© 2022. Published by The Company of Biologists Ltd | Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

Deep phenotyping for precision medicine in Parkinson's disease

Ann-Kathrin Schalkamp, Nabila Rahman, Jimena Monzón-Sandoval and Cynthia Sandor*

ABSTRACT

A major challenge in medical genomics is to understand why individuals with the same disorder have different clinical symptoms and why those who carry the same mutation may be affected by different disorders. In every complex disorder, identifying the contribution of different genetic and non-genetic risk factors is a key obstacle to understanding disease mechanisms. Genetic studies rely on precise phenotypes and are unable to uncover the genetic contributions to a disorder when phenotypes are imprecise. To address this challenge, deeply phenotyped cohorts have been developed for which detailed, fine-grained data have been collected. These cohorts help us to investigate the underlying biological pathways and risk factors to identify treatment targets, and thus to advance precision medicine. The neurodegenerative disorder Parkinson's disease has a diverse phenotypical presentation and modest heritability, and its underlying disease mechanisms are still being debated. As such, considerable efforts have been made to develop deeply phenotyped cohorts for this disorder. Here, we focus on Parkinson's disease and explore how deep phenotyping can help address the challenges raised by genetic and phenotypic heterogeneity. We also discuss recent methods for data collection and computation, as well as methodological challenges that have to be overcome.

KEY WORDS: Genetics, Phenotyping, Precision medicine

Introduction

To elucidate the genetic and molecular processes that contribute to disease, the research community has made considerable efforts to develop large case/control genetic studies. These have been remarkably successful in identifying common genetic risk variants associated with various disorders. Over 6000 genome-wide association studies (GWAS; see Glossary, Box 1) have been published for over 1000 traits that report on tens of thousands of genetic risk variants (Watanabe et al., 2019; https://www.ebi.ac.uk/gwas/). However, they do not fully explain why some carriers of risk alleles do not develop the associated disorder and why people who carry similar risk alleles develop distinct phenotypes.

A critical challenge in medicine is to understand why patients diagnosed with the same disorder vary in their clinical presentation. This is especially true for Parkinson's disease (PD), for which the age of onset, rate of progression, and type and severity of symptoms differ among the 9.3 million people worldwide who live with this disorder (Maserejian et al., 2020). As the frequency of the misdiagnosis of PD is particularly high, ∼30% (Beach and Adler, 2018), and its consequences are dramatic, it is crucial to identify the aetiology of this clinical heterogeneity. One of the challenges of studying PD is that direct access to the relevant tissue, the brain, is limited. In addition, a long prodromal phase (Box 1) precedes the first clinical symptoms, and 90% of cases are considered sporadic, with an assumed genetic heritability of ∼30% (Keller et al., 2012).

Precision medicine investigates the plethora of pathophysiologies that are associated with a disorder (Robinson, 2012). Its goal is to offer the best medical care tailored to a patient at a given time. Precision medicine is thus contrasted by the traditional one-size-fits-all approach, whereby a certain treatment is given to all patients suffering from a certain disorder. Oncology was one of the first clinical specialities to adopt this approach (Kupstas et al., 2020; Nowakowski and Czuczman, 2015; Punt et al., 2017), by analysing the genomic landscape of cancer cells to identify cancer subtypes that respond well to certain treatments (Berland et al., 2019; Schmitz et al., 2018; van der Velden et al., 2019). Detailed data are gathered throughout a patient's life, which enables early risk stratification and monitoring of high-risk patients. Prevention and early intervention thus become possible. Furthermore, these detailed data allow the selection of the best available treatment approach for each patient. The prospect of treating PD through a precision medicine approach requires knowledge about the disease mechanisms and treatment targets. Deep phenotyping may aid in the acquisition of such knowledge by guiding clinical trial design and providing insights into disease stratification (Dorsey et al., 2020).

Recently, we have seen the emergence of large, deeply phenotyped cohorts for various disorders, in which valuable clinical, imaging, genetic and biometric data have been collected, often together with longitudinal monitoring, for example the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jones-Davis and Buckholtz, 2015) and the Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2018). Such datasets allow researchers to investigate disease aetiologies and the biomarkers of disease progression, and to identify risk factors. In particular, for PD and other complex disorders, with frequent misdiagnoses, unclear disease mechanisms and diverse clinical presentations, deep phenotyping presents the opportunity to fill these gaps in knowledge. Numerous deeply phenotyped PD cohorts are currently available to researchers; nothing comparable has as yet been developed for other genetic disorders.

In this Review, we therefore use PD as a paradigm to introduce deep phenotyping and demonstrate how it can advance precision medicine, in which treatments are tailored to genetically and phenotypically heterogeneous patients. We further discuss recent methodological advances that have allowed us to utilise and understand the large and ever-increasing amount of available data.

UK Dementia Research Institute at Cardiff University, Division of Psychological Medicine and Clinical Neuroscience, Haydn Ellis Building, Maindy Road, Cardiff CF24 4HQ, UK.

*Author for correspondence (sandorc@cardiff.ac.uk)

C.S., 0000-0002-8905-1052

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

Box 1. Glossary

Apathy: a lack of motivation.
Bradykinesia: a slowness of movement, one of the clinical hallmark symptoms of Parkinson's disease.
Classification: a group of supervised machine learning methods that predict a discrete outcome (category) based on a set of variables.
Clustering: a group of unsupervised machine learning methods that groups data points into clusters. Objects in the same cluster are similar to one another and less similar to objects from other clusters.
Convolutional neural network (CNN): a type of neural network often used to analyse visual imagery where neighbouring inputs are considered together.
Deep learning (DL): a subdivision of machine learning in which neural networks, which are computational methods inspired by biological neural networks, with multiple layers extract increasingly high-level features from raw input.
Dimensionality reduction: a group of unsupervised machine learning methods that transform high-dimensional data into a low-dimensional representation that retains most of the information.
Diplopia: simultaneous perception of two images from one object.
Dosage effect: Change in a phenotype due to alternations in the dose/amount of the product of a gene.
Dyskinesia: involuntary, uncontrolled muscle movements.
Dysphagia: difficulty swallowing.
Genome-wide association studies (GWAS): statistical, hypothesis-free methods to test for the association of genetic loci and phenotypic traits.
Hyposmia: reduced ability to smell.
Latent class: a group of unsupervised machine learning methods that relate observations to latent factors that are assumed to cause the observations.
Mendelian randomisation: a method to test for putative causal relationships between modifiable risk factors and diseases.
Molecular neuroimaging: techniques to visualise molecular or cellular processes in the brain through a probe or imaging agent that creates a signal through the interaction with the event of interest.
Neuronopathy: a subgroup of disorders of the peripheral nervous system that occur as a result of neuron degeneration.
Orthostatic hypotension: low blood pressure when standing up.
Polygenic risk score (PRS): a metric of disease risk given by the combined contribution of multiple genetic variants calculated from GWAS statistics.
Prodromal phase: a latent time period preceding the clinical diagnosis, in which symptoms appear but clinical diagnostic criteria are not yet met.
Quantitative trait: a measurable phenotype that varies between individuals on a continuous scale.
Regression: a group of supervised machine learning methods that predict a continuous outcome (real-valued) based on a set of variables.
Supervised machine learning: algorithms that learn a relationship between predictors and outcomes based on labelled data.
Swarm network: a conglomerate of individual sites with private data (nodes) that exchange model parameters.
Unsupervised machine learning: algorithms that identify patterns in unlabelled data.

In particular, we focus on the emergence of deeply phenotyped cohorts in response to advancements in genetic research.

The imprecise diagnosis and complex genetics of PD

Precision medicine emerged because the traditional one-size-fits-all approach has proven unsuccessful for many disorders. One such disorder is PD, which presents a unique challenge, as its diagnosis remains difficult and its genetic background is diverse. Conventional treatment approaches have thus far been unsuccessful, and a more targeted and personalised approach is required.

Imprecise diagnosis of PD

PD symptoms result from the progressive loss of dopaminergic neurons in a brain region called the substantia nigra, the primary function of which is motor control. A definitive diagnosis is often challenging, as PD can be confounded with other Parkinsonian syndromes (Williams and Litvan, 2013); however, PD can be distinguished from these by its prolonged response to dopaminergic medication (Williams and Litvan, 2013). Nevertheless, misdiagnosis occurs up to 30% of the time (Schrag et al., 2002). The consequences of these diagnostic errors are dramatic. A recent survey of 2000 people, conducted by the Parkinson's UK charity, revealed that 50% of misdiagnosed individuals with PD receive treatment for a non-existent condition and 6% undergo unnecessary operations or procedures (https://www.parkinsons.org.uk/news/poll-finds-quarter-people-parkinsons-are-wrongly-diagnosed).

In an effort to improve the accuracy of PD diagnoses, the diagnostic criteria for PD in 2015 were updated to include non-motor symptoms (Postuma et al., 2015). The new criteria further include guidance on the use of neuroimaging to rule out PD when no presynaptic dopaminergic deficiency is found. A molecular neuroimaging (Box 1) technique commonly used for this purpose is DaTscan, which involves the injection of a radioactive tracer (Ioflupane, 123-I-FP-CIT) that attaches itself to dopamine transporters on dopaminergic neurons (Djang et al., 2012). DaTscan can discriminate PD from essential tremors and from other non-degenerative tremors (Benamer et al., 2000) and can distinguish PD from healthy controls with high accuracy (Tagare et al., 2017). However, DaTscan cannot differentiate between PD and atypical Parkinsonian disorders, such as multiple system atrophy (MSA) or progressive supranuclear palsy (PSP), which show similar degenerative characteristics. Nevertheless, new magnetic resonance imaging (MRI) techniques, such as neuromelanin-sensitive MRI, and iron-sensitive MRI, are showing promising results that will help to make this distinction and that will further refine PD stratification and prognosis (Prange et al., 2019). Despite their utility for ruling out PD, no neuroimaging techniques are currently recommended for the routine diagnosis of PD. This might be because of the phenotypic heterogeneity of PD, even in brain imaging. For example, a group of patients with a clinical diagnosis of PD but no sign of a dopaminergic deficit in DaTscan has been identified, called scans without evidence of dopaminergic deficit (SWEDD). To date, it is debated whether SWEDD is an early form of PD, a misdiagnosis of clinical PD, or whether it is a distinct movement disorder (Erro et al., 2016; Lee et al., 2021; Marek et al., 2014). At present, the only certain means of diagnosis is the discovery, at autopsy, of depleted brainstem pigmented neurons with Lewy bodies, which are abnormal aggregations of α-synuclein and can be detected histologically. Hence, there is a strong clinical need for the development of accurate, *in vivo* tests at the earliest stages of the disease, for example molecular neuroimaging tracers to visualise α-synuclein (Shah et al., 2020), and these have recently been successfully developed such that future research in this area will soon clarify the distribution of α-synuclein in the brain (Capotosti et al., 2021)

One reason why PD is difficult to diagnose is because of the broad variation in its early clinical manifestations (Foltynie et al., 2002) (Fig. 1). Rapid eye movement sleep behaviour disorder (RBD) is a sleep condition characterised by the physical enactment of dreams that are vivid, intense and often violent. RBD often precedes PD, but not systematically before the first clinical motor symptoms (Postuma, 2014). At diagnosis, over 50% of dopaminergic neurons are already lost (Lang and Lozano, 1998),

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

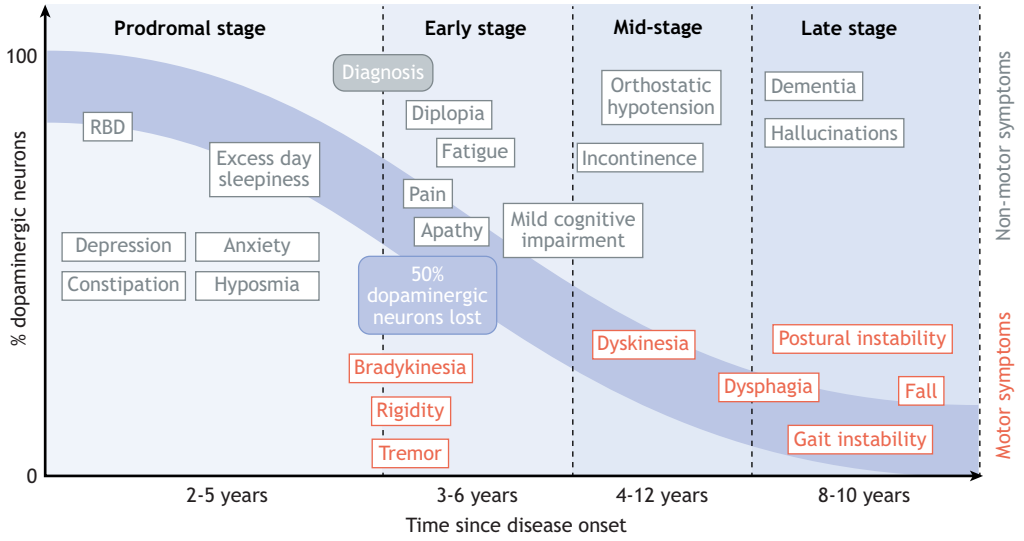

Fig. 1. Parkinson's disease (PD) is characterised by a high degree of heterogeneity. At diagnosis, >50% of dopaminergic neurons are already lost, and patients can show any combination of motor, neuropsychiatric and autonomic symptoms of differing severity. The blue area indicates the variability in the loss of dopaminergic neurons over time. RBD, rapid eye movement sleep behaviour disorder. See Glossary (Box 1) for descriptions of the symptoms.

and patients present with a diverse array of neurological, motor and autonomic impairments, each of which also demonstrate variable severity (Foltynie et al., 2002). Disease progression is also highly variable, not only in the rate of decline, but also in the development of additional impairments, such as dementia (Braak et al., 2005; Emre et al., 2007). Clinicians treat dementia with Lewy bodies (DLB) and Parkinson's disease dementia (PDD) as two distinct disease entities. DLB is diagnosed when cognitive impairment precedes Parkinsonian motor signs or begins within 1 year of its onset, whereas PDD develops within the setting of well-established PD (Fig. 5, see '1 year rule') (Jellinger, 2018; Jellinger and Korczyn, 2018). Although this timing distinction is often considered arbitrary, recent neuroimaging and post-mortem studies have demonstrated differences in the quantity and distribution patterns of Lewy bodies and α-synuclein between DLB and PDD, which suggest that these conditions have distinct aetiologies (Jellinger, 2018).

