

Review Article

The impact of artificial intelligence on biomarker discovery

Hira Javaid¹, Constantin Cezar Petrescu¹, Lisa J. Schmunk¹, Jack M. Monahan², Paul O'Reilly³, Manik Garg⁴, Leona McGirr¹, Mahmoud T. Khasawneh⁵, Mustafa Al Lail⁵, Deepak Ganta⁵, Thomas M. Stubbs¹, Benjamin B. Sun⁶, Dimitrios Vitsios⁴ and Daniel E. Martin-Herranz¹

¹Hurdle.bio / Chronomics Ltd., London, UK; ²Biomodal Ltd, The Trinity Building, Chesterford Business Park, Cambridge, UK; ³Sonrai Analytics Ltd, Queens Road, Belfast, Northern Ireland, UK; ⁴Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; ⁵Texas A&M International University, Laredo, Texas, U.S.A.; ⁶Bristol Myers Squibb, New York, U.S.A.

Correspondence: Hira Javaid (hira.javaid@hurdle.bio) and Daniel E. Martin-Herranz (daniel@hurdle.bio)



Artificial intelligence (AI) is transforming many fields, including healthcare and medicine. In biomarker discovery, AI algorithms have had a profound impact, thanks to their ability to derive insights from complex high-dimensional datasets and integrate multi-modal datatypes (such as omics, electronic health records, imaging or sensor and wearable data). However, despite the proliferation of AI-powered biomarkers, significant hurdles still remain in translating them to the clinic and driving adoption, including lack of population diversity, difficulties accessing harmonised data, costly and time-consuming clinical studies, evolving AI regulatory frameworks and absence of scalable diagnostic infrastructure. Here, we provide an overview of the AI toolkit available for biomarker discovery, and we discuss exciting examples of AI-powered biomarkers across therapeutic areas. Finally, we address the challenges ahead of us to ensure that these technologies reach patients and users globally and unlock a new era of fast innovation for precision medicine.

Introduction

Artificial intelligence (AI) can be broadly defined as ‘any technique that enables computers to mimic human behaviour and reproduce or excel over human decision-making to solve complex tasks independently or with minimal human intervention’ [1]. AI has revolutionised many domains that had traditionally been difficult to tackle computationally. More recently, transformer-based large language models (LLMs) have made newspaper headlines with their abilities to generalise across many natural language and computer vision tasks [2,3]. Consequently, the power of LLMs has triggered an AI race across technology companies.

AI-powered solutions are paving the way towards a significant impact of AI in healthcare. Some examples include protein structure prediction for drug discovery [4], applications in personalised medicine [5] and medicine-specific foundational models [6] (which can help to automate and increase efficiency in clinical workflows). With the rapid increase in large multi-omics and multi-modal healthcare datasets and deeply phenotyped biobanks, AI algorithms have become a cornerstone of engaging with big data to discover new biomarkers [7–9]. This review uses a broad definition of AI, including both classical machine learning (ML) and more recent deep learning (DL) algorithms. We discuss how researchers have used AI in the context of biomarker discovery in humans and provide an industry view on the areas where it has had more impact, while discussing some of the challenges to translate these biomarkers into everyday clinical practices.

Biomarker nomenclature and classification

A biomarker is defined as ‘a characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention, including therapeutic interventions’ [10,11]. The FDA-NIH Biomarkers, EndpointS and other Tools (BEST) biomarker classification defines seven categories of biomarkers [10], which we have summarised below [12]:

1. susceptibility/risk biomarkers, which indicate a person’s risk for developing a specific disease;
2. diagnostic biomarkers, which detect the presence of a disease or disease subtype;
3. monitoring biomarkers, which assess the status of a disease;
4. prognostic biomarkers, which give the likelihood of a clinical event;

Received: 20 December 2024

Revised: 30 July 2025

Accepted: 30 July 2025

Version of Record

Published: 9 September 2025

5. predictive biomarkers, which identify individuals more likely to have a favourable response to treatment;
6. pharmacodynamic/response biomarkers, which show that a response has occurred due to a treatment; and
7. safety biomarkers, which indicate the likelihood of or the presence of toxicity due to a medical product.

These biomarker categories may be useful at different stages of the disease journey ([Figure 1](#)). Furthermore, companion diagnostics are medical devices that are essential for the safe and effective use of a corresponding drug [13]. This includes predictive biomarkers of drug response (used to stratify patients into responders and non-responders or to identify individuals at increased risk of side effects) and safety biomarkers (to monitor a safe response to treatment and adjust dose accordingly) [13].

The toolbox for AI-powered biomarker discovery

Datatypes used in biomarker discovery

Traditional biomarkers can be expressed as a single number of a single datatype, such as HbA1c (glycated haemoglobin) or cholesterol. AI algorithms become useful in the presence of highly dimensional datasets, which have surged in the last decade [14]. There are distinct types of biological data, which can include different modalities:

- Omics data: Molecular data such as genomics, epigenomics, transcriptomics, proteomics, glycomics, metabolomics or microbiomics. When more than one of these omics datatypes is leveraged, this is generally referred to as multi-omics [15]. This has gained particular attention in the context of single-cell approaches, where multiple omics layers are profiled at the same time from a single cell [16,17].
- Electronic health records (EHRs) and questionnaire-based data: They can contain structured data (such as lab test results, demographic and lifestyle/environmental data) or unstructured data (such as text from clinical notes).
- Imaging data: Such as radiology data [like computed tomography (CT) scans], histological images (digital pathology), electrocardiograms (ECG), electroencephalography (EEG), magnetic resonance imaging (MRI) or positron emission tomography (PET) scans.
- Sensor data: Such as that derived from wearable devices.

AI algorithms can be used for batch correction, data harmonisation and dimensionality reduction within one datatype (e.g. when dealing with high-dimensional single-cell transcriptomic data) [18]; to integrate different datatypes of the same modality (e.g. multi-omics approaches that map proteomics signatures to epigenomic data) [19]; or to combine different data modalities, which is referred to as multi-modal (e.g. combining genetics and imaging data to discover new disease subtypes) [20].

Multi-modal data sources

Biobanks now hold vast amounts of multi-modal data that can be accessed for biomarker discovery ([Table 1](#)). Disease-agnostic biobanks are set up with the aim of surveying the health of a population over time. They often follow a prospective cohort design and recruit participants to perform a series of baseline tests and store tissue and data for future research use. This cross-sectional data can be enhanced by regular follow-ups with questionnaires, imaging and re-sampling, as well as linking to national healthcare records. While especially omics data generation has depended on collaborations between pharmaceutical industry, governments and private research funders (e.g. genomics and proteomics data in the UK Biobank [34]), the data are now available for general use.

Beyond disease-agnostic national or private biobanks, data valuable for biomarker discovery can be found in disease-focused biobanks and cohort studies, commercial research datasets (e.g. clinical trials), as well as real-world data (e.g. from EHRs, insurance records and consumer testing companies). Longitudinal datasets with repeat sampling of multi-modal datatypes provide opportunities for direct assessment of health progression [35]. Family studies (e.g. trios, parent–child and multigenerational) have not been widely used in biomarker discovery yet [36]. Biases introduced by minimal phenotyping and the self-reporting of certain phenotypes and pathologies may lower specificity for biomarker discovery [37].

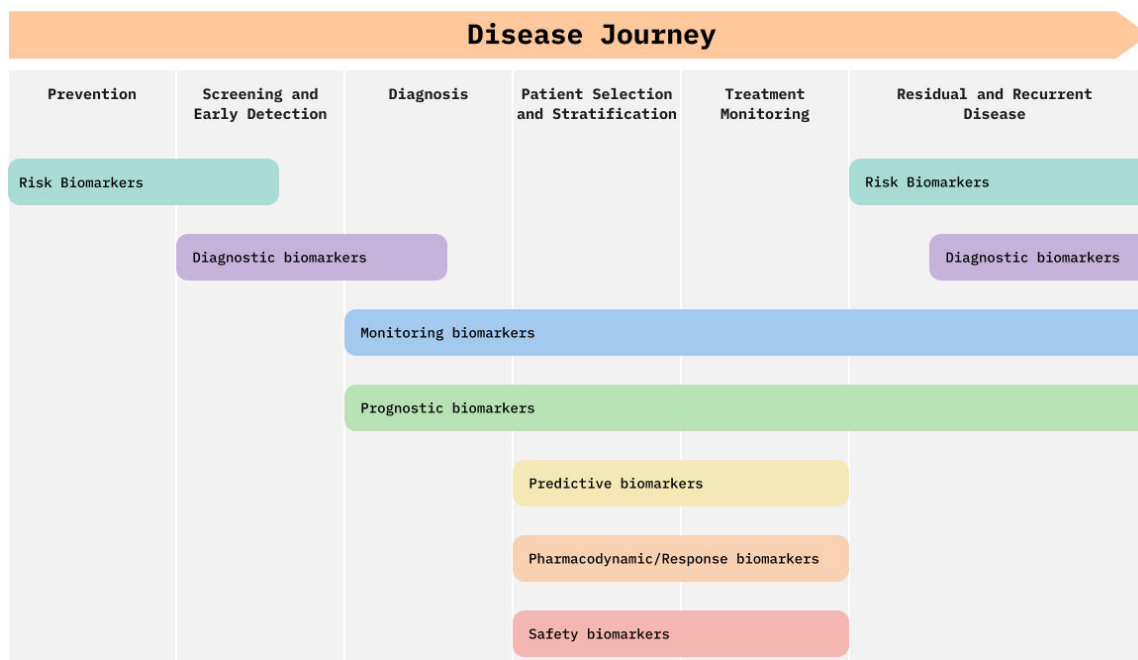


Figure 1: Biomarker categories along the disease journey.

