Towards enabling precision medicine in Alzheimer's disease and Parkinson's disease

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Colin Birkenbihl

aus San Diego, Vereinigte Staaten von Amerika

Bonn, 05. Juni 2023

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Honorar Prof. Dr. rer. nat. Holger Fröhlich
2. Gutachter: Univ.-Prof. Dr. rer. nat. Thomas Schultz

Tag der Promotion: 19. Februar 2024
Erscheinungsjahr: 2024

Abstract

Alzheimer's disease and Parkinson's disease are prominent progressive neurodegenerative diseases, with a significant clinical and economic impact on patients, their families, and society as a whole. Despite numerous clinical trials over the last two decades, no disease-modifying treatment is available for either disease. The past trial failures are attributed in part to the heterogeneity of the diseases with respect to clinical presentation and pathological manifestation and the late timing of interventions in the course of the disease. The emerging healthcare paradigm of precision medicine aims to address these challenges by bringing the right drug to the right patient at the right time.

The research presented in this thesis relies on artificial intelligence and machine learning to advance precision medicine in the context of Alzheimer's disease and Parkinson's disease. We contribute to a deeper understanding of these complex diseases through patient subtyping and present novel predictive models for patient stratification. Furthermore, we make a new patient-level dataset openly accessible for research purposes and present a web application that facilitates the exploration of large cohort datasets in the Alzheimer's domain. Additionally, we investigate systematic biases in commonly used data resources, proposing methods to assess and understand them. By doing so, we empower researchers to make informed decisions about data selection, enhancing the reliability, generalizability, and utility of their findings. Finally, we introduce a novel artificial intelligence-based approach for generating synthetic patient-level data, which can help overcoming limitations in real patient data.

In conclusion, the scientific advancements presented in this thesis collectively support robust data-driven research in the context of Alzheimer's disease and Parkinson's disease. They provide further insight into the heterogeneity of these debilitating diseases that could be leveraged to pave the way towards more effective clinical trials that are guided by the principles of precision medicine.

Acknowledgments

I want to express my deepest gratitude towards my PhD supervisor and mentor Prof. Dr. Holger Fröhlich. Your scientific rigour, strive for perfection, and relentless dedication to research are an inspiration to me. You introduced me to the field of AI, data science, and advanced statistics that I became passionate about. The knowledge I owe to our countless, lively discussions is invaluable to me. You significantly shaped the vision I have for my future career through your lessons and actions. Thank you.

I thank Prof. Dr. Thomas Schultz for agreeing to be my second PhD supervisor and reviewer of this thesis. Additionally, I want to thank him for his lessons and our scientific discussions.

To Prof. Dr. Martin Hofmann-Apitius, I extend my gratitude for many valuable lessons and opportunities that I was fortunate enough to be presented with. Despite my initial juniority, you saw the potential scientist that I could grew to become and I genuinely appreciate your trust and support. Only few would have done what you have done for me.

Thank you to all my colleagues and collaborators. It was an honor to research alongside you. I want to address a special thank you to all the students I had the pleasure to supervise.

Finally, I want to thank my family for their unconditional support. My father, who sparked my interest in science already as a young kid. My mother, who taught me the value of education. My stepfather, who formed my understanding of communication and human interaction. Kai and my daughter Cleo, who mean the world to me.

Declaration

I hereby certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Colin Birkenbihl

Publications

Thesis Publications

† Equal contribution

1. **Birkenbihl, C.†**, Salimi, Y.†, Fröhlich, H., Japanese Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Neuroimaging Initiative. (2022). Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimer's & Dementia*, 18(2), 251-261. https://doi.org/10.1002/alz.12387

2. **Birkenbihl, C.**, Emon, M. A., Vrooman, H., Westwood, S., Lovestone, S., AddNeuroMed Consortium, , Hofmann-Apitius M., Fröhlich, H., and Alzheimer's Disease Neuroimaging Initiative. (2020). Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA Journal*, 11, 367-376. https://doi.org/10.1007/s13167-020-00216-z

3. **Birkenbihl, C.**, Ahmad, A., Massat, N. J., Raschka, T., Avbersek, A., Downey, P., Armstrong, M., and Fröhlich, H. (2023). Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories. *Scientific Reports*, 13(1), 2897. https://doi.org/10.1038/s41598-023-30038-8

4. **Birkenbihl, C.**, Salimi, Y., Domingo-Fernández, D., Lovestone, S., AddNeuroMed Consortium, Fröhlich, H., Hofmann-Apitius M., and Alzheimer's Disease Neuroimaging Initiative. (2020). Evaluating the Alzheimer's disease data landscape. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1), e12102. https://doi.org/10.1002/trc2.12102

5. Salimi, Y., Domingo-Fernández, D., Bobis-Álvarez, C., Hofmann-Apitius, M., and **Birkenbihl, C.**, for the Alzheimer's Disease Neuroimaging Initiative, the Japanese Alzheimer's Disease Neuroimaging Initiative, for the Aging Brain: Vasculature, Ischemia, and Behavior Study, the

Alzheimer's Disease Repository Without Borders Investigators, for the European Prevention of Alzheimer's Disease (EPAD) Consortium. (2022). ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets. *Alzheimer's Research & Therapy*, 14(1), 69. https://doi.org/10.1186/s13195-022-01009-4

6. **Birkenbihl, C.**, Westwood, S., Shi, L., Nevado-Holgado, A., Westman, E., Lovestone, S., Hofmann-Apitius, M., and AddNeuroMed Consortium. (2021). ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *Journal of Alzheimer's Disease*, 79(1), 423-431. https://doi.org/10.3233/JAD-200948

7. Wendland, P.[†], **Birkenbihl, C.[†]**, Gomez-Freixa, M., Sood, M., Kschischo, M., and Fröhlich, H. (2022). Generation of realistic synthetic data using multimodal neural ordinary differential equations. *NPJ Digital Medicine*, 5(1), 122. https://doi.org/10.1038/s41746-022-00666-x

Other Publications

8. Golriz Khatami, S., Salimi, Y., Hofmann-Apitius, M., Oxtoby, N. P.[†], and **Birkenbihl, C.[†]** (2022). Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease. *Alzheimer's Research & Therapy*, 14(1), 1-14. https://doi.org/10.1186/s13195-022-01001-y

9. Maheux, E., Koval, I., Ortholand, J., **Birkenbihl, C.**, Archetti, D., Bouteloup, V., Epelbaum, S., Dufouil, C., Hofmann-Apitius, M., and Durrleman, S. (2023). Forecasting individual progression trajectories in Alzheimer's disease. *Nature Communications*, 14(1), 761. https://doi.org/10.1038/s41467-022-35712-5

10. Golriz Khatami, S., Robinson, C., **Birkenbihl, C.**, Domingo-Fernández, D., Hoyt, C. T., and Hofmann-Apitius, M. (2020). Challenges of integrative disease modeling in Alzheimer's disease. *Frontiers in molecular biosciences*, 6, 158. https://doi.org/10.3389/fmolb.2019.00158

11. Golubnitschaja, O., Liskova, A., Koklesova, L., Samec, M., Biringer, K., Büsselberg, D., Podbielska A., Kunin, A. A., Evsevyeva, M. E., Shapira, N., Friedemann, P., Erb, C., Dietrich, D. E., Felbel, D., Karabatsiakis, A., Bubnov, R., Polivka, J., Polivka J. Jr., **Birkenbihl, C.**, Fröhlich,

H., Hofmann-Apitius, M., and Kubatka, P. (2021). Caution,“normal” BMI: health risks associated with potentially masked individual underweight — EPMA Position Paper 2021. *EPMA Journal*, 12(3), 243-264. https://doi.org/10.1007/s13167-021-00251-4

12. Kunin, A.[†], Sargheini, N.[†], **Birkenbihl, C.[†]**, Moiseeva, N., Fröhlich, H., and Golubnitschaja, O. (2020). Voice perturbations under the stress overload in young individuals: phenotyping and suboptimal health as predictors for cascading pathologies. *EPMA Journal*, 11, 517-527. https://doi.org/10.1007/s13167-020-00229-8

13. Balabin, H., Hoyt, C. T., **Birkenbihl, C.**, Gyori, B. M., Bachman, J., Kodamullil, A. T., Plöger, P. G., Hofmann-Apitius, M., and Domingo-Fernández, D. (2022). STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs. *Bioinformatics*, 38(6), 1648-1656. https://doi.org/10.1093/bioinformatics/btac001

14. Evsevieva, M., Sergeeva, O., Mazurakova, A., Koklesova, L., Prokhorenko-Kolomoytseva, I., Shchetinin, E., **Birkenbihl, C.**, Costigliola, V., Kubatka, P., and Golubnitschaja, O. (2022). Pre-pregnancy check-up of maternal vascular status and associated phenotype is crucial for the health of mother and offspring. *EPMA Journal*, 13(3), 351-366. https://doi.org/10.1007/s13167-022-00294-1

15. Bharadhwaj, V. S., Ali, M., **Birkenbihl, C.**, Mubeen, S., Lehmann, J., Hofmann-Apitius, M., Hoyt, C. T., and Domingo-Fernández, D. (2021). CLEP: a hybrid data-and knowledge-driven framework for generating patient representations. *Bioinformatics*, 37(19), 3311-3318. https://doi.org/10.1093/bioinformatics/btab340

16. Wegner, P., Schaaf, S., Uebachs, M., Domingo-Fernández, D., Salimi, Y., Gebel, S., Sargsyan, A., **Birkenbihl, C.**, Springstubbe, S., Klockgether, T., Fluck, J., Hofmann-Apitius, M., and Kodamullil, A. T. (2022). Integrative data semantics through a model-enabled data stewardship. *Bioinformatics*, 38(15), 3850-3852. https://doi.org/10.1093/bioinformatics/btac375

Contents

1 Introduction

1.1 Neurodegenerative diseases

neurodegenerative diseases (NDD) are characterized by neuronal loss and gradual deterioration of the brain and nervous system [1]. Alzheimer's disease (AD) and Parkinson's disease (PD) are the two most common neurodegenerative diseases world-wide [2, 3]. They share a progressive course of the disease, clinical symptoms such as cognitive impairment, and pathological developments like abnormal aggregations of amyloid beta plaques. There is currently no disease modifying treatment for either of the diseases. While there are inheritable forms of AD and PD [4, 5], this thesis will focus on their idiopathic, sporadic manifestations.

1.1.1 Alzheimer's disease

AD is a progressive neurodegenerative disorder characterized by worsening symptoms of cognitive impairment and perturbation of everyday life which often culminates in a premature death [3]. AD is the leading cause of dementia, accounting for 75% of the dementia cases world-wide. It is also the sixth most common cause of death in the United States of America [6]. With progressing symptoms, patients become dependent on intensive full-time care which contributes to the economical burden for society. In 2019, providing healthcare, long-term care, and hospice services to dementia patients older than 65 years costed an estimated 290 billion USD in the US alone [6]. Additional services provided by family members and other unpaid caregivers are estimated at more than 234 billion USD annually.

The predominant form of AD is sporadic late-onset AD and the underlying etiology is hypothesized to be multifactorial and remains unknown [7]. The characteristic pathological changes in the brain include the formation of amyloid beta plaques and tau neurofibrillary tangles [8]. Both processes have been connected to neurodegeneration observed in patients suffering from AD dementia. Further factors that are believed to contribute towards an AD phenotype include perturbations in neuroinflammatory response [9], cardiovascular health [10], and apolipoprotein E (APOE) associated cholesterol metabolism [11].

Clinically, AD is commonly ordered into three progressive stages that group patients according to their increasing symptom severity, namely 1) pre-clinical AD or cognitively unimpaired, 2) mild cognitive impairment (MCI) due to AD, and 3) dementia due to AD for which ’AD’ has been often been used synonymously [6]. Pre-clinical AD is characterized by the presence of AD pathology in the form of amyloid beta aggregation and tau fibrillary tangle formation, while symptomatically patients are indistinguishable from other cognitively unimpaired individuals. MCI and clinical AD patients exhibit cognitive symptoms of increasing severity. In the past, AD was diagnosed purely based on cognitive assessments [12], which was later revised to include emerging biomarkers of AD pathology [13], and is on the verge of transitioning towards a fully biological diagnosis [14]. Considering AD a purely biological entity while disregarding patients’ symptoms is, as of now, not clinically applicable [3, 15] and challenged by the fact that many amyloid positive patients never develop symptoms that exceed what is expected given their age [16–18].

As of now, no disease modifying treatment is available for AD and until 2021 only four drugs for symptomatic treatment had been approved by the U.S. Food and Drug Administration (FDA), the last of which dated back to 2003 [19, 20]. All of these approved drugs offer only limited cognitive protection and can not halt the progression of the disease [19]. This dilemma was also recognized by the G8 in 2013, when they declared the development of a treatment for AD a primary goal until 2025 [21]. Between 2004 and 2021, about 550 phase II and phase III trials with cognitive endpoints were registered for AD [20]. None of these trials succeeded either due to the occurrence of adverse events [22] or lack of efficacy [23, 24]. Assignment to treatment arms in AD trials had even been linked to a net negative impact on patient health [25].

Recently, however, clinical trials have led to partial successes: beyond the four already mentioned drugs, two new treatments have been approved by the FDA over the last two years. In 2021, a controversial accelerated FDA approval was granted for aducanumab, a monoclonal antibody targeting amyloid beta plaques inside the brain [26]. However, due to conflicting evidence about treatment efficacy the approval was rejected by the European Medical Agency (EMA). Meanwhile, the US Medicaid system refused treatment coverage outside of clinical trials until more evidence was generated, due to the high treatment costs. The manufacturer consequently stopped the marketing of aducanumab [27]. In January 2023, lecanemab, another antibody targeting amyloid beta, was approved by the FDA, again in an accelerated procedure [28]. While cognitive decline was statistically significantly slowed compared to a placebo group, researchers have questioned the clinical significance of the outcome reduction [29, 30]. Lecanemab treatment eligibility requires a positron emission tomography (PET) scan to confirm the presence of amyloid plaques and adverse events are tracked via repeated magnetic resonance imaging (MRI). These factors make the treatment unaffordable for low and mid-income countries, where the majority of AD cases reside [31]. In conclusion, while progress has been made in the search for a treatment, the current situation remains critical.

1.1.2 Parkinson's disease

Second to AD, PD is the most common neurodegenerative disease [2] with more than six million affected patients in 2016 [32]. Approximately 3% of the global population above 80 years of age suffer from PD and 5 to 35 new cases emerge per 100,000 individuals yearly. Its cardinal symptoms include resting tremor, bradykinesia, and rigidity [33]. Besides these so-called motor-symptoms, a wide variety of non-motor symptoms, such as depression, psychosis, sleep disturbance, and cognitive impairment can be observed among PD patients [34]. These symptoms become increasingly dominant during disease progression and are a major detriment for quality of life and functional daily living. Non-motor symptoms connected to sleep or mental health, such as idiopathic rapid eye movement-sleep behaviour disorder (RBD) and depression, often even pre-date the cardinal motor-symptoms. Other symptoms such as cognitive impairment and psychosis arise very frequently in later stages of PD. A longitudinal study reported that dementia developed in 80%, and hallucinations in 74% of patients 20 years after their initial PD diagnosis [35]. PD patients with dementia often share neuropathological patterns with AD patients [6, 36].

Similar to AD, PD is a disease likely promoted by complex interactions between genetic predisposition, environmental exposures, lifestyle factors, and pathological processes [2, 32]. The hallmark neurological pathology includes neuronal loss in the dopaminergic system of the substantia nigra and accumulation of alpha-synuclein throughout the brain. Typically, brain atrophy is contained to specific areas. Multiple pathways are known to be involved in sporadic PD, examples include alpha-synuclein proteostasis, oxidative stress, mitochondrial function, axonal transport, and neuroinflammation [2].

Clinical diagnosis of PD is based on motor-symptoms, requiring bradykinesia and at least one additional cardinal motor-symptom [32, 33, 37]. However, error rates of a diagnosis performed solely on motor-symptoms reach 24% even in specialized clinics [38]. Emerging imaging markers derived from MRI or dopamine-transporter-scan (DaTSCAN) can support the diagnosis and discrimination between sporadic PD and other parkinsionisms [39].

While there also exists no disease modifying treatment for PD [40], symptomatic treatment is available. Motor-symptoms of PD can be reduced remarkably through the administration of L-DOPA, the precursor amino acid of dopamine [41]. L-DOPA serves as an exogenous dopamine replacement and virtually every PD patient receives this treatment at some point during their disease course [41, 42]. The correct dosage and administration frequency of L-DOPA is crucial to avoid common drug-induced side-effects such as dyskinesias [43]. Additionally, L-DOPA treatment often conflicts with apparent non-motor symptoms and their treatment [34, 44, 45]. The processes that give rise to adverse reactions remain largely unknown [2]. Nonetheless, some motor-symptoms remain stable despite dopamine replacement therapy including freezing of gait, postural instability, and falls. The management of non-motor symptoms remains challenging due lack of efficacy for many drugs, adverse events, or due to drug interference with motor-symptom treatments [34].

While the success rate of clinical trials has been dire in the AD domain, the situation is only marginally better in PD [40, 46]. L-DOPA treatment was established in 1969 and since then most approvals occurred in its context, for example, to treat its side-effects, or in new forms of L-DOPA administration [47]. The majority of currently ongoing trials still focuses on symptomatic treatments [48]. All conducted trials for potentially disease modifying treatments that could alter the disease trajectory of patients have been futile either due to a lack of efficacy [49], conflicting results [50], or presence of adverse-events [51].

Triggered by novel insights into the genetic processes involved in PD, new avenues in gene therapy are currently explored [52]. Furthermore, motivated by the amyloid beta-targeting trials in AD, new clinical trials emerged aiming for a clearance of alpha synuclein aggregates [53]. However, leading researchers in the PD domain call for tempered expectations towards these trials [46, 54].

1.1.3 Challenges faced in clinical trials for AD and PD

The abundance of unsuccessful trials in the areas of AD and PD prompt inquiries into the reasons behind these failures. Indeed, the AD and PD domains face similar challenges in their search for an effective disease modifying treatment. Both fields are united in their believe that the abundant trial failures can be attributed to two primary factors that have been rarely addressed in previous clinical trials: the exhibited heterogeneity of the disease [20, 55, 56] and the optimal timing of interventions in patients' disease trajectories [20, 40, 57, 58].

Both AD and PD are highly heterogeneous in their symptomatic manifestation and multifactorial in their disease underlying pathology. Upstream pathogenic molecular mechanisms and their downstream effects likely differ across patients, even if the final pathological outcomes (e.g., amyloid beta or alpha-synuclein aggregation) are shared. This concept has been supported by numerous publications, in which distinct patient subgroups have been identified with each subgroup sharing similar pathological patterns that differ from those exhibited in other subgroups [59–61]. These subgroups, often referred to as disease subtypes, are believed to be underpinned by different biological mechanisms. Drugs typically have a hypothesized mechanism of action (MOA) that targets specific molecular players (usually proteins) and biological pathways. However, the fact that multiple, independent, and/or intertwined pathways likely contribute to the pathogenesis of neurodegenerative diseases can hinder the efficacy of any specific MOA [2, 20]. Assuming that multiple mechanisms of pathogenesis are involved in AD and PD, drugs targeting specific mechanisms may only be effective for patients whose condition manifests via the mechanism in question, without showing efficacy in other patients [55]. Such subtle signals are challenging to detect in heterogeneous trial cohorts, as the signal-to-noise ratio is low and the variance is high. Consequently, the disease heterogeneity leads to increased uncertainty and wider confidence intervals for treatment effect estimates. For example, trial simulations based on patient-level AD data have shown that there is low probability to even detect treatments with 80% efficacy in slowing disease progression within a trial running for five years that employ the number of AD diagnoses during trial run time as the clinical endpoint [56].

Another obstacle in trial design arises from the variability of commonly used trial endpoints across patients. Choosing primary and secondary outcomes for a trial is a challenging decision, as various options such as symptomatic scales, biomarkers, or phenoconversions from one disease stage to another can be used to determine treatment efficacy and trial success [62]. Additionally, digital readouts become an emerging option [62–64]. So far, mainly clinical assessments of disease symptoms have served as primary outcomes in NDD related trials. For PD, the motor-symptom scores of the Unified Parkinson's Disease Rating Scale (UPDRS) [65] were predominantly used [40]. In AD trials, assessments of cognitive function such as the Alzheimer's Disease Assessment Scale–Cognitive subscale (ADAS-Cog) [66], Clinical Dementia Rating (CDR), and Clinical Dementia Rating Sum of Boxes (CDRSB) [67, 68] have been widely applied as primary endpoints [20]. Clinical assessments, however, are often times subjective, a shortcoming that could be overcome by quantitative biomarkers and digital readouts [69]. As our biological understanding of disease-relevant molecular pathways grows, these markers become a promising alternative to the clinical perspective. In this context, PET scans measuring amyloid beta plaques in AD [70] and dopamine transporter imaging in PD [71] have gained significant interest as surrogate trial endpoints. The previously discussed controversial aducanumab is one example where a biomarker-guided AD trial aiming for amyloid beta clearance reached approval by the FDA [26]. However, even if the biomarker endpoint is reached, it remains possible that no clinical improvement of patients is observed [24] as biomarkers can be remote from quality of life impacting symptoms [62].

Both AD and PD are progressive disorders that set on several years prior to the appearance of first symptoms [2, 3]. There is a strong consensus in both indication areas that potentially disease-modifying treatments are likely futile in patients in advanced disease stages suffering from severe neuronal loss [20, 40, 57]. Neurodegeneration and brain atrophy are thought to be largely irreversible once they occurred. This hampers the observation of any cognitive improvements in clinical trials involving patients residing in moderate to advanced disease stages [57, 62]. Consequentially, treatments that may have a beneficial effect in delaying symptom onset or slowing disease progression before the onset of advanced neurodegeneration could be dismissed as ineffective. Therefore, it becomes increasingly important to identify individuals at risk during the early stages of the disease, before they receive a symptom-based clinical diagnosis of either AD or PD [2, 3].

In conclusion, for both AD and PD, the essential challenge in finding disease modifying drugs through clinical trials lies in bringing the *right drug* to the *right patient* at the *right time*. In recent decades, a new paradigm of medicine has emerged that deliberately aims at solving this challenge. This paradigm is called precision medicine.

1.2 The precision medicine paradigm

Precision medicine, also referred to as personalized medicine [72] or stratified medicine [73], represents a healthcare approach that strives to customize treatment decisions for each patient, using therapies that are expected to be beneficial for that specific individual and administering treatments at the optimal time for intervention [74]. Traditional treatment regimes typically rely on the 'average treatment effect' observed in randomized control trials, with patients diagnosed with the same condition receiving equal treatment for a standard duration (i.e., a "one-size-fits-all" approach; Figure 1A). It is commonly observed, however, that some patients respond to a treatment while others do not [75]. Precision medicine founds on the concept that this variability in treatment response can at least partially be explained by the variability in human biology and environmental exposures observed across individuals, even when they were clinically diagnosed with the same disease [69]. Such biological inter-patient variation can manifest across multiple domains, for example, covering 'omics such as genomic and transcriptomics [74, 76], but also extends to lifestyle choices, medical history (e.g., electronic health records (EHR) and medical-claims), and environmental exposures (i.e., the exposome) [69]. The core concept behind precision medicine lies in stratifying patients based on their individual characteristics and matching them to the intervention with the highest likelihood for an optimal outcome (Figure 1B)[77, 78]. While this opens the opportunity to provide patients with the adequate treatment, it simultaneously reduces the occurrence of adverse-events by avoiding treatments that are unlikely to be effective for a given patient [72, 74]. The full collection of measurements by which stratification is guided is often referred to as a biomarker signature or patient profile.