Imprecise genetics of PD

Most PD cases are currently thought to be sporadic but likely have a genetic component. The heritability of PD is estimated to be ∼26%, with over 90 common variants associated with sporadic PD (Nalls et al., 2019). Individually, these variants have low effects and no clinical utility. However, the overall genetic risk of developing PD can be calculated with polygenic risk scores (PRSs; Box 1). Although people in the highest decile of the PRS distribution are six times more likely to have PD compared to the rest of the population (Nalls et al., 2019), the distributions of PRSs for PD cases and controls are highly overlapping. This means that PRSs for PD currently have a low predictive value for diagnosis and are therefore of limited value for precision medicine.

Around 10% of PD patients have monogenic forms of the disease. The identification of these rare genetic forms was a key step in understanding PD mechanisms. The first identified monogenic cause of PD was a missense mutation within *SNCA*, which causes the p.A53T amino-acid substitution in the α-synuclein protein (Polymeropoulos et al., 1997). This missense is involved in the formation of Lewy bodies, the main hallmark of PD, but is not unique in causing PD. Rare duplications and triplications of *SNCA* (Singleton et al., 2003) also cause PD with a 'dosage effect' (Box 1): greater numbers of *SNCA* copies, causing increasing endogenous levels of α-synuclein, have been associated with earlier and more severe clinical symptoms (Eriksen et al., 2005). Moreover, several other PD-causing missense and multiplication mutations have been identified in the *SNCA* gene (Appel-Cresswell et al., 2013; Kruger et al., 1998; Rosborough et al., 2017; Zarranz et al., 2004). To understand the disease mechanisms associated with different genetic variants, we need to capture the full spectrum of variants associated with disease severity and identify the types of symptoms presented by individuals.

The most important and common risk factor for PD is loss-of-function mutations in the glucocerebrosidase gene (*GBA*). These mutations cause lysosomal accumulation of glucocerebroside due to a deficiency in the glucocerebrosidase enzyme, which leads to lysosomal dysfunction (Holleran et al., 1993), resulting in increased levels of α-synuclein via inhibition of the autophagic pathway (Du et al., 2015). However, a key challenge in developing drugs to target *GBA* impairments is that this same mutation can cause multiple disorders, and disease models do not accurately mimic the clinical effects of these mutations in humans (Fig. 2). In Gaucher's disease (GD), a multi-systemic metabolic disorder that typically manifests by adolescence, both homozygous and heterozygous *GBA* mutations increase the risk of developing PD (Riboldi and Di Fonzo, 2019). GD is categorised into three main subtypes with patients exhibiting varied clinical presentations. The most common subtype is non-neuronopathic (Box 1) Type I. The neuronopathic Type II subtype has an earlier onset and is more severe with acute neurological involvement, whereas the neuronopathic Type III subtype has a more chronic presentation (Alaei et al., 2019). There are over 300 pathogenic mutations in *GBA* that affect the structural stability of glucocerebrosidase and reduce its enzymatic activity (Smith et al., 2017). GD Type I patients are frequently associated with N370S *GBA* mutations, while Type II and Type III are typically associated with L444P mutations (Riboldi and Di Fonzo, 2019). However, both heterozygous and homozygous N370S mutations in *GBA* have been found among PD patients with no GD symptoms (Aharon-Peretz et al., 2004). Moreover, an elevated frequency of disease-associated *GBA* alleles has been found among individuals with RBD (Beavan et al., 2015; Gamez-Valero et al., 2018).

The phenotypic heterogeneity induced by these pathogenic variants is uniquely observed in humans. N370S and the L444P are the most frequent *GBA* variants linked to PD (Lesage et al., 2011).

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

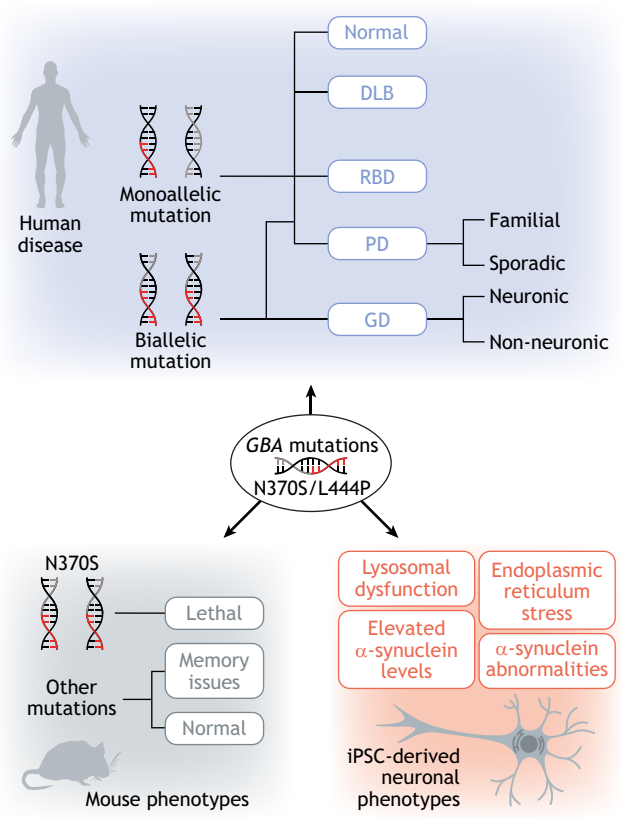

Fig. 2. *GBA* mutations: genetic heterogeneity in human versus disease models. *GBA* mutations are the most common genetic risk factor for Parkinson's disease (PD). However, *GBA* mutations are also found in different human disorders, such as Gaucher's disease (GD), dementia with Lewy bodies (DLB) and rapid eye movement sleep behaviour disorders (RBDs), and in healthy individuals. Different models of *GBA* mutations, including mouse and human induced pluripotent stem cell (iPSC)-derived neuron models, develop the same observable phenotypes. Together, this suggests that other genetic or non-genetic factors contribute to *GBA*-mutant PD in the human population.

In human induced pluripotent stem cell (iPSC) models, these variants consistently cause cellular abnormalities, such as dysfunctional autophagy in the endolysosomal pathway (Fernandes et al., 2016). However, in mouse models, N370S homozygosity is lethal at the neonatal stage of development (Xu et al., 2003), while most other mouse models of *GBA* mutations do not exhibit Parkinsonian phenotypes unless combined with a second risk factor, such as α-synuclein overexpression (Do et al., 2021; Farfel-Becker et al., 2019). This suggests that genetics alone is unable to explain the disease, because other genetic and non-genetic factors, including ageing, oxidative stress and epigenetics, modulate the clinical phenotype associated with *GBA* mutations in the human population.

LRRK2 encodes leucine-rich repeat kinase 2, also known as dardarin/PARK8. G2019S is the most common mutation associated with PD; in some populations, it can be found in 40% of people with this disorder (Lesage et al., 2010). This gain-of-function mutation in *LRRK2* has been associated with a higher risk of PD, and consequently LRRK2 inhibitors have been pursued as a potential avenue for PD treatment (Zhao and Dzamko, 2019). However, whether LRRK2 inhibition in patients is sufficient to reverse or to potentially prevent PD manifestation is currently debated. One reason for this scepticism is the incomplete genetic penetrance of the G2019S mutation, which suggests that other risk factors alter disease risk in carriers of this variant. Another reason is that some G2019S carriers exhibit the clinical manifestations of PD without developing Lewy bodies (Kalia et al., 2015).

Although PD patients with *LRRK2* and *GBA* mutations represent a small fraction of PD cases, they are nevertheless crucial for precision medicine. This is because by understanding the genetic and phenotypic heterogeneity of these mutations we can hopefully develop successful therapies to reverse their effects in PD patients. As sporadic patients represent the majority of PD cases, genomic data might also play an important role in extending the application of these therapies to sporadic patients with *GBA*/*LRRK2*-associated mutations, should these mutations be identified in this cohort. Indeed, some GBA mutations have already been reported to exacerbate disease outcome in sporadic patients and were associated with accelerated development of dementia and a more aggressive motor course (Stoker et al., 2020).

This illustrates that improving our understanding of genetic heterogeneity and how it corresponds to clinical variability should increase our ability to both predict disease and define subtypes by their aetiology, thus paving the way for more precise treatments (Hennekam and Biesecker, 2012). To achieve the aim of providing patients with tailored treatment that considers their unique genetic and phenotypic presentation, a deep understanding of the phenotypes and genetics of a disorder is needed. Precision medicine hence requires such fine-grained, deep data.

Precision medicine requires deep genomic and phenotypic data

As exemplarily presented for PD, a clinical disorder can be associated with diverse clinical phenotypes that render diagnosis difficult and with a diverse genetic background that renders treatment development and selection difficult. The one-size-fits-all approach does not account for such heterogeneity. Precision medicine could provide more tailored treatments, but achieving this requires deep understanding of the disorder, which necessitates in-depth data collection and analysis.

Why do we need better phenotypes?

The genetic and phenotypic heterogeneity observed within complex disorders complicates research. If a heterogeneous patient group is described by a single label, as often occurs in case-control studies, any subsequent analysis of this group will inherit the uncertainty and confounders from this broad diagnostic label. This has a negative impact on clinical practice, which relies on insights gained from such studies. To improve research outcomes, broad clinical labels should be replaced by sensitive, objective and detailed phenotypes.

Medical intervention research relies heavily on clinical trials, in which the effectiveness of a treatment is compared between groups. Given that the aforementioned broad diagnostic labels can capture multiple aetiologies, we may well see heterogeneous responses to treatment in a clinical trial, with only a small subset of patients showing a benefit (Fig. 3). The overall verdict in such cases would be that the treatment is not effective, despite its efficacy on a particular subset. This one-size-fits-all approach may partly explain why many clinical trials investigating disease-modifying drugs for PD have failed (Athauda and Foltynie, 2016). As such, clinical trials could greatly benefit from more granular stratifications of PD and from more personalised approaches.

When several treatments for a disorder successfully pass clinical trials, treatment selection becomes a difficult task for clinicians. Clinicians mostly rely on clinical expertise and general treatment

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

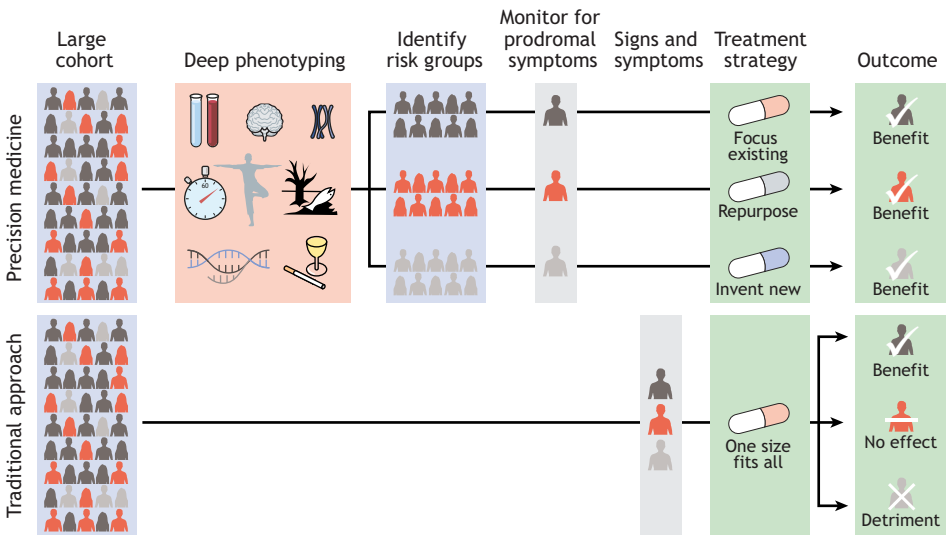

Fig. 3. The merit of deep phenotyping for precision medicine. In traditional clinical practice, the same treatment strategy is applied to anyone diagnosed with a disorder. This means that a diagnosis is made and a treatment given based on a predefined set of signs and symptoms. The outcome, including treatment effectiveness, can thus be varied due to heterogeneity within that disorder. Precision medicine uses fine-grained information gained through deep phenotyping and genetics to match the best treatment to an individual patient. In addition, it can aid in the monitoring and identification of at-risk individuals and enable preventative interventions.

guidelines when deciding which treatment to prescribe, creating a long journey of trial-and-error until a suitable treatment is found. Treatment selection could thus benefit from the insights of subtype-sensitive clinical trials and a general investigation into the association of treatment effect and biotypes. For example, clinical trials assessing treatments that target genetic forms of PD are starting to become more common (Mullin et al., 2020; Schneider and Alcalay, 2020), despite these forms being much rarer than sporadic PD. As precision medicine aims to consider the pathophysiological uniqueness of an individual as well as the genetic background, understanding the genetic basis of a disorder is an important first step in this direction.

