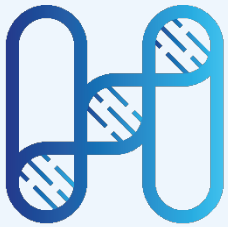

HEREDITARY

HetERogeneous sEmantic Data integratIon for the guT-bRain interplaY

Deliverable 2.16

DESIGN OF NEURODEGENERATIVE USE CASES

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

EXECUTIVE SUMMARY

This report presents the definition of the Neurodegenerative Use Cases that constitute the core of HEREDITARY.

Neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS) and other related disorders, are complex and challenging to treat due to their multifaceted pathophysiology. The current approach to neurodegenerative diseases focuses on symptom-based classification, which often fails to capture the underlying biological heterogeneity. There is therefore a pressing need for innovative approaches that leverage advanced data integration and analytics to uncover deeper insights into these diseases.

HEREDITARY aims to address these challenges by integrating multimodal data, and particularly combining genetic, clinical, biomarker and imaging data to provide a comprehensive understanding of neurodegenerative diseases.

The tasks presented in this deliverable will employ the multimodal data framework in a research setting, aiming to refine prediction and patient classification in neurodegenerative disorders and uncover novel disease-related biomarkers and etiological mechanisms. The development of a multimodal data framework will enhance predictive analysis and patient classification, Improving the discovery of novel disease-related biomarkers or etiological mechanisms.

HEREDITARY will integrate diverse data types from different centres using a federated learning infrastructure. This approach allows for secure, large-scale data analysis without compromising patient privacy by implementing privacy-preserving federated analytics. The effectiveness of the HEREDITARY framework within the realm of neurodegenerative diseases will be assessed primarily in two use cases designed to (a) advance the understanding of ALS by identifying endophenotypes through the integration of comprehensive data modalities and (b) detect biologically informed clusters within neurodegenerative diseases. These two use cases will integrate with three additional clinical applications of HEREDITARY, which will be the focus of future tasks. These additional use cases will concentrate on the early detection of Parkinson's disease using multimodal data and exploring the gut-brain axis in health and disease. Together, these five use cases will demonstrate the broad applicability and effectiveness of the HEREDITARY framework.

The outcomes of the Neurodegenerative Use Cases of HEREDITARY will improve our understanding of neurodegenerative disorders. This approach will pave the way for improved patient classification and the development of effective treatment strategies, uncovering deeper insights into the underlying mechanisms of these diseases and ultimately leading to more targeted and personalized interventions.

DOCUMENT INFORMATION

Deliverable ID	D2.16
Deliverable Title	Design of neurodegenerative use cases
Work Package	WP2
Lead Partner	UNITO
Due date	30.06.2024
Date of submission	28.06.2024
Type of deliverable	R
Dissemination level	PU

AUTHORS

Name	Organisation
Adriano Chiò	UNITO
Maurizio Grassano	UNITO
Manfredo Atzori (Reviewer)	HESSO

REVISION HISTORY

Version	Date	Author	Document history/approvals
1.0	07.06.24	Adriano Chiò	First Draft
1.1	14.06.24	Manfredo Atzori	Revision
2.0	25.06.24	Adriano Chiò	Final

Contents

1 INTRODUCTION 6

2 STATE OF THE ART 7

2.1 Multimodal data analysis in neurodegenerative disorders 7

2.2 The advantages of federated learning in this research setting 10

3 NEURODEGENERATIVE USE CASES DESCRIPTION 11

3.1 Use Case 1: Neurodegenerative diseases phenotyping and prognosis evaluation 11

3.1.1 Participating partners 11

3.1.2 Overall aim of the Use Case 11

3.1.3 Background 11

3.1.4 Rationale 11

3.1.5 Scientific Approach 12

3.1.6 Challenges and Risks: 14

3.2 Use Case 2: Next-generation diagnosing and treatment response for neurodegenerative diseases 15

3.2.1 Participating partners 15

3.2.2 Overall aim of the Use Case 15

3.2.3 Background 15

3.2.4 Rationale 15

3.2.5 Scientific Approach 16

3.2.6 Challenges and Risk 17

4 INTEGRATION OF THE USE CASES WITHIN HEREDITARY 18

4.1 Use Case 3: *Signs of Parkinson's disease in multimodal data* 18

4.1.1 Overview of the Use Case 18

4.1.2 Scientific Approach 18

4.1.3 Integration with Use Cases 1 and 2 19

4.2 Use Cases 4 and 5: *Phenotyping of the gut-brain axis in healthy individuals to understand deviations in disorders* and *Gut-Brain linkage and disease relevance* .. 19

4.2.1 Overview of the Use Cases 19

4.2.2 Scientific Approach 19

4.2.3 Integration with Use Cases 1 and 2 20

REFERENCES 21

1 INTRODUCTION

This report presents the Neurodegenerative Use Cases of HEREDITARY. In these Use Cases, HEREDITARY seeks to uncover novel disease-associated mechanisms and improve patients' classification, thus enabling significant advancement in the research of neurodegenerative diseases. The project's objectives will be achieved through the development of a robust, interoperable, trustworthy, and secure framework that integrates multimodal health data, including genetic information, while ensuring compliance with cross-national privacy-preserving policies.

The Neurodegenerative Use Cases were designed in collaboration with stakeholders to align with HEREDITARY's core objectives:

- Multimodal data integration and Interoperability
- Utilization of a multimodal semantic ontology: by enabling efficient and meaningful data access and querying, HEREDITARY addresses the need for integrating diverse linguistic and multimodal data (e.g., text, imaging, genetic data) for comprehensive analysis, ensuring the platform meets the varied requirements of policy-makers, researchers, clinicians, and patients.
- Enhancement of predictive analysis and support for complex data integration: by implementing privacy-preserving federated analytics and learning methods, HEREDITARY will ensure that data remains distributed and secure while still being accessible for global analysis and learning, which is crucial for handling sensitive medical data.
- Integration of genetic and clinical data to uncover deeper insights into disease mechanisms and patient outcomes.

Through the utilization of advanced federated analytics and learning workflows, HEREDITARY will empower clinicians, researchers, and policymakers to better understand neurodegenerative diseases and develop more effective treatment strategies.

This report begins by introducing the state of the art regarding multimodal data integration in neurodegenerative diseases. Section 3 describes Use Cases 1 and 2 in detail. Finally, Section 4 briefly explores how Use Cases 1 and 2 integrate with the other use cases from Work Package 2.

2 STATE OF THE ART

One of the fundamental challenges in addressing neurodegenerative diseases is their complex pathophysiology, which complicates treatment development. Neurodegeneration begins at the molecular level, leading to cellular death. However, the physiological onset of the disease and the manifestation of symptoms occur only after significant and irreversible neuronal loss. Consequently, neurodegenerative diseases are age-related, often taking years to decades to manifest. By this time, the mechanisms that initiated the disease may have been overtaken by those that sustain the neurodegenerative process. Therefore, targeting the initial causal mechanisms may have limited efficacy in slowing or stopping disease progression. The primary challenge in developing treatments lies in identifying all pathological events leading to cell death and targeting these to rescue afflicted neurons.

