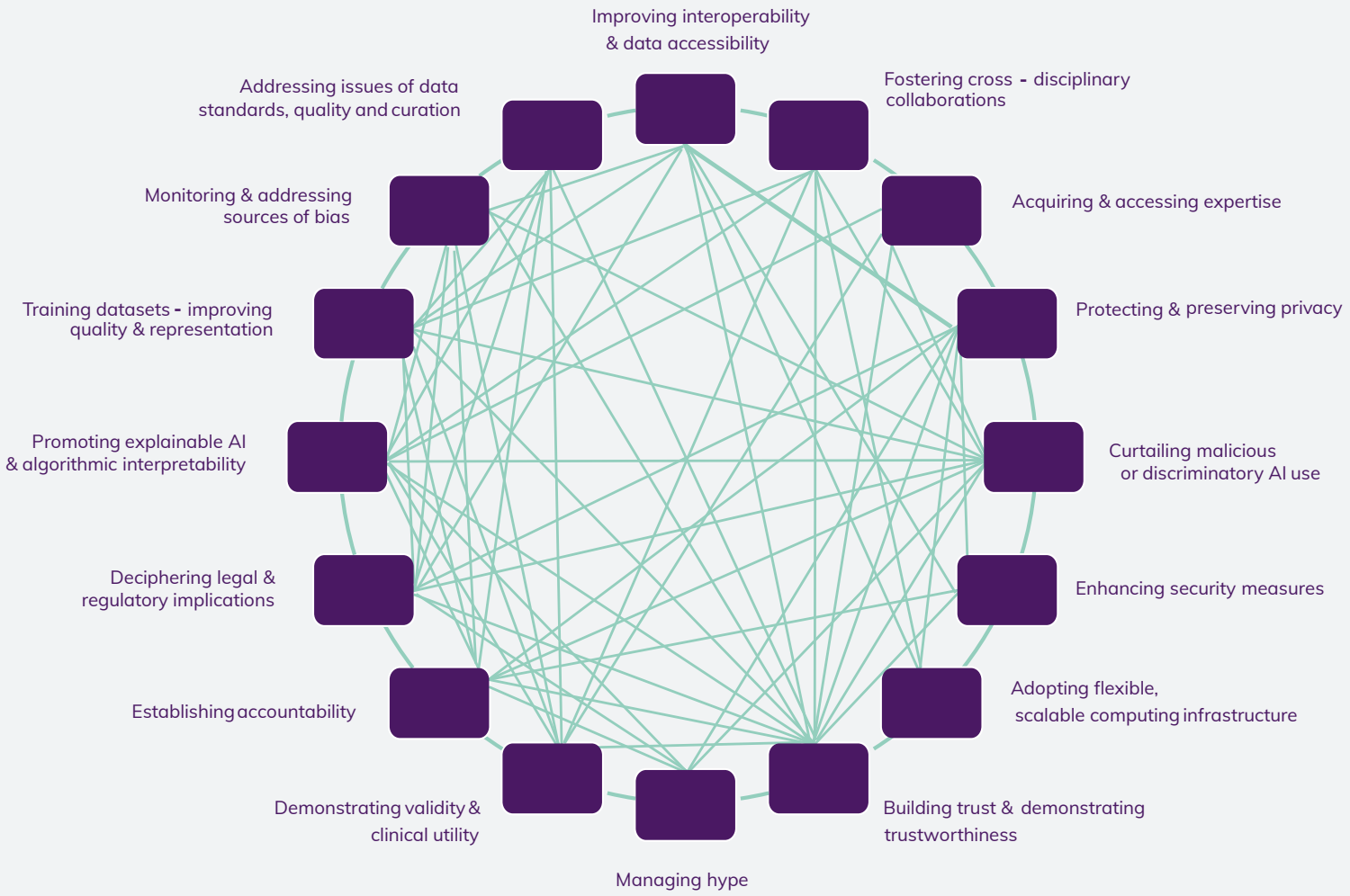
The effective use of machine learning in genomic medicine creates a complex web of interrelated issues (figure 7).