Besides the original concept of individualized patient treatment, the domain of precision medicine further includes the associated research that paves the way for comprehensive disease understanding, translational discoveries, and targeted clinical trials [72, 79]. One way of improving the understanding of multifactorial diseases resides in identifying biological pathways and mechanisms that contribute to the exhibited heterogeneity on a clinical and

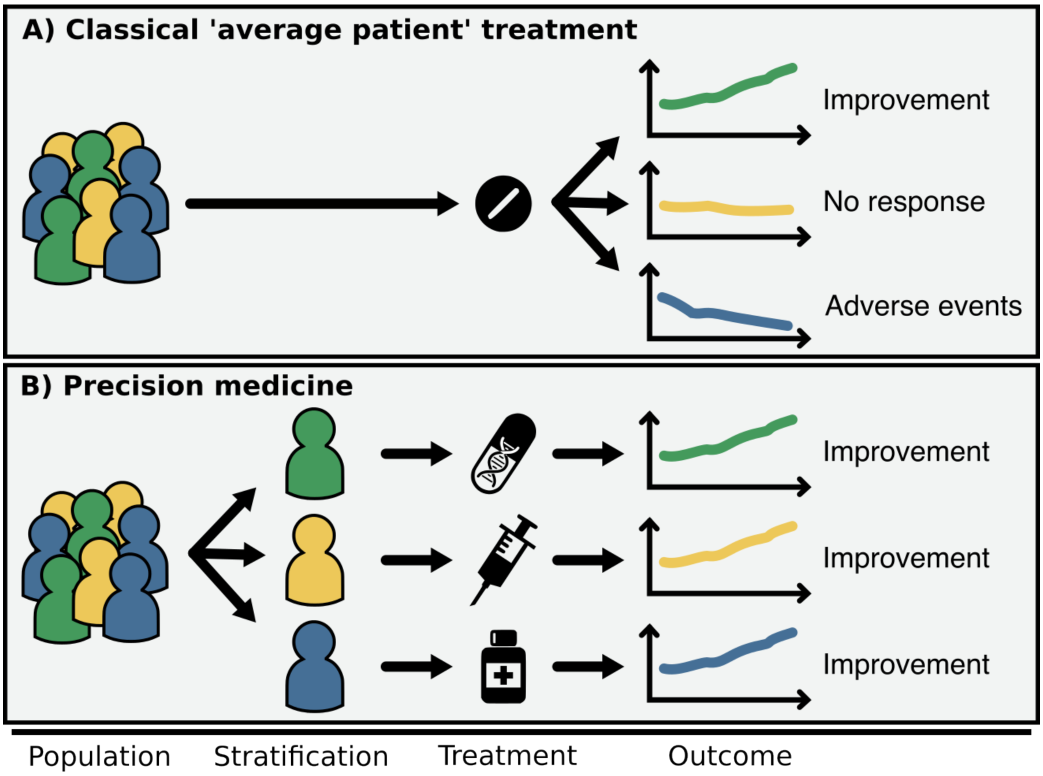


Figure 1: The conceptual idea behind precision medicine. A) The one-size-fits-all approach of treating every patient with the same drug, irrespective of their individual differences. **B)** The precision medicine approach in which patients are first stratified and then treated with targeted therapies that promises the optimal outcome considering the specific patient signatures.

pathological level [32, 61]. These distinct pathways could be connected to different etiologies and subtypes underlying the same clinical condition. Studying the key molecular players in these pathways may provide insights into new, patient subgroup-specific targets for intervention [69]. Furthermore, biomarkers that are associated with the identified molecular players could serve as potential strata allowing to group patients according to their disease subtype.

In early patient profiling approaches, mono-genetic mutations were primarily employed [80]. However, single molecular markers are often insufficient to capture the heterogeneity of complex multifactorial diseases [72], and their predictive power is often limited [81]. With the emergence of Genome Wide Association Studies (GWAS), polygenic risk scores combining several genomic variations into a single composite score gained popularity for disease prediction and stratification [82]. Nowadays, patient profiles are developed based on a more holistic view of the patient, encompassing multimodal signatures that include not only genomic markers, but also demographic variables, clinical assessments [81], and biomarkers measured through a wide range of techniques, such as imaging (e.g., MRI or PET) [70] or fluid-based assays [83]. These clinical and biological data can then be integrated into a single stratum, for example, risk of an event occurrence (e.g. death or disease progression) [84], or the probability of a patient to belong to a specific group (e.g. a disease subtype) [59].

In practice, precision medicine has found most of its success in the area of oncology, primarily due to the substantial impact of genetics on cancer development and progression [85, 86]. The decreasing costs of genotyping have facilitated the identification of specific mutations, enabling the development of targeted treatments tailored to particular cancer types [87]. Examples of genotype-guided cancer treatments are vemurafenib and dabrafenib, which target the v-Raf murine sarcoma viral oncogene homolog B (BRAF) and are prescribed to metastatic melanoma patients after testing positive for BRAF mutations via a companion diagnostic assay [77]. Precision medicine has also found application in clinical oncology practice. Tumor boards comprising several clinicians are built to make informed decisions on targeted treatment options for a particular patient based on that patient's comprehensive data [88]. Albeit less common, genotype-guided selection of treatments has also been implemented for other diseases. For example, HIV patients and individuals suffering from seizures are not prescribed certain drugs if they carry specific mutations due to a high risk of adverse event occurrence [74]. Nonetheless, there are numerous diseases for which treatment has not yet been impacted by precision medicine, often due to a lack of efficacious targeted interventions in the first place.

1.2.1 Advanced clinical trial design through precision medicine

To shift from a one-size-fits-all treatment regime to precision medicine, targeted therapies must be developed and their net benefit over standard care must be established. Accordingly, clinical trials have emerged that embrace patient stratification and explore novel treatments with deliberate MOAs specifically tailored towards certain (sub)groups of patients [89–91].

The potential of trials employing a stratification scheme can be illustrated by an example. Suppose a new link has been discovered between a set of rare mutations and a specific disease. Clinical trials can be designed specifically for drugs affecting the biological pathway that is perturbed by the mutated genes. In a traditional randomized control trial, patients would be enrolled based on a shared phenotype, neglecting their underlying biology. Assuming a multifactorial condition underlying the shared phenotype, the tested intervention would only show efficacy in a minority of patients, namely those carrying the rare mutations. Although there would be a real treatment effect for mutation carriers, it would remain hidden as the observed variability in the outcome would overshadow the true signal. A precision medicine-based trial would avoid these obstacles by incorporating genotyping into the screening procedure during patient enrollment [75]. Ignoring stratification errors, only patients affected by the mutations would be included in the study and, consequently, a significant treatment effect would be discovered. Such a stratified trial-based unmasking of an effective therapeutic agent, which was previously overshadowed by the noise introduced through non-responders, happened for instance with trastuzumab and gefitinib [75, 92]. Exactly this idea of stratification has also sparked the development of novel master protocols for clinical trials. One such protocol is the basket trial, where patients with different clinical phenotypes are recruited based on shared affected biological pathways. They are then treated with the same drug, assuming that its mechanism of action could be beneficial to all of these clinically heterogeneous patients [90].

In reality, the associations between strata and disease phenotypes are often correlative rather than entirely causal, and not every patient who matches the strata will necessarily show the expected outcome. Additionally, every stratification method (e.g. measuring biomarkers using an assay) is subject to some technical error. Therefore, stratification approaches are almost always probabilistic, rather than perfect. Trials that utilize such stratification approaches are referred to as "enrichment trials," as their aim is to enrich the trial cohort with patients who have a high probability of experiencing a certain outcome without intervention. [69]. This leads to more homogeneous trial cohorts and increases the probability of trial success [93, 94]. A meta-analysis across various cancer malignancies has found that trials employing stratification by matching patients with personalized treatments generally result in better outcomes and fewer occurrences of adverse events compared to non-stratified trials [87]. The benefits of enrichment trials in healthcare have also been acknowledged and promoted by the FDA [95].

Enrichment trials also provide substantial economical benefits over traditional trials [69, 75]. When assuming to encounter large variability, traditional trials increase their statistical power through larger sample sizes, thus, improving their chances of detecting a significant treatment effect. Precision medicine, on the other hand, addresses this challenge by reducing the variability through enrolling a more homogeneous, stratified cohort consisting of likely responders. As a result, the observed effect size is increased and, therefore, statistical analysis requires lower sample sizes to maintain the same statistical power. Accordingly, enrichment trials have to enroll fewer patients which can greatly reduce the trial costs connected to patient treatment and follow-up. Furthermore, adaptive trial designs have been developed that leverage biomarkers for longitudinal monitoring such that trial arms can be terminated timely if the intervention is foreseen to fail [90, 96]. The saved resources can then be redistributed towards the development of alternative interventions, which speeds up the drug developing pipeline as a whole [75].

1.2.2 Artificial intelligence and data-driven research in precision medicine

Precision medicine is at its core a data-driven discipline [74]. Achieving its goals requires leveraging rich data sources that hold information about the disease in question and the underlying biological condition of patients. In this context, emerging technologies from the area of artificial intelligence (AI), including machine learning and statistical learning, provide great opportunities to enable precision medicine [72, 79, 97]. By integrating complex multimodal and longitudinal data, AI algorithms can make personalized predictions and extract obfuscated disease signals.

Supervised learning methods are well suited for solving individualized prediction tasks, such as the classification of ill and healthy patients (i.e., disease diagnosis) or time-to-event analyses, such as the prognosis of symptom onset or treatment response [76, 98]. For time-to-event analysis, various parameterizations of the classical Cox proportional hazard framework have been proposed, including boosting algorithms [84] and artificial neural networks [99]. Classifications are often performed using XGBoost [100], Random Forest [101], Support Vector Machines [102], and deep learning [103].

Unsupervised algorithms are commonly used to detect structure in heterogeneous patient-level data, such as identifying disease subtypes that drive differences in patients' clinical and pathological presentation [104]. To this end, a variety of clustering algorithms have been applied, including non-negative matrix factorization [104], Gaussian mixture models (GMMs) [105], hierarchical clustering [60], and artificial neural networks [106].

In conclusion, AI provides a means for highly granular stratification of patients by leveraging complex signals extracted from multimodal and multiscale heterogeneous data. This characteristic makes AI approaches a promising solution to achieve precision medicine, especially for multifactorial diseases like AD and PD, where relevant signals are distributed across multiple data types [69, 75]. For a detailed review of the state-of-the-art research connected to the projects presented in this thesis, please refer to Section 2.

One limitation of AI approaches is their black-box behavior which limits the interpretability and comprehension of their decision process [72]. Especially in a clinical setting, patients want to feel secure and transparency about the aspects on which the AI model built its prediction can foster patient trust. However, also in basic research, it is valuable to understand the connection between single features of a patient and the outcome predicted by the model, for example, to identify associations between highly predictive features and a clinical phenotype.

Several methods belonging to the field of 'explainable AI' have been developed that establish a deeper understanding of the model and its decision process. A well exploited concept is that of feature importance, where each individual feature is assigned a score proportional to its weight in the prediction process [107, 108]. Interpretation of estimated feature importance scores enables 1) greater transparency, increasing the trust of patients and clinicians into the prediction of the models, and 2) the identification of relevant predictive associations between individual features and the outcome in question. While such associations are not necessarily causal in nature [109], they can serve as initial, exploratory evidence for possible biomarkers or mechanistic differences [110].

In conclusion, AI methods provide tools to address several key challenges within the context of precision medicine. They open new opportunities for a precision medicine-based drug development pipeline, supporting target identification and lead compound design [69, 79] and can be used for patient stratification and comprehensive patient monitoring [111]. The potential of AI to improve healthcare and patient well-being was also affirmed by the FDA in an official statement in 2019 [112].

However, to bring AI-based tools for forecasting and stratification into clinical practice, a rigorous evaluation of their efficiency and safety has to be conducted. Such an evaluation comprises four major steps [72]: Firstly, a newly developed AI model must be internally validated on instances originating from the full training dataset that were left out during model training. This helps to assure adequate predictive power. Internal validation is commonly performed already during model training, for example, within a k-fold cross-validation. As a second step, an external validation of the model has to be performed that assesses the generalizability of the model. Here, a new data resource is used that is fully independent from the dataset originally used in model training and internal validation. Good performance on this external dataset indicates a well-behaved model that did not overfit to its training data. Thirdly, the AI tool should be tested in a prospective clinical trial, where the clinical benefit is compared against standard clinical care. Finally, a regulatory agency, such as the FDA or EMA, has to evaluate the safety and benefit-to-harm trade-off and grant final approval. As of 5th October 2022, 178 AI and machine learning tools were approved by the FDA in total, yet, only a subset of those are catered towards patient stratification.

The initial development and steps one through three of the process explained above crucially depend on patient-level data. While AI is capable of leveraging heterogeneous, multimodal data as measured for example in the context of NDDs, these data still bear many challenges. Complex signals need to be learned from a rising number of different data modalities that are all subject to distinct error sources and the signal-to-noise ratio is low given the heterogeneous patient population. Especially in these highly volatile settings, a rigorous model validation is vital [69].

1.3 Precision medicine for Alzheimer's disease and Parkinson's disease

As discussed in Section 1.1, AD and PD have devastating consequences for affected individuals, their friends and relatives, and society as a whole [2, 3]. The majority of previously conducted translational research and clinical trials have had little impact on the disease trajectories of patients. This is likely due to several challenges discussed in detail in Section 1.1.3, mainly, the heterogeneity in clinical and pathological presentation of patients and the optimal timing of intervention in the earliest stages of the diseases. Both challenges can be translated into tangible research questions that lie at the heart of precision medicine: 1) Is it possible to detect patients in their earliest, pre-symptomatic disease stages and thereby enable patient stratification? 2) Can subtypes be identified within the heterogeneous patient populations that facilitate a deeper understanding of the disease?

While the multifactorial nature of AD and PD makes them prime candidates for stratification approaches, their complex pathologies and the lack of a clear understanding of the interplay between molecular players, neurodegeneration, and clinical markers presents a significant obstacle [7, 40]. Unlike in oncology, genetic information alone provides limited leverage, and exploring large, multimodal feature spaces is necessary [4, 81]. Furthermore, the progressive nature of AD and PD mandates the inclusion of time-dependent signals into precision medicine approaches, as cross-sectional snapshots of patients will miss relevant aspects of the diseases. Integrating such multiscale, longitudinal data through AI approaches could help identifying predictive signatures for patient stratification and to identify disease subtypes [113]. The availability of reliable strata and disease subtypes could enable a shift towards enrichment trials and thus contribute to the discovery of promising interventions. In fact, enrichment trials are widely seen as a mandatory step in order to advance in the search for a disease modifying treatment for both, AD and PD [40, 55, 62, 70].

Over the recent years, many endeavors have contributed to advancing precision medicine in AD and PD. Several genetic and environmental risk factors were identified, but none of them are deterministic [2, 3]. Univariable stratifications based on biomarkers such as alpha synuclein [53] and amyloid beta [114] have been explored. However, cognitive resilience and pathological resistance of patients limits their usability [115], as approximately 30% of individuals that are cognitively stable at death exhibit AD pathology during autopsy [16]. Addressing the shortcomings of univariable stratification, polygenic risk scores [116] and composite biomarker scores have been developed [117]. Furthermore, clinical assessments, medical history and comorbidities have been integrated into stratification approaches [118, 119].

Multiple disease subtypes have been proposed in AD and PD, primarily to facilitate deeper understanding of the respective disease. They were derived from patients clinical symptoms [120], thresholding of patients' age (i.e., late versus young onset) [121], pathological patterns [59, 122], genetics [104], and combinations thereof [60]. The translational impact of these undertakings remained limited to this day [62, 123, 124].

In PD, theoretical concepts for enrichment trials have been published but, as of now, have not found application [62]. In AD, the recent biomarker-guided trials can be seen as first steps towards enrichment trials, as patients are only recruited based on a positive amyloid beta PET scan [26, 28]. As previously discussed, these trials have led to two new drug approvals (i.e., aducanumab and lecanemab), albeit that the actual clinical impact and benefit of the drugs is still put at question [29, 30].

One reason why the field has yet to witness the translational impact of precision medicine lies in the validation and replication of subtypes and stratification schemes. The vast majority of studies proposing new stratification schemes are developed on single datasets and rarely evaluated in external data [125–128]. If conducted at all, replications of biomarkers often failed [129], and performances of predictive models were highly volatile in validation data [130, 131]. The lack of validation can be attributed to the characteristics of the data on which these approaches were developed, as well as common assumptions made about the data, which are seldom met.

1.3.1 Data limitations

To enable the successful development and validation of precision medicine approaches, data sources must fulfill both content and statistical requirements. For multifactorial diseases like AD and PD, it is crucial to have access to multimodal data that capture all relevant pathological and symptomatic changes occurring in patients. These data modalities may include clinical information on diagnosis and symptoms, genotyping, imaging-based biomarkers (such as PET with multiple tracers, MRI, and, for PD, DaTscan), CSF biomarkers, lifestyle factors, medical history, family history, and post-mortem autopsy [2, 3]. To integrate the multimodal and multiscale features into stratification approaches, the corresponding data should be measured for each individual and, preferably, repeatedly over time to capture the progressive nature of the diseases. As NDD data typically have a low signal-to-noise ratio, a sufficiently large sample size is essential to achieve adequate statistical power. Furthermore, one of the fundamental assumptions of AI and machine learning [132, 133] is that the underlying data are independent and identically distributed (i.i.d.). This means that the application of novel treatment regimes, stratification models, and subtyping approaches is only valid for individuals that are i.i.d. with respect to the data on which these approaches were initially developed.

In recent years, real-world data have become increasingly accessible and have supported large sample sizes, an example being the UK Biobank [134]. Utilizing real-world data has led to the approval of new cancer treatments [89], and has also been useful for clinical trial simulations [135] and epidemiology [136]. However, for disease stratification and explaining the underlying heterogeneity of a disease, a deeply phenotyped population is necessary, and real-world data often lack crucial clinical outcomes and data modalities. Therefore, historically, most research endeavors aiming to advance precision medicine in the AD domain have relied on data from observational cohort studies [137, 138].

Observational cohort studies are specifically designed to answer certain research questions [139, 140]. They measure the data modalities that are relevant for the research questions and do so repeatedly at predefined intervals, without any specific intervention scheme [141]. In addition, they employ specific rules for participant recruitment, such as inclusion and exclusion criteria, which ensure the enrollment of the appropriate study population to achieve the desired primary research goals [53]. For example, they may include only individuals at risk of developing a disease [139], carriers of specific mutations [142], or patients diagnosed with a disease within a defined time window [143].

However, observational cohort data bear challenges when employing data-driven approaches such as AI and machine learning. While these resources provide a rich data foundation for modeling approaches, they are often limited in sample size due to the high costs of multimodal data collection. More importantly, the inclusion and exclusion criteria as well as geographic location of a study shape the population from which its participants are sampled. Consequently, participants from different observational cohort studies are not necessarily samples from the same underlying statistical distribution, given that they were recruited according to different criteria and from potentially distinct regions [144–146]. Accordingly, the i.i.d. assumption of AI would be violated which has significant consequences.

AI models are designed to learn a function that maps the input data to a desired output (e.g., predicting risk based on a biomarker signature)or to learn the statistical distribution of its training data (i.e., density estimation). A well-behaved, non-overfitted model learns the full support over the multivariate distribution of the input features encountered in the training data (Figure 2A), and performs appropriately on new, independent samples from the same distribution (i.e., i.i.d. samples, Figure 2B). If such a model is however presented with data instances that fall outside of the statistical distribution of the training data, the model can not be assumed to work adequately (Figure 2C) [133]. The new sample is considered 'out of distribution' or outside of the models learned domain. Learning a model in these settings is commonly referred to as 'transfer learning' or 'domain adaptation' [132]. An overfitted model, on the other hand, will not generalize to support the full distribution of its training data (Figure 2D), and thus also fails on data that is i.i.d. to its training data (Figure 2E). When using observational data from cohort studies targeting the same phenotype, the most likely scenario is one in which the different cohorts' domains are not completely disjoint but overlap to some extent (Figure 2F).

Distribution shifts between observational cohorts can be caused by several factors beyond sampling different demographics. For instance, disparate proportions of disease subtypes among cohorts could affect the statistical distributions of symptoms and biomarkers of disease pathology. In addition, recruiting participants from different geographic locations can lead to the inclusion of distinct ethnoracial groups, which are known to impact genetic risk for both PD [147] and AD [148]. Similarly, variations in recruitment based on risk factors such as APOE $\epsilon 4$ in AD [149] or GBA mutations in PD [150] could also result in statistical discrepancies. Furthermore, statistical deviations can be introduced through differences in data collection procedures, including imaging protocols [151], pre-analytical biases in assay measurements [152], and neurocognitive testing [153]. However, the degree of domain shifts between clinical AD cohort datasets and their impact on AI model application and validation across datasets remained largely unexplored prior to this thesis.

Another impediment for data-driven research in the NDD domain is the fragmented state of the data landscape. The availability and accessibility of single hallmark longitudinal datasets like the Alzheimer's Disease NeuroImaging Initiative (ADNI) [140] and the Parkinson's Progression Markers Initiative (PPMI) [143] have greatly facilitated research in the field, with thousands of publications making use of these datasets [138, 154–156]. However, it can be difficult to identify and access alternative datasets that are suitable for replication and validation of results established on these hallmark datasets. While there are alternative datasets available in principle [139, 142, 157], they often lack visibility and are seldom explorable before data access has been granted, making it challenging to evaluate their suitability as an independent validation dataset. To qualify as an validation dataset, all data modalities and features that found use in a predictive model must be available and longitudinal follow-up length and frequency should match the expectations of the proposed model. Additionally, as cohorts are often subject of participant drop-out over time, the mere theoretical availability of a feature does not guarantee that it was measured for all patients and at every time point.

1.3.2 Generative models and synthetic data

One promising approach to address the limitations of observational cohort data is the generation of synthetic datasets. These synthetic data are generated such that they obey the characteristics of their real-world counterpart by maintaining variable correlations and longitudinal trends [158]. At the same time, they do not represent any entities that reside in the real world. Thereby, synthetic data can function as an anonymization procedure that enables data sharing through the distribution of synthetic data instead of real patient data which are highly sensitive [159–161]. Furthermore, synthetic versions of patient-level cohort datasets are also free from limitations that are commonly observed in real cohort data. As all data is generated, there are no missing values. Additionally, depending on the data generating model, arbitrary patient follow-up intervals can be specified to ensure compatibility between independent data sources. One application example that is of rising interest in AD research [162, 163] is to enforce adequate representation of underrepresented groups by generating more instances from specific patient populations [158]. Finally, synthetic data generation can provide the foundation for simulating counterfactual scenarios by introducing interventions and conditionally generating the data [160]. All of these capabilities promise critical levers to advance the transition towards precision medicine in AD and PD research.

Once again, the tools to generate such synthetic data are provided by the field of AI. Generative models constitute a class of algorithms that can learn the multivariate statistical distribution underlying a training dataset and subsequently allow to sample from it to generate new data instances [164]. A comprehensive review of the state of the art for generative modelling is provided in Section 2.1.4.

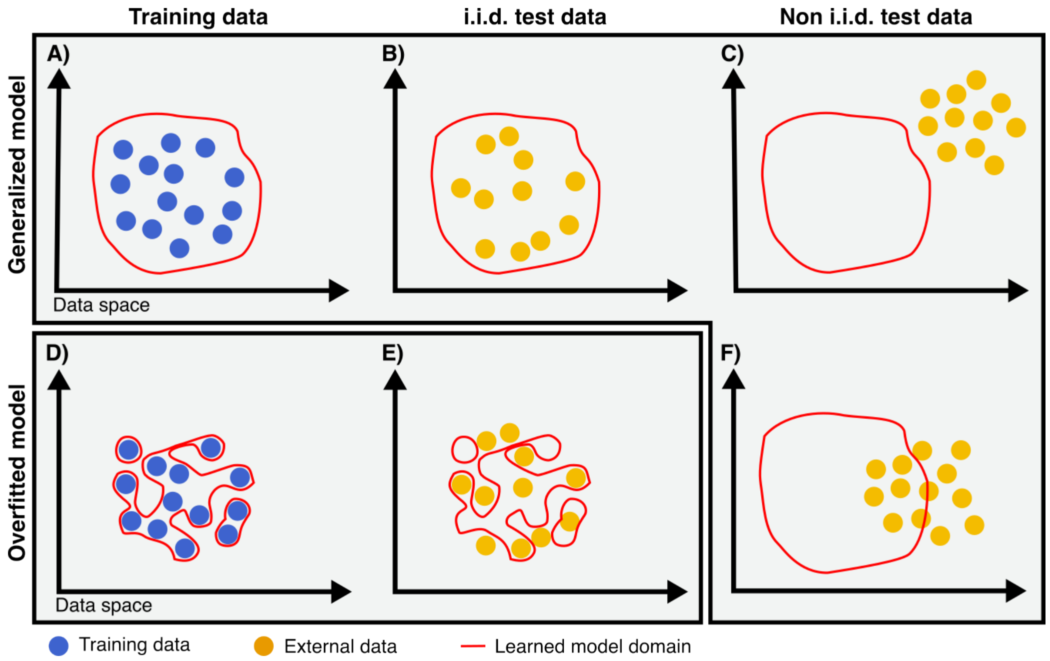


Figure 2: The i.i.d. assumption of data-driven modeling and precision medicine. The red line depicts the domain that a model has learned from the training data. Assuming good model performance, samples falling inside the domain are largely correctly predicted, while the model is expected to perform poorly on samples outside its domain (depending on the topology of the objective function over the data space). A) A well-behaved model that generalized to support the full distribution of its training data. B) A test dataset which is i.i.d. with the training data and thus the model will perform well on it. C) A non-i.i.d. test set, for which the model will fail to perform adequately, although it accurately supported the distribution of the training data. D) An overfitted model that failed to generalize. E) The overfitted model will show poor performance even when applied to a sample i.i.d. to its training data. F) A scenario more likely to be encountered across different cohort datasets recruiting the same phenotype, where samples are not i.i.d. but distributions overlap partially.

1.4 Overview about the content of this thesis

Chapter 2 introduces the major contributions of the research presented in this thesis. We discuss the current and past state of the art in context of our research and highlight the impact of our scientific efforts.

Chapter 3 presents three publications that aimed at improving the data landscape in AD by describing it, enabling the exploration of data content ahead of access approval, and contributing new open data to the field. The work described in these publications laid the foundation for publications presented in later chapters.