To accelerate progress in developing effective treatments, the field of neurodegeneration requires new and innovative approaches. A holistic investigation is urgently needed to fully capture the genetics and biology of neurodegenerative diseases, including gene-drug interactions and potential causal mechanisms. In Use Cases 1 and 2, we propose that federated learning on multimodal data can identify robust, complementary endophenotypes—closer to the underlying aetiology—leading to a deeper understanding of the biological mechanisms behind neurodegenerative diseases. This approach could help identify clusters of patients with similar biological backgrounds, moving beyond the traditional symptom-based classification.

2.1 Multimodal data analysis in neurodegenerative disorders

Recent studies have demonstrated the benefits of multimodal data analysis and the integration of different data types to achieve a better understanding of human brain health. For example, linking genomic data with magnetic resonance imaging (MRI) of the brain has been useful in investigating the mechanisms underlying brain ageing. Recent studies have focused on the brain age gap (BAG), the difference between chronological age and the apparent age of the brain estimated from imaging data, which is widely considered an indicator of brain health (Cole et al., 2017; Cole & Franke, 2017; Franke et al., 2010).

The typical approach uses one or more imaging modalities, often relying on a single structural image from each subject (Smith et al., 2019). The data is pre-processed, and features are identified for brain age prediction. Structural images may be warped into a standard space, and grey matter segmentation carried out. Alternatively, more condensed features, such as volumes of grey and white matter within multiple brain regions, may be derived. The resulting dataset, comprising multiple subjects' feature sets and their true ages, is then passed into a supervised learning algorithm that learns to predict the subjects' ages from their brain imaging features. The true age is subtracted from the estimated brain age to create a delta, with potential corrections for biases.

Combining all factors into a single estimate of brain age can provide a useful summary metric and the most accurate single estimate of a subject's age from imaging data. However, this may obscure important information about the distinctions between multiple biological factors, making it harder to understand the various causes of brain aging. Factors such as physical exercise, education, alcohol intake, smoking, dietary patterns,

hypertension, and obesity likely contribute to brain aging in different ways (Leonardsen et al., 2022; Steffener et al., 2016; Wrigglesworth et al., 2021), affecting different aspects of brain structure and function as viewed through multiple imaging modalities. Therefore, using multiple brain imaging modalities allows for the investigation of distinct biological factors related to aging, thus gaining a greater understanding of brain aging mechanisms. In this context, genome-wide association studies (GWAS) have associated BAG with common genetic variants (e.g., single nucleotide polymorphisms, SNPs) (Jonsson et al., 2019; Karolinska Schizophrenia Project (KaSP) et al., 2019; Ning et al., 2020; Smith et al., 2020). However, focusing on a single outcome does not comprehensively capture the genetic architecture of brain aging mechanisms. Machine learning methods integrating multimodal imaging data have instead improved the yield of genetic associations (Patel et al., 2024). Interindividual variations in the predicted brain age of individuals with the same chronological age and their underlying genetic factors correlate with neurological and mental disorders, such as dementia, schizophrenia, major depressive disorder, bipolar disorder, and mortality (Elliott et al., 2021; Gaser et al., 2013; Han et al., 2021; Karolinska Schizophrenia Project (KaSP) et al., 2019; Schnack et al., 2016).

These results emphasize the complex architecture of brain age and provide insights into the causal relationships between brain age and neurological and neuropsychiatric disorders. This approach can be extended to study environmental factors associated with disease. For instance, higher educational attainment may protect against Alzheimer's risk by increasing or maintaining an individual's brain reserve. Brain reserve refers to individual differences in brain structure that enable some individuals to preserve cognitive and functional status despite neuropathology (Stern, 2012; Stern et al., 2020). Structural alterations associated with Alzheimer's disease, such as cortical thinning and grey matter atrophy, can be measured in vivo using MRI, and their pre-morbid levels may serve as proxies for brain reserve capacity (Arenaza-Urquijo et al., 2013; Querbes et al., 2009; Solé-Padullés et al., 2009).

The relationship between factors that increase brain reserve capacity and their role in Alzheimer's diseases can then be further delineating using techniques like Mendelian Randomization (MR), a valuable tool to investigate cause-effect relationships between risk factors and disorders (Wen et al., 2024). For example, MR models have inferred the causally protective role of increased educational attainment on Alzheimer's risk but demonstrated that this protective effect is not mediated via an increased brain reserve or structural brain changes (Groot et al., 2018; Seyedsalehi et al., 2023). This suggests that higher education might protect through increasing cognitive reserve or through alternative biological mechanisms requiring further investigation.

Thus, the approaches discussed here can help identify numerous risk or protective lifestyle factors and neurobiological processes that may exert independent, synergistic, antagonistic, sequential, or differential influences on human brain health.

Toward a biological-based classification

A popular approach to discovering genetic variants relevant for neurodegenerative diseases is the use of whole-exome/genome sequencing to assess common and rare variants globally and then associate each variant with disease status. However, it is well established that neurodegeneration results from a combination of pathways, including inflammation, cholesterol metabolism, and endosome or ubiquitin-related functioning.

Individuals with the same symptoms can differ regarding the pathways contributing to their symptoms (Cornblath et al., 2020; Emon et al., 2020; Koretsky et al., 2023; P. Li et al., 2015; Villoslada et al., 2020). Additionally, different genetic factors may influence various aspects of the diseases. Studying disease status as a single outcome may mask genetic effects that only impact specific pathways or patient subsets. Identifying genetic variants that influence correlated traits may be key to understanding the genetic architecture of the disease.

To overcome these limitations, a pathway approach has been proposed, analysing variants in relation to distinct pathological processes reflected by different disease biomarkers. This can be achieved by applying statistical methods that identify independent clusters of biomarkers representing different biological processes. These approaches are not only conceptually appealing but may also improve the power of traditional GWAS.

However, traditional approaches for genotype-phenotype studies have limitations. While appealing, recent commentaries have highlighted inherent limitations from modelling pre-selected outcomes, necessitating a priori selections of measures that are expected to be relevant, with the risk of overlooking others. Additionally, although many studies analyse each of these phenotypes separately, the joint analysis of multivariate phenotypes has recently become popular because it can increase statistical power to detect genetic loci. Integrating association signals at a single SNP over multiple correlated dependent variables in a single comprehensive framework is not always straightforward. Therefore, the power to determine the full genetic basis of the disease could be improved by techniques that identify features from multimodal data representing different potential pathways, which could serve as outcomes for genetic (and other omics) investigations (Aschard et al., 2014).

Currently, the diagnosis of neurodegenerative diseases remains clinical, based on both inclusion and exclusion criteria. However, it is well known that the underlying degenerative process exists for years before patients present with classical clinical features. In recent years, fluid, tissue, and imaging biomarkers for these diseases have advanced to the point where a biological classification of neurodegenerative disorders can be considered. These techniques now allow for the objective identification of genetic risk, pathological processes, and neurodegeneration.

Neurodegenerative diseases are acknowledged as an extremely heterogeneous group of disorders, even regarding the underlying pathological changes. Rather than simplifying and homogenizing these disorders, any new biological approach needs to address and incorporate this heterogeneity. A biological classification is more appropriate than attempting to establish a single biological definition. Characterizing symptomatic patients based on specific biological and clinical profiles will significantly advance a broad range of research studies (e.g., epidemiology, neuroimaging, biomarker discovery, clinical trials) that are currently limited by pure clinical classifications or studies classifying patients based on a single biological criterion (e.g., a pathogenic genetic variant or single neuroimaging feature). A similar approach has been recently proposed for Parkinson's Disease (Höglinger et al., 2024).