The accuracy of a machine learning model is highly dependent on reliable, high quality training data. Fundamental to curating better training datasets are standards for interoperability and systems for better data accessibility, which in turn challenges efforts around privacy, security, and safeguarding. Connected to nearly all other considerations is the importance of building patients' and healthcare professionals' trust in AI systems. Given the extent of interconnectivity, these issues are unlikely to be addressed effectively in isolation.

Some of these challenges are not new to genomic medicine or even to data-driven healthcare. Equally many of the AI related concerns, including bias and algorithmic transparency, are not unique to its application within genomics. However, the marriage of genomics and machine learning deepens many of these pre-existing challenges, heightening the imperative to address them. In essence, there is an opportunity, indeed a necessity, to learn from different disciplines, even where sector-specific strategies are necessary.

AI holds great promise for genomic medicine, but it is worth noting that there are other clinical disciplines and aspects of healthcare delivery more likely to see greater impact from AI, at least for the foreseeable future. These include medical imaging, predicting acute illness[124], and handling repetitive administrative tasks[125].

Figure 7: Harnessing AI for genomic medicine – priority areas for action. Interconnectivity between the various issues that will need to be addressed in order to advance the benefits of AI for genomic medicine.



Priorities for policy

The initial priorities for creating an environment that facilitates the application of AI in genomic medicine and realises its near-term value are to:

Establish the right conditions for facilitating AI in genomic medicine, which includes improved digital infrastructure, data acquisition and management, access to specific technical skills, and cross-disciplinary collaborations

Prioritise the development of constructive AI tools that address well-defined, focused, and clinically relevant problems in genomics analysis and clinical genomics service delivery

Mitigate against AI bias in genomics by promoting a workforce and research environment that is representative of societal diversity, as well as monitoring and addressing sources of bias within training datasets

Facilitate research efforts to apply machine learning (including deep learning) to well-curated, high-quality genomics and biomedical datasets, and bridging the gap between knowledge discovery and clinical practice

Support efforts driven by the clinical genomics community to benchmark, review, and determine the most effective use and integration of emerging new algorithms for clinical genome analysis

Establish sector-specific strategies to address the complex challenges and limitations of AI in genomic medicine and research and place as much emphasis on addressing these challenges as on exploring the latest developments in AI research. These include:

- **Clarity on how different machine learning algorithms are applied** at various stages of clinical genome analysis
- **Strategies to advance the beneficial use of AI** applied to genomic and health data while protecting against potential harms, including understanding the potential trade-offs between interpretability and accuracy
- **Clarity about how highly adaptive and dynamic algorithms should be regulated.** Specifically, where algorithms are highly adaptive (continuously responding to new data), opaque (their inner workings are unclear) and used for high risk decisions, there is an urgent need for regulatory clarity on the requirements for transparency [126] and the implications for product certification and for liability

Establish the clinical governance arrangements for the use of specific AI applications in the practice of clinical genomics

The way ahead

For health services facing increasing demands, the most useful applications for AI in genomic medicine lie in reducing analysis times, focusing the search for disease causing variants, and improving the accuracy of *in silico* predictions made along the genomic data pipeline.

We are still far from a complete understanding of the relationship between genomic variation and many known diseases, and this is where AI techniques applied to massive multi-modal and multi-'omics datasets could provide valuable insights. Currently this form of analysis is the preserve of research and the ability to quickly, meaningfully, and routinely make sense of these massive datasets to inform patient care in the clinic is still some way off.

The application of AI, when experts in health, genomics, regulation and ethics are working in concert, presents a significant opportunity to unravel the complexity encoded in our genomes for health benefit.

Despite the anticipation and excitement, AI on its own will not advance genomic medicine. Workforce training, considered implementation, patient and public engagement, robust ethical appraisals, and other types of technologies, for example non-AI statistical techniques, will continue to be crucial.