The seven biomarker categories come into play at different stages of the patient disease journey, from prevention to diagnosis, treatment and recurrence.

Given the large-scale missingness, especially in omics and EHR data, imputation and synthetic datasets can be beneficial for biomarker discovery. Imputation of missing values across all datatypes is non-trivial, with approaches for omics and EHR data previously reviewed [38,39]. For example, genetic information can be leveraged to improve imputation of missing traits in EHRs [40], while DL-based methods can enable biomarker feature imputation [41]. Recently, generative AI approaches [such as generative adversarial networks (GANs) or transformers] have been leveraged to create synthetic datasets across EHRs (e.g. EHRSafe) [42] or epigenetic data (e.g. MethylGPT) [43].

Access to individual level data may be open access or restricted (see Table 1) and is increasingly provided in Trusted Research Environments (TREs). While this facilitates dataset maintenance and within-dataset harmonisation, it precludes training across multiple datasets and biobanks held in different systems. Federated learning approaches may be considered in these cases if data cannot be combined locally [44,45].

The data required to train an AI biomarker model will depend on the biomarker type (Figure 1). Most biomarker discovery models to date have used supervised learning, where the outcome variable needs to be defined. For example, in order to develop a susceptibility/risk biomarker, data with disease outcomes in a large population are essential (e.g. time-to-disease models), which needs prospective studies with significant follow-up data. For predictive and pharmacodynamic/response biomarkers, medication records will be needed. Sometimes the outcome variable is not clearly defined, such as in the case of new phenotypes (e.g. chronic inflammation, a hallmark of ageing) or when dealing with different disease billing codes. Importantly, biomarkers can themselves help to identify disease subtypes and fine-tune disease classification, which can lead to improvements in biomarker discovery [46].

A major consideration for biomarker discovery thus lies in the selection of the appropriate training and testing datasets. Biobank recruitment and related large-scale omics data generation still favours European samples [47,48], precluding widespread translatability of the resulting biomarkers across diverse populations and generating bias. Approaches to address this have been previously reviewed [49]. Basic characteristics like age and sex have to be taken into account and, where indicated, models should be adjusted for these. In certain cases, it may be appropriate to train distinct models for males and females when biological evidence suggests the presence of differing underlying mechanisms. Importantly, demographic and socioeconomic characteristics are often not reported and thus ignored [50], despite their extensive influence on health outcomes (e.g. social deprivation index [51]). Furthermore, cohort-to-cohort

Table 1: Examples of global multi-modal datasets that can be leveraged for AI-powered biomarker discovery.

Region	Country	Biobank name	Type	Number of participants	Recruitment years	Age at recruitment	Data modalities	Access	Longitudinal sampling	Number of years of follow-up	References
Africa	Multiple African nations	H3Africa biorepositories	Biorepository	~100,000	2012–Present	NA	Omics	Application-based	No	NA	[21]
	China	China Kadoorie Biobank	Prospective	~512,000	2004–2008	30–79 years	Omics, EHRs	Application-based; genetic data restricted	Surveys only	~15 years	[22]
Asia	Japan	Tonoku Medical Megabank	Prospective (earthquake survivor regions)	~150,000	2013–2018	~3/4 20+ years, ~1/4 < 20 years	Omics, EHRs	Application-based for academia & industry	Yes, every 5 years	~10+ years	[23]
	Japan	Biobank Japan	Disease-focused	~270,000	2003–2008 & 2012–2017	Median age: ~62 years (2003–2008)	Omics, EHRs	Application-based for academia & industry	No	~10+ years	[24]
Europe	United Kingdom	UK Biobank	Prospective	~500,000	2006–2010	40–69 years	Omics, EHRs and imaging	Application-based for academia & industry	Yes	~15+ years	[25]
	Estonia	Estonian Biobank	Prospective	~200,000	2002–2015 & 2018–2024	18+ years	Omics, EHRs	Application-based for academia & industry	Surveys only	Up to 20+ years	[26]
Middle East	Finland	FinnGen	Disease-focused (2/5) + Prospective (3/5)	~500,000	2017–Present (legacy before 2017)	Median age: ~63 years	Omics, EHRs	Application-based for academia & industry	No	~7+ years	[27]
	Qatar	Qatar Biobank	Prospective	~45,000 (target 60,000)	2012–Present	18+ years	Omics, EHRs and imaging	Application-based for medical researchers	Yes	~10+ years	[28]
North America	USA	All of Us	Prospective	~1,000,000	2018–Present	18+ years	Omics, EHRs, sensor data	Restricted; application-based for academia	Yes	~5+ years	[29]
	Canada	CanPath	Prospective	~330,000	2008–Present	30–74 years	Omics, EHRs	Application-based for academia & industry	Yes	~12+ years	[30]
South America	USA	Million Veteran Program	Prospective	~1,000,000	2011–Present	Median age: ~66 years (>90% male)	Omics, EHRs	Restricted; Application-based for academia	No	~10+ years	[31]
	Brazil	Biobank Barretos Cancer Hospital	Disease-focused (cancer)	~50,000	2006–Present	NA	Omics, EHRs	NA	NA	NA	[32]
Mexico	Mexico	Mexico City prospective study	Prospective	~110,000	1998–2004	NA	Omics, EHRs (mortality records)	Application-based	Re-survey of ~10,000 in 2015 and 2019	~20+ years	[33]

differences can limit generalisability, which needs to be considered when integrating data from biobanks and real-world evidence. This may be due to different factors, including unaccounted biological variation, differences in technology (e.g. batch effects in lab-based assays) or sample/participant selection.

AI algorithms

AI methodologies for biomarker discovery can be broadly categorised into two primary approaches: classical machine learning (ML) and deep learning (DL) (Figure 2). If multiple datatype modalities need to be considered, a multimodal infrastructure is required to integrate them, which usually leverages meta-learner algorithms trained on the output of other models that process different datatypes [52]. Selecting a suitable AI approach becomes a critical decision that affects the success and reliability of biomarker discovery, and it depends on different factors, such as the data available, the biomarker type and project constraints (like funding, know-how and infrastructure available).

Classical ML includes algorithms that are generally more interpretable, require well-crafted features and make predictions based on structured data, such as omics data, traditional blood and/or urine-based biomarkers (such as HbA1c or cholesterol), structured health records and questionnaire-based data. Some of the most popular classical ML algorithms include:

- Regression models. Elastic net regression models have been particularly useful in biomarker discovery, since they are able to deal with highly dimensional and redundant datatypes, such as those found when using epigenetic (DNA methylation) data [19]. The strength of regression models comes from their efficiency and interpretability. However, regression models struggle to capture more complex non-linear patterns [53].
- Random forest (RF) models belong to the ensemble learning category. These models use many decision trees in parallel, each trained on a subset of the data. RF models can capture non-linear patterns and process high-dimensional data with less propensity to overfitting. For instance, an RF classifier can be trained on 3.5 million genetic variants to predict Mendelian genetic disorders with more accuracy than existing methods [54]. RF models are more computationally expensive, and they require additional work for interpretability (feature importance).
- Gradient boosting tree models represent another ensemble method, with the difference that these models have the trees trained in sequence, allowing each tree to learn from the errors of the previous one. Thus, such models are effective at capturing how features interact with each other and can be used as the meta-learner for multi-modal solutions [55]. For instance, the multi-omics biomarker discovery framework MILTON uses an established and highly efficient implementation of gradient boosting (XGBoost) within a bespoke ensemble framework to build prediction models across 3000 diseases [46]. Like RF, model interpretability is possible by leveraging tree-based feature importance algorithms (e.g. split gain importance) [56] or more complex ones such as SHAP (SHapley Additive exPlanations) [57], which can be used to identify the relevant biomarkers for the outcome [58].

Survival analysis models predict the time until an event occurs, such as time-to-disease or death (mortality). Survival analysis models are efficient, lightweight and interpretable algorithms, making them favourable for building risk/susceptibility or prognostic biomarkers from omics data. The Cox proportional hazards regression model is popular and widely used to study the associations between different biomarkers and disease outcomes [19,59].