This would be an important initial step in classifying and characterizing patients at the earliest pathological stages, advancing the successful development of precision medicine-based effective disease-modifying therapies that are sorely needed.

2.2 The advantages of federated learning in this research setting

Federated Learning (FL) is a recently proposed machine learning (ML) paradigm that addresses privacy concerns by enabling advanced ML methods, such as deep learning (DL), to be trained and tested collaboratively without requiring the exchange of sensitive data between partners. FL was initially introduced as a distributed ML paradigm to train a centralized model using privacy-sensitive data from numerous clients. Given the similar privacy concerns associated with medical imaging and sequencing data, the FL paradigm has recently been applied to various life sciences applications, showing promising results. FL-based data analysis and ML methods have been proposed for analysing health records and medical imaging data, such as MRI and fMRI (X. Li et al., 2020a), the meta-analysis of biomedical data (Brisimi et al., 2018), and the analysis of genomic data, including gene expression (Zolotareva et al., 2021) and GWAS (Aziz et al., 2022).

One major challenge in studying neurodegenerative diseases, especially rare diseases like amyotrophic lateral sclerosis, is the lack of robustness due to the small sample sizes available at single centres and the high clinical and biological variability of the data. Increasing the sample size is the simplest and most effective way to address this issue. However, this is non-trivial due to the computational intensity, expense, and time required for data collection, as well as limited sample availability and the inability to share and pool existing data due to personal data protection laws. This concern is particularly significant for next-generation sequencing data and neuroimaging data, from which subjects can be identified under certain conditions. Although several open resources are now publicly available, their utility in clinical and phenotypic analysis is often limited due to inherent privacy issues, as they require relevant clinical metadata (e.g., patient sex, age, and disease characteristics) that may be identifying when combined.

These limitations are well exemplified in the case of genomic data. One major limitation of traditional GWAS is that it can only perform association tests on local data. If multiple cohorts want to conduct collaborative GWAS to benefit from larger sample sizes, they must pool their data for joint analysis. The field has also established methods for meta-analysis of individual studies, where only the results and summary statistics of the individual analyses are exchanged. Aggregated analysis requires cohorts to pool their private data for joint analysis, while meta-analysis aggregates the summary statistics from cohorts to estimate combined p-values.

The federated framework holds instead the potential to combine the advantages of aggregated analysis and meta-analysis, providing robustness against heterogeneous data while enhancing the privacy of cohorts' data (Cho et al., 2018; Raimondi et al., 2023; Wu et al., 2021).

3 NEURODEGENERATIVE USE CASES DESCRIPTION

3.1 Use Case 1: Neurodegenerative diseases phenotyping and prognosis evaluation

3.1.1 Participating partners

UNITO, UNIPD, AAU and SURF

3.1.2 Overall aim of the Use Case

The objective of this Use Case is to advance the understanding of amyotrophic lateral sclerosis (ALS) by identifying and characterizing endophenotypes by integrating a comprehensive array of data modalities.

By employing advanced machine learning techniques, this study aims to detect combinations of genomic variants and biological pathways that correlate with specific clinical features (including patient survival) and endophenotypes in ALS. This integrated multimodal framework could detect novel mechanisms involved in the disease or clinical features important for patients' prognosis and stratification, ultimately leading to the discovery of new biomarkers and potential therapeutic strategies for ALS.

3.1.3 Background

Amyotrophic lateral sclerosis (ALS) is a complex and fatal neurodegenerative disease that affects motor neurons, leading to progressive muscle weakness and death from respiratory paralysis. Despite advances in understanding the genetic basis of ALS, the diverse pathogenic mechanisms and the variability in disease progression among patients have represented a paramount challenge, complicating the identification of effective therapeutic strategies. Consequently, there is an urgent need for novel approaches that can both enhance our understanding of the disrupted biological pathways in ALS and uncover biomarkers that could improve our phenotypic categorization of patients. These approaches should aim to detect structural, functional, and metabolic changes that constitute novel biomarkers for accurate phenotyping and prognosis evaluation of ALS patients, thus enabling the discovery of novel genetic factors contributing to ALS.

3.1.4 Rationale

Understanding the basis of amyotrophic lateral sclerosis (ALS) requires novel models capable of disentangling the multitude of molecular interactions and clinical phenotypes that characterize the disease. The combined effects of genetic variants, gene-gene interactions, and the multifaceted ALS phenotype have often been overlooked.

Traditional approaches frequently fail due to their omission of genetic data integration with biological network information, such as gene or protein interactions. This integration could provide more holistic and accurate insights into the underlying biological processes. Gene-gene interactions can play significant roles in ALS, yet single-locus tests and genome-wide association studies (GWAS) are often blind to these interactions, as they typically rely on additive genetic models.

Advancements in machine learning have enhanced our ability to study the biological processes underlying ALS. Multi-omics analyses have identified distinct molecular

signatures in ALS linked to mechanisms such as stress responses and neuroinflammation. Additionally, machine-learning applications to ALS patient data have uncovered specific transcriptional and epigenetic patterns associated with different ALS genotypes and have been used to stratify sporadic ALS patients into molecular subgroups. These findings highlight the potential of advanced methods to dissect ALS complexity and facilitate the identification of therapeutic candidates targeting specific biological processes.

Another limitation hindering our progress in understanding ALS is that genetic association findings based on ALS risk do not fully explain the genetic architecture of ALS, as the complexity underlying the disease is not entirely captured by disease status only. A strategy to overcome this limitation involves studying the genetic underpinnings of ALS-related endophenotypes based on clinical data and biomarkers.

A deeper and more comprehensive understanding of neurodegenerative diseases could emerge by integrating genetics with other biomarkers, potentially uncovering previously undetected pathomechanisms. Neuroimaging, for example, can investigate the impact of genetic variations on brain structure and help delineate the molecular mechanisms induced by both common and rare genetic variants linked to ALS. The goal of integrating multimodal phenotypes with genomic approaches is not only to identify the effects of gene variants on disease risk but also to characterize the systems that directly influence ALS pathophysiology.

Efforts in this direction have been constrained by small sample sizes, lack of reproducibility across cohorts, and unclear definitions of the disease features to study. These limitations have hindered the identification of novel ALS-risk genes and endophenotypes associated with ALS variants. Multimodal learning approaches hold the potential to overcome these limitations and expand the repository of genetic variants and pathways involved in ALS, thereby enhancing the study of genotypic-phenotypic correlations.

3.1.5 Scientific Approach

Study Population

- Patients with ALS from:
  - Piemonte and Aosta Register for ALS, a population-based register (UNITO)
  - an Italian ALS Centre (UNIPD)
- open data (Answer ALS).

Data modalities

- Genomics (WGS, GWAS)
- Clinical (Longitudinal)
- Neuroimaging (Brain 18F-FDG PET, Brain MRI)
- Laboratory (Blood Exams)
- Neurophysiological (Electromyography)

Variable selection

First, the variables to include in the analysis will be evaluated. This could prove necessary in light of the characteristics of genomic data (millions of variants with different

frequencies) and the collinearity among clinical variables. For example, genetic variants will be filtered to retain only those with a higher probability of association with the disease. This includes variants in genes already known to be associated with neurodegenerative processes, variants with known functional effects on biological pathways, or rare variants with a higher likelihood of altering protein function.