Considering the significant financial investment[14] and policy work already underway to deliver AI in health and care, it is vital to address the above priorities early as part of wider efforts to accelerate the adoption of proven AI technologies[127].

AI is yet to transform clinical genomics, but offers considerable potential which will only be achieved with effective policy prioritisation and action to achieve these priorities.

Acknowledgements

The individuals listed below kindly agreed to share their experience and insight into the subjects covered in this report.

Responsibility for the content of the report rests with the author.

- Prof Atul Butte – MD, PhD. Priscilla Chan and Mark Zuckerberg Distinguished Professor of Pediatrics, Bioengineering and Therapeutic Sciences, and Epidemiology and Biostatistics at UCSF. Director, Bakar Computational Health Sciences Institute, UCSF. Chief Data Scientist, University of California Health System (UC Health).
- Dr Lizzie Dorfman – Chief of Staff for Research & Innovations, Google Health
- Dekel Gelbman – CEO FDNA
- Carla Leibowitz – Healthcare partnerships, NVIDIA (*has since changed position)
- Dr Riccardo Miotto – Assistant Professor and Data Scientist, Icahn School of Medicine at Mount Sinai, New York
- Moran Snir – Co-founder and CEO, Clear Genetics (acquired by Invitae Corp.)
- Prof Christopher Yau - Professor of Artificial Intelligence, Division of Informatics, Imaging and Data Sciences, Faculty of Biology Medicine and Health, The University of Manchester, and Turing Fellow, The Alan Turing Institute, London, UK
- Dr Wenyu Zhou – Bioinformatics Scientist, formally Postdoctoral Fellow at Snyder Lab, Stanford University

Appendices

Table 1: Definitions – Artificial intelligence, machine learning, and deep learning

Artificial intelligence (AI)	■ AI is the development and use of computing systems concerned with making machines work in an intelligent way ■ AI can be categorised as 'narrow' or 'general' whereby • **Narrow AI** focuses on performing one specific task e.g. playing chess or filtering spam emails. Whilst 'narrow' by name, the individual uses of the technology can be broad ranging and sophisticated. Narrow AI is the focus of most current AI developments within healthcare and other industries. • **General AI** refers to the concept of a sentient machine and one which can perform different types of intelligent tasks and 'human' reasoning. Many consider the possibility of 'general' AI to be decades away, and others consider it unfeasible altogether.
Machine Learning (ML)	■ Machine learning is a subset of artificial intelligence that uses algorithms which learn from data rather than being 'explicitly programmed'. The general objective of ML algorithms is to perform predictions, classifications, estimations or similar tasks. ■ Statistical models and computer science are core to machine learning. Advances in powerful computing systems have helped to propel the field in recent years. ■ Training data – are the datasets used to develop and improve the performance of ML algorithms. ■ Learning approaches can be classified as supervised, unsupervised, or semi-supervised where • Supervised learning uses 'labelled' training data and prior knowledge to learn a function that can then be used to make predictions about new data-points • Unsupervised learning is concerned with uncovering structure within a dataset without prior knowledge of how the data are organised • Semi-supervised learning is a hybrid of the above two

Table 1: Definitions – Artificial intelligence, machine learning, and deep learning

Deep learning (DL)	■ Deep learning is a branch of machine learning which uses artificial neural networks (loosely inspired by the biology of the brain) to learn from large datasets. ■ A deep neural network consists of digitised inputs, e.g. electrocardiogram images, which are processed through multiple layers of the neural network that progressively detect features within the data and then provide an output, e.g. prediction of cardiac arrhythmia. ■ 'Deep' is a reference to the numerous layers of the neural networks that enable learning.