Phenotypes that are accurate, sensitive and robust are important, as the quality of the measurement determines the utility of the analysis, especially so for genetic studies (O'Sullivan and Ioannidis, 2021). In a genetic association analysis, if the case group contains control and/or misdiagnosed subjects, disease-associated genetic loci may not be identified (Manchia et al., 2013). In addition, the selection of controls must be monitored closely as some controls, despite being healthy at study onset, may go on to develop a disorder at a later stage. In such a situation, we need better defined and more precise phenotypes to identify genetic associations rather than solely prioritising larger sample sizes for better statistical power (Manchia et al., 2013). Pastor (2012) identified two objectives for improving phenotype quality and hence the quality of genetic analysis: (1) the confirmation of clinical diagnoses through long-term follow up or additional biomarker tests and scans; and (2) the differentiation of sub-phenotypes or the usage of traits that more accurately reflect the disease spectrum. Quantitative traits (Box 1) have been shown to have better reproducibility in GWAS compared to binary traits, which often encompass broad diagnostic classes with inherent heterogeneity (O'Sullivan and Ioannidis, 2021). For example, the genetic associations for height measured in centimetres are more likely to be reproducible across different cohorts compared to genetic associations for the binary trait of being taller than 180 cm. Recent efforts to investigate the genetic basis of more precise phenotypes include studies that reveal the heritability of image-derived phenotypes (IDPs), like brain region volumes or cortical thickness measurements (Elliott et al., 2018). Thus, unbiased, objective and sensitive measures are needed to describe phenotypes.

With the emergence of next-generation sequencing, Hennekam and Biesecker (2012) foresaw the need for next-generation phenotyping back in 2012. A decade later, deeply phenotyped cohorts have become a major subject of interest for clinicians and medical geneticists, as we discuss next.

What is deep phenotyping?

In clinical practice, a phenotype is a label assigned to a specific set of observable traits, including, among others, morphological, physiological and/or behavioural traits (Robinson, 2012). Such traits can be inferred from medical history, questionnaires, clinical tests, blood tests, imaging and/or physical examinations (Fig. 4). Instead of reducing this highly complex set of traits into one disease label, deep phenotyping aims to retain this information. It tries to capture an individual's phenotypic presentation in a precise and comprehensive manner by leveraging information gained from different data sources (Robinson, 2012). These metrics are also monitored over time, instead of focusing on a single time point when a diagnosis is made (Weng et al., 2020). As a result, an individual's specific phenotype is described in all of its dimensions. Deep phenotyping thus offers a more complete picture of a disorder so that its nature, treatment and subtypes can be better understood (Dorsey et al., 2020) (Fig. 3).

Deep phenotyping also provides measures at different scales, such that the journey of a given protein can be followed from the level of genetics through omics and all the way through to manifestations in behaviour. A phenotypic assessment on different scales thus gives a better understanding of disease manifestations, their impact on daily life and their relation to pathophysiology. Combined with genetic data, which are becoming increasingly accessible due to falling costs, we can explore the heritability and true genetic basis of objective and precise phenotypes (O'Sullivan and Ioannidis, 2021). Such deep understanding can guide drug discovery and advance precision medicine in an objective and effective manner.

How can we analyse and utilise deep phenotyping data?

Traditional research primarily relies on hypothesis-driven approaches, using which specific data are gathered to answer one question. Data-driven approaches have gained popularity following advances in data collection, storage and computing. Deep phenotyping produces an abundance of high-dimensional data that can be leveraged to answer a multitude of questions. Such an abundance of data also allows for data-driven approaches (Goecks et al., 2020).

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

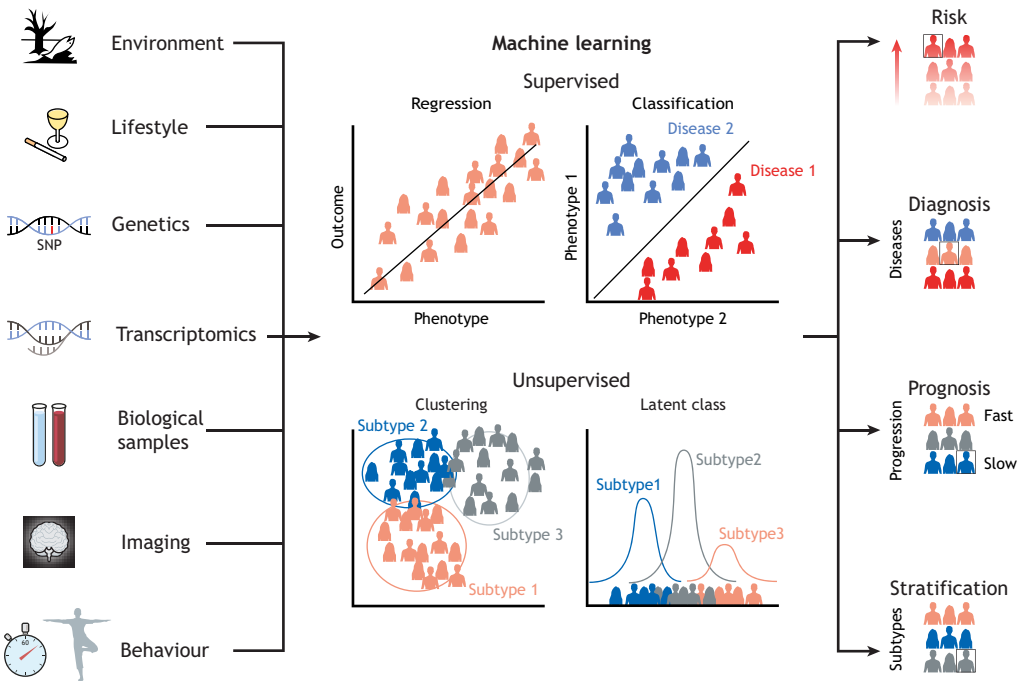

Fig. 4. Towards precision medicine by integrating multi-modal biomedical data. A number of the research objectives can be explored with genetic data and deep phenotyping. Deep phenotyping provides data on many different scales, such as environmental factors, lifestyle, multi-omics, diverse biological samples, imaging, behaviour, etc. Such complex data benefit from the advent of machine learning, such that large-scale, heterogeneous, multi-modal phenotypic and genetic data can be translated into meaningful information about risk, diagnosis, prognosis and stratification.

The abundant data generated by deep phenotyping, however, pose unique challenges to statistical methods. Deeply phenotyped cohorts offer a large amount of data for a comparably small number of individuals. Despite global efforts, including the sharing of data, to obtain a more representative number of participants, cohorts remain small in relation to the number of collected features; for example, the PPMI study collected 2442 measures on 1683 individuals. Therefore, special care must be taken when performing statistical tests and building models. Additionally, large amounts of missing, highly correlated or multi-modal data pose further challenges to conventional statistics in these cohorts. This is because standard statistical tools often fail to account for such data characteristics (Johnstone and Titterington, 2009).

Machine learning (ML) approaches are, therefore, required to handle such challenging data (Goecks et al., 2020). There are four broad applications of ML that are relevant for medicine: risk analysis, diagnosis, stratification and prognosis (Fig. 4). A major aim of precision medicine is to identify people at risk early, such that preventive measures can be applied. Risk analysis, as well as diagnosis, can be achieved through supervised ML (Box 1) techniques like regression (Box 1) and classification (Box 1) methods that reveal potential risk and protective factors (Solana-Lavalle and Rosas-Romero, 2021). Another aim of precision medicine is to provide tailored treatment to individual patients, which can be achieved by classifying patients into finely grained disease subtypes that respond better to certain treatments. Unsupervised ML (Box 1) methods like clustering (Box 1) approaches and latent class (Box 1) can reveal such subtypes (Brendel et al., 2021). Precision medicine also aims to identify treatment tailored to a specific stage of a disorder. Knowledge about these different aspects can be gained through disease progression modelling (Oxtoby et al., 2021).

Data generated through deep phenotyping therefore pose challenges; however, powerful methods, mainly from the field of ML, exist and are being developed to handle them. When such methods are successfully applied to rich and valuable datasets, we can answer important questions about disorders and thus advance precision medicine.

Data collected for deep phenotyping and the insights gained

Precision medicine requires deep phenotyping, and methods exist to handle and analyse such data to provide useful insights into disorders. Here, we discuss how such valuable data can be collected and what information can be gained through each modality.

Deep clinical phenotyping

Clinical phenotyping in PD often uses information from clinical tests, questionnaires or subjective descriptions of an individual's tremor to assign a disease label. Deep clinical phenotyping begins with a traditional clinical examination, in which such data are gathered, and then expands on this information with sensors that monitor patients over longer periods in real-life situations (Dorsey et al., 2020).

Traditional clinical examinations already provide information about phenotypic heterogeneity that can be leveraged to study subtypes. Collected data can include clinical tests and examinations for motor impairment, autonomic function and cognitive abilities, as well as questionnaires about mental health, sleep quality and problems with the activities of daily living. The identification of PD subgroups has been a research focus since 1990 (Jankovic et al., 1990). Instead of studying differences between cases and controls, differences between cases can be studied through data gathered in clinical examinations that are then analysed with ML methods. Early efforts focus on motor symptoms of PD assessed with the Unified Parkinson Disease Rating Scale (UPDRS) and differentiate three subtypes based on the ratio of the summed scores of specific domains: tremor dominant, postural instability and gait difficulty, or akinetic rigid and intermediate (Kang et al., 2005; Schiess et al., 2000). The inclusion of non-motor symptoms increases the stability and consistency of these subtypes (Ren et al., 2021). Efforts to include a broader range of clinical examinations and apply clustering techniques have identified discrete PD clinical subgroups, each displaying a characteristic set and degree of symptoms (Fereshtehnejad et al., 2015; Lawton et al., 2015). The subtypes revealed by Fereshtehnejad et al. (2015) have been subsequently shown to predict disease progression

(De Pablo-Fernandez et al., 2019). As Fereshtehnejad et al. (2017) noted, several studies reveal clinical subtypes identified through ML approaches, but no consensus has been established, nor have these methods been incorporated into clinical practice. This might be because the proposed subtypes and the methods used to define them lack consistency and stability (von Coelln et al., 2021). The inconsistency in such methods could be due to the selection of different variables for the model and the instability could be due to a selection bias introduced via data cleaning (Fereshtehnejad et al., 2017). Furthermore, the longitudinal aspects of PD have thus far been disregarded in most of these approaches, such that snapshots of patients at different stages of the disorder have been used. Owing to these issues, more fine-grained and consistently collected data that better represent the clinical phenotype or incorporation of other phenotype modalities and the application of methods that combine clustering and progression modelling (Young et al., 2018) could improve stratification efforts.

Data-driven PD diagnosis does not focus on traditional clinical examinations; instead, it uses these as the prediction target. As PD is a clinical diagnosis, a diagnosis based on clinical examinations is straightforward and does not require ML. However, a recent study showed that data gathered through other means, e.g. voice recordings, gait analysis, etc., have resulted in good prediction of PD using ML methods (Mei et al., 2021).

Digital sensors

Clinical phenotypes gathered through traditional clinical examinations have several limitations. First, detailed clinical tests and questionnaires have been criticised for their lack of precision (Regnault et al., 2019). Second, clinical tests are conducted at specific time points and only reveal snapshots of a person's phenotype. Such snapshots can be confounded by the increased phenotypic variability observed with ageing and with disease onset and progression (Sheridan et al., 2003). Third, detailed investigations by clinicians are time consuming and expensive. These specifically designed tests can take several hours and are conducted by trained staff, meaning that participants must attend clinics or be assessed at home. If an impairment is too advanced, patients may drop out due to the time investment and strain of the procedures (Dorsey et al., 2020). Finally, clinical tests are conducted in an artificial environment and do not accurately reflect real-world circumstances (Dorsey et al., 2020).

Digital sensors, which collect data and convert and transmit them digitally, can address these limitations (Brognara et al., 2019). Such devices tend to be more sensitive and accurate than traditional approaches. They also enable long-term data collection, which can provide a clearer picture of phenotypes and their trends by averaging measures over longer periods of observation (Hayes et al., 2008). In addition, these digital sensors allow automatic, non-disruptive data collection, in a real-world setting that does not depend on experienced staff. For example, in addition to extensive biannual assessments, the Personalized Parkinson Project (PPP) collects day-to-day real-world data through a wearable smart device known as the Verily Study Watch (Bloem et al., 2019). Participants are asked to wear the device all day throughout the 2 years of the study. This multi-sensor device collects data about acceleration, pulse rate, electrodermal activity, electrocardiogram, relative humidity, environmental temperature and ambient light level. Preliminary analyses show that such digital data have promising features that discriminate healthy controls from PD patients and that sensitively describe motor symptom progression (Schlachetzki et al., 2017; Shah et al., 2020).