The genetic analysis will focus on single nucleotide polymorphisms and samples passing standard whole-genome sequencing-type quality controls. As rare variants can be problematic, a separate analytical framework will be adopted for common (minor allele frequency > 1% or > 5%) and rarer variants. The specific methodologies and relative threshold will also be selected according to the phenotype investigated and the relative sample size. The association of the variants with ALS will be evaluated based on the summary statistics of genome-wide association studies (GWAS) or rare variant burden analysis (Nicolas et al., 2018; van Rheenen et al., 2021). Genetic variations will be integrated with relevant expression data (expression quantitative trait loci, eQTL) (Zhu et al., 2016; Xu et al., 2021). Rare variants will instead be prioritised based on the genes (gene expression levels in the central nervous system, protein network, involvement in neurodegenerative disease) (Li et al., 2020b; Katsonis et al., 2022), in silico computational analysis and prediction on the effect of the protein. To harmonize data from different Centres or different genome-sequencing protocols, variants will be aligned to the same Reference Genome, and identical annotation tools and variant filtering criteria will be adopted.

Similarly, neuroimaging analysis will focus on specific features extracted from the data, such as cortical volume or signal intensities. Anatomical images will be divided into distinct brain regions (cortical, subcortical and brainstem); cortical thickness and regional volumes will be measured using standard automatic procedures for volumetric measurements (van der Burgh et al., 2016; Kuan et al., 2023). Similarly, intensity values will be estimated for atlas-based regions of interest (ROIs) using automated pipelines. Diffusion tensor imaging metrics will be derived to reconstruct white matter tracts. Brain networks will be reconstructed by selecting the interconnecting tracts. Neuroimaging data provided to the consortium have already been acquired and quality control procedures have already been employed in each Centre. Additional quality control measures and standardized preprocessing pipeline will be adopted; similarly, standardized harmonization techniques and statistical methods to adjust for batch effects and site-specific variations will be adopted to ensure data harmonization.

Several clinical features will be included: the ALS Functional Rating Scale-Revised (ALSFRS-R) score, neurological examination, neuropsychological testing and cognitive classification (including the Edinburgh Cognitive and Behavioural ALS Screen), symptom duration, region of disease onset, years of education, comorbidities, handedness, age at disease onset, survival and sex.

For neurophysiological data (electromyography, EMG), semiquantitative scores for acute and chronic denervation will be calculated for the bulbar, cervical and lumbosacral regions. Different body regions and muscles will be evaluated for score calculation: bulbar region, right arm, left arm, right lower limb, and left lower limb. For each of these body regions a score will be assigned according to the amount of positive sharp waves and fibrillation potentials for active denervation and characteristics of motor unit action potentials for chronic denervation (Colombo et al., 2023). Laboratory data will be

selected to include markers of muscle loss (Hertel et al., 2022), inflammation (Murdock et al., 2021) and lipid metabolism (Ingre et al., 2020).

If possible, raw data will be used as input for the model in additional analyses.

Unsupervised Learning:

Unsupervised learning models will then be used to detect (a) endophenotypes from multimodal clinical data and (b) combinations of genomic variants and biological pathways related to clinical features in ALS. Particular focus will be given to genetic variants and clinical features correlating with patient survival.

Association Analysis:

Standard association methods, such as genome-wide association studies (GWAS) and rare variant association studies, will be employed to identify or validate the genetic basis of the identified endophenotypes. Additional analysis will be performed to assess whether the genetic and clinical features extracted from the unsupervised learning improve prognostic prediction and thus could be translated in the clinical setting,

Secondary analysis will then be performed to explore whether the biological pathways or endophenotypes identified in the previous steps are associated with specific environmental exposures. If relevant biological pathways are identified, this information will be used to prioritize potential therapeutic targets.

3.1.6 Challenges and Risks:

The complexity of the genetic basis of ALS, likely resulting from a combination of common and rare genetic variants, is one of the key challenges which this Use Case aims to tackle. Moreover, the interpretation of the results (both regarding endophenotypes, potential novel biomarkers or genetic signals) and their meaningful translation in a clinical setting will require careful consideration.

ALS is a rare and heterogeneous disease, and despite the inclusion of two large and well-characterized cohorts, the sample size may be insufficient to obtain significant results. Finally, it is worth noting that the two study cohorts come from the same ancestry, and therefore, the findings of the use case will necessitate replication in other cohorts before being considered generalizable.

3.2 Use Case 2: Next-generation diagnosing and treatment response for neurodegenerative diseases

3.2.1 Participating partners

UNITO, UNIPD, CRG, SURF, HESSO

3.2.2 Overall aim of the Use Case

The primary objective of this Use Case is to develop a system to identify biologically informed, biomarker-driven classification for neurodegenerative diseases. Leveraging machine learning techniques and multidimensional data (genetic, clinical, and biomarkers) could prompt a change in how we stratify patients with neurodegenerative disorders, integrating clinical diagnosis with putative aetiology mechanisms and molecular or anatomical markers. These methods will help delineate clusters of patients with shared biological backgrounds, paving the way towards precision-medicine approaches, including the development of disease-modifying therapies.

3.2.3 Background

Advances in the establishment of the genetic basis and in vivo biomarkers of neurodegenerative disease have placed the field in the crucial position of shifting from largely clinically based diagnostic criteria to an emphasis on the biological underpinnings of a disease. Such biological classification could be invaluable as a framework for future biomarker-based stratification and staging systems that will allow the implementation of precision medicine approaches to disease modification. Biomarker-based stratification can enhance the detection of effective drugs by ensuring that therapeutic interventions are tested on more homogeneous patient populations, thereby reducing variability in treatment response. Moreover, this approach can facilitate the identification of responder patients, enabling personalized treatment strategies that target the specific pathophysiological processes active in each subgroup.

3.2.4 Rationale

Growing evidence is establishing that neurodegenerative diseases — to date considered as distinct and uniform clinicopathological entities — have various genetic or environmental causes that initiate the disease along different, only partly overlapping pathways. Genetic and multi-omics advances have helped to clarify the differences between neurodegenerative diseases in the same clinical spectrum but, at the same time, have also challenged the traditional boundaries between disorders related to neurodegeneration. For instance, large genomic studies have demonstrated that even established pathogenic mutations are often neither sufficient nor necessary for the development of a clinically defined disease, as some patients carrying a mutation may not have neurological symptoms or manifest features overlapping between diseases.

In this regard, genetic, fluid, and imaging biomarkers are attracting growing interest, as they potentially allow for objective identification of genetic risk, pathological processes, and neurodegeneration. However, despite these advances, the current diagnostic criteria and clinical subgrouping of neurodegenerative disease are based entirely on the identification of clinical features. Furthermore, there is no single neurobiologically based disease construct. A biologically informed classification and diagnosis of

neurodegenerative disease could serve as the basis for objective clinical diagnosis and staging and for accurate subdivision of patients according to pathogenic mechanisms. A biological diagnosis could also advance research in multiple fields, such as epidemiology, biomarker discovery, and precision medicine, including the development of disease-modifying therapies. Indeed, one of the main reasons behind the difficulties in discovering disease-modifying therapies might reside in the exclusive reliance on clinical diagnosis without adequate biological stratification.

Machine learning techniques can be particularly effective in this context. These methods can group patients based on multi-dimensional data without prior knowledge of the outcomes, allowing for the discovery of novel subtypes driven by genetic, clinical, imaging, and environmental interactions. By reconstructing the network of interactions among these variables, researchers can gain a comprehensive understanding of the disease mechanisms at play in different patient subgroups.