On the other hand, DL algorithms use large artificial neural network architectures to model and interpret complex data, which can be large, unstructured and noisy [1], such as unstructured temporal health data, clinical notes, sensor or imaging data. Some of the most popular DL algorithms include:

- Convolutional neural networks (CNNs) are DL models that can process multi-dimensional data structures, such as two-dimensional images (height × width) with multiple channels (like red, green and blue) or three-dimensional data (such as volumetric medical imaging). They allow automatic identification of spatial hierarchies of features (unlike in the case of classical ML, which normally requires more feature engineering). These models are widely used to train biomarkers from imaging data, for example, for the early prediction of Alzheimer's disease (AD) from hippocampal MRI data [60]. The disadvantages of CNNs are that they primarily capture local dependencies within the data, they are computationally expensive, difficult to interpret and require large amounts of labelled training data.

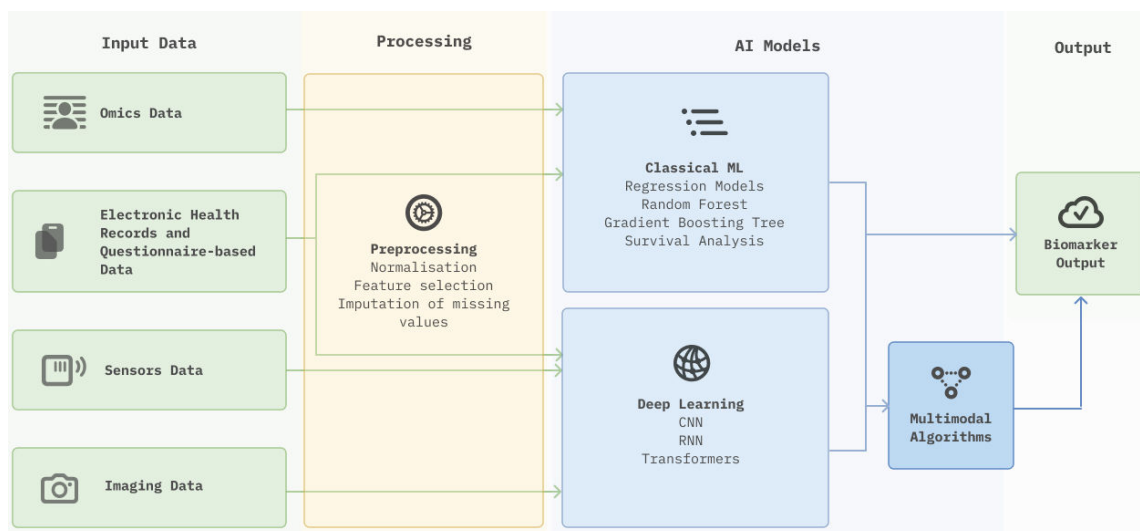


Figure 2: Flow diagram for AI-powered biomarker discovery.

This figure illustrates which datatypes are typically processed using different AI approaches: classical machine learning (ML) and deep learning (DL) algorithms. CNN, convolutional neural network; RNN, recurrent neural network.

- Recurrent neural networks (RNNs) are models used for processing sequential data. These models preserve a memory of inputs, which allows them to capture the temporal dependencies of the data (such as longitudinal EHR or sensor data). For example, RNNs have been leveraged to predict disease progression in the context of neurodegenerative disease [61]. However, a significant challenge for these models is their limited memory span, which makes it challenging for them to effectively utilise information from long-term dependencies in sequences.
- Transformers are models that use self-attention mechanisms to capture distant dependencies between data. LLMs like GPT (generative pre-trained transformers) [62] and BERT (bidirectional encoder representations from transformers) [63] are part of the transformer class and have recently increased in popularity. Transformer architectures have been used to train transcriptomic and epigenomic models that are able to both predict biological age and identify potential drug targets for age-related diseases [64]. One interesting recent application of transformers in the digital pathology domain has been in the development of foundation models. These foundation models are typically created using self-supervised learning techniques from extremely large cohorts of pathology images and have been shown to be the potential basis of state-of-the-art biomarker detection algorithms [65]. Furthermore, since transformers can use unsupervised learning, another application is processing unstructured textual data to construct a base knowledge of the context of information. For instance, data from PubMed, PMC and Wikipedia were processed to identify relevant information about rare diseases to aid their diagnosis and treatments [66]. Transformers face limitations such as high demands for large datasets and computational resources for training, along with ethical and privacy challenges.

Other AI algorithms frequently used to integrate multi-omics data include joint dimensionality reduction methods [67–69], generative adversarial networks (GANs) [20], graph neural networks [70,71] or graph transformer networks [72].

Examples of AI-powered biomarker discovery

While we have tried to cover many of the commonly used AI algorithms in biomarker discovery, the literature on the topic is immense, and a systematic analysis is beyond the scope of this review. Some recent examples of the innovative application of AI in biomarker research are highlighted in Table 2, which includes different data modalities, therapeutic areas and AI tools.

While this table is not all-inclusive, several key trends in AI-powered biomarker discovery can be observed. In general, single-omic and multi-omics biomarkers (such as ProteinScores, EpiSign, InflammAge, Nightingale plasma-based biomarkers, Duet evoC, GMW12 or MILTON) seem to

Table 2: Examples of innovative AI-based biomarkers in various therapeutic areas and the toolkit used for their discovery.

Biomarker or framework name	Therapeutic area(s)	Datatype(s)	Biomarker categories	AI algorithm(s)	Regulatory status	Highlights	References
PRSmix	Many therapeutic areas	Genomics (germline)	Risk	Elastic net	Research use only	Improve accuracy of polygenic risk scores (PRS) across complex diseases and populations by aggregating existing PRS using ML.	[73]
AI-MARRVEL (AIM)	Mendelian genetic disorders	Genomics (germline)	Diagnostic	RF classifier	Research use only	Better and faster diagnosis of genetic Mendelian disorders by training a classifier on more than 3.5 million variants from thousands of diagnostic cases.	[54]
EpSign	Mendelian genetic disorders	Epigenomics (bulk DNA methylation)	Diagnostic	Support vector machine (SVM)	Laboratory developed test	Identifies epigenotypes (methylation patterns) associated with Mendelian disorders) from peripheral blood samples. It leverages an SVM to analyse DNA methylation and generate a methylation variant pathogenicity (MVP) score to classify patients into 43 epigenotype groups.	[74]
CytoRACE2	Oncology	Transcriptomics (single-cell RNA-seq)	Prognostic, predictive	Gene set binary network (GSBN)	Research use only	Tool to analyse and interpret RNA-seq data using a novel interpretable deep learning framework (GSBN). It can be used to find gene signatures predictive of disease or response to drugs such as immune checkpoint inhibitors and chemotherapy.	[75]
ProteinScores	Age-related diseases and mortality	Proteomics	Risk	Cox proportional hazards regression and elastic net	Research use only	By integrating multiple protein markers, it is possible to build biomarkers that predict disease onset (ten year risk) better than using more traditional markers (such as age, sex and a comprehensive set of 24 lifestyle factors, clinically relevant biomarkers and physical measures); including for different age-related diseases (type 2 diabetes, chronic obstructive pulmonary disease, Alzheimer's dementia, ischaemic heart disease, Parkinson's disease) and mortality (death).	[76]
CSP-based AD biomarkers	Alzheimer's disease (AD)	Proteomics	Risk, diagnostic, prognostic, monitoring	Subtype and Stage Inference (SuStain) algorithm	Research use only	Establish a model of AD progression using cerebrospinal fluid (CSF) biomarkers and the novel Subtype and Stage Inference (SuStain) algorithm. This could become a cost-effective tool that complements/eliminates the need for PET scanning.	[77]

Continued

Table 2: Continued.