Chapter 4 describes three articles in which the heterogeneity exhibited by AD and PD patients is explored. New approaches for patient stratification based on disease risk, symptom progression subtypes, and disease progression patterns are proposed. We also investigate whether observational AD cohort datasets fulfill the i.i.d. assumption and how its violation can be assessed and understood.

Chapter 5 presents a new deep learning-based AI method for synthetic data generation and its application to generate synthetic patient-level data for AD and PD.

The thesis concludes with a brief summary and discussion of potential avenues for future research in precision medicine for AD and PD.

2 Research contributions of this thesis

In this chapter, we will discuss the contributions of this thesis to the field of precision medicine in the context of AD and PD. We position our contributions relative to the state of the art in the field, as well as discuss their impact and subsequent developments that occurred after our work was published. Our contributions include methodological developments as well as novel scientific insights.

The research presented throughout this thesis has made significant contributions towards advancing precision medicine and facilitating robust data-driven science in the NDD domain:

- Thorough assessment and comparison of AD cohort datasets to describe the AD data landscape [165, 166]
- Investigation of the i.i.d. assumption in the context of data-driven AD research and its implications for the generalizability of data-derived results and predictive models [165–167]
- Development of approaches to evaluate the impact of systematic cohort biases on data-driven results [166, 167]
- Modeling of AD progression across 6 independent cohorts, estimating expected progression times, covariate effects, and disease risk over time for different cohort populations [167]
- AD risk-based stratification with successful external validation for early detection of prodromal AD dementia [166]
- Identification of distinct PD symptom progression subtypes and their associated, potentially underlying, biological processes, as well as highlighting differences in treatment response [168]
- Making the AD data landscape explorable and foster a research culture that looks beyond single data resources for robust results [169]
- Contribution to the AD data landscape by providing a comprehensive, multimodal dataset that we made accessible to the research community [170]

- Development of a novel generative AI approach that can synthesize time-continuous, multimodal, longitudinal patient-level data that retain the signals of their real world training data [171]

In Figure 3, we position the research contributions presented in this thesis along a value chain leading towards enrichment trials and novel treatments.

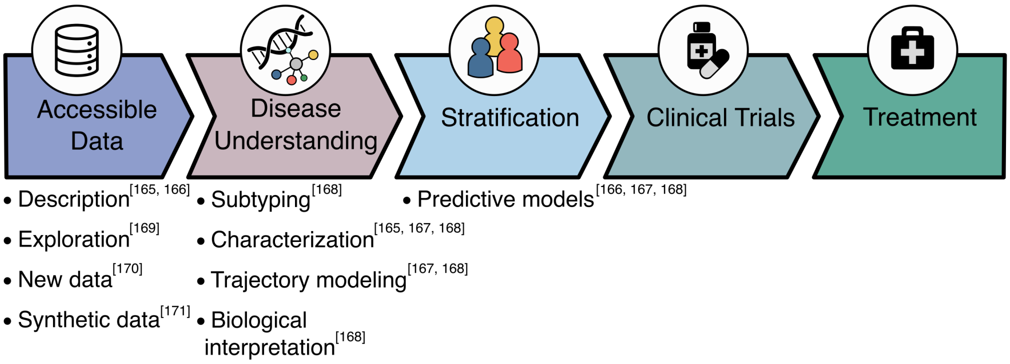


Figure 3: The value chain of data-driven precision medicine leading towards improved clinical trials with better chances of identifying efficacious treatments. The bullet points below each segment represent the research carried out in this thesis.

2.1 Facilitating data-driven research in NDDs

2.1.1 Investigating the i.i.d. assumption across AD cohort data

In previous studies, researchers have raised doubts about how representative observational AD cohort studies are of the general patient population [146, 172, 173]. Whitwell *et al.* compared demographic variables and MRI measurements from ADNI patients to a population-based cohort study, the Mayo Clinic Study of Aging [173]. They observed that ADNI patients were on average significantly higher educated, younger, and performed better in assessments of cognitive performance than the population-based sample. Further, they found that hippocampal atrophy progressed faster in ADNI for cognitively normal and MCI patients than in the Mayo clinic study of Aging. Similarly, Ferreira *et al.* conducted a study in which they compared observational cohort data to population-based cohorts and, again, found significant differences across the same demographic features that propagated into MRI signals [172]. Our own evaluation of the AD data landscape presented in Chapter 3.1, and an in-depth comparison of two major AD cohort datasets (presented in Chapter 4.2) lead to similar conclusions about deviations among AD datasets [165]. Furthermore, many studies have reported on systematic biases in fluid-based and imaging biomarkers across cohort studies [174]. Conclusively, these research endeavors provide ample evidence that observational AD cohorts are likely not i.i.d. and thus a fundamental assumption of precision medicine and AI would be violated. How significant domain shifts between clinical AD cohort datasets actually are and, more importantly, their impact on data-driven results remained largely unexplored before this thesis.

In our studies presented in Chapter 4.1 and 4.2, we went beyond investigation of established differences across demographic features and assessed the impact of systematic deviations between cohorts on data-driven disease progression modeling and predictive AI model performance [166, 167]. In both studies, we exposed the presence of significant cohort-specific biases that manifested in differences across univariable feature comparisons, extracted progression patterns, and AI model performance. To the best of our knowledge, these endeavors mark the first systematic assessments of the implications of a violation of the i.i.d. assumption for data-driven research in the AD domain. Our findings underline that the generalizability of data-driven results and models must be investigated across multiple datasets to ensure robust scientific insights.

Aiming to increase their coverage of the AD population and sample size, researchers have tried to pool data from different cohort datasets [175, 176]. However, naive pooling of cohorts poses no remedy for cohort-specific biases, as they persist in the data and are proportional to the number of patients included from each respective cohort. Consequently, the systematic biases of the largest cohorts will have the strongest influence on the achieved results.

Our findings put previous studies at question that have proclaimed that their findings would be generally applicable across AD patients. In 2019, for example, Vermunt *et al.* published a highly cited, impactful analysis that investigated the duration of the preclinical, prodromal and dementia stages of AD [175]. This analysis was conducted on a dataset pooled from different cohorts and employed multistate models to estimate the sojourn time for the respective disease states. Using the same data-driven method (i.e., multistate models), our study presented in Chapter 4.1, showcases that such estimates are highly volatile across cohorts and can not be generalized [167]. We also presented evidence that data pooling will not eradicate cohort-specific biases. The same considerations hold true for predictive models.

Numerous models have been trained and validated solely on ADNI data [125]. Whether they generalize beyond ADNI and if they are applicable to other cohorts was seldom evaluated. While a thorough validation of predictive models and data-derived results is imperative for robust science [72], it presents a challenging undertaking in the AD domain due to two main reasons: The previously outlined implications of non-i.i.d. data and identifying a dataset for validation that shares all the relevant predictors a model relies upon (see Chapter 3.1) [165, 177].

Most observational AD cohorts were most likely not sampled from completely disjoint distributions but overlap to a varying degree depending on their recruitment criteria and patient characteristics (see Figure 2F). AddNeuroMed and the Japanese Alzheimer's Disease Neuroimaging Initiative, for example, both implemented study protocols that were closely aligned to those of ADNI [178, 179]. Additionally, many cohorts employed the same NINCDS-ADRDA criteria [12] when classifying patients into the three clinical stages of cognitively unimpaired, MCI, and AD [165]. Consequently, the extent to which cohorts' distributions overlap or deviate will proportionally determine the transferability of data-driven results across them.

Proposed methods to evaluate the i.i.d. assumption across cohorts

To quantify and understand the impact of distribution shifts between cohort datasets, we proposed two workflows: In Chapter 4.2, we used propensity score matching to select a subset of patients from a validation cohort that was closer to the distribution of the training data with respect to variables that are commonly used as inclusion and exclusion criteria in cohort studies [166]. The difference in model performance between the matched group and the unmatched, complete validation cohort provides an estimate for the strength with which systematic deviations among cohorts affect predictive models. We emphasize that we do not recommend cherry-picking participants based on a broad selection of features used as predictors in the model but see our approach as a method for post-hoc estimation of cohort effects.

Along the same line, we proposed a clustering approach based on disease patterns extracted from cohort datasets (presented in Chapter 4.1) [167]. The clustering provides a quantitative overview about the likeness of cohorts with respect to their exhibited disease patterns and, thereby, goes beyond conducting univariable statistical tests on demographic variables that most publications rely upon [172, 173]. The approach can also support a post-hoc analysis of AI model performance across cohorts, as models of disease progression would be expected to decrease in performance the more distant the progression trends of cohorts are from each other.

Together these two approaches facilitate an understanding of the domain that was learned by a model and its limitations. They can provide explanations on whether generalizability of results and transfer of models across cohorts were impeded by overfitting or systematic cohort biases.

2.1.2 Evaluation of AD cohort datasets and enabling their exploration

To work across datasets, a good understanding of the available resources is necessary to identify and select cohort datasets appropriate for the envisioned research. Large consortia have been formed to organize the AD data landscape, describe it, make it searchable, and thereby assist researchers in identifying data. For example, the European Medical Information Framework (EMIF) built the EMIF-Catalog, a web-based application that stores metadata on cohort datasets and is explorable after an access application has been approved [180]. The metadata were collected by providing data owners with a questionnaire in which they reported the variables accessible within their datasets. Similarly, the Real world Outcomes across the Alzheimer's Disease spectrum for better care: Multi-modal data Access Platform (ROADMAP) project generated their ROADMAP Data Cube, a web-based visualization in which the availability of clinical outcomes and data modalities in several European AD cohort datasets is depicted [181]. The displayed content is again founded on metadata that partially originated from the EMIF-Catalog. Lawrence et al., on the other hand, opted for a literature-based review to compare metadata of distinct AD cohort studies [146].

In contrast to these previous endeavors, we assessed the AD data landscape exclusively based on the data that were factually shared after data access was granted (see Chapter 3.1 [165]. In this process, we found several mismatches between the reports on available dataset content provided through the EMIF-Catalog and ROADMAP Data Cube and the factual content of the datasets after we accessed them as third-party researchers. These observations underlined that metadata-based approaches for presenting dataset content to researchers are error prone and can cause futile expectations.

In our data-driven assessment of the AD cohort data landscape presented in Chapter 3.1, we described major AD cohort datasets on feature-level, their interoperability among each other, and existing biases in cohort populations [165]. As previously mentioned, and in concordance with previous findings by other researchers [146, 172, 173], this work contributed further evidence that AD cohort datasets often violate the i.i.d. assumption. Additionally, it highlighted shortcomings in current sampling of AD populations in the form of patient recruitment, as we observed an underrepresentation of non-white individuals in AD cohorts, and imbalances in sex distributions. Over recent years, evidence amounted that race might play an important role in AD [148, 162, 182], and extensive research has been conducted on sex differences in AD [183–186]. In the light of these findings, an accurate representation of both sexes and all races is crucial to understand the heterogeneity in AD. Our publication exposed a lack of diversity in AD cohorts and has found wide recognition in the field, for example, being cited by the ADNI consortium on several occasions [163, 187–191], among others to motivate their newest phase of ADNI which focuses on the enrollment of a more diverse cohort.

To allow researchers to make similar assessments of the available data and thus make informed decisions on which datasets to access, we further developed ADataViewer (presented in Chapter 3.2) [169]. ADataViewer is an interactive web application that enables researchers to explore AD cohort datasets on a feature-level. In this, it goes beyond previous approaches which presented metadata on included cohorts, but allows users to plot empirical distributions of variables, assess the available patient follow-up per feature and study, and provides overviews on dataset interoperability. To assists researchers in working across datasets, we published a mapping table that harmonizes the name spaces of 20 distinct AD cohort studies covering more than 1000 features. Additionally, ADataViewer provides tools that can suggest cohorts suited for replication of results based on a user-specified list of required features.

The ADataViewer itself does not provide access to the underlying datasets. To this task, two prominent data initiatives are committed in the AD domain: the Dementia Platform UK (DPUK) and Alzheimer's Disease Data Initiative (ADDI). DPUK presents a centralized access point for cohort datasets to which researchers can apply for access [192]. ADDI is a non-profit organization that was founded in 2018 to connect researchers with AD data. Both of these resources, however, lack functionalities that allow researchers to thoroughly explore their content, allowing to evaluate whether the available data suits their research designs without applying for data access first. Here, ADataViewer provides an example on how data exploration can be facilitated without requiring registration and data access applications which consequently results in significant time savings.

2.1.3 Contribution of AD cohort data

Beyond enabling exploration of the AD data landscape, we also contributed to it. AddNeuroMed was a large multi-center cohort study that collected multimodal, patient-level, and longitudinal data [157]. In its design, it closely followed the ADNI study [140] and aligned its data collection protocols to those of ADNI [178]. However, despite being a rich data source and at least to some extent comparable to the most used AD dataset (i.e., ADNI)[125], it did not see as much use. While ADNI provides an easy-to-use ADNIMERGE table that comprises commonly analyzed features in a comprehensive form, AddNeuroMed consisted of a large collection of disjoint data tables. It was missing interoperability among those tables, was incomplete, subject to multiple errors, and lacked in documentation. We provided the community with ANMerge, a new version of AddNeuroMed that corrected its previous shortcomings and added data of additional participants and more follow-up visits (presented in Chapter 3.3 [170]. We made ANMerge accessible to third-party researchers and provided easily analyzable tables that follow the example of ADNIMERGE.

Since our publication of ANMerge, the data has seen increased usage. Our department leveraged the data in several studies [167, 169, 193], for example, as a validation dataset for an AI approach built on ADNI data [166]. By third-party researchers, it has been utilized in studies concerning, for example, neuropsychological testing [194], AD subtyping [176], validation of blood-based proteomic biomarker networks [195], machine learning-based diagnosis [196], MRI-based prediction of brain age [197], and measurement of brain atrophy via automatic tools [198].

2.1.4 Generation of synthetic patient-level time-series data

Synthetic patient-level data promises to support overcoming data limitations often encountered in biomedical research. Assuming a well-behaved generative model, generated synthetic data retain the signals contained in their real world counterpart. Furthermore, they are no subject to missing values. Depending on the algorithm used, they also allow for the generation of longitudinal data in continuous or discrete time. Given the sensitivity of healthcare data, synthetic data could simplify data sharing as they are not connected to real individuals [199–201].

About a decade ago, multiple new generative AI models have emerged from the field of deep learning such as the generative adversarial networks [202], variational autoencoders [203], and normalizing flows [204]. Apart from variational autoencoders, the initial success of generative neural networks was largely found on image data [205, 206], and transferred to medical imaging [207–209]. For time-series data, new methods have been developed extending on the aforementioned architectures such as dynamic normalizing flows [210], time-series generative adversarial networks [211], and neural ordinary differential equations (neural ODEs) [212]. However, while synthetic data generated via these methods could address some of the discussed shortcomings connected real clinical data, the methods themselves are not built to handle the specifics of clinical data accurately during model training. Clinical data consist of multiple modalities containing features that are discrete and continuous, as well as time-dependent and static. Moreover, they are often subject to missing data in the first place.

New methods have been developed specifically for synthesizing clinical patient-level data [213–215]. One of these approaches is (VAMBN) which was previously published by our group [160]. VAMBN utilizes an extension of variational autoencoders [216] to encode distinct data modalities independently from each other and then fuses these autoencoder modules using a Bayesian network. It is also capable of generating longitudinal data, however, only in discrete time where generated time intervals mimic those of its training data.

During our work presented in Chapter 5.1, we developed multimodal neural ordinary differential equations (MultiNODEs) [171], an extension of neural ODEs [212]. Our major methodological contribution with MultiNODEs consists in its capability to handle static input features alongside time-dependent ones. To achieve this, we fused the core architecture of neural ODEs with an additional variational autoencoder designed for heterogeneous, incomplete data [216] and concatenate the latent representation of both modalities to form an initial condition for an ODE module.

Through this new model design, we enable the application on multimodal clinical data and, subsequently, its generation. Here, another major advantage of MultiNODEs emerges as it is capable of generating data in continuous time. This allows for synthetic data with arbitrary follow-up intervals and both, interpolation and extrapolation of the originally sampled time points. Synthetic data generated using MultiNODEs proved to retain the longitudinal patterns, marginal distributions, and correlation structure of its real world counterpart. We tested its capabilities on clinical patient-level data from AD cohorts as well as PD cohorts. Especially with respect to learning the correlation structure of the input data, MultiNODEs surpassed the VAMBN approach.

Since the publication of our work, generative models have advanced considerably, however, mainly with respect to textual data modeled via transformers [217], as well as image generation using diffusion models [218]. Both approaches quickly found their way into the biomedical field as well [219–221]. However, their use-cases differ from the scenarios in which MultiNODEs can be applied. Recently, a new library was released that combines several approaches for generation of tabular data that are commonly found in healthcare [222].

2.2 Advancing precision medicine in AD and PD

2.2.1 Diagnosis of prodromal clinical AD patients

As discussed in detail in Section 1.3, one major obstacle in recent AD trials lies in the early timing of an intervention within patients' disease trajectories [20, 57, 223]. It is widely believed that an intervention prior to the onset of severe cognitive symptoms is vital to discover significant treatment effects. Accordingly, a reliable method diagnosing prodromal AD is needed [3].

Currently, there are two dominant strategies explored to facilitate an early diagnosis of AD. The first resides in redefining AD from a clinical condition characterized by apparent symptoms towards a biological entity defined purely based on pathological developments. This redefinition has been proposed in form of the amyloid deposition, pathologic tau, and neurodegeneration (ATN) classification system [14]. The ATN system categorizes patients along its three dimensions, A, T, and N, which are operationalized through biomarker measurements derived mainly from CSF or PET imaging. Along each dimension a threshold is defined according to which individuals are considered either normal or abnormal. The thresholds are usually determined based on patient-level cohort data [224]. The ATN system became widely adapted in AD research [225–227] and several studies report that it is capable of predicting the patients' conversion to AD [228–230]. One limitation of the ATN scheme is that data-derived thresholds are seldom interchangeable between distinct cohorts due to non-standardized assays [231] and data distribution shifts [224]. Furthermore, potentially useful information from other modalities such as genomics are disregarded.

An alternative to pathology-derived thresholds is the prediction of clinical AD conversion for prodromal AD subjects using data-driven models leveraging comprehensive biomarker signatures. From a modeling perspective, predicting AD conversion represents a prognosis task which is commonly addressed via time-to-event models, as the outcome is time-dependent and subject to censoring [84, 130, 131, 166]. In the AD literature, however, also traditional classifiers for conversion at concrete time points have been used [232, 233]. Apart from AD conversion, also pathological changes have been used as outcomes for early AD detection [234] which is more aligned with the emerging biological perspective on AD [14]. Only a minority of previously conducted studies however performed an external validation of their predictive models [125].

Our research on prodromal AD prediction involved the revision and external validation of an AD risk model that was previously published by our group [84]. As presented in Chapter 4.2, we extended this previous work by training a new model on a set of predictors that were common between the original training data of the model, ADNI, and a new dataset that we prepared and published, ANMerge (described in Chapter 3.3). The new model was then trained on ADNI and externally validated on ANMerge, which marked it as one of a few approaches for AD risk prediction that had been externally validated. Other examples were risk models have been externally validated include [131] and [130]. With respect to model performance however, our risk model outperformed other externally validated models with a Harrell's C index of 0.81 compared to 0.74 presented in [131] and 0.72 achieved in [130], when validating their models on ADNI. In conclusion, at the time of its publication, and to the best of our knowledge, our risk model represented the highest performing stratification approach targeting AD conversion risk as an outcome.

Since the publication of our risk models, additional approaches for predicting AD onset have been proposed. These endeavors approached AD progression not as a time-dependent process but a binary classification task with 'conversion to AD at any time point during follow-up' as the outcome [235–237]. These approaches accordingly also ignored censoring. A review published 2 years after our risk model assessed the literature for models predicting AD conversion and noted that most models were still not validated externally [125] Conclusively, this indicates that our study still represents the state of the art for AD risk prediction.

Related to this, we later also published another approach for patient stratification in a collaborative effort [238]. However, this approach did not target AD conversion risk but directly performed a forecast of clinical outcomes ahead of time which could then be used for trial enrichment.

2.2.2 Modeling of AD patient disease trajectories

To understand the progressive course of AD, it is vital to comprehend how individuals traverse along the different stages of AD. One way to facilitate this understanding lies in modeling the disease trajectories of patients. To this end, several methodologies exist.

Specific to the neurology field, so called event-based models (EBMs) have emerged that extract a discrete sequence of biomarker events from cross-sectional patient-level data and describe the order in which biomarkers turn from a normal to an abnormal state [239]. These models found wide application for AD and other neurological diseases [240–245]. Lately, they have been expanded to also include time-series data and infer disease subtypes [59, 246]. We have also tested them across multiple datasets and designed an algorithm to combine partially overlapping event sequences [193]. One major drawback of these models is that they are time agnostic and thus do not provide any explanation on the temporal course of the disease beyond the discrete order of biomarker changes. Another discrete perspective on AD progression resides in the AD continuum proposed in the ATN-framework [14, 247].

An alternative, more widely established approach can be found in constructing state space models. State space models represent a class of data-driven algorithms that can model the time-dependent traversal of patients along a defined system of states [248]. One specific class of state space models that have found success in biomedicine are multistate models [249]. In the past, they have been leveraged to model and understand the effects of covariates on dementia progression [175, 250–254]. Herein, multistate models were designed with state spaces of different complexity and focus. The majority included the traditional three clinical AD stages, namely cognitively unimpaired, MCI, and dementia [252, 254, 255]. Brookmeyer *et al.* further incorporated transient states representing amyloidosis and neurodegeneration, respectively [250]. Vermunt *et al.* expanded the clinical AD state space by additionally introducing prodromal AD [175]. Apart from Robitaille et al., none of these endeavors replicated their results in external data [252].

We employed multistate models to model the clinical progression of AD in six independent cohort dataset (presented in Chapter 4.1) [167]. We extracted multiple progression patterns from each individual dataset including transition probabilities between the different states, the expected duration of staying within a particular state before progressing in the disease, and the probability of staying AD diagnosis free. Moreover, we assessed hazard ratios for multiple covariates, such as, age, sex, cognitive performance, APOE $\epsilon4$ genotype, and education. Across most cohorts and progression patterns, we observed large variation. For hazard ratios, the direction of the influence of covariates was consistent across cohorts but their magnitude differed significantly. Our results highlight the particularities of AD progression and their deviation across different AD patient populations. Causing factors of the variation observed across cohorts and the implications for data-driven modeling were further discussed in Section 2.1.1 and represent another important contribution of this work. To the best of our knowledge, our publication was the first to specifically explore and compare AD progression across such a variety of progression patterns and cohorts.

Since our publication, in the dementia field, multistate models have mainly been used for estimating hazard ratios of specific covariates such as amyloid PET and APOE $\epsilon4$ status [256] and olfaction [257]. None of these studies replicated their results in external cohort data.

2.2.3 Identification of PD symptom progression subtypes

PD is a highly heterogeneous disease and it has been widely suggested that multiple disease subtypes underlie the clinical condition [32, 37, 46, 55]. To uncover the subtypes of PD, numerous research projects have been conducted. First endeavors applied primarily univariable thresholds to clinical characteristics and demographic variables, dividing patients into distinct subgroups like tremor-dominant versus postural instability and gait disorder-dominant, or early-onset versus late-onset PD [123]. These categorizations were criticized as unreliable and confounded by the disease stage of patients [258, 259].

With the rise of data-driven methods in biomedical research, clustering approaches became increasingly popular to discover new PD subtypes [60, 120, 123, 260, 261]. Methodologically, these subtyping endeavors employed classical clustering techniques such as hierarchical clustering [60, 120, 260], Gaussian mixture models [261], and k-means [260, 262, 263].

The contribution of such subtypings towards the identification of a disease modifying treatment has been put at question, largely because they commonly neglected the genetic underpinning of PD [55, 124] and disregarded its progressing nature [46, 123]. The former often limited insights into the biological pathways potentially causing the heterogeneity and thus identification of potential drug targets. Accounting for the latter warrants a time-dependent perspective on PD subtypes, for example, through clustering patients based on their progressive disease trajectories rather than cross-sectionally. Previously, studies that investigated progression in the context of PD subtypes first defined the subtypes and then compared their progression rather than including progression signals into the clustering itself [260, 262, 264].

In 2018, researchers from our group designed a novel deep learning-based clustering approach that was specifically designed with clinical data in mind [106]. The Variational Deep Embeddings with Recurrence (VADER) approach fuses variational autoencoders [265], long short-term memory networks [266], and Gaussian mixtures to cluster relatively short time-series data with missing values. Longitudinal clinical patient data are often subject to both of these characteristics. Furthermore, VADER allows for a multivariate clustering covering several dimensions, which is vital for complex diseases with patients progressing along more than one scale. Consequently, VADER provides the capabilities required for meaningful subtyping of complex, progressive disease such as PD.