Table 2: Companies engaged in some form of AI and genomics activity (non-exhaustive)

Company	Broad AI use	General reported use of AI
Ardigen	Knowledge discovery	Various discovery platforms incorporating AI (including biomarkers, microbiome analysis)
BenevolentAI	Drug discovery / development	Drug discovery platform that draws on mined and inferred biomedical data
BostonGene	Therapeutic decision support	Software for personalised therapy decision making in cancer
Cambridge Cancer Genomics	Decision support	AI platform intended to support oncologists to provide personalised cancer treatment
Clear Genetics	Counselling / reporting	AI chatbot for conversing with patients about genetics
Congenica	Variant annotation / prioritisation	The Sapientia™ decision support tool incorporates third-party machine learning algorithms for genomic variant annotation and prioritisation
Deep Genomics	Drug discovery	AI driven drug discovery platform
Desktop Genetics	Gene editing	Deskgen AI platform uses machine learning to optimise CRISPR gene editing sequence libraries
Fabric Genomics	Variant interpretation	AI-driven interpretation platform for genomic tests; incorporates automated phenotype based interpretation

Table 2: Examples of companies engaged in some form of AI and genomics activity

Company	Broad AI use	General reported use of AI
FDNA	Phenotyping	Deep learning driven facial analysis software for rare disease phenotyping and phenotype driven variant prioritisation
Freenome	Cancer early detection / treatment	AI-Genomics for early cancer detection & treatment
Google (Brain)	Variant calling	DeepVariant tool for detecting genomic variants
Healx	Drug repurposing	AI for drug matching in rare diseases
IBM	Literature mining	Watson for Genomics uses AI to extract data from peer-reviewed literature
Lantern Pharma	Drug discovery / repurposing	AI driven platform for precision oncology therapeutics
Literome	Literature mining	Automated curation system to extract genomic knowledge from PubMed
OptraHealth	Counselling / reporting	AI chatbot and digital assistant for conversing with patients about genetics
Perthera	Therapeutic decision support	Combines AI analysis into cancer therapeutic matching
Philips	Knowledge discovery	'IntelliSpace Genomics' platform combines customizable pipelines with deep learning for new insights

Table 2: Examples of companies engaged in some form of AI and genomics activity (continued)

Company	Broad AI use	General reported use of AI
Sequana Health	Genome editing	AI-designed CRISPR genome editing systems
SOPHiA Genetics	Variant interpretation	The Alamut Genova genome browser integrates several missense variant pathogenicity prediction tools and algorithms
Verge Genomics	Drug discovery	Machine learning driven drug discovery company focused on neurodegenerative diseases
WuXi NextCODE	Knowledge discovery	Domain-specific AI algorithms for biological understanding

References

1. Market Research Report: Genomics. Fortune Business Insights. 2018. www.fortunebusinessinsights.com/industry-reports/genomics-market-100941
2. Collier M, Fu R, Yin L. Artificial Intelligence (AI): Healthcare's new nervous system. accenture. 2019. www.accenture.com/fi-en/insight-artificial-intelligence-healthcare
3. Acumen Research and Consulting Artificial Intelligence in Healthcare Market Size Worth US$ 8 Bn by 2026. GlobeNewswire. 2019. www.globenewswire.com/news-release/2019/07/08/1879644/0/en/Artificial-Intelligence-in-Healthcare-Market-Size-Worth-US-8-Bn-by-2026.html
4. House of Lords Select Committee on Artificial Intelligence. AI in the UK: ready, willing and able?. 2018. publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf
5. House of Commons Science and Technology Committee. Genomics and genome editing in the NHS - Third report of Session 2017-19. 2018. publications.parliament.uk/pa/cm201719/cmselect/cmsctech/349/349.pdf
6. Topol EJ. The Topol Review: Preparing the healthcare workforce to deliver the digital future. NHS Health Education England. 2019. topol.hee.nhs.uk/
7. Department of Health and Social Care Health minister: NHS must lead the world in genomic healthcare. 2019. www.gov.uk/government/news/health-minister-nhs-must-lead-the-world-in-genomic-healthcare
8. NIH US National Library of Medicine What are the benefits and risks of direct-to-consumer testing? . 2019. ghr.nlm.nih.gov/primer/dtcgenetictesting/dtcrisksbenefits
9. Tandy-Connor S, Guiltinan J, Krempely K, *et al*. False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. Genet Med. 2018. 20(12): pp. 1515-1521.
10. Weedon MN, Jackson L, Harrison J W. *et al*. Assessing the analytical validity of SNP-chips for detecting very rare pathogenic variants: implications for direct-to-consumer genetic testing. bioRxiv. 2019. p. 10.1101/696799.
11. Shendure J, Findlay GM, Snyder MW. Genomic Medicine-Progress, Pitfalls, and Promise. Cell. 2019. 177(1): pp. 45-57.
12. Davies SC. Generation genome: Annual report of the chief medical officer 2016. 2017. assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/631043/CMO_annual_report_generation_genome.pdf.
13. Davies SC. Health 2040 - Better health within reach. Annual report of the chief medical officer 2018. 2018. assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/767549/Annual_report_of_the_Chief_Medical_Officer_2018_-_health_2040_-_better_health_within_reach.pdf