Digital sensors provide large amounts of data and thus power for statistical analyses: a considerable number of observations are acquired per person per second over a long time. Such sensors can also be worn by anyone and are relatively inexpensive and non-invasive. For comparison, polysomnography monitors the sleep of a single person over a single night in an artificial sleep laboratory, which is both costly and an inconvenience for the participant. By contrast, wrist-worn accelerometers can provide data about sleep for many participants over several nights at home (Sundararajan et al., 2021). Although the sleep features assessed by wearable sensors do not match polysomnographies perfectly, they provide valuable and valid information about numerous clinical features about sleep, steps taken, physical activity, distance, etc. for many people (Evenson et al., 2015), and thus help us gain longitudinal insights into impairments in everyday life (Johansson et al., 2018).

Biomarkers and intermediate phenotypes

One strength of deep phenotyping is that it captures phenotypes at different scales and enables the study of biomarkers, which are endogenous, measurable, characteristics that mark either the risk for, or the manifestation of, a disease. Biomarkers allow deeper understanding of ongoing changes in disease pathology, from the molecular to the behavioural level. For example, changes in the brain can be detected via medical imaging, while cellular perturbations can be detected through omics measures.

These quantitative traits can be used to study the differences between clinically defined groups. However, like in genetic analyses, the inherent uncertainty and imprecision of binary disease labels affect such studies, especially in neurodegenerative disease research (Mattsson-Carlgren et al., 2020). An alternative approach is to objectively identify homogeneous groups based on biomarkers and then explore the association between these groups and clinical phenotypes (Espay et al., 2017) (Fig. 5). Methods like Mendelian randomisation (Box 1) can help identify causal links between genes and environmental factors or biomarkers (Noyce and Nalls, 2016). Thus far, a limited number of biomarkers have shed light on the neuropathophysiology of disease subtypes and have been helpful for monitoring disease progression and predicting its course. For example, the cerebrospinal fluid (CSF) biomarkers amyloid-β (Aβ42), total tau and phosphorylated tau can serve as early markers of Alzheimer's disease and thus provide clinically relevant diagnostic information (Blennow and Zetterberg, 2018). Other biomarker modalities that assess molecular markers, like CSF and blood, or positron-emission tomography (PET) have shown great prospects in understanding disease mechanisms and spreading of pathologies in neurodegeneration (Lashley et al., 2018).

Blood and CSF biomarkers

The Alzheimer's disease examples highlighted above (Blennow and Zetterberg, 2018; Lashley et al., 2018) show that blood and CSF can be useful sources of biomarkers for neurodegenerative disorders. To aid prognostic and diagnostic decision-making, biochemical markers of early PD have also been extensively studied. However, no single marker has so far been sufficient to accurately diagnose PD. For example, astrocytic cell death in PD can be detected by elevated blood and CSF levels of glial fibrillary acidic protein (Ding et al., 2021). However, this signature is also observed in MSA, PSP and corticobasal degeneration, thus complicating the differentiation of typical and atypical Parkinsonian disorders (Constantinescu et al., 2010). Conversely, neurofilament light protein levels can

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

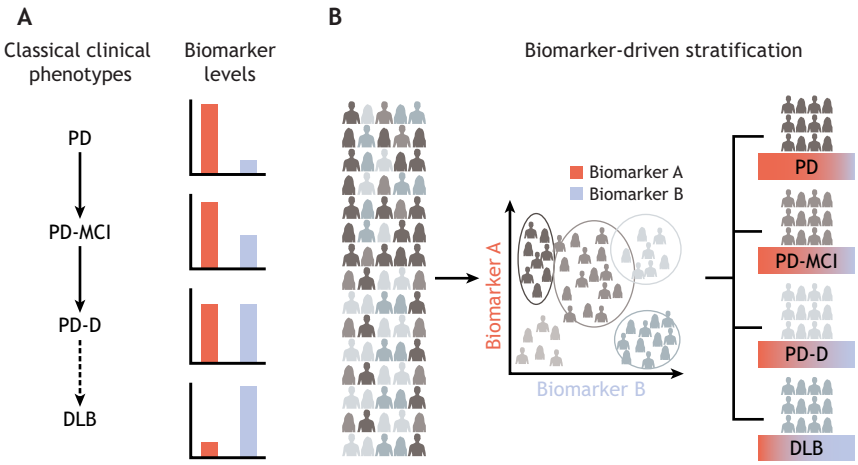

Fig. 5. Clinical phenotype-driven versus biomarker-driven research. (A) Biomarkers are useful for identifying differences between clinical phenotypes and clinical subgroups, and in providing a differential diagnosis. (B) Biomarkers can also differentiate disease subtypes, which can then be associated with clinical phenotypes and behaviour. For example, in patients with pure synucleinopathy, we expect to only see PD-specific biomarkers (red), whereas in those with AD co-pathologies, we expect abnormalities in AD-specific biomarkers as well (blue). The distinction between DLB and PDD is defined by the '1-year rule': if the onset of dementia symptoms is within 1 year of parkinsonism, the disorder is called DLB; if parkinsonism is present for more than 1 year before the onset of dementia, the disorder is called PDD. AD, Alzheimer's disease; DLB, dementia with Lewy bodies; MCI, mild cognitive impairment; PD, Parkinson's disease; PD-D, Parkinson's disease dementia.

distinguish PD from PSP and MSA, but are unable to distinguish between PSP and MSA (Constantinescu et al., 2010). As such, a repertoire of biomarkers needs to be studied simultaneously to aid the diagnosis of PD (Schapira, 2013).

The loss of dopaminergic neurons in PD likely involves inflammation, either as a cause or a consequence (Appel, 2012). A recent study reported the appearance of an α-synuclein-reactive T-cell population in the blood 10 years prior to diagnosis with motor PD (Lindestam Arlehamn et al., 2020). This result suggests that it might be possible to identify modifications in the blood of individuals prior to developing symptomatic PD or another α-synucleinopathy, such as DLB and MSA. As such, changes in the blood's transcriptome, as obtained by RNA profiling, could identify novel PD biomarkers.

Brain imaging biomarkers

Brain imaging offers rich, detailed *in vivo* data that can assist with differential diagnosis, prognosis and subtyping (Pagano et al., 2016). Various imaging modalities exist that can investigate structural, functional and molecular changes in diseased brains.

Structural MRI with T1 weighting is the most commonly available standard brain imaging resource in deeply phenotyped

Table 1. Overview of deeply phenotyped cohorts for Parkinson's disease

Study		PPMI	PPP	Luxembourg Parkinson's Study	OPDC, Discovery	CCBP	Fox Insight
Sample size	*n* cases	1400	650	800	900	4000 (also AD)	22205
	n controls	200		800	200	1000	8231
Timing	Duration (years)	>10	2	4	>5	>5	
	Frequency (months)	3-12	12	12	18	12	3-12
Clinical measures	Motor	✓	✓	✓	✓	✓	✓
	Non-motor	✓	✓	✓	✓	✓	✓
	Neuro-psychological	✓	✓	✓	✓	✓	✓
	Daily living (ADLs)		✓	✓	✓	✓	✓
	Gait			✓			
	Voice				✓		
Digital tools	Accelerometer	✓	✓		✓		
	Pulse rate	✓	✓			✓	
	ECG	✓	✓				
	Gait sensor			✓			
	Microphone				✓		
	Sleep					✓	
	Touch screen				✓		
Biospecimen	CSF	✓	✓				
	Blood	✓	✓	✓	✓	✓	
	Stool		✓			✓	
	Urine	✓		✓			
Genomics	Genotype	✓	✓	✓	✓	✓	✓
	WES/WGS	✓			✓		
	RNA sequencing	✓					
Imaging	MRI	✓	✓		✓	✓	
	DaTscan	✓			✓		

Various studies with different goals have collected a rich amount of data to study PD. These studies share many data modalities that can be merged in data-sharing efforts. Longitudinal PD cohorts that incorporate clinical measures, digital tools, biological samples and imaging are highlighted here. AD, Alzheimer's disease; ADLs, activities of daily living; CSF, cerebrospinal fluid; ECG, electrocardiography; MRI, magnetic resonance imaging; WES, whole-exome sequencing; WGS, whole-genome sequencing.

Data sources: Parkinson Progression Marker Initiative (PPMI) (Marek et al., 2018), the Personalized Parkinson Project (PPP) (Bloem et al., 2019), the Luxembourg Parkinson's Study (Hipp et al., 2018), the Oxford Parkinson Discovery Centre (OPDC), Discovery cohort (Griffanti et al., 2020), the Cincinnati Cohort Biomarker Program (CCBP) (Sturchio et al., 2020) and the Fox Insight Study (Smolensky et al., 2020).

cohorts. Structural imaging measures from such cohorts have revealed neuroanatomical PD subgroups that correspond to clinical subtypes and that can predict disease progression (Shu et al., 2021; Wang et al., 2020). Some cohorts offer molecular imaging data that can be used to research the spreading patterns of proteins in the brain. Some cohorts (Table 1) include DaTscan imaging, which can shed light on the SWEDD subgroup (Choi et al., 2017). In general, molecular imaging has been used to investigate the spreading pattern of α-synuclein in PD (Horsager et al., 2020) and to identify distinct subtypes. Such insights are valuable for precision medicine as the identification of biotypes, which are clusters of individuals that share biological signatures, can inform treatment responses in several disorders, such as cancer and Alzheimer's disease (Cattaneo et al., 2016; Machado et al., 2020).

Imaging data from deeply phenotyped cohorts are becoming increasingly available to non-imaging experts in the form of IDPs, which provide a great tool for studying the brain. IDPs summarise high-dimensional data as informative, subject-level measures of thickness, volume, connectivity or protein levels. These measures can either be curated from expert knowledge or acquired in a data-driven objective manner (Gong et al., 2021). IDPs can be linked to genetics to reveal the genetic contributions to brain abnormalities relevant for psychiatric and neurological disorders, as well as for ageing. For example, a GWAS with IDPs using data from the UK Biobank has shown that many brain characteristics are heritable and some genes, such as *EGF*, are associated with brain lesions (Elliott et al., 2018).

Cohorts and data sharing

Several PD cohorts now exist that provide data from diverse sources, enabling research into the complexity of the disease. Such cohorts include the aforementioned PPMI (Marek et al., 2018), the Oxford Parkinson Discovery Centre Discovery cohort (Griffanti et al., 2020), the PPP (Bloem et al., 2019), the Cincinnati Cohort Biomarker Program (Sturchio et al., 2020), the Luxembourg Parkinson's Study (Hipp et al., 2018) and the Fox Insight Study (Smolensky et al., 2020) (Table 1). Although these cohorts follow different objectives, they share a vast amount of common data modalities that could be merged to increase the sample size, e.g. for GWAS. Data-sharing efforts are needed to create larger, more unbiased population samples that better capture heterogeneity, especially in terms of genetics. Platforms like the Dementia Platforms UK (Koychev et al., 2020) are set up to combine data from several cohorts into a standardised framework. A similar tool for PD is still required, despite several efforts and calls for it, for example by the BioLoC-PD working group (Heinzel et al., 2017). However, some efforts to combine and harmonise PD cohorts do exist, such as the Accelerating Medicines Partnership Parkinson's Disease platform (Iwaki et al., 2021).

As medical data are sensitive and require protection, indirect ways in which to securely share such data are being explored. Instead of defining data-sharing agreements between study sites, decentralised approaches, such as swarm learning (SL), can be followed. SL does not require data exchange or a central structure. Instead, parameters are trained by local models on local data and are included in a swarm network (Box 1) that consists of multiple local sites (Warnat-Herresthal et al., 2021).

In addition to the technical and legal challenges of data sharing, data storage, analysis and transferability issues are also a concern. These are discussed in the following section.

Challenges and opportunities

New techniques may enable the faster and easier collection of large amounts of data, but they also pose new challenges in the curation, integration, sharing and interpretation of the data (Fig. 6). Data collection and storage require agreed-upon standards and global efforts. The analysis of these data is complicated by the sheer heterogeneity, multi-modality and scale of the data. Furthermore, one of the biggest challenges lies in transforming these complex data into medically actionable resources with clinical utility. Here, we focus on the data analysis aspect, as the other elements have recently been reviewed elsewhere (Matrana and Campbell, 2020; Weng et al., 2020).

How to merge data modalities?