Our clustering of PD patient trajectories using VADER revealed three subtypes with disparate symptomatic progression profiles (presented in Chapter 4.3)[168]. We thereby addressed shortcomings of previous subtyping endeavors by modeling and integrating information about longitudinal disease progression into the clustering itself. The main impact of our analysis resides in the associations we discovered between the identified subtypes and biological pathways, as they facilitate a deeper understanding of the biological processes that might contribute to the heterogeneity in disease progression. Furthermore, the pathways point towards key molecular players that could present promising targets for subtype-specific interventions. Additionally, our results highlighted that patients from different progression clusters diverted in their response to symptomatic treatment, which further emphasizes the need for subtype-specific interventions. One limitation of our study was that we did not find an external dataset that fulfilled the requirement of our modeling strategy and could act as a source for validation. Thus, the external validation of our findings remains important future work. Since the publication of our article, to the best of our knowledge, no new advancements have been made regarding PD subtypes.

3 Enabling robust data-driven modeling in Alzheimer's disease research

Especially when most research is conducted on observational cohort datasets, it is essential to understand the datasets' particularities and expose their biases and limitations. Moreover, it is critical to have external data sources available that can serve as validation datasets for developed approaches and data-mined disease patterns. These factors emphasize the need for in-depth knowledge of the available data in the field. The projects presented in this chapter aim to contribute to a rich, organized, and well-understood data landscape in AD research, which enables the subsequent development and application of precision medicine approaches.

3.1 Evaluating the Alzheimer's disease data landscape

In this section, we summarize our publication presented fully in **Appendix A.1**).

> **Birkenbihl, C.**, Salimi, Y., Domingo-Fernández, D., Lovestone, S., AddNeuroMed Consortium, Fröhlich, H., Hofmann-Apitius M., and Alzheimer's Disease Neuroimaging Initiative. (2020). Evaluating the Alzheimer's disease data landscape. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1), e12102. https://doi.org/10.1002/trc2.12102

Summary

Multiple AD cohort studies have collected datasets, each with their own recruitment criteria and study protocols which can introduce distinct cohort-specific biases into the acquired data. To achieve reproducible, robust data-driven results an evaluation of the present AD data landscape is vital to highlight systematic differences between cohorts that can limit the generalizability of scientific insights across patient populations [146, 172, 173]. To this end, previous efforts relied exclusively on metadata and literature [146, 180, 181], thereby neglecting the data content itself.

In the study *Evaluating the Alzheimer's disease data landscape*, we evaluated the AD data landscape by directly investigating the factually accessible data of nine major clinical AD cohort studies [139, 140, 170, 179, 223, 267–270], spanning a combined 60,004 participants and 13 distinct data modalities.

The investigated cohorts differed in key characteristics, such as the distribution of participant age and education level. Substantial deviations among cohorts were also found in the statistical distributions of key AD biomarkers like CSF amyloid beta and tau. Comparing assessments of participants' cognitive performance between cohorts underlined differences in disease severity. Analyzing the ethnoracial diversity displayed across cohorts revealed a strong bias towards White individuals with 79.3% of enrolled participants counting towards this group. Black/African descendants formed the second largest group with 11.5%. We further exposed discrepancies between previous studies reporting on the content in our investigated cohorts [180] and our results which emphasized the necessity to evaluate the data that is actually shared after data access was granted instead of relying on metadata only.

The systematic deviations we observed across AD cohort datasets can impede the application of data-driven methods across multiple datasets and limit generalizability of results achieved on single datasets. Additionally, our investigation asserted the importance of recruiting more ethnoracially diverse cohorts, a trend that was picked up by cohort studies over the following years [163]. Finally, comparison of our results to those gained by metadata-based approaches emphasized that an investigation of the accessible patient-level data is imperative to assess a data landscape.

Authors' contributions

Colin Birkenbihl and Martin Hofmann-Apitius conceived the project. Colin Birkenbihl and Yasamin Salimi collected the datasets. Colin Birkenbihl, Yasamin Salimi, and Daniel Domingo-Fernandéz performed the analysis. Holger Fröhlich provided guidance for the data analysis. Colin Birkenbihl and Daniel Domingo-Fernandéz wrote the manuscript. Martin Hofmann-Apitius and Holger Fröhlich revised the manuscript.

3.2 ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets

This section presents our following publication (see **Appendix A.2** for the full article):

> Salimi, Y., Domingo-Fernández, D., Bobis-Álvarez, C., Hofmann-Apitius, M., and **Birkenbihl, C.**, for the Alzheimer's Disease Neuroimaging Initiative, the Japanese Alzheimer's Disease Neuroimaging Initiative, for the Aging Brain: Vasculature, Ischemia, and Behavior Study, the Alzheimer's Disease Repository Without Borders Investigators, for the European Prevention of Alzheimer's Disease (EPAD) Consortium. (2022). ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets. *Alzheimer's Research & Therapy*, 14(1), 69. https://doi.org/10.1186/s13195-022-01009-4

Summary

With the advent of data-driven methods in biomedical science, the availability of large, deep-phenotyped patient-level datasets has become increasingly important [69]. Such datasets are crucial for both discovering and validating scientific insights [72]. However, in the AD domain, primarily the same data sources have been analyzed by researchers [125, 137], which can be explained by the comparably lower findability and accessibility of alternative datasets [170]. Furthermore, working across datasets is hampered by missing interoperability on the feature-level [177]. These aspects impede the advancement of AD research through emerging data-driven approaches such as machine learning and artificial intelligence and can bias current data-driven findings towards the few commonly used, well-explored AD cohorts.

Our publication titled *ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets* describes an online platform that enables the exploration of 20 AD cohort datasets with respect to longitudinal follow-up, demographics, ethnoracial diversity, measured modalities, and statistical properties of individual variables. We further conducted a semantic harmonization of the variable name spaces of the 20 cohorts and published the resulting mapping catalog which contains 1196 unique variables. The StudyPicker tool, which is build into ADataViewer, facilitates the identification of AD cohorts that meet user-specified requirements regarding available variables, sample sizes, and longitudinal follow-up.

By providing researchers with detailed information on available cohort datasets, we aim to promote robust data-driven research through enabling the identification of datasets for the discovery and validation of results. Furthermore, exploring the available data through ADataViewer can result in cumulative time savings by reveal potential data limitations that would otherwise remain hidden until researchers have completed the data access procedures. In addition, ADataViewer supports the design of research proposals by highlighting available resources during the drafting phase of research projects.

Authors' contributions

Colin Birkenbihl conceived and supervised the project. Yasamin Salimi and Colin Birkenbihl collected the datasets. Yasamin Salimi prepared the data for ADataViewer. Daniel Domingo-Fernandéz implemented the platform. Yasamin Salimi and Carlos Bobis-Álvarez curated the variable mappings. Colin Birkenbihl drafted the manuscript. Daniel Domingo-Fernandéz, Yasamin Salimi, and Martin Hofmann-Apitius revised the manuscript. Martin Hofmann-Apitius acquired the funding.

3.3 ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset

The summary provided below addresses our publication that is presented in **Appendix A.3**:

> **Birkenbihl, C.**, Westwood, S., Shi, L., Nevado-Holgado, A., Westman, E., Lovestone, S., Hofmann-Apitius, M., and AddNeuroMed Consortium. (2021). ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *Journal of Alzheimer's Disease*, 79(1), 423-431. https://doi.org/10.3233/JAD-200948

Summary

For pursuing precision medicine in AD research, accessible patient-level datasets are vital to develop and validate AI models [72]. The majority of data-driven AD research relied on the ADNI dataset, partially because alternative datasets are difficult to find and lack appropriate preprocessing to be actionable [125, 138, 156].

Following ADNI's example [140], in 2005, the AddNeuroMed consortium started to collect multimodal, longitudinal patient-level AD cohort data [157, 178]. The studies original aim was to discover novel AD biomarkers and the data was planned to be published and shared with researchers world-wide. A version of AddNeuroMed that was eventually uploaded on a data sharing platform, however, was erroneous, not interoperable between its distinct modalities, and lacked appropriate preprocessing to facilitate its analysis.

In our work titled *ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset*, we present an updated version of the AddNeuroMed data named ANMerge, in which the before mentioned shortcomings were corrected and additional data from two sister cohorts of AddNeuroMed were merged into one single dataset.

ANMerge contains data for 1,702 unique patients that stem from the original AddNeuroMed study, the Maudsley BRC Dementia Case Registry at King's Health Partners cohort (DCR), and the Alzheimer's Research

Trust UK cohort (ART) [271]. The longest patient follow-up spanned 12 years. The measured data modalities include clinical assessments, structural MRI, genotyping, transcriptomic profiling, and blood plasma proteomics. In ANMerge, all data modalities are fully interoperable, with unified patient identifiers and feature names. Furthermore, a detailed description of its content is provided through the corresponding publication and the data is accessible for third party researchers after successful data access application.

By making a ANMerge accessible, we provided the AD research community with a comprehensive alternative to previously published cohort datasets, and thereby support the discovery and robust validation of scientific insights. This work built the foundation for the research projects presented in Chapters 4.1 and 4.2, as well as other projects performed by our group and others not presented in this thesis [169, 176, 193–198, 272].

Authors' contributions

Colin Birkenbihl and Martin Hofmann-Apitius conceived the project. Simon Lovestone, Sarah Westwood, Eric Westman, Liu Shi, and Alejo Nevado-Holgado acquired and provided the original raw data. Colin Birkenbihl preprocessed the data and assembled and harmonized the ANMerge dataset. Colin Birkenbihl made the new dataset accessible. Colin Birkenbihl wrote the manuscript. Martin Hofmann-Apitius, Simon Lovestone, Liu Shi, Sarah Westwood, and Eric Westman revised the manuscript.

4 Data-driven analysis of the heterogeneity in AD and PD

In Chapter 3, we established a foundation for robust data-driven modeling across AD datasets. In this chapter, we build on those efforts and present work that models patients' disease trajectories (4.1), predicts individual disease risk (4.2 and 4.1), derives disease progression subtypes through clustering (4.3), and highlights the statistical implications of working across potentially non-i.i.d. patient-level datasets (4.2 and 4.1). These approaches and models offer new ways to stratify patients and promote a better understanding of the disease.

4.1 Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling

This section presents our publication (see **Appendix A.4**):

> **Birkenbihl, C.**[1], Salimi, Y.[1], Fröhlich, H., Japanese Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Neuroimaging Initiative. (2022). Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimer's & Dementia*, 18(2), 251-261. https://doi.org/10.1002/alz.12387

[1] Joint first authors.

Summary

The modeling of disease progression in AD is a critical aspect of understanding its dynamics and identifying opportunities for early intervention and recruitment of pre-symptomatic patients into clinical trials [20, 57]. Cohort study data are often used as a basis for these endeavors [138, 175]. However, the use of different inclusion and exclusion criteria across AD cohorts can lead to a violation of the i.i.d. assumption, which can hinder the generalizability of the results obtained.

In our study titled *'Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling'*, we modeled the progression of AD in six independent cohort datasets, compared the extracted progression patterns, and assessed their robustness and concordance across cohorts. Additionally, we proposed a clustering approach to identify the similarity of cohorts based on their exhibited progression trends.

To extract progression patterns from the cohort data, we utilized multistate models [273] with states representing the three clinical stages of AD: cognitively unimpaired, MCI, and clinical AD (i.e., dementia). Specifically, we trained one multistate model for each cohort to estimate the state transition probabilities, the probability of remaining AD diagnosis-free over time, covariate hazard ratios, and sojourn times (i.e., the expected time a participant stays in a state). Notably, we observed substantial differences in all estimated patterns across the six independent cohorts.

In a second set of analyses, we investigated whether the models had learned cohort-specific biases. We looked at the relationships between model covariates and disease progression and found that hazard ratios for the same covariates differed significantly across cohorts. Furthermore, we applied each fitted model to a combined dataset comprising all cohorts' patients, and found that progression estimates made by each model for the same data differed significantly. This indicated that the models had indeed learned cohort-specific biases from their respective training data.

We proposed a clustering approach to assess the similarity of cohort datasets based on their exhibited progression patterns. To achieve this, we first constructed a cohort similarity matrix containing the likelihoods of each cohort's observations under the fitted model of all other cohorts, respectively. We then transformed the similarity matrix into a distance matrix, which subsequently served as the basis for hierarchical clustering.

The identified differences in progression patterns and cohort-specific biases suggest that AD cohort datasets do not necessarily represent i.i.d. samples. Additionally, the findings of our study highlight that results obtained on single AD cohorts do probably not generalize to the general AD population. Our proposed clustering approach can serve as a valuable post-hoc method to quantify the similarity of data-mined patterns across different data sources. Furthermore, our results underscore the need for rigorous validation and replication of data-driven models and results in the AD domain.

Authors' contributions

Colin Birkenbihl and Holger Fröhlich designed the study. Colin Birkenbihl and Yasamin Salimi implemented the methods and ran the experiments. Colin Birkenbihl wrote the manuscript. Holger Fröhlich and Yasamin Salimi revised the manuscript. Holger Fröhlich supervised the project.

4.2 Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice

Below, we summarize our publication that is presented in **Appendix A.5**:

> **Birkenbihl, C.**, Emon, M. A., Vrooman, H., Westwood, S., Lovestone, S., AddNeuroMed Consortium, , Hofmann-Apitius M., Fröhlich, H., and Alzheimer's Disease Neuroimaging Initiative. (2020). Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA Journal*, 11, 367-376. https://doi.org/10.1007/s13167-020-00216-z

Summary

In 2018, our group proposed an AI model for early detection of patients at risk of AD that predicts a clinical AD diagnosis based on a multimodal feature signature [84]. The model was trained on the ADNI cohort [140] and internal validation achieved high prediction performance indicated by a C-index of 0.86. However, the model was not externally validated.

Through our previous work on the ANMerge dataset (presented in 3.3, [157, 170, 178]), the possibility of external validation became apparent. However, recruitment procedures can introduce systematic biases into the collected data, violating the i.i.d. assumption needed for external validation [133](see also Sections 4.1 and 3.1). With the publication of *Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice*, we aimed to 1) systematically assess differences between two landmark AD cohorts, namely ADNI and ANMerge, 2) externally validate the proposed risk model on ANMerge, and 3) evaluate the impact of systematic cohort differences on applying AI approaches across AD cohorts.

Based on our analysis of 200 shared features between ADNI and ANMerge, we found significant differences across many of them using descriptive statistics and hypothesis testing. These features included demographics, neuroimaging, and clinical assessments. Our findings suggest that a considerable number of ANMerge participants may fall outside the distribution of ADNI, and thus outside the domain of our risk model.

We revised the originally proposed risk model to limit its predictors to the intersection of the originally incorporated features and the those available in ANMerge. Afterwards, we trained the new model on ADNI again and performed an internal validation that yielded a slightly lower C-index of 0.83, as was expected given less information was included in the model. In external validation of the model on ANMerge, it achieved a C-index of 0.81.

In a second analysis, we investigated the impact of the identified cohort differences on model validation. Since participants in ANMerge and ADNI were likely not i.i.d., we performed propensity score matching to identify ANMerge participants who were similar to ADNI ones in terms of a small subset of features commonly used as recruitment criteria: sex, age, years of education, APOE $\epsilon4$ status, and MMSE. After matching, many of the initially observed differences between the two cohorts were no longer significant. When we applied the risk model to the matched ANMerge participants, we obtained a C-index of 0.88. This result indicates that the validation performance of an AI model is influenced by the proportion of participants in the validation cohort that fall outside the model domain learned on the training cohort.

To the best of our knowledge, this was one of the first times that a AD risk model was externally validated, which marked an important proof of concept that early detection of AD is possible by leveraging personalized multimodal data signatures. Furthermore, by putting the results in perspective of the data differences, the article raised awareness about a critical yet underexplored aspect of data-driven modeling in AD research. Additionally, we proposed a post-hoc evaluation strategy that allows to investigate a potential out-of-distribution effect when validating models externally.

Authors' contributions

Colin Birkenbihl, Martin Hofmann-Apitius, Holger Fröhlich designed the project. Holger Fröhlich supervised the project. Colin Birkenbihl, Sarah Westwood, Simon Lovestone retrieved the data. Colin Birkenbihl, Mohamed Asif Emon, Henri Vrooman, Holger Fröhlich analyzed the data. Colin Birkenbihl and Holger Fröhlich wrote the manuscript.

4.3 Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories

The following summary addresses our publication printed in **Appendix A.6**).

> **Birkenbihl, C.**, Ahmad, A., Massat, N. J., Raschka, T., Avbersek, A., Downey, P., Armstrong, M., and Fröhlich, H. (2023). Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories. *Scientific Reports*, 13(1), 2897. https://doi.org/10.1038/s41598-023-30038-8

Summary

PD is a highly heterogeneous disease likely comprising multiple disease subtypes [2, 37, 46, 55]. Various studies have attempted to identify these subtypes in order to disentangle the heterogeneity of the disease [60, 120, 260]. However, most of these studies neglected the progressive nature of PD by relying only on cross-sectional data. Additionally, insights into the biological pathways potentially causing the heterogeneity remained limited.

In our publication titled *Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories*, we identified and characterized three distinct PD progression clusters using longitudinal patient-level data. For clustering the multivariate trajectories of PD patients along six clinical outcomes covering both motor and non-motor symptoms, we employed the VADER approach that was previously published by our group [106]. The data used in this study originated from PPMI and included only *de-novo* PD patients who received their diagnosis at most two months prior to their first data collection visit [143].

The three identified clusters divide PD patients into subgroups experiencing 'slow', 'moderate', and 'fast' progression of the disease. We did not find significant differences between clusters for potential confounders such as time since diagnosis, distribution of Hoehn and Yahr stages at study baseline, and biological sex. Statistical analysis of DaTSCAN measurements revealed that patients from distinct clusters suffered from varying degrees of dopaminergic loss, with 'fast'-progressors being the most affected. When investigating the response to L-DOPA treatment, we found that motor symptoms in 'fast'-progressors increased steadily despite continuous treatment, whereas patients belonging to the 'slow' and 'moderate' cluster showed relatively stable management of motor symptoms while the disease progressed.

Using sparse group LASSO [274], we discovered associations between the progression clusters and clinical features, genetic markers, and biological pathways. Our results indicated, for example, that patients with rapid eye movement sleep behavior disorder and hallucinations were more likely to experience a 'faster' progression of PD. Moreover, we found that the cluster exhibiting 'fast' progression was associated with genetic perturbation of several pathways related to vesicle transport, Golgi fragmentation, and neuronal protection. Conclusively, this study provides insight into various types of PD progression and their association with unique clinical and biological features, which enhances our understanding of the heterogeneity of PD.

Authors' contributions

Patrick Downey, Martin Armstrong, and Holger Fröhlich designed the project. Martin Armstrong, Holger Fröhlich, Andreja Avbersek., Patrick Downey supervised the project. Colin Birkenbihl, Tamara Raschka, Nathalie J. Massat, Ashar Ahmad. analysed the data and implemented algorithms. Colin Birkenbihl, Holger Fröhlich, Martin Armstrong, Patrick Downey, Andreja Avbersek., Nathalie J. Massat drafted the manuscript.

5 Synthetic data: addressing the limitations of patient-level clinical data through generative models

In Chapter 3 we covered the limitations of patient-level clinical data. Chapter 4 delved into their implications for data-driven modelling in NDD research. Especially missing values, irregularities of assessment intervals across cohort studies, and data availability pose re-occurring challenges when applying data-driven approaches to biomedical data. However, generative modeling (a paradigm explained in more detail in Section 1.3.2) provides a potential solution to these limitations by enabling the generation of realistic synthetic patient-level data.

5.1 Generation of realistic synthetic data using multimodal neural ordinary differential equations

This section presents the following publication **(see Appendix A.7)**:

> Wendland, P.[1], **Birkenbihl, C.**[1], Gomez-Freixa, M., Sood, M., Kschischo, M., and Fröhlich, H. (2022). Generation of realistic synthetic data using multimodal neural ordinary differential equations. *NPJ Digital Medicine*, 5(1), 122. https://doi.org/10.1038/s41746-022-00666-x

[1] Joint first authors.

Summary

Authentic synthetic data can overcome several limitations encountered when dealing with patient-level data. They are highly regular with respect to patient follow-up and are not subject to missing values, allowing for more reliable and consistent analyses. Synthetic data can also support the simulation of counterfactual scenarios, enabling researchers to explore “what-if” scenarios without collecting new data. Additionally, synthetic data can serve as an anonymization approach, allowing for the sharing of data while protecting patient privacy [160, 199–201].

In our publication titled *Generation of realistic synthetic data using multimodal neural ordinary differential equations*, we introduce multimodal neural ordinary differential equations (MultiNODEs) a new generative AI model specifically designed for the generation of longitudinal patient-level data.

Methodologically, MultiNODEs represent an extension of neural ordinary differential equations [212] that enables the generation of multimodal data, which includes both time-dependent and static variables. To incorporate static variables, we utilized a variational autoencoder that was specifically designed for heterogeneous incomplete data (HI-VAE) [216] to embed the static variables into a latent space. The latent representation of the static variables is then concatenated with the latent representation of the time-dependent variables, which forms the initial condition for an ODE. To appropriately handle missing values in the training data, we further adopted a specific imputation layer first proposed in [106].

A significant advantage of MultiNODEs compared to other multimodal generative methods [160] lies in their capability to generate data in continuous time, allowing for smooth interpolation and extrapolation of trajectories and sampling of arbitrary time intervals. During data generation, latent representations of static and longitudinal variables are randomly sampled from a Bayesian network that models the interdependencies between them. The time-dependent variables are then generated from a latent ODE given the sampled initial condition while static variables are directly decoded from their latent representation using the HI-VAE.

We demonstrate the performance of MultiNODEs by applying them to patient-level clinical data from AD patients [269] and PD patients [143], respectively. For both datasets, MultiNODEs generated authentic synthetic data that retained the longitudinal dynamics, marginal distributions, and correlation structure of the real-world data. We further investigate the generative properties and robustness of MultiNODEs by evaluating their performance on data simulated using a non-linear epidemiological model.

Authors' contributions

Holger Fröhlich, Colin Birkenbihl, Philipp Wendland, and Maik Kschischo conceived the project. Philipp Wendland implemented the method. Philipp Wendland, Colin Birkenbihl, Marc Gomez-Freixa, and Meemansa Sood performed the experiments. Colin Birkenbihl and Holger Fröhlich drafted the manuscript. Philipp Wendland, Maik Kschischo, and Meemansa Sood revised the manuscript. Colin Birkenbihl and Holger Fröhlich supervised the work.

6 Conclusion

Despite extensive efforts to identify disease-modifying treatments for AD and PD, none have been found to date [2, 3]. This is believed to be due to the heterogeneity among patients and the late timing of interventions within patients' disease trajectories [57]. With the research presented in this thesis, we contributed to the vision of precision medicine in AD and PD through data-driven approaches aiming to mitigate the aforementioned challenges. We proposed new methods for patient stratification to identify AD in its earliest pre-symptomatic form, disentangled the heterogeneity in PD symptom progression to deepen disease understanding, and modeled AD progression across multiple independent patient populations. Beyond this, we promote robust data-driven modeling in the AD and PD domain by exploring the i.i.d. assumption across AD cohorts, describing the AD data landscape and providing tools for its exploration, contributing openly accessible data to researchers, and publishing a novel deep learning architecture for generating multimodal, time-continuous synthetic patient-level data. Many of these endeavors included or enabled thorough validation and replication of data-driven findings. Without adequate validation, research will not advance from a proof-of-concept stage to closing the translational gap and actually improving patients' lives [72].

There are various effective ways to achieve precision medicine through patient stratification. In this thesis, we present risk-based time-to-event models, trajectory modeling, and trajectory clustering. From a data-driven modeling perspective, each approach has its own advantages and disadvantages in different scenarios. Nevertheless, the stratification approaches that are best suited for clinical use will eventually prevail and lead to a transformation in healthcare. Initially, their impact will likely be seen in the context of enrichment trials. Once efficacious interventions have been identified, stratification can be applied to match individual patients with the optimal treatments based on their biomarker signatures. Matching patients with the right intervention becomes even more relevant since amyloid beta antibodies have been successful in phase III trials. Given the significant amount of adverse events connected to these drugs [275], identifying the right patients is crucial. In this context, it will become even more important to investigate and eventually predict the cognitive reserve and pathological resistance of patients, to avoid unnecessary treatments.

We do not anticipate that our presented stratification approaches will directly be implemented in clinical practice in their current form. As they have not been prospectively validated and lack regulatory approval, it would be illegal to apply them to patients for clinical decision support [72]. Nevertheless, we believe that they underline the necessity, complexity, challenges, and feasibility of data-driven research in the NDD domain. We are confident that our approaches represent significant strides towards a better comprehension of the current limitations in the field and, as a result, contribute to eventually bridging the translational gap.