14. Department of Health and Social Care Health Secretary announces £250 million investment in artificial intelligence. 2019. www.gov.uk/government/news/health-secretary-announces-250-million-investment-in-artificial-intelligence

15. Ream M, Woods T, Joshi I, *et al*. Accelerating Artificial Intelligence in health and care: results from a state of the nation survey. Academic Health Sciences Network. 2018. http://ai.ahsnnetwork.com/about/aireport

16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019. 25(1): pp. 44-56.

17. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med. 2018. 378(11): pp. 981-983.

18. Ross J, Webb C, Rahman F. Artificial intelligence in healthcare. Academy of Royal Medical Colleges. 2019. www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf

19. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018. 2(10): pp. 719-731.

20. Bioethics briefing note: Artificial intelligence (AI) in healthcare and research. Nuffield Council on Bioethics. 2018. http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf

21. Fenech M, Strukelj N, Buston O. Future Advocacy: Ethical, social and political challenges of artificial intelligence in health. Wellcome Trust. 2018. wellcome.ac.uk/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf

22. De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018. 24(9): pp. 1342-1350.

23. Hosny A, Parmar C, Quackenbush J, *et al*. Artificial intelligence in radiology. Nat Rev Cancer. 2018. 18(8): pp. 500-510.

24. Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. Nat Med. 2019. 25(1): pp. 14-15.

25. Ordish J, Murfet H, Hall A. Algorithms as medical devices. PHG Foundation. 2019. www.phgfoundation.org/documents/algorithms-as-medical-devices.pdf

26. Artificial Intelligence and Machine Learning in Software as a Medical Device. U.S. Food & Drug Administration. 2019. www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

27. Department of Health and Social Care: Code of conduct for data-driven health and care technology. 2018. www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology

28. Machine learning: the power and promise of computers that learn by example. The Royal Society. 2017. royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

29. Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. Nat Med. 2019. 25(1): pp. 24-29.

30. Zou J, Huss, M, Abid, A, *et al*. A primer on deep learning in genomics. Nat Genet. 2019. 51(1): pp. 12-18.

31. Stanford Medicine 2017 Health Trends Report: Harnessing the power of data in health. Stanford University. 2017. med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf

32. Miotto R, Wang F, Wang S, *et al*. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018. 19(6): pp. 1236-1246.

33. Stephens ZD, Lee SY, Faghri F. *et al*. Big Data: Astronomical or Genomical? PLoS Biol. 2015. 13(7): p. e1002195.

34. Ching T, Himmelstein DS, Beaulieu-Jones, BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018. 15(141): p. 20170387.

35. Beaulieu-Jones, BK, Greene CS. Pooled Resource Open-Access, A. L. S. Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform. 2016. 64: pp. 168-178.

36. Basile AO, Ritchie MD. Informatics and machine learning to define the phenotype. Expert Rev Mol Diagn. 2018. 18(3): pp. 219-226.

37. Ferry Q, Steinberg J, Webber C. *et al*. Diagnostically relevant facial gestalt information from ordinary photos. Elife. 2014. 3: p. e02020.