Deep phenotyping produces different data modalities that have to be studied together. The classical method used to integrate different data sources is to merge different modalities into a single matrix. However, this can introduce a bias, as higher-dimensional data (e.g. imaging), unlike lower-dimensionality data (e.g. demographics), will often be preferred by algorithms simply because of their size and thus larger influence. Markello et al. (2021) have proposed

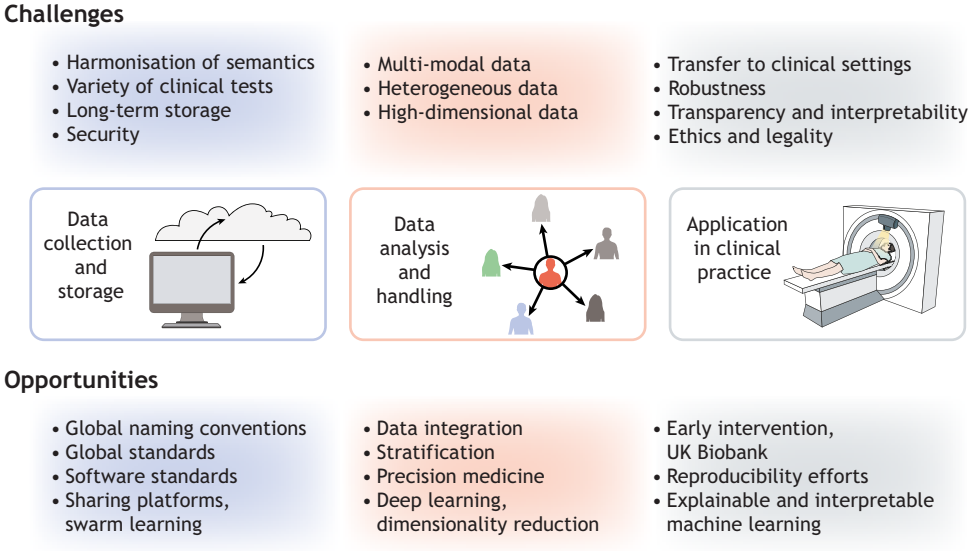

Fig. 6. Overview of challenges and resulting opportunities. Deep phenotyping produces large amounts of data, which present various challenges in three domains: data storage, analysis and application. However, these challenges give us the opportunity to set global standards and, once the infrastructure is in place, to gain valuable novel insights into disorders that can guide us to precision medicine.

creating a similarity network for each data modality separately and then iteratively fusing these together. This similarity network fusion approach has the advantage of overcoming the dimensionality bias, plus it generates a low-dimensional representation of each source that can be interpreted.

Another issue is the noise associated with each modality. Gene network methods are widely used to identify perturbed molecular pathways that underlie complex genetic disorders. A common limitation of all these approaches stems from the functional datasets themselves, as no single dataset or data modality can provide a complete picture of the functional association between genes. Although it is possible to merge different datasets and to build a more comprehensive and unique network, different levels of noise from each data source get incorporated into the result, which can lead to false associations. Honti et al. proposed a powerful method to address this issue by weighting functional similarities between genes according to their likelihood of influencing the same mammalian phenotype(s) (Honti et al., 2014; Sandor et al., 2017).

Analysing imaging and genetic data

New fields of study have emerged that deal with the combination of different data modalities. Imaging genetics integrates analysis of brain imaging and genomics to gain insights into the genetic impact on brain function and structure. In early efforts, a single genetic marker (e.g. a single-nucleotide polymorphism) and a single imaging trait were studied, and then PRSs and multiple IDPs were studied together (Shen and Thompson, 2020). Today, methods that integrate multiple single-nucleotide polymorphisms and multiple traits exist that model the influence of genetic variation on several IDPs. To decrease the amount of parameters needed to fit such univariate models, sparse multivariate models have emerged in which all features are integrated into a single large model. This further allows us to model the relationship between genetics and phenotypes while accounting for dependencies between phenotypes (Nathoo et al., 2019).

Analysing imaging and transcriptomic data

Analogously to imaging genetics, imaging transcriptomics deals with the integrated analysis of brain imaging and gene expression data to gain insights into the molecular changes associated with neurodegeneration. As omics offer a dynamic dimension, disease progression can be followed by using imaging transcriptomics (Katrib et al., 2016). A common approach is to correlate gene expression with IDPs through shared defined regions of interest (Mroczek et al., 2021). This means that a discrete map is applied to the brain in which all measures in one region are summarised to represent that region. This is largely made possible through the publicly available dataset from the Allen Human Brain Atlas (https://human.brain-map.org/), which holds gene expression data for 102 brain regions and for 20,000 genes of six post-mortem brains from healthy donors (Hawrylycz et al., 2012). Guidelines to handle this dataset have been proposed to standardise the research in this emerging field (Arnatkeviciute et al., 2019). Recent efforts have attempted to utilise the spatial resolution of both data sources (Zarkali et al., 2021).

Heterogeneity as an opportunity

In disease research, the problems arising from high heterogeneity should also be viewed as an opportunity. Past efforts to investigate PD have led neither to a successful understanding of the disorder nor to a disease-modifying treatment. Thus embracing data heterogeneity and investigating it could provide new insights. Several methods for uncovering heterogeneity in large datasets have been proposed. We can classify these approaches as subtype and stage models: subtype models focus on finding homogeneous subgroups while ignoring the disease stage; stage models ignore subtype heterogeneity but investigate the disease stage. The SuStaIn model (Young et al., 2018) combines both of these efforts by integrating clustering and disease progression modelling. It has successfully shed light on Alzheimer's disease subtypes based on the spreading of phosphorylated tau (Vogel et al., 2021) and could inform spreading patterns and progression subgroups in PD as well.

How to handle large data?

Deep phenotyping combined with genetic data has led to the generation of unprecedented amounts of data, which comes with its own set of challenges. First, the storage and handling of high-dimensional data is very computationally demanding. This issue is typically addressed through the use of high-performance computing systems, by sharing resources among research institutes and by using cloud-based systems (Bauermeister et al., 2020). Second, the analysis of high-dimensional data requires large sample sizes to provide sufficient statistical power. Data-sharing efforts and the decreasing costs of data collection are helping in this regard (Iwaki et al., 2021). Third, appropriate methods to process such data need to be developed and applied. Typically, dimensionality reduction (Box 1) techniques are used to extract meaningful features that can be interpreted (Tao et al., 2017). An alternative approach is deep learning (DL; Box 1). Despite its debated role in medicine due to a lack of transparency and model interpretation, DL is gaining popularity for its ability to handle high-dimensional datasets. Especially in medical imaging, convolutional neural networks (CNNs; Box 1) are often applied with good results (Choi et al., 2017). The concerns regarding transparency are being addressed through the branch of interpretable and explainable artificial intelligence that, for example, generates visualisations of the decision process, such that physicians can review the decision made by the model (Magesh et al., 2020).

From deeply phenotyped cohorts to the general population

Although deeply phenotyped cohorts constitute a unique opportunity for precision medicine, the collection and analysis of certain data modalities are time consuming for clinicians, patients and researchers alike. This means that participation cannot be extended to the general population nor to cohorts that include hundreds of thousands of individuals. One such resource-heavy data modality is the definitive diagnosis of RBD using polysomnography, which monitors various body functions during sleep in a specialised clinic (Hogl and Stefani, 2017). Identifying RBD in the general population is crucial as RBD patients that carry severe *GBA* variants show faster transition to PD and dementia (Krohn et al., 2020). Fortunately, *in vitro* models have also highlighted a possible alternative diagnostic tool for RBD based on findings that implicate lysosomal storage dysfunction as an early marker of GBA deficiency (Bae et al., 2015). These insights, combined with data resources such as the UK Biobank, offer a unique opportunity to identify severe *GBA* variant carriers or individuals with sleep disorders. The UK Biobank provides a broad range of phenotypic data, including cognitive and sleep measures, digital markers like movement recorded by smartwatch accelerometers, genetic information and, more recently, blood proteomic profiles for over 500,000 adults aged 37-73 years (Bycroft et al., 2018). To expand the search for individuals in the prodromal phase of RBD in the general population, we need to

Disease Models & Mechanisms (2022) 15, dmm049376. doi:10.1242/dmm.049376

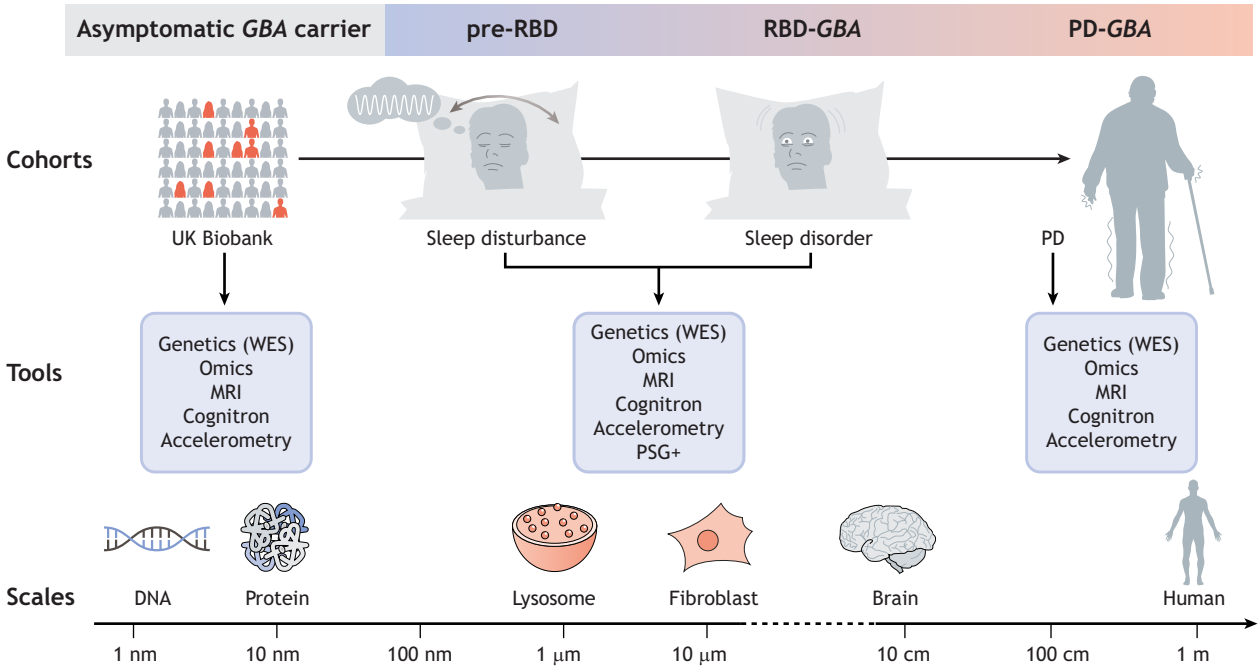

Fig. 7. A model for precision medicine in diagnosing and treating PD. To evaluate how pharmacological interventions might reverse the early (pre)clinical symptoms of PD, the features of *in vitro* disease models, such as lysosomal dysfunction in fibroblasts with *GBA* mutations, need to be linked to the phenotypes of RBD/PD patients, such as their sleep and biomarker profiles. The UK Biobank population could also be profiled to identify the earliest features of RBD and thus help more at-risk patients. Cognitron is an artificial intelligence tool to evaluate mental skills of an individual (Hampshire et al., 2021; https://www.cognitron.co.uk/). *GBA*, glucocerebrosidase gene; MRI, magnetic resonance imaging; PD, Parkinson's disease; PSG, polysomnography; RBD, rapid eye movement sleep behaviour disorder; WES, whole-exome sequencing.

assemble such diverse sources of data across different scales, spanning *in vitro* cellular models and clinical cohorts, as well as the general population (Fig. 7). Additionally, using novel computational approaches, features of *in vitro* cellular models have to be linked to biomarker profiles of deeply phenotyped patients on an individual level and expanded to the UK Biobank population to identify the earliest features of disease.

Precision medicine has been criticised for its lack of clinical utility and its failure to address the demands of public health (Ramaswami et al., 2018). With key exceptions, such as the UK Biobank, which combines principles of deep phenotyping with general health aims (Bycroft et al., 2018), most deeply phenotyped cohorts focus on one disorder and explore it in depth with a strong emphasis on clinical interventions and drug research. On the surface, this focus on one disorder can be regarded as a financial investment without much benefit for the general population. However, methods developed for deeply phenotyped cohorts, and the research insights into disease mechanisms and risks, provide valuable information for the general public. Furthermore, identifying disorders earlier could achieve a shift from treatment to prevention, which would greatly benefit the general population.

Future perspectives

Deeply phenotyped cohorts offer tremendous opportunities to advance our understanding of complex disorders. The observed phenotypic and genetic heterogeneity of such diseases must be addressed to understand their underlying mechanisms and to provide targeted treatments. High-throughput sequencing technologies provide insights into genetic heterogeneity, while deep phenotyping provides insights into phenotypic heterogeneity via clinical (and intermediate biological) phenotypes. These complex data challenge traditional statistical methods, but advances in ML and data-sharing efforts show how such data can be translated into meaningful and clinically valuable information.