6.1 Perspective future work

In the bigger context of the value chain that precision medicine can offer to NDD research (presented in Figure 3 in Chapter 2), the next steps point towards testing stratification approaches in the context of adaptive enrichment trials to improve the chances of finding efficacious treatments. But also with respect to our work presented in this thesis, there are numerous opportunities to extend them:

Our proposed AD risk model has been externally validated in one independent cohort, representing the second step in the validation process to eventually apply an AI approach in healthcare [72]. However, given the large feature space employing this model is costly both in monetary expenses and time spent collecting the necessary data. While genotyping has become less of an obstacle as became evident in oncology [86], our risk model additionally relies on MRI which remains costly. Giving the success of recent amyloid beta antibodies when treating early AD [26, 28, 275], we believe that PET scan-based features will be more regularly assessed instead of MRI, as they are necessary to receive the treatment. Therefore, it would be preferable to reduce the feature space to the minimally required predictors that achieve reasonable predictive performance and include PET-measured amyloid burden. Finally, the next step would consist of prospectively validating the model during a clinical trial and seeking regulatory approval.

With respect to our PD progression subtyping, it is important to note that it primarily captured the motor symptom progression of patients. Future work could explore the incorporation of additional non-motor symptoms, such as cognitive decline and neuropsychiatric symptoms, into the subtyping approach, as these symptoms also have a significant impact on patients' quality of life and disease management. Furthermore, the identified subtypes still require external validation. Due to the complex requirements on the data, we were unable to identify an adequate cohort for external validation. Here, new avenues need to be explored that enable leveraging data from patients that have not been aligned temporally. Temporal alignment of heterogeneous clinical trajectory could, for example, be achieved by learning relative time shifts for patients based on their observed disease trajectories. Moreover, we also see great opportunities to transfer the approach to the AD domain. Finally, it would be of great interest and economical impact to understand how an enrichment of subtypes using AI models would affect the statistical power of clinical trials.

Our contributions to the AD data landscape have found wide recognition in the field (see Chapter 2 for details). However, as more cohorts are published and others continue to collect data, constant updating will be required. Furthermore, our data harmonization efforts were mainly restricted to the semantic mapping of feature names across distinct datasets. To allow for modeling seamlessly across datasets, a statistical harmonization is also required. To this end, a global data model would be beneficial that harmonizes the entirety of data representations in the field. To the best of our knowledge, previously published data models have not made a significant impact so far [276]. Moreover, an automation of the mapping process and mathematical transformations to harmonize data representations would accelerate robust cohort data-driven analyses considerably. While we have published tools to support the semantic mapping process [277], there remains a long way to go. Finally, we believe that the extension of these efforts to the PD domain would be vital.

Considering synthetic data, there has been an ongoing debate on the trade-off between their fidelity and authenticity on one hand and the re-identification risk of patients in the training data on the other. Although several approaches have been developed, there is a need for agreed-upon best practices to enable comparable benchmarks [278, 279]. In addition, incorporating unstructured data, such as doctoral letters, electronic health records and medical claims data, into new generative models and combine them with the clinical modalities we utilized in our MultiNODEs represents a promising direction for synthetic patient-level data in the future.

Acronyms

AD Alzheimer's disease. 1

ADAS-Cog Alzheimer's Disease Assessment Scale–Cognitive subscale. 6

ADNI Alzheimer's Disease NeuroImaging Initiative. 17

AI artificial intelligence. 11

APOE apolipoprotein E. 2

CDR Clinical Dementia Rating. 6

CDRSB Clinical Dementia Rating Sum of Boxes. 6

DaTSCAN dopamine-transporter-scan. 4

EHR electronic health records. 7

EMA European Medical Agency. 2

FDA U.S. Food and Drug Administration. 2

GMMs Gaussian mixture models. 12

i.i.d. independent and identically distributed. 15

MCI mild cognitive impairment. 2

MOA mechanism of action. 5

MRI magnetic resonance imaging. 3

NDD neurodegenerative diseases. 1

PD Parkinson's disease. 1

PET positron emission tomography. 3

PPMI Parkinson's Progression Markers Initiative. 17

RBD rapid eye movement-sleep behaviour disorder. 3

UPDRS Unified Parkinson's Disease Rating Scale. 6

References

1. Dugger, B. N. & Dickson, D. W. Pathology of Neurodegenerative Diseases. *Cold Spring Harbor Perspectives in Biology* **9,** a028035. ISSN: , 1943-0264. (2023) (July 2017).
2. Poewe, W. *et al.* Parkinson Disease. *Nature Reviews Disease Primers* **3,** 1–21. ISSN: 2056-676X. (2023) (Mar. 2017).
3. Scheltens, P. *et al.* Alzheimer's Disease. *The Lancet* **397,** 1577–1590. ISSN: 0140-6736, 1474-547X. (2023) (Apr. 2021).
4. Bandres-Ciga, S., Diez-Fairen, M., Kim, J. J. & Singleton, A. B. Genetics of Parkinson's Disease: An Introspection of Its Journey towards Precision Medicine. *Neurobiology of Disease* **137,** 104782. ISSN: 0969-9961. (2023) (Apr. 2020).
5. Tang, M. *et al.* Neurological Manifestations of Autosomal Dominant Familial Alzheimer's Disease: A Comparison of the Published Literature with the Dominantly Inherited Alzheimer Network Observational Study (DIAN-OBS). *The Lancet Neurology* **15,** 1317–1325. ISSN: 1474-4422, 1474-4465. (2023) (Dec. 2016).
6. Association, A. 2019 Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia* **15,** 321–387. ISSN: 1552-5279. (2023) (2019).
7. Musiek, E. S. & Holtzman, D. M. Three Dimensions of the Amyloid Hypothesis: Time, Space and 'Wingmen'. *Nature Neuroscience* **18,** 800–806. ISSN: 1546-1726. (2023) (June 2015).
8. Selkoe, D. J. & Hardy, J. The Amyloid Hypothesis of Alzheimer's Disease at 25 Years. *EMBO Molecular Medicine* **8,** 595–608. ISSN: 1757-4676. (2023) (June 2016).
9. Heneka, M. T. *et al.* Neuroinflammation in Alzheimer's Disease. *The Lancet Neurology* **14,** 388–405. ISSN: 1474-4422. (2023) (Apr. 2015).
10. de Bruijn, R. F. & Ikram, M. A. Cardiovascular Risk Factors and Future Risk of Alzheimer's Disease. *BMC Medicine* **12,** 130. ISSN: 1741-7015. (2023) (Nov. 2014).

11. Puglielli, L., Tanzi, R. E. & Kovacs, D. M. Alzheimer's Disease: The Cholesterol Connection. *Nature Neuroscience* **6,** 345–351. ISSN: 1546-1726. (2023) (Apr. 2003).

12. McKhann, G. *et al.* Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group* under the Auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34,** 939–939. ISSN: 0028-3878, 1526-632X. (2023) (July 1984).

13. Jack Jr., C. R. *et al.* Introduction to the Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's & Dementia* **7,** 257–262. ISSN: 1552-5279. (2023) (2011).

14. Jack Jr., C. R. *et al.* NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease. *Alzheimer's & Dementia* **14,** 535–562. ISSN: 1552-5279. (2023) (2018).

15. van der Flier, W. M. & Scheltens, P. The ATN Framework—Moving Preclinical Alzheimer Disease to Clinical Relevance. *JAMA Neurology* **79,** 968–970. ISSN: 2168-6149. (2023) (Oct. 2022).

16. Aizenstein, H. J. *et al.* Frequent Amyloid Deposition Without Significant Cognitive Impairment Among the Elderly. *Archives of Neurology* **65,** 1509–1517. ISSN: 0003-9942. (2023) (Nov. 2008).

17. Tiwari, M. K. & Kepp, K. P. β-Amyloid Pathogenesis: Chemical Properties versus Cellular Levels. *Alzheimer's & Dementia* **12,** 184–194. ISSN: 1552-5279. (2023) (2016).

18. Kepp, K. P. Ten Challenges of the Amyloid Hypothesis of Alzheimer's Disease. *Journal of Alzheimer's Disease* **55,** 447–457. ISSN: 1387-2877. (2023) (Jan. 2017).

19. Masters, C. L. *et al.* Alzheimer's Disease. *Nature Reviews Disease Primers* **1,** 1–18. ISSN: 2056-676X. (2023) (Oct. 2015).

20. Kim, C. K. *et al.* Alzheimer's Disease: Key Insights from Two Decades of Clinical Trial Failures. *Journal of Alzheimer's Disease* **87,** 83–100. ISSN: 1387-2877. (2023) (Jan. 2022).

21. Wallace, L., Walsh, S. & Brayne, C. The Legacy of the 2013 G8 Dementia Summit: Successes, Challenges, and Potential Ways Forward. *The Lancet Healthy Longevity* **2,** e455–e457. ISSN: 2666-7568. (2023) (Aug. 2021).

22. Doody, R. S. *et al.* A Phase 3 Trial of Semagacestat for Treatment of Alzheimer's Disease. *New England Journal of Medicine* **369,** 341–350. ISSN: 0028-4793. (2023) (July 2013).

23. Honig, L. S. *et al.* Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. *New England Journal of Medicine* **378,** 321–330. ISSN: 0028-4793. (2023) (Jan. 2018).

24. Wang, G. *et al.* Evaluation of Dose-Dependent Treatment Effects after Mid-Trial Dose Escalation in Biomarker, Clinical, and Cognitive Outcomes for Gantenerumab or Solanezumab in Dominantly Inherited Alzheimer's Disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **14,** e12367. ISSN: 2352-8729. (2023) (2022).

25. Feustel, A. C., MacPherson, A., Fergusson, D. A., Kieburtz, K. & Kimmelman, J. Risks and Benefits of Unapproved Disease-Modifying Treatments for Neurodegenerative Disease. *Neurology* **94,** e1–e14. ISSN: 0028-3878, 1526-632X. (2023) (Jan. 2020).

26. Budd Haeberlein, S. *et al.* Two Randomized Phase 3 Studies of Aducanumab in Early Alzheimer's Disease. *The Journal of Prevention of Alzheimer's Disease* **9,** 197–210. ISSN: 2426-0266. (2023) (Apr. 2022).

27. Roy, M. & Leo, L. Biogen CEO to Step down; Drugmaker Pulls Back on Alzheimer's Drug Aduhelm. *Reuters.* (2023) (May 2022).

28. van Dyck, C. H. *et al.* Lecanemab in Early Alzheimer's Disease. *New England Journal of Medicine* **388,** 9–21. ISSN: 0028-4793. (2023) (Jan. 2023).

29. Lancet, T. Lecanemab for Alzheimer's Disease: Tempering Hype and Hope. *The Lancet* **400,** 1899. ISSN: 0140-6736, 1474-547X. (2023) (Dec. 2022).

30. Thambisetty, M. & Howard, R. Lecanemab Trial in AD Brings Hope but Requires Greater Clarity. *Nature Reviews Neurology* **19,** 132–133. ISSN: 1759-4766. (2023) (Mar. 2023).

31. Prince, M. *et al.* World Alzheimer Report 2015, The Global Impact of Dementia: An Analysis of Prevalence, Incidence, Cost and Trends (Sept. 2015).

32. Armstrong, M. J. & Okun, M. S. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA* **323,** 548–560. ISSN: 0098-7484. (2023) (Feb. 2020).

33. Postuma, R. B. *et al.* MDS Clinical Diagnostic Criteria for Parkinson's Disease. *Movement Disorders* **30,** 1591–1601. ISSN: 1531-8257. (2023) (2015).

34. Weintraub, D. & Mamikonyan, E. The Neuropsychiatry of Parkinson Disease: A Perfect Storm. *The American Journal of Geriatric Psychiatry: Official Journal of the American Association for Geriatric Psychiatry* **27,** 998–1018. ISSN: 1545-7214 (Sept. 2019).

35. Hely, M. A., Reid, W. G., Adena, M. A., Halliday, G. M. & Morris, J. G. The Sydney Multicenter Study of Parkinson's Disease: The Inevitability of Dementia at 20 Years. *Movement Disorders* **23,** 837–844. ISSN: 1531-8257. (2023) (2008).

36. Irwin, D. J. *et al.* Neuropathological and Genetic Correlates of Survival and Dementia Onset in Synucleinopathies: A Retrospective Analysis. *The Lancet Neurology* **16,** 55–65. ISSN: 1474-4422, 1474-4465. (2023) (Jan. 2017).

37. Kalia, L. V. & Lang, A. E. Parkinson's Disease. *The Lancet* **386,** 896–912. ISSN: 0140-6736, 1474-547X. (2023) (Aug. 2015).

38. Tolosa, E., Wenning, G. & Poewe, W. The Diagnosis of Parkinson's Disease. *The Lancet Neurology* **5,** 75–86. ISSN: 1474-4422, 1474-4465. (2023) (Jan. 2006).

39. Stoessl, A. J., Lehericy, S. & Strafella, A. P. Imaging Insights into Basal Ganglia Function, Parkinson's Disease, and Dystonia. *The Lancet* **384,** 532–544. ISSN: 0140-6736, 1474-547X. (2023) (Aug. 2014).

40. Kalia, L. V., Kalia, S. K. & Lang, A. E. Disease-Modifying Strategies for Parkinson's Disease. *Movement Disorders* **30,** 1442–1450. ISSN: 1531-8257. (2023) (2015).

41. PD MED Collaborative group. Long-Term Effectiveness of Dopamine Agonists and Monoamine Oxidase B Inhibitors Compared with Levodopa as Initial Treatment for Parkinson's Disease (PD MED): A Large, Open-Label, Pragmatic Randomised Trial. *The Lancet* **384,** 1196–1205. ISSN: 0140-6736, 1474-547X. (2023) (Sept. 2014).

42. LeWitt, P. A. & Fahn, S. Levodopa Therapy for Parkinson Disease: A Look Backward and Forward. *Neurology* **86,** S3–S12. ISSN: 0028-3878, 1526-632X. (2023) (Apr. 2016).

43. Wu, J., Lim, E.-C., Nadkarni, N. V., Tan, E.-K. & Kumar, P. M. The Impact of Levodopa Therapy-Induced Complications on Quality of Life in Parkinson's Disease Patients in Singapore. *Scientific Reports* **9,** 9248. ISSN: 2045-2322. (2023) (June 2019).

44. Chaudhuri, K. R. & Schapira, A. H. Non-Motor Symptoms of Parkinson's Disease: Dopaminergic Pathophysiology and Treatment. *The Lancet Neurology* **8,** 464–474. ISSN: 1474-4422, 1474-4465. (2023) (May 2009).

45. Storch, A. *et al.* Nonmotor Fluctuations in Parkinson Disease: Severity and Correlation with Motor Complications. *Neurology* **80,** 800–809. ISSN: 0028-3878, 1526-632X. (2023) (Feb. 2013).

46. Jankovic, J. & Tan, E. K. Parkinson's Disease: Etiopathogenesis and Treatment. *Journal of Neurology, Neurosurgery & Psychiatry* **91,** 795–808. ISSN: 0022-3050, 1468-330X. (2023) (Aug. 2020).

47. McFarthing, K. *et al.* Parkinson's Disease Drug Therapies in the Clinical Trial Pipeline: 2020. *Journal of Parkinson's Disease* **10,** 757–774. ISSN: 1877-7171. (2023) (Jan. 2020).

48. McFarthing, K. *et al.* Parkinson's Disease Drug Therapies in the Clinical Trial Pipeline: 2022 Update. *Journal of Parkinson's Disease* **12,** 1073–1082. ISSN: 1877-7171. (2023) (Jan. 2022).

49. Pagano, G. *et al.* Trial of Prasinezumab in Early-Stage Parkinson's Disease. *New England Journal of Medicine* **387,** 421–432. ISSN: 0028-4793. (2023) (Aug. 2022).

50. Olanow, C. W. *et al.* A Double-Blind, Delayed-Start Trial of Rasagiline in Parkinson's Disease. *New England Journal of Medicine* **361,** 1268–1278. ISSN: 0028-4793. (2023) (Sept. 2009).

51. Olanow, C. W. *et al.* A Double-Blind Controlled Trial of Bilateral Fetal Nigral Transplantation in Parkinson's Disease. *Annals of Neurology* **54,** 403–414. ISSN: 1531-8249. (2023) (2003).

52. Axelsen, T. M. & Woldbye, D. P. D. Gene Therapy for Parkinson's Disease, An&nbsp;Update. *Journal of Parkinson's Disease* **8,** 195–215. ISSN: 1877-7171. (2023) (Jan. 2018).

53. Merchant, K. M. *et al.* A Proposed Roadmap for Parkinson's Disease Proof of Concept Clinical Trials Investigating Compounds Targeting Alpha-Synuclein. *Journal of Parkinson's Disease* **9,** 31–61. ISSN: 1877-7171. (2023) (Jan. 2019).

54. Espay, A. J., Hauser, R. A. & Armstrong, M. J. The Narrowing Path for Nilotinib and Other Potential Disease-Modifying Therapies for Parkinson Disease. *JAMA Neurology* **77,** 295–297. ISSN: 2168-6149. (2023) (Mar. 2020).

55. Lang, A. E. & Espay, A. J. Disease Modification in Parkinson's Disease: Current Approaches, Challenges, and Future Considerations. *Movement Disorders* **33,** 660–677. ISSN: 1531-8257. (2023) (2018).

56. Anderson, R. M., Hadjichrysanthou, C., Evans, S. & Wong, M. M. Why Do so Many Clinical Trials of Therapies for Alzheimer's Disease Fail? *The Lancet* **390,** 2327–2329. ISSN: 0140-6736, 1474-547X. (2023) (Nov. 2017).

57. Sperling, R. A., Jack, C. R. & Aisen, P. S. Testing the Right Target and Right Drug at the Right Stage. *Science Translational Medicine* **3,** 111cm33–111cm33. (2023) (Nov. 2011).

58. Espay, A. J. *et al.* Disease Modification and Biomarker Development in Parkinson Disease: Revision or Reconstruction? *Neurology* **94,** 481–494. ISSN: 0028-3878, 1526-632X. (2023) (Mar. 2020).

59. Vogel, J. W. *et al.* Four Distinct Trajectories of Tau Deposition Identified in Alzheimer's Disease. *Nature Medicine* **27,** 871–881. ISSN: 1546-170X. (2023) (May 2021).

60. Fereshtehnejad, S.-M., Zeighami, Y., Dagher, A. & Postuma, R. B. Clinical Criteria for Subtyping Parkinson's Disease: Biomarkers and Longitudinal Progression. *Brain* **140,** 1959–1976. ISSN: 0006-8950. (2023) (July 2017).

61. Murray, M. E. *et al.* Neuropathologically Defined Subtypes of Alzheimer's Disease with Distinct Clinical Characteristics: A Retrospective Study. *The Lancet Neurology* **10,** 785–796. ISSN: 1474-4422. (2023) (Sept. 2011).

62. Macklin, E. A., Coffey, C. S., Brumm, M. C. & Seibyl, J. P. Statistical Considerations in the Design of Clinical Trials Targeting Prodromal Parkinson Disease. *Neurology* **99,** 68–75. ISSN: 0028-3878, 1526-632X. (2023) (Aug. 2022).

63. Powers, R. *et al.* Smartwatch Inertial Sensors Continuously Monitor Real-World Motor Fluctuations in Parkinson's Disease. *Science Translational Medicine* **13,** eabd7865. (2023) (Feb. 2021).

64. Rusz, J. *et al.* Speech Biomarkers in Rapid Eye Movement Sleep Behavior Disorder and Parkinson Disease. *Annals of Neurology* **90,** 62–75. ISSN: 1531-8249. (2023) (2021).

65. Goetz, C. G. *et al.* Movement Disorder Society-sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results. *Movement Disorders* **23,** 2129–2170. ISSN: 1531-8257. (2023) (2008).

66. Kueper, J. K., Speechley, M. & Montero-Odasso, M. The Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness In&nbsp;Pre-Dementia Populations. A Narrative Review. *Journal of Alzheimer's Disease* **63,** 423–444. ISSN: 1387-2877. (2023) (Jan. 2018).

67. Morris, J. C. The Clinical Dementia Rating (CDR): Current Version and Scoring Rules. *Neurology* **43,** 2412–a. ISSN: 0028-3878, 1526-632X. (2023) (Nov. 1993).

68. O'Bryant, S. E. *et al.* Staging Dementia Using Clinical Dementia Rating Scale Sum of Boxes Scores: A Texas Alzheimer's Research Consortium Study. *Archives of Neurology* **65,** 1091–1095. ISSN: 0003-9942. (2023) (Aug. 2008).

69. Hartl, D. *et al.* Translational Precision Medicine: An Industry Perspective. *Journal of Translational Medicine* **19,** 245. ISSN: 1479-5876. (2023) (June 2021).

70. Sevigny, J. *et al.* The Antibody Aducanumab Reduces Aβ Plaques in Alzheimer's Disease. *Nature* **537,** 50–56. ISSN: 1476-4687. (2023) (Sept. 2016).

71. Seibyl, J. P. & Kuo, P. What Is the Role of Dopamine Transporter Imaging in Parkinson Prevention Clinical Trials? *Neurology* **99,** 61–67. ISSN: 0028-3878, 1526-632X. (2023) (Aug. 2022).

72. Fröhlich, H. *et al.* From Hype to Reality: Data Science Enabling Personalized Medicine. *BMC Medicine* **16,** 150. ISSN: 1741-7015. (2023) (Aug. 2018).

73. Lonergan, M. *et al.* Defining Drug Response for Stratified Medicine. *Drug Discovery Today* **22,** 173–179. ISSN: 1359-6446. (2023) (Jan. 2017).

74. Ginsburg, G. S. & Phillips, K. A. Precision Medicine: From Science to Value. *Health affairs (Project Hope)* **37,** 694–701. ISSN: 0278-2715. (2023) (May 2018).

75. Seyhan, A. A. & Carini, C. Are Innovation and New Technologies in Precision Medicine Paving a New Era in Patients Centric Care? *Journal of Translational Medicine* **17,** 114. ISSN: 1479-5876. (2023) (Apr. 2019).

76. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: Multi-Omics Late Integration with Deep Neural Networks for Drug Response Prediction. *Bioinformatics* **35,** i501–i509. ISSN: 1367-4803. (2023) (July 2019).

77. Hey, S. P., Gerlach, C. V., Dunlap, G., Prasad, V. & Kesselheim, A. S. The Evidence Landscape in Precision Medicine. *Science Translational Medicine* **12,** eaaw7745. (2023) (Apr. 2020).

78. Kosorok, M. R. & Laber, E. B. Precision Medicine. *Annual Review of Statistics and Its Application* **6,** 263–286. (2023) (2019).

79. Ching, T. *et al.* Opportunities and Obstacles for Deep Learning in Biology and Medicine. *Journal of The Royal Society Interface* **15,** 20170387. (2023) (Apr. 2018).

80. Antoniou, A. *et al.* Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *The American Journal of Human Genetics* **72,** 1117–1130. ISSN: 0002-9297, 1537-6605. (2023) (May 2003).

81. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic Health Records and Polygenic Risk Scores for Predicting Disease Risk. *Nature Reviews Genetics* **21,** 493–502. ISSN: 1471-0064. (2023) (Aug. 2020).

82. Torkamani, A., Wineinger, N. E. & Topol, E. J. The Personal and Clinical Utility of Polygenic Risk Scores. *Nature Reviews Genetics* **19,** 581–590. ISSN: 1471-0064. (2023) (Sept. 2018).

83. Lamb, J. R., Jennings, L. L., Gudmundsdottir, V., Gudnason, V. & Emilsson, V. It's in Our Blood: A Glimpse of Personalized Medicine. *Trends in Molecular Medicine* **27,** 20–30. ISSN: 1471-4914, 1471-499X. (2023) (Jan. 2021).

84. Khanna, S. *et al.* Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms. *Scientific Reports* **8,** 11173. ISSN: 2045-2322. (2023) (July 2018).

85. Buchbinder, E. I. & Hodi, F. S. Impact of Precision Medicine in Oncology: Immuno-oncology. *The Cancer Journal* **29,** 15. ISSN: 1528-9117. (2023) (2023).

86. Dugger, S. A., Platt, A. & Goldstein, D. B. Drug Development in the Era of Precision Medicine. *Nature Reviews Drug Discovery* **17,** 183–196. ISSN: 1474-1784. (2023) (Mar. 2018).

87. Schwaederle, M. *et al.* Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *Journal of Clinical Oncology* **33,** 3817–3825. ISSN: 0732-183X. (2023) (Nov. 2015).

88. Dalton, W. B. *et al.* Personalized Medicine in the Oncology Clinic: Implementation and Outcomes of the Johns Hopkins Molecular Tumor Board. *JCO Precision Oncology,* 1–19. (2023) (Nov. 2017).

89. Fountzilas, E., Tsimberidou, A. M., Vo, H. H. & Kurzrock, R. Clinical Trial Design in the Era of Precision Medicine. *Genome Medicine* **14,** 101. ISSN: 1756-994X. (2023) (Aug. 2022).

90. Park, J. J. H. *et al.* Systematic Review of Basket Trials, Umbrella Trials, and Platform Trials: A Landscape Analysis of Master Protocols. *Trials* **20,** 572. ISSN: 1745-6215. (2023) (Sept. 2019).