The identification of clusters based on clinical features and biomarkers holds significant promise in revealing different pathophysiological profiles of patients manifesting similar symptoms, as well as grouping together patients with different clinical manifestations but sharing the same biological background and drug-response likelihood.

3.2.5 Scientific Approach

Study Population

- Patients with ALS from UNITO and UNIPD (see Use Case 1)
- Patients with Frontotemporal Dementia (FTD) from UNITO (optional)
- Patients with Multiple Sclerosis (MS) from UNITO and UNIPD (optional)
- AD individuals from public access cohort (optional)

Data modalities

- For ALS: see *Use Case 1*
- For MS: Imaging (Brain & spinal cord MRI), clinical scales, immunological data; Optical Coherence Tomography (OCT)
- For FTD: Imaging (Brain 18F-FDG-PET), whole-exome sequencing (FTD)

Variable selection

Variable selection will adopt a similar approach to *Use Case 1.*

For genomic data on ALS, FTD, and MS, genetic analysis will concentrate on SNPs from whole-genome sequencing and GWAS data. Given the challenges posed by rare variants, a separate analytical framework will be used for common and rare variants. The methodologies and thresholds will be tailored to the specific phenotype under investigation and the sample size. Variant associations with ALS will be evaluated using summary statistics from genome-wide association studies (GWAS) or rare variant burden analysis. Genetic variations will be integrated with relevant expression data, such as eQTL. Rare variants will be prioritized based on factors like gene expression levels in the central nervous system, involvement in neurodegenerative diseases, protein-network analysis, and in silico computational protein impact predictions.

For neuroimaging analysis in ALS and FTD, the focus will be on extracting specific features such as cortical volume and signal intensities. Anatomical images will be segmented into distinct brain regions (cortical, subcortical, and brainstem), with cortical thickness and regional volumes measured using standard automated. Intensity values will be estimated for atlas-based regions of interest (ROIs) using automated pipelines. Diffusion tensor imaging (DTI) metrics will be used to reconstruct white matter tracts, and brain networks will be mapped by selecting the interconnecting tracts.

For MS, additional measures will be included (Eshagi et al., 2016; Andorra et al., 2024), such as evidence of disease activity evaluated through lesion activity (presence of gadolinium-enhancing lesions) and lesion load (new or enlarging T2 lesions and T2 lesion volume). Total lesion volume and the evaluation of normal-appearing white matter for each ROI will also be considered.

Section 3.1.5 details the clinical and neurophysiological features of ALS to be included in the analysis. For MS, clinical progression will be assessed through neurological examinations and the Expanded Disability Status Scale (EDSS), with the use of disease-modifying drugs recorded. The presence and pattern of oligoclonal Ig bands in cerebrospinal fluid will serve as an immunological biomarker for MS. Additionally, retinal atrophy monitored by OCT will be included in the analysis (Martinez-Lapiscina et al., 2017; Lin et al., 2021).

Unsupervised learning

Clinical data, neuroimaging, biomarkers and functional properties of disease-associated genetic variants (linked to biological and anatomical entities) are integrated to identify clusters of patients. This approach will be initially tested in ALS. Secondary analyses will be performed to validate the approach in other neurodegenerative disease (FTD or AD) or to a neurological disease of different nature (MS).

Comparison with clinical data

The integrated classification is compared to traditional clinical-based classification. Patients sharing a similar biological basis but different phenotypes will also be prioritized for the research or potential environmental modifiers, including treatment.

3.2.6 Challenges and Risk

Other than the challenges already described in Use Case 1, it should be noted that the available data may not adequately capture the effect of environmental exposures and treatment responses as they change over time, and the number of patients may be insufficient to study these aspects comprehensively.

Furthermore, the identification and validation of novel endophenotypes depend on sufficiently large and representative sample sizes. Inadequate sample sizes can limit the statistical power and generalizability of the findings, potentially hindering the discovery of meaningful biological subgroups. Finally, the success of a biologically informed classification system relies on the availability of comprehensive and high-quality genetic, clinical, imaging, and biomarker data. Gaps in critical data elements may impede the development of robust classification models and reduce the accuracy of the identified endophenotypes.

4 INTEGRATION OF THE USE CASES WITHIN HEREDITARY

4.1 Use Case 3: *Signs of Parkinson's disease in multimodal data*

4.1.1 Overview of the Use Case

This Use Case enhances early diagnosis and monitoring of Parkinson's Disease (PD) using advanced retinal imaging techniques, including Optical Coherence Tomography (OCT). By identifying and validating retinal biomarkers that reflect PD-related neurodegeneration, the project seeks to predict the disease before clinical symptoms manifest. Early detection through retinal biomarkers can significantly improve disease management and the development of disease-modifying treatments.

Then, retinal imaging will be integrated with other biomarkers such as neurological findings and neurophysiological data. Combining these diverse data sources allows for a more comprehensive characterization of PD, providing insights into the disease's progression and underlying pathophysiological mechanisms.

This Use Case will leverage federated learning to ensure the robustness and generalizability of the developed models across multiple clinical sites. This collaborative approach helps validate the findings in diverse populations and clinical settings, enhancing the reliability of the retinal biomarkers identified. Ultimately, the goal is to create a comprehensive, non-invasive tool that integrates various biomarkers for early PD detection and management, potentially transforming the clinical approach to this neurodegenerative disorder.

4.1.2 Scientific Approach

Data modalities

- Clinical (Longitudinal - (Clinical Scales, Cognitive testing)
- Retinal imaging (fundus photos and OCT)
- Neuroimaging (Brain MRI/PET/SPECT)
- Neurophysiological data (EEG and Deep Brain Recording)
- Biomarkers (plasma, cerebrospinal fluid)

Variable selection

To standardize the cohort, inclusion/exclusion criteria for the diagnosis of PD and related disorders will be adopted based on existing well-established criteria. Individuals with conditions that could interfere with the study outcomes will be excluded to maintain the integrity of our results. Collected clinical data include detailed records of disease history (disease duration, treatment history and disease severity using standardized), and neurological and ophthalmologic. Imaging data will be derived from high-quality and reproducible techniques. Strict quality control measures will be implemented to ensure that only high-quality images are included, and available tools will be used to harmonize data, ensuring consistency and accuracy in the data collected.

4.1.3 Integration with Use Cases 1 and 2

This Use Case focuses on Parkinson's Disease, a common neurodegenerative disorder, and thus shares similar goals and methodologies to Use Cases 1 and 2. Therefore, we expect the three use cases to reciprocally inform each other on tasks such as feature extraction, multimodal data integration (including imaging, neurophysiological, cognitive, and clinical data), and machine-learning methods. The collaboration among the partners involved in the Use Cases will provide valuable insights into how to combine multimodal data for disease prediction and classification. Additionally, they will share methodologies for harmonizing, sharing, and analysing data from multiple centres, fostering a unified approach to early disease detection and progression monitoring across the project.

4.2 Use Cases 4 and 5: Phenotyping of the gut-brain axis in healthy individuals to understand deviations in disorders and Gut-Brain linkage and disease relevance

4.2.1 Overview of the Use Cases

These Use Cases aim to establish a comprehensive understanding of the gut-brain axis (GBA), a bidirectional communication system linking the central and enteric nervous systems, and its influence on health and disease. Recent research has highlighted the importance of the gut microbiota in these interactions, yet the extent of its impact on brain function remains unclear. By mapping the healthy gut-brain axis, these Use Cases will identify deviations associated with various diseases, providing a reference for future research and potential therapeutic interventions targeting the microbiome.