Biomarker or framework name	Therapeutic area(s)	Datatype(s)	Biomarker categories	AI algorithm(s)	Regulatory status	Highlights	References
Nightingale Health plasma-based metabolic biomarkers	Many therapeutic areas	Metabolomics	Risk	Cox proportional hazards regression	Conformité Européenne (CE)-marked	Create a large-scale metabolomics biomarker-disease association atlas across >700 diseases and >100k individuals and identify signatures for disease risk.	[59]
Gut Microbiome Wellness Index2 (GMW12)	Wellness	Microbiome (genomics)	Diagnostic, monitoring	LASSO-penalised logistic regression	Research use only	Developing a disease-agnostic universal biomarker of gut health which can distinguish between healthy and unhealthy individuals. It was trained using 8069 metagenomic samples from 54 studies and 26 countries.	[78]
Prothrombin Time International Normalised Ratio (PT INR)	Cardiovascular disease	EHR	Predictive, monitoring	Dense neural networks and recurrent neural networks (RNNs)	Research use only	Deep learning algorithms are used to predict patient response to warfarin, a common anticoagulant which shows high response variability among individuals, thus enabling more optimised dose adjustments.	[79]
Brainomix 360 e-ASPECTS	Stroke	Imaging (CT)	Diagnostic	Deep learning algorithms	FDA 510(k) and CE-marked	It uses DL algorithms to analyse CT scans and quantify ischaemic damage for acute ischaemic stroke patients. It is able to identify subtle changes which may be missed by human readers and therefore improve decisions on treatment.	[80]
MSInuit	Colorectal cancer	Imaging (digital pathology)	Diagnostic	Convolutional neural networks (CNNs)	CE-marked	Pre-screening tool that identifies microsatellite instability (MSI) status of colorectal cancer patients from haematoxylin-eosin (H&E) stained slides. By ruling out ~50% of non-MSI cases, it is a cost-effective way to alleviate MSI testing burden in clinical practice.	[81]
Quantitative continuous scoring	Lung cancer	Imaging (digital pathology)	Predictive	Deep learning algorithms	Research use only	AI-based immunohistochemical imaging biomarker for scoring TROP2 expression in non-small cell lung cancer (NSCLC) and predicting response to a TROP2-directed antibody-drug conjugate.	[82]
Stratipath breast	Breast cancer	Imaging (digital pathology)	Prognostic	DL algorithms	CE-marked	Prognostic risk stratification of breast cancer patients from H&E-stained tissue sample images. This is particularly relevant for the intermediate-risk ER+/HER2- patients, since	[83]

Continued

Table 2: Continued.

Biomarker or framework name	Therapeutic area(s)	Datatype(s)	Biomarker categories	AI algorithm(s)	Regulatory status	Highlights	References
Virchow2	Oncology	Imaging (digital pathology)	Diagnostic, prognostic, predictive	Transformer-based foundational model	Research use only	improved prognosis can provide guidance relating to treatment decisions for adjuvant chemotherapy, and thus reduce over- and undertreatment of patients. A self-supervised learning foundational model trained on 3.1 million histopathology whole slide images, with diverse tissues, originating institutions and stains. It can perform many tasks in digital pathology. [65]	
Alzheimer's Disease progression model	Alzheimer's disease (AD)	Imaging (MRI)	Risk, prognostic	Convolutional neural networks (CNNs), LASSO-regularised Cox proportional hazards regression	Research use only	Building time-to-event models from MRI features [60] to predict progression of mild cognitive impairment (MCI) to Alzheimer's disease (AD), which can help to stratify risk of MCI patients.	
Apple Watch AFib history	Atrial fibrillation (AF)	Wearables (smartwatch)	Diagnostic	DL algorithms	FDA 510(k)	Using the photoplethysmography sensor which measures changes in blood flow, it is possible to detect irregular pulses and atrial fibrillation. This demonstrates a real-world application of wearable technology for health continuous monitoring. [84]	
MILTON	Many therapeutic areas	Multi-omics (genomics, proteomics, EHR)	Risk, diagnostic	Gradient boosting (XGBoost) within a bespoke ensemble framework	Research use only	This ensemble machine-learning framework leverages multi-omics data to predict >3000 diseases, outperforming available polygenic risk scores and augmenting genetic association analyses. [46]	
Duet ewoC	Oncology	Multi-omics (genomics, epigenomics)	Diagnostic, monitoring	LASSO-regularised logistic regression	Research use only	By combining measurements of 5mC and 5hmC in [85] cDNA from liquid biopsies, it enhances diagnostic accuracy in colorectal cancer detection (AUC = 0.95) compared with traditional approaches that conflate these markers (modified C, AUC = 0.66). The assay can also detect genetics if used at higher sequencing depth. Together, this method can improve the sensitivity of liquid biopsy tests for early cancer detection.	

Continued

Table 2: Continued.

Biomarker or framework name	Therapeutic area(s)	Datatype(s)	Biomarker categories	AI algorithm(s)	Regulatory status	Highlights	References
PreciousTGP	Ageing and age-related diseases	Multi-omics (transcriptomics, epigenomics)	Risk	TabTransformer	Research use only	This model predicts biological age using RNA-seq and DNA methylation, and it also identifies potential therapeutic targets for age-related diseases.	[64]
InflammAge	Systemic chronic inflammation and ageing	Multi-omics (epigenomics, proteomics)	Risk	Elastic net	Laboratory developed test	It leverages a framework that maps blood proteomics to saliva DNA methylation data. This can help to build accessible biomarkers for preventative healthcare, including those required to quantify hallmarks of ageing, such as systemic chronic inflammation.	[19]
Gene-SGAN	Neurological disease	Multi-modal (genomics, MRI)	Risk, diagnostic	Deep weakly-supervised clustering method based on generative adversarial networks (GAN) and variational inference (VI)	Research use only	Multi-modal integration of neuroimaging phenotypes and genetic variants, which allows identification of disease subtypes and associated endophenotypic signatures.	[20]
High-grade serous ovarian cancer (HGSO) prognostic model	Ovarian cancer	Multi-modal (genomics, transcriptomics, proteomics, digital pathology)	Prognostic	RF classifier	Research use only	Multi-modal data (histopathological image features, genomics, transcriptomics and proteomics) are used to predict survival for ovarian cancer patients (five year AUC = 0.911) and stratify them into high- and low-risk groups with high accuracy (HR = 18.23).	[86]
ArteraAI	Prostate cancer	Multi-modal (EHRs, digital pathology)	Predictive	Convolutional neural networks (CNNs) and multimodal deep learning architecture	Laboratory developed test	By combining digital pathology images from pretreatment prostate tissue and clinical data, the algorithm is able to predict which prostate cancer patients will benefit from androgen deprivation therapy, therefore avoiding unnecessary toxicity and side effects.	[87]

ER+/HER2–, oestrogen receptor positive/ human epidermal growth factor receptor 2 negative.

leverage more classical ML algorithms, such as LASSO, elastic net regression or tree-based models [19,46,59,74,76,78,85]. When it comes to imaging or multi-modal biomarkers, DL approaches are more suited due to the complexity of the data. For example, Brainomix uses DL algorithms to analyse CT scans of ischaemic stroke patients (which are able to detect subtle changes that may be missed by clinical experts) [80,88]. DL is also widely leveraged for tissue-based digital pathology biomarkers for cancer [89,90]. These biomarkers may be alternatives to existing ones such as microsatellite instability (MSI) or, more interestingly, be novel imaging biomarkers which are not possible using traditional pathological tests [82,87].

It can be seen that a great number of published studies generate susceptibility/risk and diagnostic biomarkers, which have applications in preventative medicine and early disease detection. This is particularly important in areas such as oncology, where early detection is crucial for prognosis and patient survival [87]. Neurology and cardiology biomarkers are other hot areas for the use of AI tools (such as cerebrospinal fluid-based AD biomarkers or prediction of AD progression) as they can aid clinicians in better understanding disease risk, progression and improve decisions on treatment, as well as dosage selection [60,76,77,79,80,84,88]. Biomarkers for wellness and health monitoring are also becoming a focal point, such as gut health monitoring from microbiome data [78] and quantifying ageing and its hallmarks using epigenetic data [19]. Continuous monitoring and real-time tracking are also gaining traction through the power of AI to derive insights from longitudinal data, such as from wearables like smartwatches [84]. This demonstrates a shift in reactive healthcare to a more proactive preventative approach, brought on by the use of AI, which will reduce disease burden and improve population health.

AI-powered biomarker discovery is also changing the drug discovery process. This includes applications to improve novel target identification and design biomarker-informed patient recruitment strategies for clinical trials, which can improve trial success rate and deliver on the promise of precision medicine. Specifically, recent AI-based models [46] help us better identify misdiagnosed and/or undiagnosed cases, leading to augmented case-control cohorts and thus higher statistical power to detect strong genetic signals (previously undetected in baseline association studies). These signals can not only elucidate novel gene targets for a disease area but also help us triage among an existing pool of candidates, supported by additional types of evidence. Other AI frameworks that leverage multi-omics data for novel target identification have also been reported, such as PandaOmics, which was used to identify potential therapeutic targets for amyotrophic lateral sclerosis [91]. Moreover, biomarker-informed patient recruitment is particularly critical in prevention trials, especially in oncology, as it enables the selection of patients with homogeneous disease profiles. This approach enhances the likelihood of identifying populations that are more responsive to specific treatments, thereby maximising the drug's efficacy potential [92].

Unfortunately, population bias is still commonplace, with most biomarkers being trained in European populations and with limited data from diverse ethnic groups [59,78]. Nevertheless, there is an increased focus on generalisability and external validation of biomarkers in more diverse cohorts. This is evidenced in studies such as GMW12, which has been validated across 54 studies in 26 countries [78], or in recent efforts using DL approaches to improve polygenic risk scores across different ancestries [93].