38. Gurovich Y, Hanani Y, Bar O , *et al*. DeepGestalt - Identifying Rare Genetic Syndromes Using Deep Learning. arXiv. 2018. p. 1801.07637

39. Mishima H, Suzuki H, Doi M, *et al*. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. J Hum Genet. 2019. 64(8): pp. 789-794.

40. Gurovich Y, Hanani Y, Bar O, *et al*. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019. 25(1): pp. 60-64.

41. Hsieh TC, Mensah MA, Pantel JT, *et al*. PEDIA: prioritization of exome data by image analysis. Genet Med. 2019. pp. doi:10.1038/s41436-019-0566-2.

42. Bera K, Schalper KA, Rimm DL, *et al*. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019. 16(11): pp. 703-715.

43. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal. 2016. 33: pp. 170-175.

44. Mobadersany P, Yousefi S, Amgad M, *et al*. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 2018. 115(13): pp. E2970-E2979.

45. Yu KH, Berry GJ, Rubin DL, *et al*. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. Cell Syst. 2017. 5(6): pp. 620-627 e3.

46. Zhang JX, Fang JZ, Duan W, *et al*. Predicting DNA hybridization kinetics from sequence. Nat Chem. 2018. 10(1): pp. 91-98.

47. Teng H, Cao MD, Hall MB, *et al*. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. Gigascience. 2018. 7(5): p. doi: 10.1093/gigascience/giy037.

48. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019. 20(1): p. 129.

49. Boza, V, Brejova, B, Vinar, T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One. 2017. 12(6): p. e0178751.

50. Poplin, R, Chang, P. C, Alexander, D, *et al*. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018. 36(10): pp. 983-987.

51. Kolesnikov A, Chang P-C, ChinJ, *et al*. Highly Accurate SNP and Indel Calling on PacBio CCS with DeepVariant. Inside DNAnexus. 2019. blog.dnanexus.com/2019-01-14-highly-accurate-snp-indel-calling-pacbio-ccs-deepvariant

52. Yun T, McLean C, Chang P-C, *et al*. Improved non-human variant calling using species-specific DeepVariant models. google.github.io. 2018. google.github.io/deepvariant/posts/2018-12-05-improved-non-human-variant-calling-using-species-specific-deepvariant-models

53. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018. 16: pp. 15-24.

54. Ainscough BJ, Barnell EK, Ronning P, *et al*. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nat Genet. 2018. 50(12): pp. 1735-1743.

55. Sahraeian SME, Liu R, Lau B, *et al*. Deep convolutional neural networks for accurate somatic mutation detection. Nat Commun. 2019. 10(1): p. 1041.

56. Pounraja VK Jayakar G, Jensen M, *et al*. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res. 2019. 29(7): pp. 1134-1143.

57. Zarrei M, MacDonald JR, Merico D, *et al*. A copy number variation map of the human genome. Nat Rev Genet. 2015. 16(3): pp. 172-83.

58. Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? Genome Biol. 2013. 14(5): p. 205.

59. Leung MKK, Delong A, Alipanahi B, *et al*. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. Proceedings of the IEEE. 2016. 104(1): pp. 176-97.

60. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. Nat Methods. 2010. 7(4): pp. 248-9.

61. Schwarz JM, Rodelsperger C, Schuelke M, *et al*. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010. 7(8): pp. 575-6.

62. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014. 46(3): pp. 310-5.

63. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017. 18(1): p. 225.

64. Mahmood K, Jung CH, Philip G, *et al*. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics. 2017. 11(1): p. 10.

65. AlQuraishi M. End-to-End Differentiable Learning of Protein Structure. Cell Syst. 2019. 8(4): pp. 292-301 e3.

66. AlphaFold: Using AI for scientific discovery. DeepMind. 2019. deepmind.com/blog/alphafold

67. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015. 12(10): pp. 931-4.

68. McKinney BA, Reif DM, Ritchie MD, *et al*. Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics. 2006. 5(2): pp. 77-88.