91. Renfro, L. A. & Sargent, D. J. Statistical Controversies in Clinical Research: Basket Trials, Umbrella Trials, and Other Master Protocols: A Review and Examples. *Annals of Oncology* **28,** 34–43. ISSN: 0923-7534, 1569-8041. (2023) (Jan. 2017).

92. Lee, H. J. *et al.* Prognostic and Predictive Values of EGFR Overexpression and EGFR Copy Number Alteration in HER2-positive Breast Cancer. *British Journal of Cancer* **112,** 103–111. ISSN: 1532-1827 (Jan. 2015).

93. Freidlin, B. & Korn, E. L. Biomarker Enrichment Strategies: Matching Trial Design to Biomarker Credentials. *Nature Reviews Clinical Oncology* **11,** 81–90. ISSN: 1759-4782. (2023) (Feb. 2014).

94. Sicklick, J. K. *et al.* Molecular Profiling of Cancer Patients Enables Personalized Combination Therapy: The I-PREDICT Study. *Nature Medicine* **25,** 744–750. ISSN: 1546-170X (May 2019).

95. FDA. *Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enrichment-strategies-clinical-trials-support-approval-human-drugs-and-biological-products. Apr. 2019. (2023).

96. Mills, E. J. & Nsanzimana, S. Have Clinical Trials in HIV Finally Matured? *The Lancet HIV* **6,** e561–e563. ISSN: 2352-3018. (2023) (Sept. 2019).

97. Huang, S., Yang, J., Fong, S. & Zhao, Q. Artificial Intelligence in Cancer Diagnosis and Prognosis: Opportunities and Challenges. *Cancer Letters* **471,** 61–71. ISSN: 0304-3835. (2023) (Feb. 2020).

98. Koelzer, V. H., Sirinukunwattana, K., Rittscher, J. & Mertz, K. D. Precision Immunoprofiling by Image Analysis and Artificial Intelligence. *Virchows Archiv* **474,** 511–522. ISSN: 1432-2307. (2023) (Apr. 2019).

99. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research* **20,** 1–30. ISSN: 1533-7928. (2023) (2019).

100. Dinh, A., Miertschin, S., Young, A. & Mohanty, S. D. A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning. *BMC Medical Informatics and Decision Making* **19,** 211. ISSN: 1472-6947. (2023) (Nov. 2019).

101. Jhee, J. H. *et al.* Prediction Model Development of Late-Onset Preeclampsia Using Machine Learning-Based Methods. *PLOS ONE* **14,** e0221202. ISSN: 1932-6203. (2023) (Aug. 2019).

102. Lampe, L. *et al.* Comparative Analysis of Machine Learning Algorithms for Multi-Syndrome Classification of Neurodegenerative Syndromes. *Alzheimer's Research & Therapy* **14,** 62. ISSN: 1758-9193. (2023) (May 2022).

103. Courtiol, P. *et al.* Deep Learning-Based Classification of Mesothelioma Improves Prediction of Patient Outcome. *Nature Medicine* **25,** 1519–1525. ISSN: 1546-170X. (2023) (Oct. 2019).

104. Emon, M. A. *et al.* Clustering of Alzheimer's and Parkinson's Disease Based on Genetic Burden of Shared Molecular Mechanisms. *Scientific Reports* **10,** 19097. ISSN: 2045-2322. (2023) (Nov. 2020).

105. Kasa, S. R., Bhattacharya, S. & Rajan, V. Gaussian Mixture Copulas for High-Dimensional Clustering and Dependency-Based Subtyping. *Bioinformatics* **36,** 621–628. ISSN: 1367-4803. (2023) (Jan. 2020).

106. de Jong, J. *et al.* Deep Learning for Clustering of Multivariate Clinical Patient Trajectories with Missing Values. *GigaScience* **8,** giz134. ISSN: 2047-217X (Nov. 2019).

107. Lundberg, S. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* Nov. 2017. arXiv: `1705.07874 [cs, stat]`. (2023).

108. Molnar, C. *et al. Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process* Sept. 2021. arXiv: `2109.01433 [cs, stat]`. (2023).

109. Janzing, D., Minorics, L. & Bloebaum, P. *Feature Relevance Quantification in Explainable AI: A Causal Problem* in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (PMLR, June 2020), 2907–2916. (2023).

110. Shmueli, G. To Explain or to Predict? *Statistical Science* **25,** 289–310. ISSN: 0883-4237, 2168-8745. (2023) (Aug. 2010).

111. Vamathevan, J. *et al.* Applications of Machine Learning in Drug Discovery and Development. *Nature Reviews Drug Discovery* **18,** 463–477. ISSN: 1474-1784. (2023) (June 2019).

112. FDA. *Statement from FDA Commissioner Scott Gottlieb, M.D. on Steps toward a New, Tailored Review Framework for Artificial Intelligence-Based Medical Devices* https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-steps-toward-new-tailored-review-framework-artificial. Apr. 2019. (2023).

113. Ahmad, A. & Fröhlich, H. Integrating Heterogeneous Omics Data via Statistical Inference and Learning Techniques. *Genomics and Computational Biology* **2,** e32–e32. ISSN: 2365-7154. (2022) (Sept. 2016).

114. Willemse, E. A. J. *et al.* Comparing CSF Amyloid-Beta Biomarker Ratios for Two Automated Immunoassays, Elecsys and Lumipulse, with Amyloid PET Status. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **13,** e12182. ISSN: 2352-8729. (2023) (2021).

115. Ramanan, V. K. *et al.* Coping with Brain Amyloid: Genetic Heterogeneity and Cognitive Resilience to Alzheimer's Pathophysiology. *Acta Neuropathologica Communications* **9,** 48. ISSN: 2051-5960. (2023) (Mar. 2021).

116. Nalls, M. A. *et al.* Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson's Disease: A Meta-Analysis of Genome-Wide Association Studies. *The Lancet Neurology* **18,** 1091–1102. ISSN: 1474-4422. (2023) (Dec. 2019).

117. Shi, L. *et al.* Discovery and Validation of Plasma Proteomic Biomarkers Relating to Brain Amyloid Burden by SOMAscan Assay. *Alzheimer's & Dementia* **15,** 1478–1488. ISSN: 1552-5279. (2023) (2019).

118. Arnaldi, D. *et al.* Stratification Tools for Disease-Modifying Trials in Prodromal Synucleinopathy. *Movement Disorders* **37,** 52–61. ISSN: 1531-8257. (2023) (2022).

119. Jennings, D. *et al.* Conversion to Parkinson Disease in the PARS Hyposmic and Dopamine Transporter–Deficit Prodromal Cohort. *JAMA Neurology* **74,** 933–940. ISSN: 2168-6149. (2023) (Aug. 2017).

120. Zhou, Z. *et al.* Subtyping of Early-Onset Parkinson's Disease Using Cluster Analysis: A Large Cohort Study. *Frontiers in Aging Neuroscience* **14,** 1040293. ISSN: 1663-4365. (2023) (Nov. 2022).

121. Marder, K. *et al.* Familial Aggregation of Early- and Late-Onset Parkinson's Disease. *Annals of Neurology* **54,** 507–513. ISSN: 1531-8249. (2023) (2003).

122. Ferreira, D., Nordberg, A. & Westman, E. Biological Subtypes of Alzheimer Disease: A Systematic Review and Meta-Analysis. *Neurology* **94,** 436–448. ISSN: 0028-3878, 1526-632X. (2023) (Mar. 2020).

123. Fereshtehnejad, S.-M. & Postuma, R. B. Subtypes of Parkinson's Disease: What Do They Tell Us About Disease Progression? *Current Neurology and Neuroscience Reports* **17,** 34. ISSN: 1534-6293. (2023) (Mar. 2017).

124. Marras, C. & Lang, A. Parkinson's Disease Subtypes: Lost in Translation? *Journal of Neurology, Neurosurgery, and Psychiatry* **84,** 409–415. ISSN: 1468-330X (Apr. 2013).

125. Chen, Y. *et al.* Prediction Models for Conversion From Mild Cognitive Impairment to Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Frontiers in Aging Neuroscience* **14.** ISSN: 1663-4365. (2023) (2022).

126. Frölich, L. *et al.* Incremental Value of Biomarker Combinations to Predict Progression of Mild Cognitive Impairment to Alzheimer's Dementia. *Alzheimer's Research & Therapy* **9,** 84. ISSN: 1758-9193. (2023) (Oct. 2017).

127. Lei, B. *et al.* Deep and Joint Learning of Longitudinal Data for Alzheimer's Disease Prediction. *Pattern Recognition* **102,** 107247. ISSN: 0031-3203. (2023) (June 2020).

128. Kruthika, K. R., Rajeswari & Maheshappa, H. D. Multistage Classifier-Based Approach for Alzheimer's Disease Prediction and Retrieval. *Informatics in Medicine Unlocked* **14,** 34–42. ISSN: 2352-9148. (2023) (Jan. 2019).

129. Shi, L. *et al.* A Decade of Blood Biomarkers for Alzheimer's Disease Research: An Evolving Field, Improving Study Designs, and the Challenge of Replication. *Journal of Alzheimer's Disease* **62,** 1181–1198. ISSN: 1387-2877. (2023) (Jan. 2018).

130. Licher, S. *et al.* Development and Validation of a Dementia Risk Prediction Model in the General Population: An Analysis of Three Longitudinal Studies. *American Journal of Psychiatry* **176,** 543–551. ISSN: 0002-953X. (2023) (July 2019).

131. Van Maurik, I. S. *et al.* Biomarker-Based Prognosis for People with Mild Cognitive Impairment (ABIDE): A Modelling Study. *The Lancet Neurology* **18,** 1034–1044. ISSN: 1474-4422, 1474-4465. (2023) (Nov. 2019).

132. Farahani, A., Voghoei, S., Rasheed, K. & Arabnia, H. R. *A Brief Review of Domain Adaptation* in *Advances in Data Science and Information Engineering* (eds Stahlbock, R. *et al.*) (Springer International Publishing, Cham, 2021), 877–894. ISBN: 978-3-030-71704-9.

133. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* (2023) (Springer, New York, NY, 2013).

134. Bycroft, C. *et al.* The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **562,** 203–209. ISSN: 1476-4687. (2023) (Oct. 2018).

135. Chen, Z. *et al.* Exploring the Feasibility of Using Real-World Data from a Large Clinical Data Research Network to Simulate Clinical Trials of Alzheimer's Disease. *npj Digital Medicine* **4,** 1–9. ISSN: 2398-6352. (2023) (May 2021).

136. Ponjoan, A. *et al.* Is It Time to Use Real-World Data from Primary Care in Alzheimer's Disease? *Alzheimer's Research & Therapy* **12,** 60. ISSN: 1758-9193. (2023) (May 2020).

137. Yao, X. *et al.* Mapping Longitudinal Scientific Progress, Collaboration and Impact of the Alzheimer's Disease Neuroimaging Initiative. *PLOS ONE* **12,** e0186095. ISSN: 1932-6203. (2023) (Nov. 2017).

138. Weiner, M. W. *et al.* Impact of the Alzheimer's Disease Neuroimaging Initiative, 2004 to 2014. *Alzheimer's & Dementia* **11,** 865–884. ISSN: 1552-5279. (2023) (2015).

139. Solomon, A., Kivipelto, M., Molinuevo, J. L., Tom, B. & Ritchie, C. W. European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): Study Protocol. *BMJ Open* **8,** e021017. ISSN: 2044-6055, 2044-6055. (2023) (Dec. 2018).

140. Mueller, S. G. *et al.* The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics* **15,** 869–877. ISSN: 1052-5149, 1557-9867. (2023) (Nov. 2005).

141. Koychev, I. *et al.* Deep and Frequent Phenotyping Study Protocol: An Observational Study in Prodromal Alzheimer's Disease. *BMJ Open* **9,** e024498. ISSN: 2044-6055, 2044-6055. (2023) (Mar. 2019).

142. Moulder, K. L. *et al.* Dominantly Inherited Alzheimer Network: Facilitating Research and Clinical Trials. *Alzheimer's Research & Therapy* **5,** 48. ISSN: 1758-9193. (2023) (Oct. 2013).

143. Marek, K. *et al.* The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology. Biological Markers for Neurodegenerative Diseases* **95,** 629–635. ISSN: 0301-0082. (2023) (Dec. 2011).

144. Canevelli, M. *et al.* "Real World" Eligibility for Aducanumab. *Journal of the American Geriatrics Society* **69,** 2995–2998. ISSN: 1532-5415. (2023) (2021).

145. Rothwell, P. M. External Validity of Randomised Controlled Trials: "To Whom Do the Results of This Trial Apply?" *The Lancet* **365,** 82–93. ISSN: 0140-6736, 1474-547X. (2023) (Jan. 2005).

146. Lawrence, E. *et al.* A Systematic Review of Longitudinal Studies Which Measure Alzheimer's Disease Biomarkers. *Journal of Alzheimer's Disease* **59,** 1359–1379. ISSN: 1387-2877. (2023) (Jan. 2017).

147. Foo, J. N. *et al.* Genome-Wide Association Study of Parkinson's Disease in East Asians. *Human Molecular Genetics* **26,** 226–232. ISSN: 0964-6906. (2023) (Jan. 2017).

148. Qin, W. *et al.* Race-Related Association between APOE Genotype and Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Journal of Alzheimer's Disease* **83,** 897–906. ISSN: 1387-2877. (2023) (Jan. 2021).

149. Hunsberger, H. C., Pinky, P. D., Smith, W., Suppiramaniam, V. & Reed, M. N. The Role of APOE4 in Alzheimer's Disease: Strategies for Future Therapeutic Interventions. *Neuronal Signaling* **3,** NS20180203. ISSN: 2059-6553. (2023) (Apr. 2019).

150. Marek, K. *et al.* PPMI 2.O New Science/New Cohorts - Transforming PPMI (2490). *Neurology* **94.** ISSN: 0028-3878, 1526-632X. (2023) (Apr. 2020).

151. Properzi, M. J. *et al.* Nonlinear Distributional Mapping (NoDiM) for Harmonization across Amyloid-PET Radiotracers. *NeuroImage* **186,** 446–454. ISSN: 1053-8119. (2023) (Feb. 2019).

152. Ruiz-Godoy, L. *et al.* Identification of Specific Pre-Analytical Quality Control Markers in Plasma and Serum Samples. *Analytical Methods* **11,** 2259–2271. ISSN: 1759-9679. (2023) (Apr. 2019).

153. Vonk, J. M. J. *et al.* Cross-National Harmonization of Cognitive Measures across HRS HCAP (USA) and LASI-DAD (India). *PLOS ONE* **17,** e0264166. ISSN: 1932-6203. (2023) (Feb. 2022).

154. PPMI. *PPMI — Publications* Apr. 2023. (2023).

155. ADNI. *ADNI — Publications* Apr. 2023. (2023).

156. Veitch, D. P. *et al.* Understanding Disease Progression and Improving Alzheimer's Disease Clinical Trials: Recent Highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia* **15,** 106–152. ISSN: 1552-5279. (2023) (2019).

157. Lovestone, S. *et al.* AddNeuroMed—The European Collaboration for the Discovery of Novel Biomarkers for Alzheimer's Disease. *Annals of the New York Academy of Sciences* **1180,** 36–46. ISSN: 1749-6632. (2023) (2009).

158. Myles, P., Ordish, J. & Tucker, A. The Potential Synergies between Synthetic Data and in Silico Trials in Relation to Generating Representative Virtual Population Cohorts. *Progress in Biomedical Engineering* **5,** 013001. ISSN: 2516-1091. (2023) (Jan. 2023).

159. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic Data in Machine Learning for Medicine and Healthcare. *Nature Biomedical Engineering* **5,** 493–497. ISSN: 2157-846X. (2023) (June 2021).

160. Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M. & Fröhlich, H. Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. *Frontiers in Big Data* **3.** ISSN: 2624-909X. (2023) (2020).

161. Sood, M. *et al.* Realistic Simulation of Virtual Multi-Scale, Multi-Modal Patient Trajectories Using Bayesian Networks and Sparse Auto-Encoders. *Scientific Reports* **10,** 10971. ISSN: 2045-2322. (2023) (July 2020).

162. Chin, A. L., Negash, S. & Hamilton, R. Diversity and Disparity in Dementia: The Impact of Ethnoracial Differences in Alzheimer Disease. *Alzheimer Disease & Associated Disorders* **25,** 187. ISSN: 0893-0341. (2023) (July 2011).

163. Weiner, M. W. *et al.* Increasing Participant Diversity in AD Research: Plans for Digital Screening, Blood Testing, and a Community-Engaged Approach in the Alzheimer's Disease Neuroimaging Initiative 4. *Alzheimer's & Dementia* **19,** 307–317. ISSN: 1552-5279. (2023) (2023).

164. Bond-Taylor, S., Leach, A., Long, Y. & Willcocks, C. G. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44,** 7327–7347. ISSN: 0162-8828, 2160-9292, 1939-3539. arXiv: `2103.04922 [cs, stat]`. (2023) (Nov. 2022).

165. Birkenbihl, C. *et al.* Evaluating the Alzheimer's Disease Data Landscape. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **6,** e12102. ISSN: 2352-8737. (2023) (2020).

166. Birkenbihl, C. *et al.* Differences in Cohort Study Data Affect External Validation of Artificial Intelligence Models for Predictive Diagnostics of Dementia - Lessons for Translation into Clinical Practice. *EPMA Journal* **11,** 367–376. ISSN: 1878-5085. (2023) (Sept. 2020).

167. Birkenbihl, C., Salimi, Y., Fröhlich, H., Initiative, f. t. J. A. D. N. & Initiative, t. A. D. N. Unraveling the Heterogeneity in Alzheimer's Disease Progression across Multiple Cohorts and the Implications for Data-Driven Disease Modeling. *Alzheimer's & Dementia* **18,** 251–261. ISSN: 1552-5279. (2023) (2022).

168. Birkenbihl, C. *et al.* Artificial Intelligence-Based Clustering and Characterization of Parkinson's Disease Trajectories. *Scientific Reports* **13,** 2897. ISSN: 2045-2322. (2023) (Feb. 2023).

169. Salimi, Y. *et al.* ADataViewer: Exploring Semantically Harmonized Alzheimer's Disease Cohort Datasets. *Alzheimer's Research & Therapy* **14,** 69. ISSN: 1758-9193. (2023) (May 2022).

170. Birkenbihl, C. *et al.* ANMerge: A Comprehensive and Accessible Alzheimer's Disease Patient-Level Dataset. *Journal of Alzheimer's Disease* **79,** 423–431. ISSN: 1387-2877. (2023) (Jan. 2021).

171. Wendland, P. *et al.* Generation of Realistic Synthetic Data Using Multimodal Neural Ordinary Differential Equations. *npj Digital Medicine* **5,** 1–10. ISSN: 2398-6352. (2023) (Aug. 2022).

172. Ferreira, D. *et al.* The Interactive Effect of Demographic and Clinical Factors on Hippocampal Volume: A Multicohort Study on 1958 Cognitively Normal Individuals. *Hippocampus* **27,** 653–667. ISSN: 1098-1063. (2023) (2017).

173. Whitwell, J. L. *et al.* Comparison of Imaging Biomarkers in the Alzheimer Disease Neuroimaging Initiative and the Mayo Clinic Study of Aging. *Archives of Neurology* **69,** 614–622. ISSN: 0003-9942. (2023) (May 2012).

174. Thibeau-Sutre, E., Couvy-Duchesne, B., Dormont, D., Colliot, O. & Burgos, N. *MRI Field Strength Predicts Alzheimer's Disease: A Case Example of Bias in the ADNI Data Set* in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (Mar. 2022), 1–4.

175. Vermunt, L. *et al.* Duration of Preclinical, Prodromal, and Dementia Stages of Alzheimer's Disease in Relation to Age, Sex, and APOE Genotype. *Alzheimer's & Dementia* **15,** 888–898. ISSN: 1552-5279. (2023) (2019).

176. Poulakis, K. *et al.* Multi-Cohort and Longitudinal Bayesian Clustering Study of Stage and Subtype in Alzheimer's Disease. *Nature Communications* **13,** 4566. ISSN: 2041-1723. (2023) (Aug. 2022).

177. Golriz Khatami, S. *et al.* Challenges of Integrative Disease Modeling in Alzheimer's Disease. *Frontiers in Molecular Biosciences* **6.** ISSN: 2296-889X. (2023) (2020).

178. Lovestone, S., Francis, P. & Strandgaard, K. Biomarkers for Disease Modification Trials–the Innovative Medicines Initiative and AddNeuroMed. *The Journal of Nutrition, Health & Aging* **11,** 359–361. ISSN: 1279-7707 (2007).

179. Iwatsubo, T. *et al.* Japanese and North American Alzheimer's Disease Neuroimaging Initiative Studies: Harmonization for International Trials. *Alzheimer's & Dementia* **14,** 1077–1087. ISSN: 1552-5279. (2023) (2018).

180. Oliveira, J. L., Trifan, A. & Bastião Silva, L. A. EMIF Catalogue: A Collaborative Platform for Sharing and Reusing Biomedical Data. *International Journal of Medical Informatics* **126,** 35–45. ISSN: 1386-5056. (2023) (June 2019).

181. Janssen, O. *et al.* Real-World Evidence in Alzheimer's Disease: The ROADMAP Data Cube. *Alzheimer's & Dementia* **16,** 461–471. ISSN: 1552-5279. (2023) (2020).

182. Weiner, M. F. Perspective on Race and Ethnicity in Alzheimer's Disease Research. *Alzheimer's & Dementia* **4,** 233–238. ISSN: 1552-5260. (2023) (July 2008).

183. Mazure, C. M. & Swendsen, J. Sex Differences in Alzheimer's Disease and Other Dementias. *The Lancet Neurology* **15,** 451–452. ISSN: 1474-4422, 1474-4465. (2023) (Apr. 2016).

184. Coughlan, G. T. *et al.* Association of Age at Menopause and Hormone Therapy Use With Tau and β-Amyloid Positron Emission Tomography. *JAMA Neurology.* ISSN: 2168-6149. (2023) (Apr. 2023).

185. Nebel, R. A. *et al.* Understanding the Impact of Sex and Gender in Alzheimer's Disease: A Call to Action. *Alzheimer's & Dementia* **14,** 1171–1183. ISSN: 1552-5260. (2023) (Sept. 2018).

186. Ferretti, M. T. *et al.* Sex Differences in Alzheimer Disease — the Gateway to Precision Medicine. *Nature Reviews Neurology* **14,** 457–469. ISSN: 1759-4766. (2023) (Aug. 2018).

187. Mindt, M. R. *et al.* The Community Engaged Digital Alzheimer's Research (CEDAR) Study: A Digital Intervention to Increase Research Participation of Black American Participants in the Brain Health Registry. *The Journal of Prevention of Alzheimer's Disease.* ISSN: 2426-0266. (2023) (Mar. 2023).

188. Veitch, D. P. *et al.* Using the Alzheimer's Disease Neuroimaging Initiative to Improve Early Detection, Diagnosis, and Treatment of Alzheimer's Disease. *Alzheimer's & Dementia* **18,** 824–857. ISSN: 1552-5279. (2023) (2022).

189. Weiner, M. W. *et al.* Increasing Participant Diversity in AD Research: Plans for Digital Screening, Blood Testing, and a Community-Engaged Approach in the Alzheimer's Disease Neuroimaging Initiative 4. *Alzheimer's & Dementia* **19,** 307–317. ISSN: 1552-5279. (2023) (2023).

190. Ashford, M. T. *et al.* Screening and Enrollment of Underrepresented Ethnocultural and Educational Populations in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia* **18,** 2603–2613. ISSN: 1552-5279. (2023) (2022).

191. Mindt, M. R. *et al.* Improving Generalizability and Study Design of Alzheimer's Disease Cohort Studies in the United States by Including under-Represented Populations. *Alzheimer's & Dementia* **19,** 1549–1557. ISSN: 1552-5279. (2023) (2023).

192. Bauermeister, S. *et al.* The Dementias Platform UK (DPUK) Data Portal. *European Journal of Epidemiology* **35,** 601–611. ISSN: 1573-7284. (2023) (June 2020).

193. Golriz Khatami, S. *et al.* Comparison and Aggregation of Event Sequences across Ten Cohorts to Describe the Consensus Biomarker Evolution in Alzheimer's Disease. *Alzheimer's Research & Therapy* **14,** 55. ISSN: 1758-9193. (2023) (Apr. 2022).

194. Fillenbaum, G. G. & Mohs, R. CERAD (Consortium to Establish a Registry for Alzheimer's Disease) Neuropsychology Assessment Battery: 35 Years and Counting. *Journal of Alzheimer's Disease* **Preprint,** 1–27. ISSN: 1387-2877. (2023) (Jan. 2023).

195. Dammer, E. B. *et al.* Multi-Platform Proteomic Analysis of Alzheimer's Disease Cerebrospinal Fluid and Plasma Reveals Network Biomarkers Associated with Proteostasis and the Matrisome. *Alzheimer's Research & Therapy* **14,** 174. ISSN: 1758-9193. (2023) (Nov. 2022).