The challenges to go from biomarker discovery to the clinic

AI has notably contributed towards progress in biomarker discovery; yet significant challenges remain to translate these biomarkers to the clinic. High-level calculations show that only between 1% and 2% of all the biomarkers published in the literature actually achieve clinical adoption by healthcare providers and patients [94,95]. Furthermore, it has been estimated that in order to develop and bring to market a typical in vitro diagnostic (e.g. those that will require FDA review), it can take between 20 and 100 USD million and 3–7 years [96], normally followed by a long period before widespread adoption in clinical guidelines.

There are many challenges that need to be considered and addressed in order to see the fruits of AI in biomarker discovery having an impact on the real world (Figure 3). Challenges that affect diagnostics more broadly include cumbersome and expensive clinical trials (and access to samples), meeting new regulatory requirements or unlocking reimbursement pathways [97].

Importantly, there are certain challenges that particularly affect those biomarkers that may be discovered using AI. For example, accessing real-world data and real-world evidence from different populations (beyond biobank data, such as from hospitals) is paramount to avoid bias during model training, but this data is still heavily siloed and lacks harmonisation and interoperability [98]. In order to

	 Biomarker discovery	 Biomarker development	 Clinical studies	 Regulatory submission	 Manufacturing/ infrastructure deployment	 Distribution/ adoption
Challenges	Real-world data access and quality	Sample access	Clinical trial design and recruitment	Evolving AI regulations	Infrastructure scalability	Unlocking reimbursement pathways
	Population bias	Finding a cost-effective assay	Lack of access to diverse populations	Time-consuming approval pathways	Supply chain coordination (labs, assays, bioinformatics pipelines)	Adoption of AI-biomarkers in healthcare
	Data harmonisation					
	Omics data generation and computing costs		High costs and time-consuming			

Figure 3: Key challenges that need to be considered to bring a biomarker from its early research stages to wide clinical adoption.

solve this problem, different global federated real-world data platforms are being developed, which can give access to large amounts of diverse and standardised data while protecting patient privacy and ensuring ethical use [99]. Furthermore, open-source common data models, such as OMOP (Observational Medical Outcomes Partnership), are paving the way towards data standardisation [100].

Additionally, multi-omics technologies are generally expensive, so it is important to consider the trade-off between accuracy and cost for a specific use case and disease during biomarker discovery, so that a viable health economics model can be proposed at the end. For instance, in the multi-omics biomarker discovery framework MILTON, researchers were able to demonstrate that adding a small number of proteomics biomarkers to a set of traditional clinical biomarkers is sufficient to significantly improve model performance in the prediction of specific diseases [46].

Infrastructure can become a bottleneck as well. Computational costs to train large DL models (particularly when there is a need to access GPUs) are significant, which is increasing the research gap between academia and industry, with fewer academics training models from scratch [101]. Furthermore, the diagnostics infrastructure becomes critical when different omics layers are combined in a single biomarker. This may require orchestrating a complex supply chain of different laboratories, assays and bioinformatic pipelines to be able to offer the final diagnostic at scale.

Despite these challenges, it is worth noting that the number of biomarkers with AI is increasing over time. When focusing on AI/ML-enabled medical devices, there are ~950 examples that have already passed FDA review [102], showing a bright future ahead for AI-powered biomarkers.

The future of AI-powered biomarker discovery

AI has become a cornerstone of biomarker discovery, leading to the development of AI-powered biomarkers. In the coming years, we expect AI to also accelerate scientific discovery by enabling new forms of scientific collaboration and automation. One promising concept is that of a ‘Virtual Lab’. This is a system composed of multiple AI agents (powered by LLMs), each taking on the role of a virtual scientist (e.g. principal investigator, immunologist, computational biologist or machine learning specialist), and working together in a co-ordinated fashion to solve complex biomedical problems. For instance, a recent study demonstrated how such a Virtual Lab could design and implement a computational pipeline for nanobody development for new SARS-CoV-2 variants. This involved deciding on a strategy to modify existing nanobodies (rather than designing them *de novo*), generating candidate sequences, modelling nanobody-antigen structures and evaluating binding affinity *in silico*, all orchestrated by AI agents. The nanobody candidates proposed by the AI system were subsequently validated in wet-lab experiments and shown to bind emerging SARS-CoV-2 variants [103]. Therefore, it is not unthinkable to imagine a near future where multi-modal biomarker discovery is highly automated by AI agents. They could help to find the best biomarkers available in the literature for a given therapeutic area, suggest biomarker hypotheses (e.g. the best datatype for a particular application), generate code to train and test biomarker models faster than existing methods or guide doctors to accomplish better diagnoses by integrating all the available biological data and models for a given patient. This will require increased collaboration between experts from different fields, such as biologists, computational biologists, AI/ML engineers and doctors.

Clinicians are used to the concept of multi-modal diagnosis (e.g. when they review MRI results in the context of medical history). With the broader use of AI technologies to assist and automate this process,

it will be critical that AI-powered biomarkers are interpretable and explainable, so they are adopted in clinical practice [104]. Advances in continuous biomarker monitoring through new wearable and sensor technology will lead to at-home real-time health monitoring, adding another dimension to quantify health and disease [105]. Importantly, these technologies need to be democratised and made accessible to all populations, so that the health equity gap and health inequalities are reduced [106]. AI has truly the potential to move biomarker discovery from the hardware layer (developing new ‘wet-lab’ assays) to the software layer (new insights from existing data and assays), therefore accelerating innovation cycles and the path towards precision medicine.

Summary bullet points

- Artificial intelligence (AI) is driving innovation in healthcare research and biomarker discovery owing to its ability to derive insights from complex high-dimensional datasets and integrate multi-modal datatypes (such as omics, electronic health records, imaging or sensor and wearable data).
- To ensure that AI-powered biomarkers are generalised and unbiased, large datasets from diverse populations are required, such as biobanks, longitudinal studies, real-world data and data from clinical trials.
- Classical machine learning models (like elastic net regression and random forests) are more commonly utilised for single-omics and multi-omics biomarkers, whereas deep learning models (like convolutional neural networks or transformers) are important in handling datatypes like imaging, large unstructured electronic health records or multi-modal data.
- Key challenges to translating AI-powered biomarkers to the clinic and driving adoption need to be addressed, including lack of population diversity, difficulties accessing harmonised data, costly and time-consuming clinical studies, evolving AI regulatory frameworks and absence of scalable diagnostic infrastructure.
- AI agents will likely automate and enhance many steps of the biomarker discovery process, accelerating the paradigm shift towards precision medicine.

Competing Interests

HJ, CCP, LS and LM are employees at Chronomics Ltd. DEMH and TMS are co-founders and shareholders at Chronomics Ltd., a company focused on multi-omics and epigenetic biomarker development and diagnostic infrastructure. MG and DV are current employees and/or stockholders of AstraZeneca. JMM is a current employee of Biomodal and holds share options. POR is a current employee of Sonrai Analytics Ltd. and holds shares in the company. BBS is a full-time employee of Bristol Myers Squibb.

Funding

The authors received no specific funding for this work.

Open Access

This article has been published open access under our Subscribe to Open programme, made possible through the support of our subscribing institutions, learn more here: https://portlandpress.com/pages/open_access_options_and_prices#conditional

CRedit Author Contribution

HJ, CCP, LS and DEMH wrote the original draft and conceptualised the figures. All authors reviewed, edited and approved the manuscript.

Acknowledgments

We would like to thank Madhurima Dey and Augustin Amour for inviting us to submit this manuscript. We are grateful to Ian Robinson for designing the final version of the figures.

Abbreviations

AD, Alzheimer's disease; AF, atrial fibrillation; AI, artificial intelligence; AUC, Area Under the Curve; BERT, Bidirectional Encoder Representations from Transformers; BEST, Biomarkers, EndpointS, and other Tools; C, Cytosine; CE, Conformité Européenne; CE-marked, Conformité Européenne - marked; CNNs, convolutional neural networks; CSF, cerebrospinal fluid; CT, computed tomography; 2D, Two Dimensional; 3D, Three Dimensional; DL, deep learning; DNA, Deoxyribonucleic Acid; ECG, Electrocardiograms; EEG, Electroencephalography; EHRs, electronic health records; ER+/HER2−, oestrogen receptor positive/human epidermal growth factor receptor 2 negative; FDA-NIH, Food and Drug Administration - National Institutes of Health; GANs, generative adversarial networks; GMW12, Gut Microbiome Wellness Index 2; GNNs, Graph Neural Networks; GPT, Generative Pre-trained Transformer; GPUs, Graphics Processing Unit; GSB, Gene Set Binary Network; GTNs, Graph Transformer Networks; GenAI, Generative Artificial Intelligence; H&E, Haematoxylin-Eosin; H&E, haematoxylin-eosin; HGSOC, high-grade serous ovarian cancer; HR, Hazard Ratio; HbA1c, Glycated hemoglobin; LASSO, Least Absolute Shrinkage and Selection Operator; LLMs, large language models; MCI, mild cognitive impairment; ML, machine learning; MRI, magnetic resonance imaging; MSI, microsatellite instability; MVP, methylation variant pathogenicity; NSCLC, non-small cell lung cancer; OMOP, Observational Medical Outcomes Partnership; PET, positron emission tomography; PMC, PubMed Central; PRS, Polygenic Risk Score; PT INR, Prothrombin time international normalised ratio; RF, random forest; RGB, Red, Green and Blue; RNA-Seq, Ribonucleic Acid Sequencing; RNNs, recurrent neural networks; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; SHAP, SHapley Additive exPlanations; SVM, Support vector machine; SuStain, Subtype and Stage Inference Algorithm; TREs, Trusted Research Environments; UK, United Kingdom; USD, United States Dollar; VI, variational inference; cfDNA, Cell-Free Deoxyribonucleic Acid; 5hmC, 5-hydroxymethylcytosine; 5mC, 5-methylcytosine.