One of the biggest challenges is to transfer disease insight captured in deeply phenotyped cohorts to the general population and dissect the prodromal phase of a disorder. Gathering as much in-depth data from the general population is not feasible, but it is done for deeply phenotyping cohorts. Therefore, novel approaches to transfer our insights to clinical practice are needed. As most cohorts focus on specific disorders, they provide limited merit to the general population. Nevertheless, such shortcomings can be addressed by the wealth of data provided by public resources, such as the UK Biobank, that pose a unique opportunity to align clinical cohorts to the general population. This would require diverse skills and expertise in diverse areas, including clinical, cellular, genomic, pharmacological, computational and artificial intelligence, to come together and embark on interdisciplinary collaboration to push the boundaries of scientific research. Therefore, through combined efforts of industry and academia, the goal of precision medicine is reachable for PD and other complex disorders.

Acknowledgements

We thank Prof. Caleb Webber and Samuel Keat for constructive criticism of the manuscript.

Competing interests

The authors declare no competing or financial interests.

Funding

C.S. and N.R. are supported by the Ser Cymru II programme, which is part-funded by Cardiff University and the European Regional Development Fund through the Welsh Government. A.S. is supported by a PhD studentship funded by Health and Care Research Wales. C.S., N.R. and J.M.-S. are supported by the UK Dementia Research Institute, which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

References

Aharon-Peretz, J., Rosenbaum, H. and Gershoni-Baruch, R. (2004). Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. *N. Engl. J. Med.* **351**, 1972-1977. doi:10.1056/NEJMoa033277

Alaei, M. R., Tabrizi, A., Jafari, N. and Mozafari, H. (2019). Gaucher disease: new expanded classification emphasizing neurological features. *Iran J. Child Neurol.* **13**, 7-24. doi:10.17650/2073-8803-2018-13-4-7-22

Appel, S. H. (2012). Inflammation in Parkinson's disease: cause or consequence? *Mov. Disord.* **27**, 1075-1077. doi:10.1002/mds.25111

Appel-Cresswell, S., Vilarino-Guell, C., Encarnacion, M., Sherman, H., Yu, I., Shah, B., Weir, D., Thompson, C., Szu-Tu, C., Trinh, J. et al. (2013). Alpha-synuclein p.H50Q, a novel pathogenic mutation for Parkinson's disease. *Mov. Disord.* **28**, 811-813. doi:10.1002/mds.25421

Arnatkeviciute, A., Fulcher, B. D. and Fornito, A. (2019). A practical guide to linking brain-wide gene expression and neuroimaging data. *Neuroimage* **189**, 353-367. doi:10.1016/j.neuroimage.2019.01.011

Athauda, D. and Foltynie, T. (2016). Challenges in detecting disease modification in Parkinson's disease clinical trials. *Parkinsonism Relat. Disord.* **32**, 1-11. doi:10.1016/j.parkreldis.2016.07.019

Bae, E. J., Yang, N. Y., Lee, C., Lee, H. J., Kim, S., Sardi, S. P. and Lee, S. J. (2015). Loss of glucocerebrosidase 1 activity causes lysosomal dysfunction and alpha-synuclein aggregation. *Exp. Mol. Med.* **47**, e153. doi:10.1038/emm.2014.128

Bauermeister, S., Orton, C., Thompson, S., Barker, R. A., Bauermeister, J. R., Ben-Shlomo, Y., Brayne, C., Burn, D., Campbell, A., Calvin, C. et al. (2020). The dementias platform UK (DPUK) data portal. *Eur. J. Epidemiol.* **35**, 601-611. doi:10.1007/s10654-020-00633-4

Beach, T. G. and Adler, C. H. (2018). Importance of low diagnostic Accuracy for early Parkinson's disease. *Mov. Disord.* **33**, 1551-1554. doi:10.1002/mds.27485

Beavan, M., McNeill, A., Proukakis, C., Hughes, D. A., Mehta, A. and Schapira, A. H. (2015). Evolution of prodromal clinical markers of Parkinson disease in a GBA mutation-positive cohort. *JAMA Neurol.* **72**, 201-208. doi:10.1001/jamaneurol.2014.2950

Benamer, T. S., Patterson, J., Grosset, D. G., Booij, J., de Bruin, K., van Royen, E., Speelman, J. D., Horstink, M. H., Sips, H. J., Dierckx, R. A. et al. (2000). Accurate differentiation of parkinsonism and essential tremor using visual assessment of [123I]-FP-CIT SPECT imaging: the [123I]-FP-CIT study group. *Mov. Disord.* **15**, 503-510. doi:10.1002/1531-8257(200005)15:3<503::AID-MDS1013>3.0.CO;2-V

Berland, L., Heeke, S., Humbert, O., Macocco, A., Long-Mira, E., Lassalle, S., Lespinet-Fabre, V., Lalvee, S., Bordone, O., Cohen, C. et al. (2019). Current views on tumor mutational burden in patients with non-small cell lung cancer treated by immune checkpoint inhibitors. *J. Thorac. Dis.* **11**, S71-S80. doi:10.21037/jtd.2018.11.102

Blennow, K. and Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: current status and prospects for the future. *J. Intern. Med.* **284**, 643-663. doi:10.1111/joim.12816

Bloem, B. R., Marks, W. J., Jr, Silva de Lima, A. L., Kuijf, M. L., van Laar, T., Jacobs, B. P. F., Verbeek, M. M., Helmich, R. C., van de Warrenburg, B. P., Evers, L. J. W. et al. (2019). The Personalized Parkinson Project: examining disease progression through broad biomarkers in early Parkinson's disease. *BMC Neurol.* **19**, 160. doi:10.1186/s12883-019-1394-3

Braak, H., Rub, U., Jansen Steur, E. N., Del Tredici, K. and de Vos, R. A. (2005). Cognitive status correlates with neuropathologic stage in Parkinson disease. *Neurology* **64**, 1404-1410. doi:10.1212/01.WNL.0000158422.41380.82

Brendel, M., Su, C., Hou, Y., Henchcliffe, C. and Wang, F. (2021). Comprehensive subtyping of Parkinson's disease patients with similarity fusion: a case study with BioFIND data. *NPJ Parkinsons Dis.* **7**, 83. doi:10.1038/s41531-021-00228-0

Brognara, L., Palumbo, P., Grimm, B. and Palmerini, L. (2019). Assessing gait in Parkinson's disease using wearable motion sensors: a systematic review. *Diseases* **7**, 18. doi:10.3390/diseases7010018

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209. doi:10.1038/s41586-018-0579-z

Capotosti, F., Vokali, E., Molette, J., Ravache, M., Delgado, C., Kocher, J., Pittet, L., Dimitrakopoulos, I. K., Di-Bonaventura, I., Touilloux, T. et al. (**2021**). The development of [^{18}F]ACI-12589, a high affinity and selective alpha-synuclein radiotracer, as a biomarker for Parkinson's disease and other synucleinopathies. *Alzheimers Dement.* **17**, e053943. doi:10.1002/alz.053943

Cattaneo, G. M., Bettinardi, V., Mapelli, P. and Picchio, M. (2016). PET guidance in prostate cancer radiotherapy: quantitative imaging to predict response and guide treatment. *Phys. Med.* **32**, 452-458. doi:10.1016/j.ejmp.2016.02.013

Choi, H., Ha, S., Im, H. J., Paek, S. H. and Lee, D. S. (2017). Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *Neuroimage Clin.* **16**, 586-594. doi:10.1016/j.nicl.2017.09.010

Constantinescu, R., Rosengren, L., Johnels, B., Zetterberg, H. and Holmberg, B. (2010). Consecutive analyses of cerebrospinal fluid axonal and glial markers in Parkinson's disease and atypical Parkinsonian disorders. *Parkinsonism Relat. Disord.* **16**, 142-145. doi:10.1016/j.parkreldis.2009.07.007

De Pablo-Fernandez, E., Lees, A. J., Holton, J. L. and Warner, T. T. (2019). Prognosis and neuropathologic correlation of clinical subtypes of parkinson disease. *JAMA Neurol* **76**, 470-479. doi:10.1001/jamaneurol.2018.4377

Ding, Z. B., Song, L. J., Wang, Q., Kumar, G., Yan, Y. Q. and Ma, C. G. (2021). Astrocytes: a double-edged sword in neurodegenerative diseases. *Neural Regen. Res.* **16**, 1702-1710. doi:10.4103/1673-5374.306064

Djang, D. S., Janssen, M. J., Bohnen, N., Booij, J., Henderson, T. A., Herholz, K., Minoshima, S., Rowe, C. C., Sabri, O., Seibyl, J. et al. (2012). SNM practice guideline for dopamine transporter imaging with 123I-ioflupane SPECT 1.0. *J. Nucl. Med.* **53**, 154-163. doi:10.2967/jnumed.111.100784

Do, J., Perez, G., Berhe, B., Tayebi, N. and Sidransky, E. (2021). Behavioral phenotyping in a murine model of GBA1-associated parkinson disease. *Int. J. Mol. Sci.* **22**, 6826. doi:10.3390/ijms22136826

Dorsey, E. R., Omberg, L., Waddell, E., Adams, J. L., Adams, R., Ali, M. R., Amodeo, K., Arky, A., Augustine, E. F., Dinesh, K. et al. (2020). Deep phenotyping of Parkinson's disease. *J. Parkinsons Dis.* **10**, 855-873. doi:10.3233/JPD-202006

Du, T. T., Wang, L., Duan, C. L., Lu, L. L., Zhang, J. L., Gao, G., Qiu, X. B., Wang, X. M. and Yang, H. (2015). GBA deficiency promotes SNCA/alpha-synuclein accumulation through autophagic inhibition by inactivated PPP2A. *Autophagy* **11**, 1803-1820. doi:10.1080/15548627.2015.1086055

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., Marchini, J. and Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210-216. doi:10.1038/s41586-018-0571-7

Emre, M., Aarsland, D., Brown, R., Burn, D. J., Duyckaerts, C., Mizuno, Y., Broe, G. A., Cummings, J., Dickson, D. W., Gauthier, S. et al. (2007). Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Mov. Disord.* **22**, 1689-1707; quiz 1837. doi:10.1002/mds.21507

Eriksen, J. L., Przedborski, S. and Petrucelli, L. (2005). Gene dosage and pathogenesis of Parkinson's disease. *Trends Mol. Med.* **11**, 91-96. doi:10.1016/j.molmed.2005.01.001

Erro, R., Schneider, S. A., Stamelou, M., Quinn, N. P. and Bhatia, K. P. (2016). What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? New evidence and continuing controversies. *J. Neurol. Neurosurg. Psychiatry* **87**, 319-323. doi:10.1136/jnnp-2014-310256

Espay, A. J., Schwarzschild, M. A., Tanner, C. M., Fernandez, H. H., Simon, D. K., Leverenz, J. B., Merola, A., Chen-Plotkin, A., Brundin, P., Kauffman, M. A. et al. (2017). Biomarker-driven phenotyping in Parkinson's disease: a translational missing link in disease-modifying clinical trials. *Mov. Disord.* **32**, 319-324. doi:10.1002/mds.26913

Evenson, K. R., Goto, M. M. and Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int. J. Behav. Nutr. Phys. Act.* **12**, 159. doi:10.1186/s12966-015-0314-1

Farfel-Becker, T., Do, J., Tayebi, N. and Sidransky, E. (2019). Can GBA1-associated parkinson disease be modeled in the mouse? *Trends Neurosci.* **42**, 631-643. doi:10.1016/j.tins.2019.05.010

Fereshtehnejad, S. M., Romenets, S. R., Anang, J. B., Latreille, V., Gagnon, J. F. and Postuma, R. B. (2015). New clinical subtypes of parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes. *JAMA Neurol* **72**, 863-873. doi:10.1001/jamaneurol.2015.0703

Fereshtehnejad, S. M., Zeighami, Y., Dagher, A. and Postuma, R. B. (2017). Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain* **140**, 1959-1976. doi:10.1093/brain/awx118

Fernandes, H. J., Hartfield, E. M., Christian, H. C., Emmanoulidou, E., Zheng, Y., Booth, H., Bogetofte, H., Lang, C., Ryan, B. J., Sardi, S. P. et al. (2016). ER stress and autophagic perturbations lead to elevated extracellular alpha-synuclein in GBA-N370S Parkinson's iPSC-derived dopamine neurons. *Stem Cell Rep.* **6**, 342-356. doi:10.1016/j.stemcr.2016.01.013

Foltynie, T., Brayne, C. and Barker, R. A. (2002). The heterogeneity of idiopathic Parkinson's disease. *J. Neurol.* **249**, 138-145. doi:10.1007/PL00007856

Gamez-Valero, A., Iranzo, A., Serradell, M., Vilas, D., Santamaria, J., Gaig, C., Alvarez, R., Ariza, A., Tolosa, E. and Beyer, K. (2018). Glucocerebrosidase gene variants are accumulated in idiopathic REM sleep behavior disorder. *Parkinsonism Relat. Disord.* **50**, 94-98. doi:10.1016/j.parkreldis.2018.02.034

Goecks, J., Jalili, V., Heiser, L. M. and Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell* **181**, 92-101. doi:10.1016/j.cell.2020.03.022

Gong, W., Beckmann, C. F. and Smith, S. M. (2021). Phenotype discovery from population brain imaging. *Med. Image Anal.* **71**, 102050. doi:10.1016/j.media.2021.102050

Griffanti, L., Klein, J. C., Szewczyk-Krolikowski, K., Menke, R. A. L., Rolinski, M., Barber, T. R., Lawton, M., Evetts, S. G., Begeti, F., Crabbe, M. et al. (2020). Cohort profile: the oxford Parkinson's disease centre discovery cohort MRI substudy (OPDC-MRI). *BMJ Open* **10**, e034110. doi:10.1136/bmjopen-2019-034110

Hampshire, A., Trender, W., Chamberlain, S. R., Jolly, A. E., Grant, J. E., Patrick, F., Mazibuko, N., Williams, S. C., Barnby, J. M., Hellyer, P. et al.