196. Maddalena, L., Granata, I., Giordano, M., Manzo, M. & Guarracino, M. R. Integrating Different Data Modalities for the Classification of Alzheimer's Disease Stages. *SN Computer Science* **4,** 249. ISSN: 2661-8907. (2023) (Mar. 2023).

197. Dartora, C. *et al. Predicting the Age of the Brain with Minimally Processed T1-weighted MRI Data* Sept. 2022. (2023).

198. Cavedo, E. *et al.* Validation of an Automatic Tool for the Rapid Measurement of Brain Atrophy and White Matter Hyperintensity: QyScore®. *European Radiology* **32,** 2949–2961. ISSN: 1432-1084. (2023) (May 2022).

199. Tucker, A., Wang, Z., Rotalinti, Y. & Myles, P. Generating High-Fidelity Synthetic Patient Data for Assessing Machine Learning Healthcare Software. *npj Digital Medicine* **3,** 1–13. ISSN: 2398-6352. (2023) (Nov. 2020).

200. Arora, A. & Arora, A. Generative Adversarial Networks and Synthetic Patient Data: Current Challenges and Future Perspectives. *Future Healthc J* **9,** 190–193. ISSN: 2514-6645, 2514-6653. (2023) (July 2022).

201. Kokosi, T. & Harron, K. Synthetic Data in Medical Research. *BMJ Medicine* **1.** ISSN: 2754-0413. (2023) (Sept. 2022).

202. Goodfellow, I. J. *et al. Generative Adversarial Networks* June 2014. arXiv: `1406.2661 [cs, stat]`. (2023).

203. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* June 2014. arXiv: `1312.6114 [cs, stat]`. (2023).

204. Rezende, D. J. & Mohamed, S. *Variational Inference with Normalizing Flows* June 2016. arXiv: `1505.05770 [cs, stat]`. (2023).

205. Kingma, D. P. & Dhariwal, P. *Glow: Generative Flow with Invertible 1x1 Convolutions* in *Advances in Neural Information Processing Systems* **31** (Curran Associates, Inc., 2018). (2023).

206. Karras, T., Aila, T., Laine, S. & Lehtinen, J. *Progressive Growing of GANs for Improved Quality, Stability, and Variation* Feb. 2018. arXiv: `1710.10196 [cs, stat]`. (2023).

207. Singh, N. K. & Raza, K. in *Health Informatics: A Computational Perspective in Healthcare* (eds Patgiri, R., Biswas, A. & Roy, P.) 77–96 (Springer, Singapore, 2021). ISBN: 9789811597350. (2023).

208. Hirte, A. U. *et al.* Realistic Generation of Diffusion-Weighted Magnetic Resonance Brain Images with Deep Generative Models. *Magnetic Resonance Imaging* **81,** 60–66. ISSN: 0730-725X. (2023) (Sept. 2021).

209. Emami, H., Dong, M., Nejad-Davarani, S. P. & Glide-Hurst, C. K. Generating Synthetic CTs from Magnetic Resonance Images Using Generative Adversarial Networks. *Medical Physics* **45,** 3627–3636. ISSN: 2473-4209. (2023) (2018).

210. Deng, R., Chang, B., Brubaker, M. A., Mori, G. & Lehrmann, A. *Modeling Continuous Stochastic Processes with Dynamic Normalizing Flows* July 2021. arXiv: `2002.10516 [cs, stat]`. (2023).

211. Yoon, J., Jarrett, D. & van der Schaar, M. *Time-Series Generative Adversarial Networks* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019). (2023).

212. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. *Neural Ordinary Differential Equations* Dec. 2019. arXiv: `1806.07366 [cs, stat]`. (2023).

213. Walonoski, J. *et al.* Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record. *Journal of the American Medical Informatics Association* **25,** 230–238. ISSN: 1527-974X. (2023) (Mar. 2018).

214. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D. Synthetic Data Generation for Tabular Health Records: A Systematic Review. *Neurocomputing* **493,** 28–45. ISSN: 0925-2312. (2023) (July 2022).

215. Buczak, A. L., Babin, S. & Moniz, L. Data-Driven Approach for Creating Synthetic Electronic Medical Records. *BMC Medical Informatics and Decision Making* **10,** 59. ISSN: 1472-6947. (2023) (Oct. 2010).

216. Nazabal, A., Olmos, P. M., Ghahramani, Z. & Valera, I. *Handling Incomplete Heterogeneous Data Using VAEs* May 2020. arXiv: `1807.03653 [cs, stat]`. (2023).

217. Bubeck, S. *et al. Sparks of Artificial General Intelligence: Early Experiments with GPT-4* https://arxiv.org/abs/2303.12712v5. Mar. 2023. (2023).

218. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. *High-Resolution Image Synthesis with Latent Diffusion Models* Apr. 2022. arXiv: `2112.10752 [cs]`. (2023).

219. Nguyen, L. X., Sone Aung, P., Le, H. Q., Park, S.-B. & Hong, C. S. *A New Chapter for Medical Image Generation: The Stable Diffusion Method* in *2023 International Conference on Information Networking (ICOIN)* (Jan. 2023), 483–486.

220. Liu, Z. *et al. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4* Mar. 2023. arXiv: `2303.11032 [cs]`. (2023).

221. Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* **388,** 1233–1239. ISSN: 0028-4793. (2023) (Mar. 2023).

222. Qian, Z., Cebere, B.-C. & van der Schaar, M. *Synthcity: Facilitating Innovative Use Cases of Synthetic Data in Different Data Modalities* Jan. 2023. arXiv: `2301.07573 [cs]`. (2023).

223. Sperling, R. A. *et al.* The A4 Study: Stopping AD Before Symptoms Begin? *Science Translational Medicine* **6,** 228fs13–228fs13. (2023) (Mar. 2014).

224. Salimi, Y. *et al. Exploring the Intricacies and Pitfalls of the ATN Framework: An Assessment across Cohorts and Thresholding Methodologies* Dec. 2022. (2023).

225. Baldeiras, I. *et al.* Alzheimer's Disease Diagnosis Based on the Amyloid, Tau, and Neurodegeneration Scheme (ATN) in a Real-Life Multicenter Cohort of General Neurological Centers. *Journal of Alzheimer's Disease* **90,** 419–432. ISSN: 1387-2877. (2023) (Jan. 2022).

226. Duong, M. T. *et al.* Dissociation of Tau Pathology and Neuronal Hypometabolism within the ATN Framework of Alzheimer's Disease. *Nature Communications* **13,** 1495. ISSN: 2041-1723. (2023) (Mar. 2022).

227. Van de Beek, M. *et al.* Association of the ATN Research Framework With Clinical Profile, Cognitive Decline, and Mortality in Patients With Dementia With Lewy Bodies. *Neurology* **98,** e1262–e1272. ISSN: 0028-3878, 1526-632X. (2023) (Mar. 2022).

228. Cullen, N. C. *et al.* Plasma Biomarkers of Alzheimer's Disease Improve Prediction of Cognitive Decline in Cognitively Unimpaired Elderly Populations. *Nature Communications* **12,** 3555. ISSN: 2041-1723. (2023) (June 2021).

229. Ezzati, A. *et al.* Predictive Value of ATN Biomarker Profiles in Estimating Disease Progression in Alzheimer's Disease Dementia. *Alzheimer's & Dementia* **17,** 1855–1867. ISSN: 1552-5279. (2023) (2021).

230. Eckerström, C., Svensson, J., Kettunen, P., Jonsson, M. & Eckerström, M. Evaluation of the ATN Model in a Longitudinal Memory Clinic Sample with Different Underlying Disorders. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **13,** e12031. ISSN: 2352-8729. (2023) (2021).

231. Teunissen, C. E., Verwey, N. A., Kester, M. I., van Uffelen, K. & Blankenstein, M. A. Standardization of Assay Procedures for Analysis of the CSF Biomarkers Amyloid β((1-42)), Tau, and Phosphorylated Tau in Alzheimer's Disease: Report of an International Workshop. *International Journal of Alzheimer's Disease* **2010,** 635053. ISSN: 2090-0252 (Sept. 2010).

232. James, C., Ranson, J. M., Everson, R. & Llewellyn, D. J. Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients. *JAMA Network Open* **4,** e2136553. ISSN: 2574-3805. (2023) (Dec. 2021).

233. Liu, S. *et al.* A Novelty Detection Approach to Effectively Predict Conversion from Mild Cognitive Impairment to Alzheimer's Disease. *International Journal of Machine Learning and Cybernetics* **14,** 213–228. ISSN: 1868-808X. (2023) (Jan. 2023).

234. Long, J. M. *et al.* Preclinical Alzheimer's Disease Biomarkers Accurately Predict Cognitive and Neuropathological Outcomes. *Brain* **145,** 4506–4518. ISSN: 0006-8950. (2023) (Dec. 2022).

235. Bloch, L., Friedrich, C. M. & for the Alzheimer's Disease Neuroimaging Initiative. Data Analysis with Shapley Values for Automatic Subject Selection in Alzheimer's Disease Data Sets Using Interpretable Machine Learning. *Alzheimer's Research & Therapy* **13,** 155. ISSN: 1758-9193. (2023) (Sept. 2021).

236. Palmqvist, S. *et al.* Prediction of Future Alzheimer's Disease Dementia Using Plasma Phospho-Tau Combined with Other Accessible Measures. *Nature Medicine* **27,** 1034–1042. ISSN: 1546-170X. (2023) (June 2021).

237. El-Sappagh, S., Alonso, J. M., Islam, S. M. R., Sultan, A. M. & Kwak, K. S. A Multilayer Multimodal Detection and Prediction Model Based on Explainable Artificial Intelligence for Alzheimer's Disease. *Scientific Reports* **11,** 2660. ISSN: 2045-2322. (2023) (Jan. 2021).

238. Maheux, E. *et al.* Forecasting Individual Progression Trajectories in Alzheimer's Disease. *Nature Communications* **14,** 761. ISSN: 2041-1723. (2023) (Feb. 2023).

239. Fonteijn, H. M. *et al.* An Event-Based Model for Disease Progression and Its Application in Familial Alzheimer's Disease and Huntington's Disease. *NeuroImage* **60,** 1880–1889. ISSN: 1053-8119. (2023) (Apr. 2012).

240. Venkatraghavan, V., Bron, E. E., Niessen, W. J. & Klein, S. Disease Progression Timeline Estimation for Alzheimer's Disease Using Discriminative Event Based Modeling. *NeuroImage* **186,** 518–532. ISSN: 1053-8119. (2023) (Feb. 2019).

241. Firth, N. C. *et al.* Sequences of Cognitive Decline in Typical Alzheimer's Disease and Posterior Cortical Atrophy Estimated Using a Novel Event-Based Model of Disease Progression. *Alzheimer's & Dementia* **16,** 965–973. ISSN: 1552-5279. (2023) (2020).

242. Young, A. L. *et al.* A Data-Driven Model of Biomarker Changes in Sporadic Alzheimer's Disease. *Brain* **137,** 2564–2577. ISSN: 0006-8950. (2023) (Sept. 2014).

243. Scotton, W. J. *et al.* A Data-Driven Model of Brain Volume Changes in Progressive Supranuclear Palsy. *Brain Communications* **4,** fcac098. ISSN: 2632-1297. (2023) (June 2022).

244. O'Connor, A. *et al.* Quantitative Detection and Staging of Presymptomatic Cognitive Decline in Familial Alzheimer's Disease: A Retrospective Cohort Analysis. *Alzheimer's Research & Therapy* **12,** 126. ISSN: 1758-9193. (2023) (Oct. 2020).

245. Oxtoby, N. P. *et al.* Data-Driven Sequence of Changes to Anatomical Brain Connectivity in Sporadic Alzheimer's Disease. *Frontiers in Neurology* **8.** ISSN: 1664-2295. (2023) (2017).

246. Young, A. L. *et al.* Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference. *Nature Communications* **9,** 4273. ISSN: 2041-1723. (2023) (Oct. 2018).

247. Tan, M.-S. *et al.* Longitudinal Trajectories of Alzheimer's ATN Biomarkers in Elderly Persons without Dementia. *Alzheimer's Research & Therapy* **12,** 55. ISSN: 1758-9193. (2023) (May 2020).

248. Diebold, F. X. State Space Modeling of Time Series: A Review Essay. *Journal of Economic Dynamics and Control* **13,** 597–612. ISSN: 0165-1889. (2023) (Oct. 1989).

249. Hougaard, P. Multi-State Models: A Review. *Lifetime Data Analysis* **5,** 239–264. ISSN: 1572-9249. (2023) (Sept. 1999).

250. Brookmeyer, R. & Abdalla, N. Multistate Models and Lifetime Risk Estimation: Application to Alzheimer's Disease. *Statistics in Medicine* **38,** 1558–1565. ISSN: 1097-0258. (2023) (2019).

251. Brookmeyer, R. & Abdalla, N. Design and Sample Size Considerations for Alzheimer's Disease Prevention Trials Using Multistate Models. *Clinical Trials* **16,** 111–119. ISSN: 1740-7745. (2023) (Apr. 2019).

252. Robitaille, A. *et al.* Transitions across Cognitive States and Death among Older Adults in Relation to Education: A Multistate Survival Model Using Data from Six Longitudinal Studies. *Alzheimer's & Dementia* **14,** 462–472. ISSN: 1552-5279. (2023) (2018).

253. Coley, N. *et al.* A Longitudinal Study of Transitions Between Informal and Formal Care in Alzheimer Disease Using Multistate Models in the European ICTUS Cohort. *Journal of the American Medical Directors Association* **16,** 1104.e1–1104.e7. ISSN: 1525-8610. (2023) (Dec. 2015).

254. Zhang, L. *et al.* Analysis of Conversion of Alzheimer's Disease Using a Multi-State Markov Model. *Statistical Methods in Medical Research* **28,** 2801–2819. ISSN: 0962-2802. (2023) (Sept. 2019).

255. Sanz-Blasco, R. *et al.* Transition from Mild Cognitive Impairment to Normal Cognition: Determining the Predictors of Reversion with Multi-State Markov Models. *Alzheimer's & Dementia* **18,** 1177–1185. ISSN: 1552-5279. (2023) (2022).

256. Jack Jr, C. R. *et al.* Long-Term Associations between Amyloid Positron Emission Tomography, Sex, Apolipoprotein E and Incident Dementia and Mortality among Individuals without Dementia: Hazard Ratios and Absolute Risk. *Brain Communications* **4,** fcac017. ISSN: 2632-1297. (2023) (Apr. 2022).

257. Knight, J. E. *et al.* Transitions Between Mild Cognitive Impairment, Dementia, and Mortality: The Importance of Olfaction. *The Journals of Gerontology: Series A,* glad001. ISSN: 1758-535X. (2023) (Jan. 2023).

258. Nutt, J. G. Motor Subtype in Parkinson's Disease: Different Disorders or Different Stages of Disease? *Movement Disorders* **31,** 957–961. ISSN: 1531-8257. (2023) (2016).

259. Simuni, T. *et al.* How Stable Are Parkinson's Disease Subtypes in de Novo Patients: Analysis of the PPMI Cohort? *Parkinsonism & Related Disorders* **28,** 62–67. ISSN: 1353-8020, 1873-5126. (2023) (July 2016).

260. Fereshtehnejad, S.-M. *et al.* New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression: A Prospective Cohort Comparison With Other Phenotypes. *JAMA Neurology* **72,** 863–873. ISSN: 2168-6149. (2023) (Aug. 2015).

261. Dadu, A. *et al.* Identification and Prediction of Parkinson's Disease Subtypes and Progression Using Machine Learning in Two Cohorts. *npj Parkinson's Disease* **8,** 1–12. ISSN: 2373-8057. (2023) (Dec. 2022).

262. Lawton, M. *et al.* Developing and Validating Parkinson's Disease Subtypes and Their Motor and Cognitive Progression. *Journal of Neurology, Neurosurgery & Psychiatry* **89,** 1279–1287. ISSN: 0022-3050, 1468-330X. (2023) (Dec. 2018).

263. Erro, R. *et al.* Clinical Clusters and Dopaminergic Dysfunction in De-Novo Parkinson Disease. *Parkinsonism & Related Disorders* **28,** 137–140. ISSN: 1353-8020, 1873-5126. (2023) (July 2016).

264. Alves, G., Larsen, J. P., Emre, M., Wentzel-Larsen, T. & Aarsland, D. Changes in Motor Subtype and Risk for Incident Dementia in Parkinson's Disease. *Movement Disorders* **21,** 1123–1130. ISSN: 1531-8257. (2023) (2006).

265. Jiang, Z., Zheng, Y., Tan, H., Tang, B. & Zhou, H. *Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering* June 2017. arXiv: `1611.05148 [cs]`. (2023).

266. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780. ISSN: 0899-7667. (2023) (Nov. 1997).

267. Ellis, K. A. *et al.* The Australian Imaging, Biomarkers and Lifestyle (AIBL) Study of Aging: Methodology and Baseline Characteristics of 1112 Individuals Recruited for a Longitudinal Study of Alzheimer's Disease. *International Psychogeriatrics* **21,** 672–687. ISSN: 1741-203X, 1041-6102. (2023) (Aug. 2009).

268. Bos, I. *et al.* The EMIF-AD Multimodal Biomarker Discovery Study: Design, Methods and Cohort Characteristics. *Alzheimer's Research & Therapy* **10,** 64. ISSN: 1758-9193. (2023) (July 2018).

269. Besser, L. *et al.* Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Disease & Associated Disorders* **32,** 351. ISSN: 0893-0341. (2023) (Oct. 2018).

270. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *Journal of Alzheimer's Disease* **64,** S161–S189. ISSN: 1387-2877. (2023) (Jan. 2018).

271. Hye, A. *et al.* Proteome-Based Plasma Biomarkers for Alzheimer's Disease. *Brain* **129,** 3042–3050. ISSN: 0006-8950. (2023) (Nov. 2006).

272. Karaman, B. K., Mormino, E. C., Sabuncu, M. R. & Initiative, f. t. A. D. N. Machine Learning Based Multi-Modal Prediction of Future Decline toward Alzheimer's Disease: An Empirical Study. *PLOS ONE* **17,** e0277322. ISSN: 1932-6203. (2023) (Nov. 2022).

273. Jackson, C. Multi-State Models for Panel Data: The Msm Package for R. *Journal of Statistical Software* **38,** 1–28. ISSN: 1548-7660. (2023) (Jan. 2011).

274. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* **22,** 231–245. ISSN: 1061-8600. (2022) (Apr. 2013).

275. *Lilly's Donanemab Significantly Slowed Cognitive and Functional Decline in Phase 3 Study of Early Alzheimer's Disease — Eli Lilly and Company* https://investor.lilly.com/news-releases/news-release-details/lillys-donanemab-significantly-slowed-cognitive-and-functional. (2023).

276. Neville, J. *et al.* Accelerating Drug Development for Alzheimer's Disease through the Use of Data Standards. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **3,** 273–283. ISSN: 2352-8737. (2023) (June 2017).

277. Wegner, P. *et al.* Integrative Data Semantics through a Model-Enabled Data Stewardship. *Bioinformatics* **38,** 3850–3852. ISSN: 1367-4803. (2023) (Aug. 2022).

278. Alaa, A., Breugel, B. V., Saveliev, E. S. & van der Schaar, M. *How Faithful Is Your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models* in *Proceedings of the 39th International Conference on Machine Learning* (PMLR, June 2022), 290–306. (2023).

279. Jordon, J. *et al. Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-identification* in *Proceedings of the NeurIPS 2020 Competition and Demonstration Track* (PMLR, Aug. 2021), 206–215. (2023).

A Appendix

A.1 Evaluating the Alzheimer's disease data landscape

Reprinted with permission from "Birkenbihl, C., Salimi, Y., Domingo-Fernández, D., Lovestone, S., AddNeuroMed Consortium, Fröhlich, H., Hofmann-Apitius M., and Alzheimer's Disease Neuroimaging Initiative. (2020). Evaluating the Alzheimer's disease data landscape. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1), e12102.".

Received: 10 September 2020 | Accepted: 21 September 2020 | Published online: 2 December 2020

DOI: 10.1002/trc2.12102

Alzheimer's & Dementia
Translational Research
& Clinical Interventions

RESEARCH ARTICLE

Evaluating the Alzheimer's disease data landscape

Colin Birkenbihl[1,2] | Yasamin Salimi[1,2] | Daniel Domingo-Fernándéz[1,2] | Simon Lovestone[3] | AddNeuroMed consortium | Holger Fröhlich[1,2] | Martin Hofmann-Apitius[1,2] | the Japanese Alzheimer's Disease Neuroimaging Initiative* | and the Alzheimer's Disease Neuroimaging Initiative†

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

[2] Bioinformatics Group, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

[3] Department of Psychiatry, University of Oxford, Oxford, UK

Correspondence
Colin Birkenbihl, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53757 Sankt Augustin, Germany.
Email: colin.birkenbihl@scai.fraunhofer.de

*Japanese Alzheimer's Disease Neuroimaging Initiative: Data used in preparation of this article were obtained from the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) database deposited in the National Bioscience Database Center Human Database, Japan (Research ID: hum0043.v1, 2016). As such, the investigators within J-ADNI contributed to the design and implementation of J-ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of J-ADNI investigators can be found at: https://humandbs.biosciencedbc.jp/en/hum0043-j-adni-authors.

†Alzheimer's Disease Neuroimaging Initiative: Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Funding information
European Union's Seventh Framework Programme, Grant/Award Number: FP7/2007-2013

Abstract

Introduction: Numerous studies have collected Alzheimer's disease (AD) cohort data sets. To achieve reproducible, robust results in data-driven approaches, an evaluation of the present data landscape is vital.

Methods: Previous efforts relied exclusively on metadata and literature. Here, we evaluate the data landscape by directly investigating nine patient-level data sets generated in major clinical cohort studies.

Results: The investigated cohorts differ in key characteristics, such as demographics and distributions of AD biomarkers. Analyzing the ethnoracial diversity revealed a strong bias toward White/Caucasian individuals. We described and compared the measured data modalities. Finally, the available longitudinal data for important AD biomarkers was evaluated. All results are explorable through our web application ADataViewer (https://adata.scai.fraunhofer.de).

Discussion: Our evaluation exposed critical limitations in the AD data landscape that impede comparative approaches across multiple data sets. Comparison of our results to those gained by metadata-based approaches highlights that thorough investigation of real patient-level data is imperative to assess a data landscape.

KEYWORDS
Alzheimer's disease, biomarker, clinical study, cohort, cohort study, data, data access, data sharing, data viewer, data-driven, data set, dementia, disease modeling, FAIR data, magnetic resonance imaging, open-science, patient level data

1 | BACKGROUND

In the field of Alzheimer's disease (AD) research, numerous cohort studies have been conducted, and their collected data build the basis for a plethora of research projects. However, each of these studies only reflects patients of a particular subpopulation defined by inclusion and exclusion criteria. This is becoming especially relevant with respect to the increasing popularity of data-driven approaches and machine learning.[1,2] After analyzing a single cohort, it is mandatory to demonstrate that results are reproducible in independent, external data originating from distinct cohort studies. Furthermore, it is essential to conduct comparative analyses across data sets to assess whether the observed patterns are robust.[3] Such systematic data-driven approaches are, however, hampered because patient-level data are often difficult to access or entirely inaccessible. Moreover, we have limited knowledge about how the distinct cohort data sets available in our field compare to each other on a qualitative (eg, overlap of measured variables) as well as quantitative level (eg, values encountered in the data).[4,5] Thus, to leverage the full potential of collected patient-level data, it is important to characterize the clinical AD data landscape in detail.

1.1 | Metadata-driven evaluations of the Alzheimer's disease data landscape

Evaluating a data landscape involves organizing and comparing data sets to: (1) qualitatively assess their collected data modalities and variables, and (2) quantitatively describe the demographics of the study population and distributions of measured variables. Such characterization provides a detailed overview of the data accessibility and supports the design of research projects and future cohort studies. Finally, evaluating a data landscape inherently exposes potential flaws with regard to interoperability between existing data sets and underrepresentation of important disease or population characteristics.

In the AD field, previous studies have attempted to establish a comprehensive view of the AD data landscape as well as to demonstrate how cohort data sets relate to each other. For example, the European Medical Information Framework (EMIF) collected metadata of AD cohort studies by providing data owners with a questionnaire in which they could specify the variables contained in their data sets. The resulting metadata is presented through the EMIF-Catalog.[6] Similarly, the Real world Outcomes across the Alzheimer's Disease spectrum for better care: Multi-modal data Access Platform (ROADMAP) project generated an overview of clinical outcomes and data modalities that were collected in several European AD cohort studies.[7] By analyzing metadata (partially originating from the EMIF-Catalog), ROADMAP created the ROADMAP Data Cube, a web application that shows the availability of AD-related outcomes in a selected set of European dementia cohorts (https://datacube.roadmap-alzheimer.org). Lawrence et al., on the other hand, opted for a literature-based approach to assess the AD data landscape. The authors reviewed publications corresponding to AD cohort data sets and gathered the contained information.[7]

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed relevant literature through bibliographic search engines. Relevant cohort data sets have been discovered through data portals, data publications, and citations in the literature. Applications were filed for 18 cohort data collections of which 9 were successful.
2. Interpretation: The presented results illustrate the current state of the Alzheimer's disease (AD) data landscape from a patient-level data-centric perspective, whereas previous investigations relied solely on provided cohort metadata. This investigation exposes limitations in data availability and interoperability, and establishes a detailed overview on what resources current data sets provide for data-driven analyses.
3. Future directions: This work emphasizes the need for a common semantic framework for patient-level AD data to enable the community to work across cohort data sets and ultimately to generate robust scientific insights to advance AD research.