References

- Janiesch, C., Zschech, P. and Heinrich, K. (2021) Machine learning and deep learning. *Electron. Markets* **31**, 685–695 <https://doi.org/10.1007/s12525-021-00475-2>
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X, et al. (2024) Large Language Models: A Survey. <https://doi.org/https://arxiv.org/abs/2402.06196>
- Hamadi, R. (2023) Large Language Models Meet Computer Vision: A Brief Survey. <http://arxiv.org/abs/2311.16673>
- Borkakoti, N. and Thornton, J.M. (2023) AlphaFold2 protein structure prediction: Implications for drug discovery. *Curr. Opin. Struct. Biol.* **78**, 102526 <https://doi.org/10.1016/j.sbi.2022.102526> PMID: 36621153
- Vadapalli, S., Abdelhalim, H., Zeeshan, S. and Ahmed, Z. (2022) Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Brief. Bioinformatics* **23**, bbac191 <https://doi.org/10.1093/bib/bbac191> PMID: 35595537
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L, et al. (2023) Towards expert-level medical question answering with large language models. <http://arxiv.org/abs/2305.09617>
- Acosta, J.N., Falcone, G.J., Rajpurkar, P. and Topol, E.J. (2022) Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 <https://doi.org/10.1038/s41591-022-01981-2> PMID: 36109635
- Prelaj, A., Miskovic, V., Zanitti, M., Trovo, F., Genova, C., Viscardi, G, et al. (2024) Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. *Ann. Oncol.* **35**, 29–65 <https://doi.org/10.1016/j.annonc.2023.10.125> PMID: 37879443
- Ng, S., Masarone, S., Watson, D. and Barnes, M.R. (2023) The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res.* **394**, 17–31 <https://doi.org/10.1007/s00441-023-03816-z> PMID: 37498390
- (2024) BEST (Biomarkers, EndpointS, and other Tools) Resource - NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK338448>
- Garg, M. and Catharine's CollegeS.. (2022) Prognostic biomarker discovery from omics data using machine learning approaches. <https://www.repository.cam.ac.uk/handle/1810/342044>
- FDA. (2024) Biomarkers and Qualification | FDA. <https://www.fda.gov/drugs/biomarker-qualification-program/about-biomarkers-and-qualification>
- FDA. (2024) Companion Diagnostics. <https://www.fda.gov/medical-devices/in-vitro-diagnostics/companion-diagnostics>
- Ahmed, Z., Wan, S., Zhang, F. and Zhong, W. (2024) Artificial intelligence for omics data analysis. *BMC Methods* **1** <https://doi.org/10.1186/s44330-024-00004-5>
- DeGroat, W., Mendhe, D., Bhusari, A., Abdelhalim, H., Zeeshan, S. and Ahmed, Z. (2023) IntelliGenes: a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. *Bioinformatics* **39**, btad755 <https://doi.org/10.1093/bioinformatics/btad755> PMID: 38096588
- Vandereyken, K., Siffrim, A., Thienpont, B. and Voet, T. (2023) Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 <https://doi.org/10.1038/s41576-023-00580-2> PMID: 36864178
- Wu, X., Zhao, Z., Yu, W., Liu, S., Zhou, M., Jiang, N. et al. (2024) Single-cell multiomics identifies glycan epitope LacNAc as a Potential Cell-Surface effector marker of peripheral t cells in bladder cancer patients. *ACS Chem. Biol.* **19**, 2535–2547 <https://doi.org/10.1021/acscchembio.4c00635>
- Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K. et al. (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 <https://doi.org/10.1038/s41592-019-0619-0> PMID: 31740819

- 19 Schmunk, L.J., Call, T.P., McCartney, D.L., Javaid, H., Hastings, W.J., Jovicevic, V. et al. (2025) A novel framework to build saliva-based DNA methylation biomarkers: Quantifying systemic chronic inflammation as a case study. *Aging Cell* **24**, e14444 PMID: 39888134 <https://onlinelibrary.wiley.com/doi/10.1111/accel.14444>,
- 20 Yang, Z., Wen, J., Abdulkadir, A., Cui, Y., Erus, G., Mamourian, E. et al. (2024) Gene-SGAN: discovering disease subtypes with imaging and genetic signatures via multi-view weakly-supervised deep clustering. *Nat. Commun.* **15** <https://doi.org/10.1038/s41467-023-44271-2>
- 21 H3Africa. (2024) H3Africa – Human Heredity & Health in Africa. <https://h3africa.org/>
- 22 China Kadoorie Biobank (CKB). (2024) China Kadoorie Biobank. <https://www.ckbiobank.org/>
- 23 Kuriyama, S., Yaegashi, N., Nagami, F., Arai, T., Kawaguchi, Y., Osumi, N. et al. (2016) The tohoku medical megabank project: design and mission. *J. Epidemiol.* **26**, 493–511 <https://doi.org/10.2188/jea.JE20150268> PMID: 27374138
- 24 BioBank Japan. (2024) Overview of BBJ's Samples and Data. <https://biobankjp.org/en/researchers/1971#gsc.tab=0>
- 25 Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J. et al. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 <https://doi.org/10.1371/journal.pmed.1001779> PMID: 25826379
- 26 Institute of Genomics. (2024) Estonian Biobank. <https://genomics.ut.ee/en/content/estonian-biobank>
- 27 FinnGen. (2024) FinnGen: an expedition into genomics and medicine. <https://www.finnngen.fi/en>
- 28 QPHI. (2024) Qatar Precision Health Institute. <https://www.qphi.org.qa/>
- 29 National Institutes of Health. (2024) All of Us Research Program. <https://allofus.nih.gov/>
- 30 CanPath. (2024) Canadian Partnership for Tomorrow's Health. <https://canpath.ca/>
- 31 MVP Home. (2024) Veterans Affairs. <https://www.mvp.va.gov/pwa/>
- 32 Neuber, A.C., Tostes, C.H., Ribeiro, A.G., Marczyński, G.T., Komoto, T.T., Rogeri, C.D. et al. (2024) The biobank of barretos cancer hospital: 14 years of experience in cancer research. *Cell Tissue Bank.* <https://doi.org/https://link.springer.com/article/10.1007/s10561-021-09941-9>
- 33 Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU). (2024) Mexico City Prospective Study. <https://www.ctsu.ox.ac.uk/research/prospective-blood-based-study-of-150-000-individuals-in-mexico>
- 34 Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.H., Richardson, T.G. et al. (2023) Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 <https://doi.org/10.1038/s41586-023-06592-6> PMID: 37794186
- 35 Belsky, D.W., Caspi, A., Corcoran, D.L., Sugden, K., Poulton, R., Arseneault, L. et al. (2022) DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife* **11**, e73420 <https://doi.org/10.7554/eLife.73420> PMID: 35029144
- 36 Van Houtum, L.A.E.M., Baaré, W.F.C., Beckmann, C.F., Castro-Fornieles, J., Cecil, C.A.M., Ditttrich, J. et al. (2024) Running in the FAMILY: understanding and predicting the intergenerational transmission of mental illness. *Eur. Child Adolesc. Psychiatry* **33**, 3885–3898 <https://doi.org/10.1007/s00787-024-02423-9>
- 37 Cai, N., Revez, J.A., Adams, M.J., Andlauer, T.F.M., Breen, G., Byrne, E.M. et al. (2020) Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.* **52**, 437–447 <https://doi.org/10.1038/s41588-020-0594-5> PMID: 32231276
- 38 Flores, J.E., Claborn, D.M., Weller, Z.D., Webb-Robertson, B.J.M., Waters, K.M. and Bramer, L.M. (2023) Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front. Artif. Intell.* **6**, 1098308 <https://doi.org/10.3389/frai.2023.1098308> PMID: 36844425
- 39 Vahdati, A., Cotterill, S., Marsden, A. and Kontopantelis, E. (2024) Enhancing data integrity in Electronic Health Records: Review of methods for handling missing data. <https://doi.org/https://www.medrxiv.org/content/10.1101/2024.05.13.24307268v1>
- 40 Salvatore, M., Kundu, R., Du, J., Friese, C.R., Mondul, A.M., Hanauer, D. et al. (2024) Reducing Information and Selection Bias in EHR-Linked Biobanks via Genetics-Informed Multiple Imputation and Sample Weighting. 10–29 *Epidemiology*. <https://doi.org/10.1101/2024.10.28.24316286>
- 41 An, U., Pazokitoroudi, A., Alvarez, M., Huang, L., Bacanu, S., Schork, A.J. et al. (2023) Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nat. Genet.* **55**, 2269–2276 <https://doi.org/10.1038/s41588-023-01558-w> PMID: 37985819
- 42 Yoon, J., Mizrahi, M., Ghalaty, N.F., Jarvinen, T., Ravi, A.S., Brune, P. et al. (2023) EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digit. Med.* **6**, 141 <https://doi.org/10.1038/s41746-023-00888-7> PMID: 37567968
- 43 Ying, K., Song, J., Cui, H., Zhang, Y., Li, S., Chen, X. et al. (2024) MethylGPT: a foundation model for the DNA methylome. <https://www.biorxiv.org/content/10.1101/2024.10.30.621013v2>
- 44 Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P. and Duncan, J.S. (2020) Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **65**, 101765 <https://doi.org/10.1016/j.media.2020.101765> PMID: 32679533
- 45 Delfin, C., Dragan, I., Kuznetsov, D., Tajés, J.F., Smit, F., Coral, D.E. et al. (2024) A federated database for obesity research: an IMI-SOPHIA Study. *Life (Basel)*. **14**, 262 <https://doi.org/10.3390/life14020262> PMID: 38398771
- 46 Garg, M., Karpinski, M., Matelska, D., Middleton, L., Burren, O.S., Hu, F. et al. (2024) Disease prediction with multi-omics and biomarkers empowers case–control genetic discoveries in the UK Biobank. *Nat. Genet.* **56**, 1821–1831 <https://doi.org/10.1038/s41588-024-01898-1>
- 47 Fitipaldi, H. and Franks, P.W. (2023) Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022. *Hum. Mol. Genet.* **32**, 520–532 <https://doi.org/10.1093/hmg/ddac245> PMID: 36190496
- 48 Prictor, M., Teare, H.J.A. and Kaye, J. (2018) Equitable participation in biobanks: the risks and benefits of a “dynamic consent” approach. *Front. Public Health* **6** <https://doi.org/10.3389/fpubh.2018.00253>
- 49 Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D.J., Martin, I., Kullo, I.J. et al. (2024) Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25 <https://doi.org/10.1038/s41576-023-00637-2> PMID: 37620596
- 50 Watkins, S.H., Testa, C., Chen, J.T., De Vivo, I., Simpkin, A.J., Tilling, K. et al. (2023) Epigenetic clocks and research implications of the lack of data on whom they have been developed: a review of reported and missing sociodemographic characteristics. *Environ. Epigenet.* **9**, dvad005 <https://doi.org/10.1093/eeep/dvad005> PMID: 37564905