(2021). Cognitive deficits in people who have recovered from COVID-19. *EClinicalMedicine* **39**, 101044. doi:10.1016/j.eclinm.2021.101044

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., van de Lagemaat, L. N., Smith, K. A., Ebbert, A., Riley, Z. L. et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391-399. doi:10.1038/nature11405

Hayes, T. L., Abendroth, F., Adami, A., Pavel, M., Zitzelberger, T. A. and Kaye, J. A. (2008). Unobtrusive assessment of activity patterns associated with mild cognitive impairment. *Alzheimers Dement* **4**, 395-405. doi:10.1016/j.jalz.2008.07.004

Heinzel, S., Lerche, S., Maetzler, W. and Berg, D. (2017). Global, yet incomplete overview of cohort studies in Parkinson's disease. *J. Parkinsons Dis.* **7**, 423-432. doi:10.3233/JPD-171100

Hennekam, R. C. and Biesecker, L. G. (2012). Next-generation sequencing demands next-generation phenotyping. *Hum. Mutat.* **33**, 884-886. doi:10.1002/humu.22048

Hipp, G., Vaillant, M., Diederich, N. J., Roomp, K., Satagopam, V. P., Banda, P., Sandt, E., Mommaerts, K., Schmitz, S. K., Longhino, L. et al. (2018). The luxembourg Parkinson's study: a comprehensive approach for stratification and early diagnosis. *Front. Aging Neurosci.* **10**, 326. doi:10.3389/fnagi.2018.00326

Hogl, B. and Stefani, A. (2017). REM sleep behavior disorder (RBD): Update on diagnosis and treatment. *Somnologie (Berl)* **21**, 1-8. doi:10.1007/s11818-016-0048-6

Holleran, W. M., Takagi, Y., Menon, G. K., Legler, G., Feingold, K. R. and Elias, P. M. (1993). Processing of epidermal glucosylceramides is required for optimal mammalian cutaneous permeability barrier function. *J. Clin. Invest.* **91**, 1656-1664. doi:10.1172/JCI116374

Honti, F., Meader, S. and Webber, C. (2014). Unbiased functional clustering of gene variants with a phenotypic-linkage network. *PLoS Comput. Biol.* **10**, e1003815. doi:10.1371/journal.pcbi.1003815

Horsager, J., Andersen, K. B., Knudsen, K., Skjaerbaek, C., Fedorova, T. D., Okkels, N., Schaeffer, E., Bonkat, S. K., Geday, J., Otto, M. et al. (2020). Brain-first versus body-first Parkinson's disease: a multimodal imaging case-control study. *Brain* **143**, 3077-3088. doi:10.1093/brain/awaa238

Iwaki, H., Leonard, H. L., Makarious, M. B., Bookman, M., Landin, B., Vismer, D., Casey, B., Gibbs, J. R., Hernandez, D. G., Blauwendraat, C. et al. (2021). Accelerating medicines partnership: Parkinson's disease. genetic resource. *Mov. Disord.* **36**, 1795-1804. doi:10.1002/mds.28549

Jankovic, J., McDermott, M., Carter, J., Gauthier, S., Goetz, C., Golbe, L., Huber, S., Koller, W., Olanow, C., Shoulson, I. et al. (1990). Variable expression of Parkinson's disease: a base-line analysis of the DATATOP cohort. The Parkinson Study Group. *Neurology* **40**, 1529-1534. doi:10.1212/WNL.40.10.1529

Jellinger, K. A. (2018). Dementia with Lewy bodies and Parkinson's disease-dementia: current concepts and controversies. *J. Neural. Transm. (Vienna)* **125**, 615-650. doi:10.1007/s00702-017-1821-9

Jellinger, K. A. and Korczyn, A. D. (2018). Are dementia with Lewy bodies and Parkinson's disease dementia the same disease? *BMC Med.* **16**, 34. doi:10.1186/s12916-018-1016-8

Johansson, D., Malmgren, K. and Alt Murphy, M. (2018). Wearable sensors for clinical applications in epilepsy, Parkinson's disease, and stroke: a mixed-methods systematic review. *J. Neurol.* **265**, 1740-1752. doi:10.1007/s00415-018-8786-y

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philos. Trans. A Math. Phys. Eng. Sci.* **367**, 4237-4253.

Jones-Davis, D. M. and Buckholtz, N. (2015). The impact of the Alzheimer's disease neuroimaging initiative 2: what role do public-private partnerships have in pushing the boundaries of clinical and basic science research on Alzheimer's disease? *Alzheimers Dement* **11**, 860-864. doi:10.1016/j.jalz.2015.05.006

Kalia, L. V., Lang, A. E., Hazrati, L. N., Fujioka, S., Wszolek, Z. K., Dickson, D. W., Ross, O. A., Van Deerlin, V. M., Trojanowski, J. Q., Hurtig, H. I. et al. (2015). Clinical correlations with Lewy body pathology in LRRK2-related Parkinson disease. *JAMA Neurol* **72**, 100-105. doi:10.1001/jamaneurol.2014.2704

Kang, G. A., Bronstein, J. M., Masterman, D. L., Redelings, M., Crum, J. A. and Ritz, B. (2005). Clinical characteristics in early Parkinson's disease in a central California population-based study. *Mov. Disord.* **20**, 1133-1142. doi:10.1002/mds.20513

Katrib, A., Hsu, W., Bui, A. and Xing, Y. (2016). "RADIOTRANSCRIPTOMICS": a synergy of imaging and transcriptomics in clinical assessment. *Quant. Biol.* **4**, 1-12. doi:10.1007/s40484-016-0061-6

Keller, M. F., Saad, M., Bras, J., Bettella, F., Nicolaou, N., Simon-Sanchez, J., Mittag, F., Buchel, F., Sharma, M., Gibbs, J. R. et al. (2012). Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum. Mol. Genet.* **21**, 4996-5009. doi:10.1093/hmg/dds335

Koychev, I., Young, S., Holve, H., Ben Yehuda, M. and Gallacher, J. (2020). Dementias platform UK clinical studies and great minds register: protocol of a targeted brain health studies recontact database. *BMJ Open* **10**, e040766. doi:10.1136/bmjopen-2020-040766

Krohn, L., Ruskey, J. A., Rudakou, U., Leveille, E., Asayesh, F., Hu, M. T. M., Arnulf, I., Dauvilliers, Y., Hogl, B., Stefani, A. et al. (2020). GBA variants in REM sleep behavior disorder: a multicenter study. *Neurology* **95**, e1008-e1016. doi:10.1212/WNL.0000000000010042

Kruger, R., Kuhn, W., Muller, T., Woitalla, D., Graeber, M., Kosel, S., Przuntek, H., Epplen, J. T., Schols, L. and Riess, O. (1998). Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease. *Nat. Genet.* **18**, 106-108. doi:10.1038/ng0298-106

Kupstas, A. R., Hoskin, T. L., Day, C. N., Boughey, J. C., Habermann, E. B. and Hieken, T. J. (2020). Biological subtype, treatment response and outcomes in inflammatory breast cancer using data from the National Cancer Database. *Br. J. Surg.* **107**, 1033-1041. doi:10.1002/bjs.11469

Lang, A. E. and Lozano, A. M. (1998). Parkinson's disease. Second of two parts. *N. Engl. J. Med.* **339**, 1130-1143. doi:10.1056/NEJM199810153391607

Lashley, T., Schott, J. M., Weston, P., Murray, C. E., Wellington, H., Keshavan, A., Foti S. C., Foiani, M., Toombs, J., Rohrer, J. D. et al. (2018). Molecular biomarkers of Alzheimer's disease: progress and prospects. *Dis. Model. Mech.* **11,** dmm031781. doi:10.1242/dmm.031781

Lawton, M., Baig, F., Rolinski, M., Ruffman, C., Nithi, K., May, M. T., Ben-Shlomo, Y. and Hu, M. T. (2015). Parkinson's disease subtypes in the oxford parkinson disease centre (OPDC) discovery cohort. *J. Parkinsons Dis.* **5**, 269-279. doi:10.3233/JPD-140523

Lee, J. W., Song, Y. S., Kim, H., Ku, B. D. and Lee, W. W. (2021). Patients with scans without evidence of dopaminergic deficit (SWEDD) do not have early Parkinson's disease: Analysis of the PPMI data. *PLoS One* **16**, e0246881.

Lesage, S., Patin, E., Condroyer, C., Leutenegger, A. L., Lohmann, E., Giladi, N., Bar-Shira, A., Belarbi, S., Hecham, N., Pollak, P. et al. (2010). Parkinson's disease-related LRRK2 G2019S mutation results from independent mutational events in humans. *Hum. Mol. Genet.* **19**, 1998-2004. doi:10.1093/hmg/ddq081

Lesage, S., Anheim, M., Condroyer, C., Pollak, P., Durif, F., Dupuits, C., Viallet, F., Lohmann, E., Corvol, J. C., Honore, A. et al. (2011). Large-scale screening of the Gaucher's disease-related glucocerebrosidase gene in Europeans with Parkinson's disease. *Hum. Mol. Genet.* **20**, 202-210. doi:10.1093/hmg/ddq454

Lindestam Arlehamn, C. S., Dhanwani, R., Pham, J., Kuan, R., Frazier, A., Rezende Dutra, J., Phillips, E., Mallal, S., Roederer, M., Marder, K. S. et al. (2020). . alpha-Synuclein-specific T cell reactivity is associated with preclinical and early Parkinson's disease. *Nat. Commun.* **11**, 1875. doi:10.1038/s41467-020-15626-w

Machado, A., Ferreira, D., Grothe, M. J., Eyjolfsdottir, H., Almqvist, P. M., Cavallin, L., Lind, G., Linderoth, B., Seiger, A., Teipel, S. et al. (2020). The cholinergic system in subtypes of Alzheimer's disease: an in vivo longitudinal MRI study. *Alzheimers Res. Ther.* **12**, 51. doi:10.1186/s13195-020-00620-7

Magesh, P. R., Myloth, R. D. and Tom, R. J. (2020). An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Comput. Biol. Med.* **126**, 104041. doi:10.1016/j.compbiomed.2020.104041

Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R. and Alda, M. (2013). The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS One* **8**, e76295. doi:10.1371/journal.pone.0076295

Marek, K., Seibyl, J., Eberly, S., Oakes, D., Shoulson, I., Lang, A. E., Hyson, C. and Jennings, D. and **Parkinson Study Group PRECEPT Investigators** (2014). Longitudinal follow-up of SWEDD subjects in the PRECEPT Study. *Neurology* **82**, 1791-1797. doi:10.1212/WNL.0000000000000424

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C. M., Trojanowski, J. Q. et al. (2018). The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann. Clin. Transl. Neurol.* **5**, 1460-1477. doi:10.1002/acn3.644

Markello, R. D., Shafiei, G., Tremblay, C., Postuma, R. B., Dagher, A. and Misic, B. (2021). Multimodal phenotypic axes of Parkinson's disease. *NPJ Parkinsons Dis.* **7**, 6. doi:10.1038/s41531-020-00144-9

Matrana, M. R. and Campbell, B. (2020). Precision Medicine and the Institutional Review Board: Ethics and the Genome. *Ochsner J.* **20**, 98-103. doi:10.31486/toj.19.0098

Mattsson-Carlgren, N., Palmqvist, S., Blennow, K. and Hansson, O. (2020). Increasing the reproducibility of fluid biomarker studies in neurodegenerative studies. *Nat. Commun.* **11**, 6252. doi:10.1038/s41467-020-19957-6

Mei, J., Desrosiers, C. and Frasnelli, J. (2021). Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Front. Aging Neurosci.* **13**, 633752. doi:10.3389/fnagi.2021.633752

Mroczek, M., Desouky, A. and Sirry, W. (2021). Imaging transcriptomics in neurodegenerative diseases. *J. Neuroimaging* **31**, 244-250. doi:10.1111/jon.12827

Mullin, S., Smith, L., Lee, K., D'Souza, G., Woodgate, P., Elflein, J., Hallqvist, J., Toffoli, M., Streeter, A., Hosking, J. et al. (2020). Ambroxol for the treatment of patients with parkinson disease with and without glucocerebrosidase gene mutations: a nonrandomized, noncontrolled trial. *JAMA Neurol* **77**, 427-434. doi:10.1001/jamaneurol.2019.4611

Maserejian, N., Vinikoor-Imler, L. and Dilley, A. (2020). Estimation of the 2020 global population of Parkinson's disease (PD). *Mov. Disord.* **35** Suppl. 1, S1-S599. doi:10.1002/mds.27968

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D. A., Noyce, A. J., Xue, A. et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091-1102. doi:10.1016/S1474-4422(19)30320-5

Nathoo, F. S., Kong, L. and Zhu, H. (2019). A Review of Statistical Methods in Imaging Genetics. *Can. J. Stat.* **47**, 108-131. doi:10.1002/cjs.11487

Nowakowski, G. S. and Czuczman, M. S. (2015). ABC, GCB, and double-hit diffuse large B-cell lymphoma: does subtype make a difference in therapy selection? *Am. Soc. Clin. Oncol. Educ. Book* e449-57.