69. Zhang Q, Shen Z, Huang DS. Modeling *in-vivo* protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network. Sci Rep. 2019. 9(1): p. 8484.

70. Hashemifar S, Neyshabur B, Khan AA, *et al*. Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics. 2018. 34(17): pp. i802-i810.

71. Zhang Z, Pan Z, Ying Y, *et al*. Deep-learning augmented RNA-seq analysis of transcript splicing. Nat Methods. 2019. 16(4): pp. 307-310.

72. Pirooznia M, Yang JY, Yang MQ, *et al*. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008. 9 Suppl 1: p. S13.

73. Holder LB, Haque MM, Skinner MK. Machine learning for epigenetics and future medical applications. Epigenetics. 2017. 12(7): pp. 505-514.

74. Sundaram L, Gao H, Padigepati SR, *et al*. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018. 50(8): pp. 1161-1170.

75. Jaganathan K, Kyriazopoulou-Panagiotopoulou S, McRae JF, *et al*. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019. 176(3): pp. 535-548 e24.

76. Zhou J, Theesfeld CL, Yao K, *et al*. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018. 50(8): pp. 1171-1179.

77. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018. 19(5): pp. 299-310.

78. Li J, Li X, Zhang S, *et al*. Gene-Environment Interaction in the Era of Precision Medicine. Cell. 2019. 177(1): pp. 38-44.

79. Zitnik M, Nguyen F, Wang B, *et al*. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. Inf Fusion. 2019. 50: pp. 71-91.

80. Camacho DM, Collins KM, Powers RK, *et al*. Next-Generation Machine Learning for Biological Networks. Cell. 2018. 173(7): pp. 1581-1592.

81. Langelier C, Kalantar KL, Moazed F, *et al*. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci U S A. 2018. 115(52): pp. E12353-E12362.

82. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. 2018. 34(4): pp. 301-312.

83. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018. 19(9): pp. 581-590.

84. Ho DSW, Schierding W, Wake M, *et al*. Machine Learning SNP Based Prediction for Precision Medicine. Front Genet. 2019. 10: p. 267.

85. Qu K, Guo F, Liu X, *et al*. Application of Machine Learning in Microbiology. Front Microbiol. 2019. 10: p. 827.

86. Zhou YH, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet. 2019. 10: p. 579.

87. Way GP, Greene CS. Bayesian deep learning for single-cell analysis. Nat Methods. 2018. 15(12): pp. 1009-1010.

88. Auslander N, Wolf YI, Koonin EV. In silico learning of tumor evolution through mutational time series. Proc Natl Acad Sci U S A. 2019. 116(19): pp. 9501-9510.

89. Caravagna G, Giarratano Y, Ramazzotti D, *et al*. Detecting repeated cancer evolution from multi-region tumor sequencing data. Nat Methods. 2018. 15(9): pp. 707-714.

90. Mourikis TP, Benedetti L, Foxall E, *et al*. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun. 2019. 10(1): p. 3101.

91. Clark MM, Hildreth A, Batalov S, *et al*. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med. 2019. 11(489).

92. Deignan JL, Chung WK, Kearney HM, *et al*. Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2019. 21(6): pp. 1267-1270.

93. Costain G, Jobling R, Walker S, *et al*. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. Eur J Hum Genet. 2018. 26(5): pp. 740-744.

94. Vamathevan J, Clark D, Czodrowski P, *et al*. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019. 18(6): pp. 463-477.

95. Zafeiris D, Rutella S, Ball GR. An Artificial Neural Network Integrated Pipeline for Biomarker Discovery Using Alzheimer's Disease as a Case Study. Comput Struct Biotechnol J. 2018. 16: pp. 77-87.

96. Ekins S, Puhl AC, Zorn KM, *et al*. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019. 18(5): pp. 435-441.

97. Madhukar NS, Elemento O. Bioinformatics Approaches to Predict Drug Responses from Genomic Sequencing. Methods Mol Biol. 2018. 1711: pp. 277-296.