- 51 Jiang, Z., Zhang, S., Zeng, P. and Wang, T. (2023) Influence of social deprivation on morbidity and all-cause mortality of cardiometabolic multi-morbidity: a cohort analysis of the UK Biobank cohort. *BMC Public Health* **23**, 2177 <https://doi.org/10.1186/s12889-023-17008-5> PMID: 37932741
- 52 Ma, Y., Zhao, S., Wang, W., Li, Y. and King, I. (2022) Multimodality in meta-learning: A comprehensive survey. *Knowl. Based Syst.* **250**, 108976 <https://doi.org/10.1016/j.knosys.2022.108976>
- 53 De Lima Camillo, L.P., Lapierre, L.R. and Singh, R. (2022) A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* **8** <https://doi.org/10.1038/s41514-022-00085-y>
- 54 Mao, D., Liu, C., Wang, L., Al-Ouran, R., Deisseroth, C., Pasupuleti, S. et al. (2024) AI-MARRVEL — A Knowledge-Driven AI System for Diagnosing Mendelian Disorders. *NEJM AI* **1** <https://doi.org/10.1056/Aloa2300009>
- 55 Ghadekar, P., Tekade, S., Sakharwade, D., Tripathi, A., Tiwadi, S. and Zanzane, S. (2024) Multimodal PCOS Detection: Combining XG Boost for Images with Zero Shot Learning for Textual Data. 2024 Asia Pacific Conference on Innovation in Technology (APCIT), pp. 1–8, MYSORE, India
- 56 Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Statist.* **29** <https://doi.org/10.1214/aos/1013203451>
- 57 Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*.
- 58 Bifarin, O.O. (2023) Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *PLoS ONE* **18**, e0284315 <https://doi.org/10.1371/journal.pone.0284315> PMID: 37141218
- 59 Julkunen, H., Cichońska, A., Tiainen, M., Koskela, H., Nybo, K., Mäkelä, V. et al. (2023) Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat. Commun.* **14**, 604 <https://doi.org/10.1038/s41467-023-36231-7> PMID: 36737450
- 60 Li, H., Habes, M., Wolk, D.A., Fan, Y. and Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle Study of Aging (2019) A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimers. Dement.* **15**, 1059–1070 <https://doi.org/10.1016/j.jalz.2019.02.007>
- 61 Mehdi-pour Ghazi, M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S. et al. (2019) Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. *Med. Image Anal.* **53**, 39–46 <https://doi.org/10.1016/j.media.2019.01.004> PMID: 30682584
- 62 OpenAI, A.J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I. et al. (2023) GPT-4 Technical Report. <http://arxiv.org/abs/2303.08774>
- 63 Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Vol. 1 of, pp. 4171–4186
- 64 Urban, A., Sidorenko, D., Zagirova, D., Kozlova, E., Kalashnikov, A., Pushkov, S. et al. Precious1GPT: multimodal transformer-based transfer learning for aging clock development and feature importance analysis for aging and age-related disease target discovery. *Aging (Milano)* **15** <https://doi.org/10.18632/aging.204788>
- 65 Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G. et al. (2024) Virchow 2: Scaling Self-Supervised Mixed Magnification Models in Pathology. <http://arxiv.org/abs/2408.00738>
- 66 Segura-Bedmar, I., Camino-Perdones, D. and Guerrero-Aspizua, S. (2022) Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. *BMC Bioinformatics* **23**, 263 <https://doi.org/10.1186/s12859-022-04810-y> PMID: 35794528
- 67 Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C. et al. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 <https://doi.org/10.15252/msb.20178124> PMID: 29925568
- 68 Velten, B., Braunger, J.M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D. et al. (2022) Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 <https://doi.org/10.1038/s41592-021-01343-9> PMID: 35027765
- 69 Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E. et al. (2021) Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 <https://doi.org/10.1038/s41467-020-20430-7> PMID: 33402734
- 70 Valous, N.A., Popp, F., Zörnig, I., Jäger, D. and Charoentong, P. (2024) Graph machine learning for integrated multi-omics analysis. *Br. J. Cancer* **131**, 205–211 <https://doi.org/10.1038/s41416-024-02706-7>
- 71 Middleton, L., Melas, I., Vasavda, C., Raies, A., Rozemberczki, B., Dhindsa, R.S. et al. (2024) Phenome-wide identification of therapeutic genetic targets, leveraging knowledge graphs, graph neural networks, and UK Biobank data. *Sci. Adv.* **10**, ead1424 <https://doi.org/10.1126/sciadv.ad1424>
- 72 Johnson, R., Gottlieb, U., Shaham, G., Eisen, L., Waxman, J., Devons-Sberro, S. et al. Unified Clinical Vocabulary Embeddings for Advancing Precision Medicine. *medRxiv* <https://doi.org/10.1101/2024.12.03.24318322>
- 73 Truong, B., Hull, L.E., Ruan, Y., Huang, Q.Q., Hornsby, W., Martin, H. et al. (2024) Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genom.* **4**, 100523 <https://doi.org/10.1016/j.xgen.2024.100523> PMID: 38508198
- 74 Sadikovic, B., Levy, M.A., Kerkhof, J., Aref-Eshghi, E., Schenkel, L., Stuart, A. et al. (2021) Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. *Genet. Med.* **23**, 1065–1074 <https://doi.org/10.1038/s41436-020-01096-4> PMID: 33547396
- 75 Kang, M., Armenteros, J.J.A., Gulati, G.S., Gleyzer, R., Avagyan, S., Brown, E.L. et al. (2024) Biorxiv. Mapping single-cell developmental potential in health and disease with interpretable deep learning. <http://biorxiv.org/lookup/doi/10.1101/2024.03.19.585637>
- 76 Gadd, D.A., Hillary, R.F., Kuncheva, Z., Mangelis, T., Cheng, Y., Dissanayake, M. et al. (2024) Blood protein assessment of leading incident diseases and mortality in the UK Biobank. *Nat. Aging* **4**, 939–948 <https://doi.org/10.1038/s43587-024-00655-7> PMID: 38987645
- 77 Salvadó, G., Horie, K., Barthélemy, N.R., Vogel, J.W., Pichet Binette, A., Chen, C.D. et al. (2024) Disease staging of Alzheimer's disease using a CSF-based biomarker model. *Nat. Aging* **4**, 694–708 <https://doi.org/10.1038/s43587-024-00599-y> PMID: 38514824
- 78 Chang, D., Gupta, V.K., Hur, B., Cobo-López, S., Cunningham, K.Y., Han, N.S. et al. (2024) Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome taxonomic profiles. *Nat. Commun.* **15**, 7447 <https://doi.org/10.1038/s41467-024-51651-9>
- 79 Lee, H., Kim, H.J., Chang, H.W., Kim, D.J., Mo, J. and Kim, J.E. (2021) Development of a system to support warfarin dose decisions using deep neural networks. *Sci. Rep.* **11** <https://doi.org/10.1038/s41598-021-94305-2>