To achieve this, the relationship between gut microbiome alterations, environmental factors, and health-related data will be investigated in a large population sample. This includes exploring associations between specific gut microbiome changes, environmental pollutants, genetic factors, and metabolic health outcomes. The goal is to create a detailed map of the healthy gut-brain axis, offering valuable insights for understanding the mechanisms behind these interactions and their implications for neurological and neuropsychiatric health.

The approach involves the analysis of microbiome and metabolome data, histological data from colon biopsies, and functional MRI (fMRI) data from healthy individuals and psychiatric disorders. Using advanced neuroimaging techniques and neuroimaging-omics, the project aims to visualize the impact of the gut microbiota on the human brain. Techniques such as Linked Independent Component Analysis (LICA) will be employed to integrate brain connectivity maps with microbiome data, uncovering patterns of variation shared between brain networks and gut microbiota. These findings will then be linked with health-related factors such as behaviour, stress, societal influences, and environmental conditions. By linking brain network connectivity and microbiome compositions to these factors, the project aims to elucidate the complex interactions within the gut-brain axis. This comprehensive approach will provide a foundation for identifying potential therapeutic targets and interventions, enhancing our understanding of the gut-brain connection in health and disease.

4.2.2 Scientific Approach

Data modalities

- Clinical (Longitudinal)
- Microbiome and metabolome data
- Histological data from colon biopsies
- Functional MRI (fMRI)

Variable selection

Biosamples will be extracted from the Healthy Brain Study (HBS) database, which includes data from 700 individuals including repeated measures. These participants have undergone extensive phenotyping, including detailed information from various lifestyle and life-history questionnaires, cognitive tasks, saliva cortisol samples, stress-related measures. Microbiome and metabolome analysis will be performed on fecal samples. A relative abundance table will summarize the internal composition of the microbial community in each subject, while the pathway abundance table of the metabolic functionality will be used to map putative pathway activity on specific functional and metabolic activities. Functional and structural brain data will be acquired from Magnetic Resonance Imaging (MRI). Standard preprocessing, smoothing and filtering, and normalization procedures will be adopted to ensure the accuracy and comparability of the brain images.

4.2.3 Integration with Use Cases 1 and 2

Use Cases 4 and 5 will provide invaluable insights into how to integrate multiple --omics data, including metabolomics and brain imaging data. Furthermore, it will demonstrate how multimodal data integration can be applied to investigate environmental factors, thus enhancing this task in Use Cases 1 and 2. Additionally, Use Cases 4 and 5 will gain insights into the role of the gut-brain axis pathways in healthy individuals, which could be later leveraged to explore the role of the gut microbiome in neurodegenerative disease.

REFERENCES

The bibliographic entries are arranged in lexicographical order based on the key, following the APA style. This enables us to place the entries and citations in the table and text in any sequence, allowing for later sorting while ensuring consistency.

Key	Reference
Andorra et al., 2024	Andorra, M., Freire, A., Zubizarreta, I., de Rosbo, N. K., Bos, S. D., Rinas, M., Høgestøl, E. A., de Rodez Benavent, S. A., Berge, T., Brune-Ingebretse, S., Ivaldi, F., Cellerino, M., Pardini, M., Vila, G., Pulido-Valdeolivas, I., Martinez-Lapiscina, E. H., Llufriu, S., Saiz, A., Blanco, Y., Martinez-Heras, E., … Villoslada, P. (2024). Predicting disease severity in multiple sclerosis using multimodal data and machine learning. Journal of neurology, 271(3), 1133–1149. https://doi.org/10.1007/s00415-023-12132-z
Arenaza-Urquijo et al., 2013	Arenaza-Urquijo, E. M., Molinuevo, J.-L., Sala-Llonch, R., Solé-Padullés, C., Balasa, M., Bosch, B., Olives, J., Antonell, A., Lladó, A., Sánchez-Valle, R., Rami, L., & Bartrés-Faz, D. (2013). Cognitive Reserve Proxies Relate to Gray Matter Loss in Cognitively Healthy Elderly with Abnormal Cerebrospinal Fluid Amyloid-β Levels. Journal of Alzheimer's Disease, 35(4), 715–726. https://doi.org/10.3233/JAD-121906
Aschard et al., 2014	Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., & Kraft, P. (2014). Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies. The American Journal of Human Genetics, 94(5), 662–676. https://doi.org/10.1016/j.ajhg.2014.03.016
Aziz et al., 2022	Aziz, M. M. A., Anjum, M. M., Mohammed, N., & Jiang, X. (2022). Generalized genomic data sharing for differentially private federated learning. Journal of Biomedical Informatics, 132, 104113. https://doi.org/10.1016/j.jbi.2022.104113
Brisimi et al., 2018	Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. Ch., & Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. International Journal of Medical Informatics, 112, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
Cho et al., 2018	Cho, H., Wu, D. J., & Berger, B. (2018). Secure genome-wide association analysis using multiparty computation. Nature Biotechnology, 36(6), 547–551. https://doi.org/10.1038/nbt.4108
Cole & Franke, 2017	Cole, J. H., & Franke, K. (2017). Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends in Neurosciences, 40(12), 681–690. https://doi.org/10.1016/j.tins.2017.10.001
Cole et al., 2017	Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage, 163, 115–124. https://doi.org/10.1016/j.neuroimage.2017.07.059

Key	Reference
Colombo et al., 2023	Colombo, E., Doretti, A., Scheveger, F., Maranzano, A., Pata, G., Gagliardi, D., Meneri, M., Messina, S., Verde, F., Morelli, C., Corti, S., Maderna, L., Silani, V., & Ticozzi, N. (2023). Correlation between clinical phenotype and electromyographic parameters in amyotrophic lateral sclerosis. Journal of neurology, 270(1), 511–518. https://doi.org/10.1007/s00415-022-11404-4
Cornblath et al., 2020	Cornblath, E. J., Robinson, J. L., Irwin, D. J., Lee, E. B., Lee, V. M.-Y., Trojanowski, J. Q., & Bassett, D. S. (2020). Defining and predicting transdiagnostic categories of neurodegenerative disease. Nature Biomedical Engineering, 4(8), 787–800. https://doi.org/10.1038/s41551-020-0593-y
Elliott et al., 2021	Elliott, M. L., Belsky, D. W., Knodt, A. R., Ireland, D., Melzer, T. R., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T. E., & Hariri, A. R. (2021). Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. Molecular Psychiatry, 26(8), 3829–3838. https://doi.org/10.1038/s41380-019-0626-7
Emon et al., 2020	Emon, M. A., Heinson, A., Wu, P., Domingo-Fernández, D., Sood, M., Vrooman, H., Corvol, J.-C., Scordis, P., Hofmann-Apitius, M., & Fröhlich, H. (2020). Clustering of Alzheimer's and Parkinson's disease based on genetic burden of shared molecular mechanisms. Scientific Reports, 10(1), 19097. https://doi.org/10.1038/s41598-020-76200-4
Eshaghi et al., 2021	Eshaghi, A., Young, A. L., Wijeratne, P. A., Prados, F., Arnold, D. L., Narayanan, S., Guttmann, C. R. G., Barkhof, F., Alexander, D. C., Thompson, A. J., Chard, D., & Ciccarelli, O. (2021). Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. Nature communications, 12(1), 2078. https://doi.org/10.1038/s41467-021-22265-2
Franke et al., 2010	Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage, 50(3), 883–892. https://doi.org/10.1016/j.neuroimage.2010.01.005
Gaser et al., 2013	Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., & Alzheimer's Disease Neuroimaging Initiative. (2013). BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. PLoS ONE, 8(6), e67346. https://doi.org/10.1371/journal.pone.0067346
Groot et al., 2018	Groot, C., Van Loenhoud, A. C., Barkhof, F., Van Berckel, B. N. M., Koene, T., Teunissen, C. C., Scheltens, P., Van Der Flier, W. M., & Ossenkoppele, R. (2018). Differential effects of cognitive reserve and brain reserve on cognition in Alzheimer disease. Neurology, 90(2). https://doi.org/10.1212/WNL.0000000000004802
Han et al., 2021	Han, L. K. M., Dinga, R., Hahn, T., Ching, C. R. K., Eyler, L. T., Aftanas, L., Aghajani, M., Aleman, A., Baune, B. T., Berger, K., Brak,