Noyce, A. J. and Nalls, M. A. (2016). Mendelian randomization - the key to understanding aspects of Parkinson's disease causation? *Mov. Disord.* **31**, 478-483. doi:10.1002/mds.26492

O'Sullivan, J. W. and Ioannidis, J. P. A. (2021). Reproducibility in the UK biobank of genome-wide significant signals discovered in earlier genome-wide association studies. *Sci. Rep.* **11**, 18625. doi:10.1038/s41598-021-97896-y

Oxtoby, N. P., Leyland, L. A., Aksman, L. M., Thomas, G. E. C., Bunting, E. L., Wijeratne, P. A., Young, A. L., Zarkali, A., Tan, M. M. X., Bremner, F. D. et al. (2021). Sequence of clinical and neurodegeneration events in Parkinson's disease progression. *Brain* **144**, 975-988. doi:10.1093/brain/awaa461

Pagano, G., Niccolini, F. and Politis, M. (2016). Imaging in Parkinson's disease. *Clin. Med. (Lond)* **16**, 371-375. doi:10.7861/clinmedicine.16-4-371

Pastor, P. (2012). Genetic heterogeneity in Parkinson disease: the meaning of GWAS and replication studies. *Neurology* **79**, 619-620. doi:10.1212/WNL.0b013e318264e3d2

Polymeropoulos, M. H., Lavedan, C., Leroy, E., Ide, S. E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R. et al. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045-2047. doi:10.1126/science.276.5321.2045

Postuma, R. B. (2014). Prodromal Parkinson's disease–using REM sleep behavior disorder as a window. *Parkinsonism Relat. Disord.* **20** Suppl. 1, S1-S4. doi:10.1016/S1353-8020(13)00400-8

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E. et al. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* **30**, 1591-1601. doi:10.1002/mds.26424

Prange, S., Metereau, E. and Thobois, S. (2019). Structural Imaging in Parkinson's disease: new developments. *Curr. Neurol. Neurosci. Rep.* **19**, 50. doi:10.1007/s11910-019-0964-5

Punt, C. J., Koopman, M. and Vermeulen, L. (2017). From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat. Rev. Clin. Oncol* **14**, 235-246. doi:10.1038/nrclinonc.2016.171

Ramaswami, R., Bayer, R. and Galea, S. (2018). Precision medicine from a public health perspective. *Annu. Rev. Public Health* **39**, 153-168. doi:10.1146/annurev-publhealth-040617-014158

Regnault, A., Boroojerdi, B., Meunier, J., Bani, M., Morel, T. and Cano, S. (2019). Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. *J. Neurol.* **266**, 1927-1936. doi:10.1007/s00415-019-09348-3

Ren, J., Pan, C., Li, Y., Li, L., Hua, P., Xu, L., Zhang, L., Zhang, W., Xu, P. and Liu, W. (2021). Consistency and stability of motor subtype classifications in patients with de novo Parkinson's disease. *Front. Neurosci.* **15**, 637896. doi:10.3389/fnins.2021.637896

Riboldi, G. M. and Di Fonzo, A. B. (2019). GBA, gaucher disease, and Parkinson's disease: from genetic to clinic to new therapeutic approaches. *Cells* **8**, 364. doi:10.3390/cells8040364

Robinson, P. N. (2012). Deep phenotyping for precision medicine. *Hum. Mutat.* **33**, 777-780. doi:10.1002/humu.22080

Rosborough, K., Patel, N. and Kalia, L. V. (2017). . alpha-synuclein and parkinsonism: updates and future perspectives. *Curr. Neurol. Neurosci. Rep.* **17**, 31. doi:10.1007/s11910-017-0737-y

Sandor, C., Beer, N. L. and Webber, C. (2017). Diverse type 2 diabetes genetic risk factors functionally converge in a phenotype-focused gene network. *PLoS Comput. Biol.* **13**, e1005816. doi:10.1371/journal.pcbi.1005816

Schapira, A. H. (2013). Recent developments in biomarkers in Parkinson disease. *Curr. Opin. Neurol.* **26**, 395-400. doi:10.1097/WCO.0b013e3283633741

Schiess, M. C., Zheng, H., Soukup, V. M., Bonnen, J. G. and Nauta, H. J. (2000). Parkinson's disease subtypes: clinical classification and ventricular cerebrospinal fluid analysis. *Parkinsonism Relat. Disord.* **6**, 69-76. doi:10.1016/S1353-8020(99)00051-6

Schlachetzki, J. C. M., Barth, J., Marxreiter, F., Gossler, J., Kohl, Z., Reinfelder, S., Gassner, H., Aminian, K., Eskofier, B. M., Winkler, J. et al. (2017). Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLoS One* **12**, e0183989.

Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., Roulland, S., Kasbekar, M., Young, R. M., Shaffer, A. L. et al. (2018). Genetics and pathogenesis of diffuse large B-cell lymphoma. *N. Engl. J. Med.* **378**, 1396-1407. doi:10.1056/NEJMoa1801445

Schneider, S. A. and Alcalay, R. N. (2020). Precision medicine in Parkinson's disease: emerging treatments for genetic Parkinson's disease. *J. Neurol.* **267**, 860-869. doi:10.1007/s00415-020-09705-7

Schrag, A., Ben-Shlomo, Y. and Quinn, N. (2002). How valid is the clinical diagnosis of Parkinson's disease in the community? *J. Neurol. Neurosurg. Psychiatry* **73**, 529-534. doi:10.1136/jnnp.73.5.529

Shah, V. V., McNames, J., Mancini, M., Carlson-Kuhta, P., Nutt, J. G., El-Gohary, M., Lapidus, J. A., Horak, F. B. and Curtze, C. (2020). Digital biomarkers of mobility in Parkinson's disease during daily living. *J. Parkinsons Dis.* **10**, 1099-1111. doi:10.3233/JPD-201914

Shen, L. and Thompson, P. M. (2020). Brain imaging genomics: integrated analysis and machine learning. *Proc. IEEE Inst. Electr. Electron. Eng.* **108**, 125-162. doi:10.1109/JPROC.2019.2947272

Sheridan, P. L., Solomont, J., Kowall, N. and Hausdorff, J. M. (2003). Influence of executive function on locomotor function: divided attention increases gait variability in Alzheimer's disease. *J. Am. Geriatr. Soc.* **51**, 1633-1637. doi:10.1046/j.1532-5415.2003.51516.x

Shu, Z. Y., Cui, S. J., Wu, X., Xu, Y., Huang, P., Pang, P. P. and Zhang, M. (2021). Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter. *Magn. Reson. Med.* **85**, 1611-1624. doi:10.1002/mrm.28522

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R. et al. (2003). . alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841. doi:10.1126/science.1090278

Smith, L., Mullin, S. and Schapira, A. H. V. (2017). Insights into the structural biology of Gaucher disease. *Exp. Neurol.* **298**, 180-190. doi:10.1016/j.expneurol.2017.09.010

Smolensky, L., Amondikar, N., Crawford, K., Neu, S., Kopil, C. M., Daeschler, M., Riley, L., 23andMe Research Team, Brown, E., Toga, A. W. et al. (2020). Fox Insight collects online, longitudinal patient-reported outcomes and genetic data on Parkinson's disease. *Sci. Data* **7**, 67. doi:10.1038/s41597-020-0401-2

Solana-Lavalle, G. and Rosas-Romero, R. (2021). Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. *Comput. Methods Programs Biomed.* **198**, 105793. doi:10.1016/j.cmpb.2020.105793

Stoker, T. B., Camacho, M., Winder-Rhodes, S., Liu, G., Scherzer, C. R., Foltynie, T., Evans, J., Breen, D. P., Barker, R. A. and Williams-Gray, C. H. (2020). Impact of GBA1 variants on long-term clinical progression and mortality in incident Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **91**, 695-702. doi:10.1136/jnnp-2020-322857

Sturchio, A., Marsili, L., Vizcarra, J. A., Dwivedi, A. K., Kauffman, M. A., Duker, A. P., Lu, P., Pauciulo, M. W., Wissel, B. D., Hill, E. J. et al. (2020). Phenotype-agnostic molecular subtyping of neurodegenerative disorders: the cincinnati cohort biomarker program (CCBP). *Front. Aging Neurosci.* **12**, 553635. doi:10.3389/fnagi.2020.553635

Sundararajan, K., Georgievska, S., Te Lindert, B. H. W., Gehrman, P. R., Ramautar, J., Mazzotti, D. R., Sabia, S., Weedon, M. N., van Someren, E. J. W., Ridder, L. et al. (2021). Sleep classification from wrist-worn accelerometer data using random forests. *Sci. Rep.* **11**, 24. doi:10.1038/s41598-020-79217-x

Tagare, H. D., DeLorenzo, C., Chelikani, S., Saperstein, L. and Fulbright, R. K. (2017). Voxel-based logistic analysis of PPMI control and Parkinson's disease DaTscans. *Neuroimage* **152**, 299-311. doi:10.1016/j.neuroimage.2017.02.067

Tao, C., Nichols, T. E., Hua, X., Ching, C. R. K., Rolls, E. T., Thompson, P. M. and Feng, J. and **Alzheimer's Disease Neuroimaging Initiative** (2017). Generalized reduced rank latent factor regression for high dimensional tensor fields, and neuroimaging-genetic applications. *Neuroimage* **144**, 35-57. doi:10.1016/j.neuroimage.2016.08.027

van der Velden, D. L., Hoes, L. R., van der Wijngaart, H., van Berge Henegouwen, J. M., van Werkhoven, E., Roepman, P., Schilsky, R. L., de Leng, W. W. J., Huitema, A. D. R., Nuijen, B. et al. (2019). The Drug Rediscovery protocol facilitates the expanded use of existing anticancer drugs. *Nature* **574**, 127-131. doi:10.1038/s41586-019-1600-x

Vogel, J. W., Young, A. L., Oxtoby, N. P., Smith, R., Ossenkoppele, R., Strandberg, O. T., La Joie, R., Aksman, L. M., Grothe, M. J., Iturria-Medina, Y. et al. (2021). Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat. Med.* **27**, 871-881. doi:10.1038/s41591-021-01309-6

von Coelln, R., Gruber-Baldini, A. L., Reich, S. G., Armstrong, M. J., Savitt, J. M. and Shulman, L. M. (2021). The inconsistency and instability of Parkinson's disease motor subtypes. *Parkinsonism Relat. Disord.* **88**, 13-18. doi:10.1016/j.parkreldis.2021.05.016

Wang, L., Cheng, W., Rolls, E. T., Dai, F., Gong, W., Du, J., Zhang, W., Wang, S., Liu, F., Wang, J. et al. (2020). Association of specific biotypes in patients with Parkinson disease and disease progression. *Neurology* **95**, e1445-e1460. doi:10.1212/WNL.0000000000010498

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Handler, K., Pickkers, P., Aziz, N. A.

et al. (2021). Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265-270. doi:10.1038/s41586-021-03583-3

Watanabe, K., Stringer, S., Frei, O., Umicevic Mirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M. and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339-1348. doi:10.1038/s41588-019-0481-0

Weng, C., Shah, N. H. and Hripcsak, G. (2020). Deep phenotyping: Embracing complexity and temporality-Towards scalability, portability, and interoperability. *J. Biomed. Inform.* **105**, 103433. doi:10.1016/j.jbi.2020.103433

Williams, D. R. and Litvan, I. (2013). Parkinsonian syndromes. *Continuum (Minneap Minn)* **19**, 1189-1212.

Xu, Y. H., Quinn, B., Witte, D. and Grabowski, G. A. (2003). Viable mouse models of acid beta-glucosidase deficiency: the defect in Gaucher disease. *Am. J. Pathol.* **163**, 2093-2101. doi:10.1016/S0002-9440(10)63566-3

Young, A. L., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N. C., Cash, D. M., Thomas, D. L., Dick, K. M., Cardoso, J. et al. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat. Commun.* **9**, 4273. doi:10.1038/s41467-018-05892-0

Zarkali, A., McColgan, P., Leyland, L. A., Lees, A. J., Rees, G. and Weil, R. S. (2021). Organisational and neuromodulatory underpinnings of structural-functional connectivity decoupling in patients with Parkinson's disease. *Commun. Biol.* **4**, 86. doi:10.1038/s42003-020-01622-9

Zarranz, J. J., Alegre, J., Gomez-Esteban, J. C., Lezcano, E., Ros, R., Ampuero, I., Vidal, L., Hoenicka, J., Rodriguez, O., Atares, B. et al. (2004). The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia. *Ann. Neurol.* **55**, 164-173. doi:10.1002/ana.10795

Zhao, Y. and Dzamko, N. (2019). Recent developments in LRRK2-targeted therapy for Parkinson's disease. *Drugs* **79**, 1037-1051. doi:10.1007/s40265-019-01139-4