- 80 Kobeissi, H., Kallmes, D.F., Benson, J., Nagelschneider, A., Madhavan, A., Messina, S.A. et al. (2023) Impact of e-ASPECTS software on the performance of physicians compared to a consensus ground truth: a multi-reader, multi-case study. *Front. Neurol.* **14**, 1221255 <https://doi.org/10.3389/fneur.2023.1221255> PMID: 37745671
- 81 Saillard, C., Dubois, R., Tchita, O., Loiseau, N., Garcia, T., Adriansen, A. et al. (2023) Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. *Nat. Commun.* **14**, 6695 <https://doi.org/10.1038/s41467-023-42453-6>
- 82 Garassino, M.C., Sands, J., Paz-Ares, L., Lisberg, A., Johnson, M., Pérol, M. et al. (2024) PL02.11 Normalized Membrane Ratio of TROP2 by Quantitative Continuous Scoring is Predictive of Clinical Outcomes in TROPION-Lung 01. *J. Thorac. Oncol.* **19**, S2–S3 <https://doi.org/10.1016/j.jtho.2024.09.015>
- 83 Sharma, A., Lövgren, S.K., Eriksson, K.L., Wang, Y., Robertson, S., Hartman, J. et al. (2024) Breast Cancer. Validation of an AI-based solution for breast cancer risk stratification using routine digital histopathology images. <https://doi.org/10.1186/s13058-024-01879-6> DOI:10.1186/s13058-024-01879-6
- 84 Perez, M.V., Mahaffey, K.W., Hedlin, H., Rumsfeld, J.S., Garcia, A., Ferris, T. et al. (2019) Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N. Engl. J. Med.* **381**, 1909–1917 <https://doi.org/10.1056/NEJMoa1901183> PMID: 31722151
- 85 Puddu, F., Johansson, A., Modat, A., Scotcher, J., Sethi, R., Harding, N. et al. 5-methylcytosine and 5-hydroxymethylcytosine are synergistic biomarkers for early detection of colorectal cancer. *Cancer Biology*. <https://doi.org/10.1101/2024.10.30.621123>
- 86 Zeng, H., Chen, L., Zhang, M., Luo, Y. and Ma, X. (2021) Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecol. Oncol.* **163**, 171–180 <https://doi.org/10.1016/j.ygyno.2021.07.015> PMID: 34275655
- 87 Spratt, D.E., Tang, S., Sun, Y., Huang, H.-C., Chen, E., Mohamad, O. et al. (2023) Artificial Intelligence predictive model for hormone therapy use in prostate Cancer. *NEJM Evid.* **2**, EVIDoa2300023 <https://doi.org/10.1056/EVIDoa2300023> PMID: 38320143
- 88 Herweh, C., Ringleb, P.A., Rauch, G., Gerry, S., Behrens, L., Möhlenbruch, M. et al. (2016) Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. *Int. J. Stroke* **11**, 438–445 <https://doi.org/10.1177/1747493016632244> PMID: 26880058
- 89 Geaney, A., O'Reilly, P., Maxwell, P., James, J.A., McArt, D. and Salto-Tellez, M. (2023) Translation of tissue-based artificial intelligence into clinical practice: from discovery to adoption. *Oncogene* **42**, 3545–3555 <https://doi.org/10.1038/s41388-023-02857-6> PMID: 37875656
- 90 Shmatko, A., Ghaffari Laleh, N., Gerstung, M. and Kather, J.N. (2022) Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 <https://doi.org/10.1038/s43018-022-00436-4> PMID: 36138135
- 91 Pun, F.W., Liu, B.H.M., Long, X., Leung, H.W., Leung, G.H.D., Mewborne, Q.T. et al. (2022) Identification of Therapeutic Targets for Amyotrophic Lateral Sclerosis Using PandaOmics - An AI-Enabled Biological Target Discovery Platform. *Front. Aging Neurosci.* **14**, 914017 <https://doi.org/10.3389/fnagi.2022.914017> PMID: 35837482
- 92 Wong, C.H., Siah, K.W. and Lo, A.W. (2019) Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 <https://doi.org/10.1093/biostatistics/kxx069> PMID: 29394327
- 93 Gyawali, P.K., Le Guen, Y., Liu, X., Belloy, M.E., Tang, H., Zou, J. et al. (2023) Improving genetic risk prediction across diverse population by disentangling ancestry representations. *Commun. Biol.* **6**, 964 <https://doi.org/10.1038/s42003-023-05352-6> PMID: 37736834
- 94 Faculty of Medicine, Imperial College London. (2024) Biomarker Assessment, Validation, and Translation. <https://www.imperial.ac.uk/department-surgery-cancer/research/surgery/groups/biomarker-assessment-validation-and-translation/>
- 95 Martin-Herranz, D. E.. (2024) Post | LinkedIn. https://www.linkedin.com/posts/daniel-elias-martin-herranz_translation-gap-diagnostic-activity-7062121877875384321-kRMI/
- 96 Hurdle. (2024) How We Accelerate Biomarker Development: From R&D to the Clinic. <https://hurdle.bio/blog/how-we-accelerate-biomarker-development-from-rd-to-the-clinic/>
- 97 deloitte. (2022) Reforming diagnostics Turning challenges into enablers Contents. <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/deloitte-uk-future-diagnostics-companion-report-2022.pdf>
- 98 Baumfeld Andre, E., Carrington, N., Siami, F.S., Hiatt, J.C., McWilliams, C., Hiller, C. et al. (2022) The current landscape and emerging applications for real-World Data in Diagnostics and Clinical Decision Support and its Impact on Regulatory Decision Making. *Clin. Pharmacol. Ther.* **112**, 1172–1182 <https://doi.org/10.1002/cpt.2565> PMID: 35213741
- 99 Palchuk, M.B., London, J.W., Perez-Rey, D., Drebert, Z.J., Winer-Jones, J.P., Thompson, C.N. et al. (2023) A global federated real-world data and analytics platform for research. *JAMIA Open* **6**, ooad035 <https://doi.org/10.1093/jamiaopen/ooad035> PMID: 37193038
- 100 Henke, E., Zoch, M., Peng, Y., Reinecke, I., Sedlmayr, M. and Bathelt, F. (2024) Conceptual design of a generic data harmonization process for OMOP common data model. *BMC Med. Inform. Decis. Mak.* **24**, 58 <https://doi.org/10.1186/s12911-024-02458-7> PMID: 38408983
- 101 Khandelwal, A., Yun, T., Merullo, J., Bach, S.H., Sun, C., Nayak, N.V. et al. (2024) \$100K or 100 Days: Trade-offs when Pre-Training with Academic. <http://arxiv.org/abs/2410.23261>
- 102 U.S. Food and Drug Administration (FDA). (2024) Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>
- 103 Swanson, K., Wu, W., Bulaong, N.L., Pak, J.E. and Zou, J. The virtual lab: ai agents design New SARS-CoV-2 Nanobodies with experimental validation. *bioRxiv*. <http://biorxiv.org/content/early/2024/11/12/2024.11.11.623004>
- 104 Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 310 <https://doi.org/10.1186/s12911-020-01332-6> PMID: 33256715
- 105 Sempionatto, J.R., Lasalde-Ramirez, J.A., Mahato, K., Wang, J. and Gao, W. (2022) Wearable chemical sensors for biomarker discovery in the omics era. *Nat. Rev. Chem.* **6**, 899–915 <https://doi.org/10.1038/s41570-022-00439-w>
- 106 Chan, S.C.C., Neves, A.L., Majeed, A. and Faisal, A. (2024) Bridging the equity gap towards inclusive artificial intelligence in healthcare diagnostics. *BMJ* q490 <https://doi.org/10.1136/bmj.q490>