Key	Reference
	I., Filho, G. B., Carballedo, A., Connolly, C. G., Couvy-Duchesne, B., Cullen, K. R., Dannlowski, U., Davey, C. G., Dima, D., … Schmaal, L. (2021). Brain aging in major depressive disorder: Results from the ENIGMA major depressive disorder working group. Molecular Psychiatry, 26(9), 5124–5139. https://doi.org/10.1038/s41380-020-0754-0
Hertel et al., 2022	Hertel, N., Kuzma-Kozakiewicz, M., Gromicho, M., Grosskreutz, J., de Carvalho, M., Uysal, H., Dengler, R., Petri, S., & Körner, S. (2022). Analysis of routine blood parameters in patients with amyotrophic lateral sclerosis and evaluation of a possible correlation with disease progression-a multicenter study. Frontiers in neurology, 13, 940375. https://doi.org/10.3389/fneur.2022.940375
Höglinger et al., 2024	Höglinger, G. U., Adler, C. H., Berg, D., Klein, C., Outeiro, T. F., Poewe, W., Postuma, R., Stoessl, A. J., & Lang, A. E. (2024). A biological classification of Parkinson's disease: The SynNeurGe research diagnostic criteria. The Lancet Neurology, 23(2), 191–204. https://doi.org/10.1016/S1474-4422(23)00404-0
Ingre et al., 2020	Ingre, C., Chen, L., Zhan, Y., Termorshuizen, J., Yin, L., & Fang, F. (2020). Lipids, apolipoproteins, and prognosis of amyotrophic lateral sclerosis. Neurology, 94(17), e1835–e1844. https://doi.org/10.1212/WNL.0000000000009322
Jonsson et al., 2019	Jonsson, B. A., Bjornsdottir, G., Thorgeirsson, T. E., Ellingsen, L. M., Walters, G. B., Gudbjartsson, D. F., Stefansson, H., Stefansson, K., & Ulfarsson, M. O. (2019). Brain age prediction using deep learning uncovers associated sequence variants. Nature Communications, 10(1), 5409. https://doi.org/10.1038/s41467-019-13163-9
Karolinska Schizophrenia Project (KaSP) et al., 2019	Karolinska Schizophrenia Project (KaSP), Kaufmann, T., Van Der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M. K., Bøen, E., Borgwardt, S., Brandt, C. L., Buitelaar, J., Celius, E. G., Cervenka, S., … Westlye, L. T. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nature Neuroscience, 22(10), 1617–1623. https://doi.org/10.1038/s41593-019-0471-7
Katsonis et al., 2022	Katsonis, P., Wilhelm, K., Williams, A., & Lichtarge, O. (2022). Genome interpretation using in silico predictors of variant impact. Human genetics, 141(10), 1549-1577.
Koretsky et al., 2023	Koretsky, M. J., Alvarado, C., Makarious, M. B., Vitale, D., Levine, K., Bandres-Ciga, S., Dadu, A., Scholz, S. W., Sargent, L., Faghri, F., Iwaki, H., Blauwendraat, C., Singleton, A., Nalls, M., & Leonard, H. (2023). Genetic risk factor clustering within and across neurodegenerative diseases. Brain, 146(11), 4486–4494. https://doi.org/10.1093/brain/awad161
Kuan et al., 2023	Kuan, L. H., Parnianpour, P., Kushol, R., Kumar, N., Anand, T., Kalra, S., & Greiner, R. (2023). Accurate personalized survival prediction

Key	Reference
	for amyotrophic lateral sclerosis patients. Scientific reports, 13(1), 20713. https://doi.org/10.1038/s41598-023-47935-7
Leonardsen et al., 2022	Leonardsen, E. H., Peng, H., Kaufmann, T., Agartz, I., Andreassen, O. A., Celius, E. G., Espeseth, T., Harbo, H. F., Høgestøl, E. A., Lange, A.-M. D., Marquand, A. F., Vidal-Piñeiro, D., Roe, J. M., Selbæk, G., Sørensen, Ø., Smith, S. M., Westlye, L. T., Wolfers, T., & Wang, Y. (2022). Deep neural networks learn general and clinically relevant representations of the ageing brain. NeuroImage, 256, 119210. https://doi.org/10.1016/j.neuroimage.2022.119210
Li et al., 2015	Li, P., Nie, Y., & Yu, J. (2015). An Effective Method to Identify Shared Pathways and Common Factors among Neurodegenerative Diseases. PLOS ONE, 10(11), e0143045. https://doi.org/10.1371/journal.pone.0143045
Li et al., 2020a	Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., & Duncan, J. S. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Medical Image Analysis, 65, 101765. https://doi.org/10.1016/j.media.2020.101765
Li et al., 2020b	Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., ... & Lin, X. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nature genetics, 52(9), 969-983.
Lin et al., 2021	Lin, T. Y., Vitkova, V., Asseyer, S., Martorell Serra, I., Motamedi, S., Chien, C., ... & Zimmermann, H. G. (2021). Increased serum neurofilament light and thin ganglion cell–inner plexiform layer are additive risk factors for disease activity in early multiple sclerosis. Neurology: Neuroimmunology & Neuroinflammation, 8(5), e1051.
Martinez-Lapiscina et al., 2016	Martinez-Lapiscina, E. H., Arnow, S., Wilson, J. A., Saidha, S., Preiningerova, J. L., Oberwahrenbrock, T., ... & Villoslada, P. (2016). Retinal thickness measured with optical coherence tomography and risk of disability worsening in multiple sclerosis: a cohort study. The Lancet Neurology, 15(6), 574-584.
Murdock et al., 2021	Murdock, B. J., Goutman, S. A., Boss, J., Kim, S., & Feldman, E. L. (2021). Amyotrophic Lateral Sclerosis Survival Associates With Neutrophils in a Sex-specific Manner. Neurology(R) neuroimmunology & neuroinflammation, 8(2), e953. https://doi.org/10.1212/NXI.0000000000000953
Nicolas et al., 2018	Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J. A., Kenna, B. J., Nalls, M. A., Keagle, P., Rivera, A. M., van Rheenen, W., Murphy, N. A., van Vugt, J. J. F. A., Geiger, J. T., Van der Spek, R. A., Pliner, H. A., Shankaracharya, Smith, B. N., Marangi, G., … Landers, J. E. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron, 97(6), 1267–1288. https://doi.org/10.1016/j.neuron.2018.02.027