98. McCartney M. AI in medicine must be rigorously tested. BMJ. 2018. 361: p. k1752.

99. Kim HK, Min S, Song M, *et al*. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat Biotechnol. 2018. 36(3): pp. 239-241.

100. Leenay RT, Aghazadeh A, Hiatt J, *et al*. Large dataset enables prediction of repair after CRISPR-Cas9 editing in primary T cells. Nat Biotechnol. 2019. 37(9): pp. 1034-7.

101. Shen MW, Arbab, M, Hsu JY, *et al*. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature. 2018. 563(7733): pp. 646-651.

102. Listgarten J, Weinstein M, Kleinstiver BP, *et al*. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomed Eng. 2018. 2(1): pp. 38-47.

103. Harvey H. Ready...set...AI — preparing NHS medical imaging data for the future. Towards Data Science - Medium.com. 2017. towardsdatascience.com/ready-set-ai-preparing-nhs-medical-imaging-data-for-the-future-8e85ed5a2824

104. Martin AR, Kanai M, Kamatani Y, *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019. 51(4): pp. 584-591.

105. Manrai AK, Funke BH, Rehm HL, *et al*. Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med. 2016. 375(7): pp. 655-65.

106. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019. 177(1): pp. 26-31.

107. Advancing our health: prevention in the 2020s. gov.uk. 2019. www.gov.uk/government/consultations/advancing-our-health-prevention-in-the-2020s/advancing-our-health-prevention-in-the-2020s-consultation-document

108. Hutson M. Artificial intelligence faces reproducibility crisis. Science. 2018. 359(6377): pp. 725-726.

109. Gundersen OE, Kjensmo S. State of the Art: Reproducibility in Artificial Intelligence. The Thirty-Second AAAI Conference on Artificial Intelligence. 2018. AAAI-18: pp. 1644-51.

110. NHS and social care data: off-shoring and the use of public cloud services guidance. NHS Digital. 2018. digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/nhs-and-social-care-data-off-shoring-and-the-use-of-public-cloud-services/nhs-and-social-care-data-off-shoring-and-the-use-of-public-cloud-services-guidance

111. Dove ES, Joly Y, Tasse AM, *et al*. Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genet. 2015. 23(10): pp. 1271-8.

112. Evidence standards framework for digital health technologies. National Institute of Health and Care Excellence. 2019. www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

113. User guide for evidence standards framework for digital health technologies. National Institute for Health and Care Excellence. 2019. www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/user-guide.pdf

114. Towards trustable machine learning. Nat Biomed Eng. 2018. 2(10): pp. 709-710.

115. Opening the black box of machine learning. Lancet Respir Med. 2018. 6(11): p. 801.

116. Caruna R, Yin L, Gehrke J, *et al*. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 15. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. pp. 1721-30.

117. Zech J R, Badgeley MA, Liu M, *et al*. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018. 15(11): p. e1002683.

118. Yu MK, Ma J, Fisher J, *et al*. Visible Machine Learning for Biomedicine. Cell. 2018. 173(7): pp. 1562-1565.

119. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019. 1: pp. 206-15.

120. The One-Way Mirror: Public attitudes to commercial access to health data. Report prepared for the Wellcome Trust. Ipsos Mori. 2016. www.ipsos.com/sites/default/files/publication/5200-03/sri-wellcome-trust-commercial-access-to-health-data.pdf

121. Harwich E, Laycock K. Thinking on its own: AI in the NHS. Reform. 2018. reform.uk/research/making-nhs-data-work-everyone

122. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019. 10(1): p. 3069.

123. Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis. The Royal Society. 2019. royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf

124. Tomasev N, Glorot X, Rae JW, *et al*. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019. 572(7767): pp. 116-119.

125. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019. 6(2): pp. 94-98.

126. Hall A, Ordish J. Black Box Medicine and Transparency. PHG Foundation. 2020.

127. Joshi I, Morley J (eds). Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care. nhsx.nhs.uk. 2019. www.nhsx.nhs.uk/assets/NHSX_AI_report.pdf