Key	Reference
Ning et al., 2020	Ning, K., Zhao, L., Matloff, W., Sun, F., & Toga, A. W. (2020). Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants. Scientific Reports, 10(1), 10. https://doi.org/10.1038/s41598-019-56089-4
Patel et al., 2024	Patel, K., Xie, Z., Yuan, H., Islam, S. M. S., Xie, Y., He, W., Zhang, W., Gottlieb, A., Chen, H., Giancardo, L., Knaack, A., Fletcher, E., Fornage, M., Ji, S., & Zhi, D. (2024). Unsupervised deep representation learning enables phenotype discovery for genetic association studies of brain imaging. Communications Biology, 7(1), 414. https://doi.org/10.1038/s42003-024-06096-7
Querbes et al., 2009	Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., & The Alzheimer's Disease Neuroimaging Initiative. (2009). Early diagnosis of Alzheimer's disease using cortical thickness: Impact of cognitive reserve. Brain, 132(8), 2036–2047. https://doi.org/10.1093/brain/awp105
Raimondi et al., 2023	Raimondi, D., Chizari, H., Verplaetse, N., Löscher, B.-S., Franke, A., & Moreau, Y. (2023). Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of Crohn's disease patients. Scientific Reports, 13(1), 19449. https://doi.org/10.1038/s41598-023-46887-2
Schnack et al., 2016	Schnack, H. G., Van Haren, N. E. M., Nieuwenhuis, M., Hulshoff Pol, H. E., Cahn, W., & Kahn, R. S. (2016). Accelerated Brain Aging in Schizophrenia: A Longitudinal Pattern Recognition Study. American Journal of Psychiatry, 173(6), 607–616. https://doi.org/10.1176/appi.ajp.2015.15070922
Seyedsalehi et al., 2023	Seyedsalehi, A., Warrier, V., Bethlehem, R. A. I., Perry, B. I., Burgess, S., & Murray, G. K. (2023). Educational attainment, structural brain reserve and Alzheimer's disease: A Mendelian randomization analysis. Brain, 146(5), 2059–2074. https://doi.org/10.1093/brain/awac392
Smith et al., 2019	Smith, S. M., Elliott, L. T., Alfaro-Almagro, F., McCarthy, P., Nichols, T. E., Douaud, G., & Miller, K. L. (2020). Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. eLife, 9, e52677. https://doi.org/10.7554/eLife.52677
Smith et al., 2020	Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., & Miller, K. L. (2019). Estimation of brain age delta from brain imaging. NeuroImage, 200, 528–539. https://doi.org/10.1016/j.neuroimage.2019.06.017
Solé-Padullés et al., 2009	Solé-Padullés, C., Bartrés-Faz, D., Junqué, C., Vendrell, P., Rami, L., Clemente, I. C., Bosch, B., Villar, A., Bargalló, N., Jurado, M. A., Barrios, M., & Molinuevo, J. L. (2009). Brain structure and function related to cognitive reserve variables in normal aging, mild cognitive impairment and Alzheimer's disease. Neurobiology of Aging, 30(7), 1114–1124. https://doi.org/10.1016/j.neurobiolaging.2007.10.008

Key	Reference
Steffener et al., 2016	Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., & Stern, Y. (2016). Differences between chronological and brain age are related to education and self-reported physical activity. Neurobiology of Aging, 40, 138–144. https://doi.org/10.1016/j.neurobiolaging.2016.01.014
Stern, 2012	Stern, Y. (2012). Cognitive reserve in ageing and Alzheimer's disease. The Lancet Neurology, 11(11), 1006–1012. https://doi.org/10.1016/S1474-4422(12)70191-6
Stern et al., 2020	Stern, Y., Arenaza-Urquijo, E. M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., Ewers, M., Franzmeier, N., Kempermann, G., Kremen, W. S., Okonkwo, O., Scarmeas, N., Soldan, A., Udeh-Momoh, C., Valenzuela, M., Vemuri, P., Vuoksimaa, E., & and the Reserve, Resilience and Protective Factors PIA Empirical Definitions and Conceptual Frameworks Workgroup. (2020). Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimer's & Dementia, 16(9), 1305–1311. https://doi.org/10.1016/j.jalz.2018.07.219
Van der Burgh et al., 2016	van der Burgh, H. K., Schmidt, R., Westeneng, H. J., de Reus, M. A., van den Berg, L. H., & van den Heuvel, M. P. (2016). Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. NeuroImage. Clinical, 13, 361–369. https://doi.org/10.1016/j.nicl.2016.10.008
van Rheenen et al., 2021	van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., de Klein, N., Westra, H. J., Bakker, O. B., Deelen, P., Shireby, G., Hannon, E., Moisse, M., Baird, D., Restuadi, R., Dolzhenko, E., Dekker, A. M., Gawor, K., Westeneng, H. J., Tazelaar, G. H. P., … Veldink, J. H. (2021). Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nature genetics, 53(12), 1636–1648. https://doi.org/10.1038/s41588-021-00973-1
Villoslada et al., 2020	Villoslada, P., Baeza-Yates, R., & Masdeu, J. C. (2020). Reclassifying neurodegenerative diseases. Nature Biomedical Engineering, 4(8), 759–760. https://doi.org/10.1038/s41551-020-0600-3
Xu et al., 2021	Xu, Z., Wu, C., Wei, P., & Pan, W. (2017). A powerful framework for integrating eQTL and GWAS summary data. Genetics, 207(3), 893-902.
Wen et al., 2024	Wen, J., Zhao, B., Yang, Z., Erus, G., Skampardoni, I., Mamourian, E., Cui, Y., Hwang, G., Bao, J., Boquet-Pujadas, A., Zhou, Z., Veturi, Y., Ritchie, M. D., Shou, H., Thompson, P. M., Shen, L., Toga, A. W., & Davatzikos, C. (2024). The genetic architecture of multimodal human brain age. Nature Communications, 15(1), 2604. https://doi.org/10.1038/s41467-024-46796-6
Wrigglesworth et al., 2021	Wrigglesworth, J., Ward, P., Harding, I. H., Nilaweera, D., Wu, Z., Woods, R. L., & Ryan, J. (2021). Factors associated with brain

Key	Reference
	ageing—A systematic review. BMC Neurology, 21(1), 312. https://doi.org/10.1186/s12883-021-02331-4
Wu et al., 2021	Wu, X., Zheng, H., Dou, Z., Chen, F., Deng, J., Chen, X., Xu, S., Gao, G., Li, M., Wang, Z., Xiao, Y., Xie, K., Wang, S., & Xu, H. (2021). A novel privacy-preserving federated genome-wide association study framework and its application in identifying potential risk variants in ankylosing spondylitis. Briefings in Bioinformatics, 22(3), bbaa090. https://doi.org/10.1093/bib/bbaa090
Zhu et al., 2016	Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., ... & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics, 48(5), 481-487.
Zolotareva et al., 2021	Wu, X., Zheng, H., Dou, Z., Chen, F., Deng, J., Chen, X., Xu, S., Gao, G., Li, M., Wang, Z., Xiao, Y., Xie, K., Wang, S., & Xu, H. (2021). A novel privacy-preserving federated genome-wide association study framework and its application in identifying potential risk variants in ankylosing spondylitis. Briefings in Bioinformatics, 22(3), bbaa090. https://doi.org/10.1093/bib/bbaa090