1.2 | Moving beyond metadata through data-level investigations

All of the above-mentioned undertakings attempted to evaluate the AD data landscape solely on the basis of metadata and literature, without investigating the underlying patient-level data. However, reviewing study protocols can only explain the original design of a given study and thereby neglects unforeseen changes in procedures or participant recruitment throughout study runtime. The alternative approach is a patient-level and data-driven evaluation of the AD data landscape, which is a tedious and time-consuming endeavor. The first hurdle of such an approach is gaining access to a sufficient number of cohort data sets. Data access typically requires completing an application procedure with numerous legal requirements and considerations. If access is granted, intensive manual curation and investigation of data follow. Although difficult to establish, a comprehensive data-driven view on the AD data landscape is crucial, since reliance exclusively on metadata assumes that these metadata correctly describe the underlying data sets and that the data sets are complete. In contrast, a patient-level and data-driven evaluation (1) is not subject to these assumptions, (2) allows for a quantitative investigation of important cohort statistics, and (3) illustrates the amount and quality of the data accessible to the field.

1.3 | Novelty and impact of this work

In this work, we aimed at assessing the current AD data landscape through meticulous investigation and curation of accessible cohort

TABLE 1 The investigated AD cohorts and their references

Cohort	Consortium	Reference
A4	Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease	9
ADNI	The Alzheimer's Disease Neuroimaging Initiative	10
ANMerge	AddNeuroMed	11
AIBL	The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing	12
EMIF-1000	European Medical Information Framework	13
EPAD v1500	European Prevention of Alzheimer's Dementia	14
JADNI	Japanese Alzheimer's Disease Neuroimaging Initiative	15
NACC	The National Alzheimer's Coordinating Center	16
ROSMAP	The Religious Orders Study and Memory and Aging Project	17

data sets on the data level rather than solely relying on metadata and/or literature. To accomplish this task, we traced down, accessed, investigated, and compared nine of the major clinical cohort study data sets available in the AD field. Here, we comprehensively describe the acquired data and show which data modalities we found in the data sets as well as their overlaps with other studies. In addition, we assessed the longitudinal follow-up on the biomarker level and demonstrated to what extent current AD data are covering the progression of the disease. Furthermore, we compared the content we observed in these data sets with the reported findings of metadata-based approaches.[6,8] Finally, we made all results available through ADataViewer (https://adata.scai.fraunhofer.de), an interactive web-portal that allows researchers to explore the AD data landscape generated based on the investigated data sets.

2 | METHODS

2.1 | Investigated cohorts

We aimed to acquire as many major AD cohort studies as possible to allow for a thorough investigation of the data landscape. We only considered data sets that were downloadable, hereby excluding data portals with restricted data access from our investigations. Most of the data sets we accessed were shared after completing an official data request process. We applied for access to 18 distinct AD cohort data sets. Until submitting this work for publication, we were granted access to nine (Table 1). We discuss the reasons behind failed data access applications in the Supplementary Text. Notably, not all of the accessed data sets are observational cohort studies in the strict sense; for more information, please see the Supplementary Text.

It is important to be aware that not all of these studies followed the same design or goals. Each study enforced its own recruitment criteria and enrolled participants following distinct selection processes. Although some aimed for a case-control setting and included a substantial amount of AD patients in their cohort, others deliberately excluded them to focus on early disease progression. Therefore, the cohort data sets are all subject to inherent biases.

2.2 | Generating the summary statistics

To illustrate the content of the data sets, we characterized the demographics of each cohort and described the encountered statistical distributions of important AD biomarkers. The demographic variables we considered are: participant age, sex, and completed years of education. The AD biomarkers we compared between cohorts are motivated in the Supplementary Text. In addition, we assessed the diversity of ethnoracial groups in our acquired AD cohorts, since it is known that ethnoracial factors may impact AD and related findings.[19] More detailed definitions of the ethnoracial groups can be found in the Supplementary Text.

For numerical variables, we describe the encountered distributions using the 25%, 50%, and 75% quantiles of the raw measurements. For categorical ones, we describe the proportion of study participants falling into its respective categories. In some data sets, single variables were reported only numerically given that they were placed within a defined value range (eg, 400 to 1700). If the measurement appeared to be outside of this range, the exact number was not reported but replaced with a cutoff (eg, ">1700"). To allow for calculations, we considered these values to be equal to the mentioned cutoff (here, 1700).

2.3 | Generating the data availability map

While establishing a data landscape, it is of high interest to identify the data modalities that were measured in the underlying studies as well as to compare their overlaps. However, assessing the availability of data modalities in clinical cohort data sets is not straightforward. This process involves intensive and meticulous manual curation of the acquired data sets and thereby the definition of applicable curation criteria specifying under which circumstances each data modality is considered as "available." Furthermore, it is often necessary to define a gradual categorization to represent the degree of availability. For example, exclusively measuring two specific single nucleotide polymorphisms (SNPs) is not equal to conducting a genome-wide genotyping of individuals. Similarly, distributing normalized brain volumes summed over both hemispheres is less informative than providing the underlying raw magnetic resonance (MR) images. The latter would enable researchers to process the images according to their needs, whereas the former impedes interoperability to other data sets due to differences in employed image-processing pipelines. This could hamper certain analyses such as systematic comparisons across cohorts or validation approaches.

To enable a meaningful, comparable assessment of the availability of data modalities, we established criteria for categorizing the availability

TABLE 2 Description of the investigated cohorts

Cohorts	N	Healthy	MCI	AD	N with 2+ visits	Follow-up Interval (months)	Location	Diagnostic criteria AD
A4	6943	6943	0	0	0:	≈8	US, Canada, Australia	AD patients excluded
ADNI	2249	813	1016	389	1978 (88%)	6	USA, Canada	NINCDS-ADRDA
AIBL	1378	803	134	181	1019 (74%)	18	Australia	NINCDS-ADRDA
ANMerge	1702	793	397	512	1254 (74%)	12	Europe	NINCDS-ADRDA
EMIF	1221	386	526	201	0	no follow-up	Europe	NINCDS-ADRDA
EPAD v1500	1500	1410	80	3	0:	6	Europe	NINCDS-ADRDA
JADNI	537	151	233	149	518	6	Japan	NINCDS-ADRDA
NACC	40858	15894	3649	11761	27657 (68%)	12	US	UDS Form D1
ROSMAP	3627	2514	898	203	3335 (92%)	12	US	NINCDS-ADRDA

NOTE: The numbers of diagnosed subjects do not always add up to N, since patients with different dementia diagnoses (eg, Lewy body or frontotemporal dementia) were excluded. N, Total number of participants; CTL/MCI/AD, Number of participants with the respective diagnosis at study baseline; 2+ visits, Number of study participants for whom data for at least two time points are available; Follow-up Interval, Approximated regular time interval between participant visits; Longitudinal data have been collected but are not yet released.

TABLE 3 Distribution of demographic variables and key AD biomarkers encountered in each cohort

	Female %	Age	Education	APOE ε4 %	MMSE	CDR	CDR-SB	Hippocampus	A-beta	t-Tau	p-Tau
A4	57.7	68, 71, 75	14, 16, 18	34.3	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	6, 7, 7			
ADNI	47	68, 73, 78	14, 16, 18	45.6	26, 28, 29	0.0, 0.5, 0.5	0.0, 1.0, 2.0	5948, 6864, 7651	596, 854, 1396	193, 258, 350	17, 24, 34
AIBL	57.9	67, 73, 79	10, 12, 15	36	26, 28, 30	0.0, 0.0, 0.5	0.0, 0.0, 1.0	3, 3, 3	445, 567, 802	238, 366, 516	43, 64, 81
ANMerge	59.3	71, 77, 81	8, 11, 14	38.8	24, 28, 29	0.0, 0.5, 0.5	0.0, 0.5, 4.0	5311, 6270, 7142			
EMIF	46.2	62, 68, 74	9, 12, 15	46.8	25, 28, 29	0.5, 0.5, 0.5		6357, 7223, 8004	385, 525, 739	160, 278, 504	37, 52, 74
EPAD	56.9	60, 66, 71	12, 15, 17	37.7	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	4413, 4808, 5182	899, 1319, 1700	162, 201, 252	13, 17, 22
JADNI	52.7	66, 72, 77	12, 12, 16	46.1	24, 26, 29	0.0, 0.5, 0.5	0.0, 1.5, 3.0	5260, 6133, 7132	254, 315, 454	67, 101, 138	36, 48, 73
NACC	57.2	65, 72, 79	12, 16, 18	40.6	23, 27, 29	0.0, 0.5, 0.5	0.0, 1.0, 4.0	43.5%	46.5%	43.9%	43.9%
ROSMAP	72.8	73, 79, 84	14, 16, 18	25.1	27, 29, 30						

NOTE: We show the 25%, 50%, and 75% quantiles of numerical variables at baseline. Categorical variables are given as the proportion of participants falling into one respective category. *APOE* ε4%, Proportion of participants with at least one *APOE* ε4 allele; Hippocampus, A-beta, t-Tau, p-Tau, NACC values are given as the proportion of "abnormal observations".

of each modality into three discrete stages (Supplementary Table S1): stage 0, no data were available for the respective modality; stage 1, data were partially available; and stage 2, more complete data or unprocessed raw data were available.

2.4 | Investigating longitudinal follow-up across studies

To assess how far existing cohort data sets cover the time dimension of AD, we conducted a thorough investigation of their respective longitudinal follow-up. For each cohort, we evaluated how many participants were assessed at each follow-up visit and implicitly analyzed the drop-out over study runtime. Since not all measurements were performed at each visit and not every individual participated in all sample collections, we further focused on the follow-up and coverage of important AD biomarkers. Determining the amount of available longitudinal data per biomarker provides insight on how much information we can exploit to model and ultimately understand patterns of AD progression. As of publication of this article, EPAD and NACC are still subject of ongoing data collection, while ADNI received funding to extend their study and continue participant recruitment.

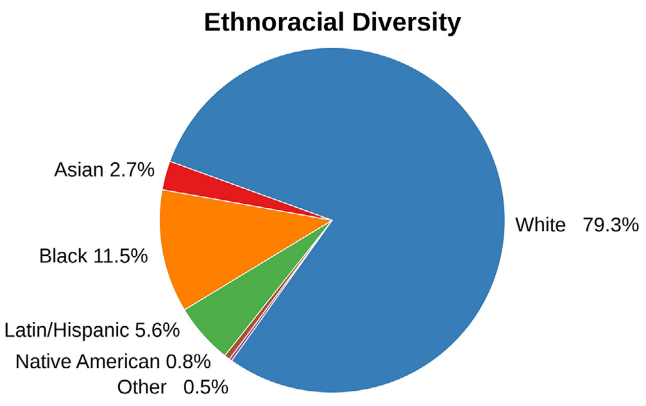


FIGURE 1 Combined ethnoracial diversity found across the investigated AD cohorts. Table S2 shows the individual compositions of each cohort

3 | RESULTS

3.1 | Investigation of the AD data landscape

Altogether, we investigated data from nine studies comprising a total of 60,004 assessed study participants. Table 2 shows how these participants were distributed among the analyzed cohorts. With NACC being the exception (n = 40,858), all studies recruited individuals in the low thousands (n = ≈1200 to 3600). According to their diagnosis, participants could be separated into three groups: cognitively healthy controls, patients with mild cognitive impairment (MCI), and patients with AD. Seven of the investigated studies based their diagnoses on the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria[20] which significantly increases the interoperability between those data sets, since AD follows the same semantic description. Depending on each study's goals, the recruitment process focused on enrolling more or fewer individuals falling into specific diagnosis groups.

Although no data are shared through our web-portal, information on how to access the data sets can be found at https://adata.scai.fraunhofer.de/cohorts.

3.2 | Characterization of the cohorts

Investigation of the cohort demographics revealed considerable differences between key demographic characteristics of the acquired cohorts. EPAD, for example, recruited a comparably young and primarily non-symptomatic cohort, whereas participants of ANMerge and ROSMAP were significantly older (Table 3). Across all cohorts, the age range spans roughly from 60 (lowest 25% quantile) to 85 years (highest 75% quantile). Theoretically, this opens the opportunity to construct a pseudo-continuum of 25 years of disease history. Furthermore, in most studies, we observed the general tendency that more female than male participants enrolled into the studies. Overall, most individuals included in the AD cohort studies were highly educated (≈14 years on average). As previously mentioned by Whitwell et al., a high level of education can act as cognitive reserve, possibly concealing a prodromal manifestation of AD.[5] Numerous demographic differences found between studies may result from distinct recruitment criteria which, again, mirror the individual study goals. Although distinct recruitment criteria lead to a broader sampling of the AD population, they reduce the direct comparability between data sets because they inevitably introduce bias into the data. One key example is recruitment specifically for participants with AD risk factors (eg, *APOE* genotype). This could significantly bias the patterns exhibited in the data in comparison to another data set with a lower amount of *APOE* ε4–positive participants.

To further highlight one potential bias in AD data, we analyzed the ethnoracial diversity encountered in the investigated AD cohorts (Figure 1). An aggregated analysis of all acquired data sets demonstrates that most of these recruited individuals come from a White/Caucasian background (79.3%). The second largest group was Black/African descendants with 11.5%, followed by participants of Latin/Hispanic heritage with 5.6%. Here, we would like to point out that these findings are heavily influenced by the study location and the number of enrolled participants per study. Because the majority of the studies have been conducted in the United States, their locally exhibited ethnoracial diversity overshadows signals from European cohorts. However, the analogous plots for each European cohort show not only a similar, but even more extreme tendency toward White/Caucasian individuals (EPAD: 99% white; ANMerge: 98,5% white; see https://adata.scai.fraunhofer.de/ethnicity).

The ethnoracial composition in the investigated cohorts relies on the diversity of populations from which the participants have been recruited. Nonetheless, our results elucidate that there is a substantial bias toward White/Caucasian in AD data sets and a severe underrepresentation of other ethnoracial groups, which, in turn, could be problematic for developing personalized treatments.

3.3 | Availability of data modalities

To analyze which modalities are available in our investigated cohorts and to explore the overlaps between them, we assigned a score of availability per data modality according to our previously described criteria (Table S1).

In Figure 2A, we show an overview of the data modalities and their availability score in all acquired cohort data sets. Commonly assessed modalities throughout all studies were demographic variables (eg, age, sex, and education) as well as clinical assessments (eg, Mini Mental State Examination [MMSE]). Regarding these two modalities, eight studies were assigned the availability score 2, with EMIF and AIBL being the only exceptions due to missing ethnoracial information. Cerebrospinal fluid (CSF) biomarker measurements were found to be present in all data sets but ANMerge. With regard to autopsy data, only ROSMAP contained a detailed collection, ranging from simple measurements such as brain weight to comprehensive brain proteomics

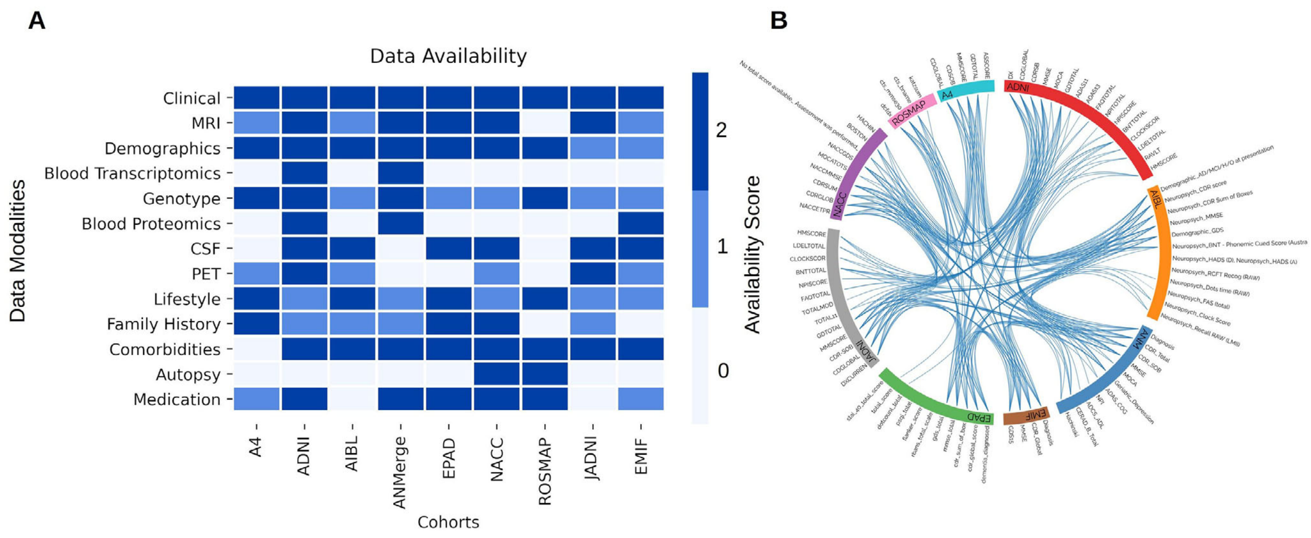


FIGURE 2 Interoperability of AD data sets. A, Availability of data modalities scored based on the defined criteria. The criteria are explained in Supplementary Table 1. B, Equivalence of clinical assessment variables across cohorts. PET = positron emission tomography

and transcriptomics. Although seven studies released some structural MRI data, three of those limited the shared data to processed MRI features (eg, brain volumes). In our case, only ADNI, NACC, JADNI, EPAD, and ANMerge granted access to the raw images.

Although the purpose of this section is to provide a comprehensive overview about the availability of data modalities, we would like to emphasize that the presented results are strongly dependent on our defined curation criteria, and different criteria could lead to deviating results. In addition, all investigated data sets could hold more information than we presented here. Due to our premise of looking exclusively into those patient-level data that have indeed been shared with us, it is possible that we missed modalities or resources that are existent but were not shared (eg, MRI images). Our results can be explored at https://adata.scai.fraunhofer.de/modality.

3.4 | Metadata investigation versus data investigations

To establish how our observations of data availability differed from results gained by solely investigating metadata, we qualitatively compared our findings to the metadata presented in the EMIF catalog.[6*] Only four of our investigated studies were listed: ADNI, ANMerge, EMIF, and EPAD. Although the majority of our findings are in concordance with the EMIF-catalog, deviations between metadata and the real data exist. We encountered variables in the data sets that are reported as absent in the catalog (eg, Global Deterioration Scale in ANMerge), or were not listed at all. Other variables and even modalities are reported to be present, yet could not be found in the respective data set. For instance, the catalog states that post-mortem brain autopsy was performed in ANMerge, for which we could not find any evidence.

Similar observations were made when comparing our findings to the review by Lawrence et al.[8] Here, for example, the reported longitudinal follow-up of ANMerge is significantly shorter than what we observed in the data (reported: 12 months, data: 84 months). In addition, the reported number of participants with at least two visits does not equal our findings (reported: 378, data: 1254 participants).

3.5 | Availability of data modalities

The finding of common modalities across cohorts does not imply that the measured variables are interoperable or even comparable on a semantic level. By mapping a variety of variables across the data sets, we established an overview of their interoperability (Figure 2B). We would like to emphasize that the current version of these mappings is not complete but a proof of concept that a semantic integration of these data sets is, in theory, possible. However, this integration is cumbersome and time-consuming, as many data sets exhibit low interoperability and distinct variable naming conventions. An in-depth view of the preliminary mappings is given at https://adata.scai.fraunhofer.de/feature_comparison.

3.6 | Disease manifestation across cohorts

To evaluate how severely patients from each cohort have been affected by AD, we compared the distributions of both cognitive outcomes and key biomarkers for the cognitively affected patient subgroups (ie, participants with an MCI or AD diagnosis). Table 3 shows the distributions for each complete cohort including healthy controls, MCI, and AD patients. Analogous tables per diagnosis subgroup can be found at https://adata.scai.fraunhofer.de/cohorts.

According to the MMSE scores, AD patients from AIBL (quantiles: 15, 20, 25), ANMerge (quantiles: 16, 21, 25), and NACC (quantiles:

[*] Accessed on February 2, 2020.

16, 21, 25) showed the worst cognitive performance. ADNI (quantiles: 21, 23, 25) contained patients with fewer cognitive symptoms. The CDR Dementia Staging Instrument (CDR) Sum of Boxes (CDR-SOB) scores slightly shift the perspective. Here, ANMerge is the most affected cohort, with its 25%, 50%, and 75% quantiles of the CDR-SOB scores being 4, 6, and 9, respectively. AIBL patients scored 3.5, 5, and 7, which slightly contradicts the image painted by the MMSE scores. Again, ADNI shows the least cognitive symptoms with its CDR-SOB quantiles being 3, 4.5, and 5.

A comparison of raw biomarker measurements between cohorts proved to be impossible, since encountered values are on different scales and may be subject to batch effects. Thus we analyzed how much measurements diverged from their respective control population in each cohort (Supplementary Text).

The prerequisite for comparative approaches involving biomarker measurements across data sets is an alignment of their underlying data models (ie, making data interoperable). In our analysis, we found that each study had defined its own data model, and variable names differed between them. This forced us to individually map variables to their corresponding counterparts in other studies to enable comparisons in the first place (eg, combine "lh_hippo_volume" and "rh_hippo_volume" and map to "Hippocampus"). Another difficulty is that numerous data sets reported values of equivalent variables in different ways. For example, CSF biomarker measurements are reported to be either normal (0) or abnormal (1) in NACC, whereas other studies provide numerical values that were capped at different thresholds between studies (eg, " >1700"). All these factors led to a severe lack of interoperability between data sets, which significantly limits comparative approaches and restricts them to more standardized variables like clinical assessment scores.

3.7 | Longitudinal follow-up

The majority of the investigated studies have collected longitudinal data in the form of repeated measurements. The intervals of data collection differed across studies (Table 2). Figure 3A displays the drop-out of study participants over time relative to the size of the cohort. In this analysis, participants were considered if at least one measurement was taken at the respective month. However, an individual's participation in some assessments does not imply that all biomarker values were acquired for the same individual on all visits. Thus we additionally investigated the amount of study participants for which select AD biomarkers were measured over time (Figure 3). Plots for all of the investigated biomarkers can be found at https://adata.scai.fraunhofer.de/follow-up.

One example biomarker that we selectively investigated is CSF amyloid beta for which Figure 3B displays the longitudinal coverage. Comparing Figure 3B with Figure 3A demonstrates that CSF samples were, if at all, taken only from a small fraction of participants consistently over time. Summed over all the investigated cohorts, only 273 participants (0.5%) have undergone CSF sampling at baseline and again 3 years after. In contrast to CSF, cognitive assessments follow the drop-out curves quite closely (Figure 3C). Although these findings are not surprising given the invasiveness of CSF sample collection, they raise severe concerns regarding the robustness of statistical analysis results obtained from CSF data. In turn, this again elucidates that comparative longitudinal approaches in the AD field are limited mainly to cognitive assessments or suffer from small sample size.

4 | DISCUSSION

In this work, we established an overview of the AD data landscape by investigating patient-level data from nine major clinical AD cohort studies. Our results demonstrate that the individual data sets vary with respect to key characteristics, such as number of enrolled participants per diagnosis, demographic composition, and distribution of important AD biomarkers. Assessing the ethnoracial diversity in the cohorts exposed a severe overrepresentation of White/Caucasian individuals compared to other ethnoracial backgrounds. To appraise the availability of modalities in each study, we categorized each modality based on the relative presence of data in each cohort. Another important remark of our findings is the limited number of longitudinal follow-up measurements for important AD biomarkers like CSF amyloid beta. Finally, we made all results explorable through ADataViewer, an interactive web application that can help researchers to identify cohort data sets that are suitable for their research.

4.1 | Achieving data set interoperability through one common data model

Our analysis exposed major challenges that severely impede comparative approaches on AD cohort data. Although there has been work on standardizing data collection[21,22] as well as on guidelines defining an AD-specific data model,[23] we still experience a deficit in interoperability across AD data sets. The investigated cohort data sets neither followed a common naming system for variables nor represented values of the same measurement in an equal manner. On top of that, some studies shared only processed values instead of the underlying raw data. This further impedes interoperability, since differences in applied processing pipelines inevitably introduce systematic biases into the data. One promising approach to increase data set interoperability could be a comprehensive, AD-specific common data model. Such a data model could support the alignment and mapping of variables by providing easy-to-follow guidelines and a dedicated interface for retrospective data harmonization.

4.2 | Data limitations hamper disease modeling

In the context of personalized medicine, training models on predominantly White/Caucasian participants can lead to biased models. It is known that exhibited patterns of biomarker measurements differ across AD patients from distinct ethnoracial groups.[25,26] Given that

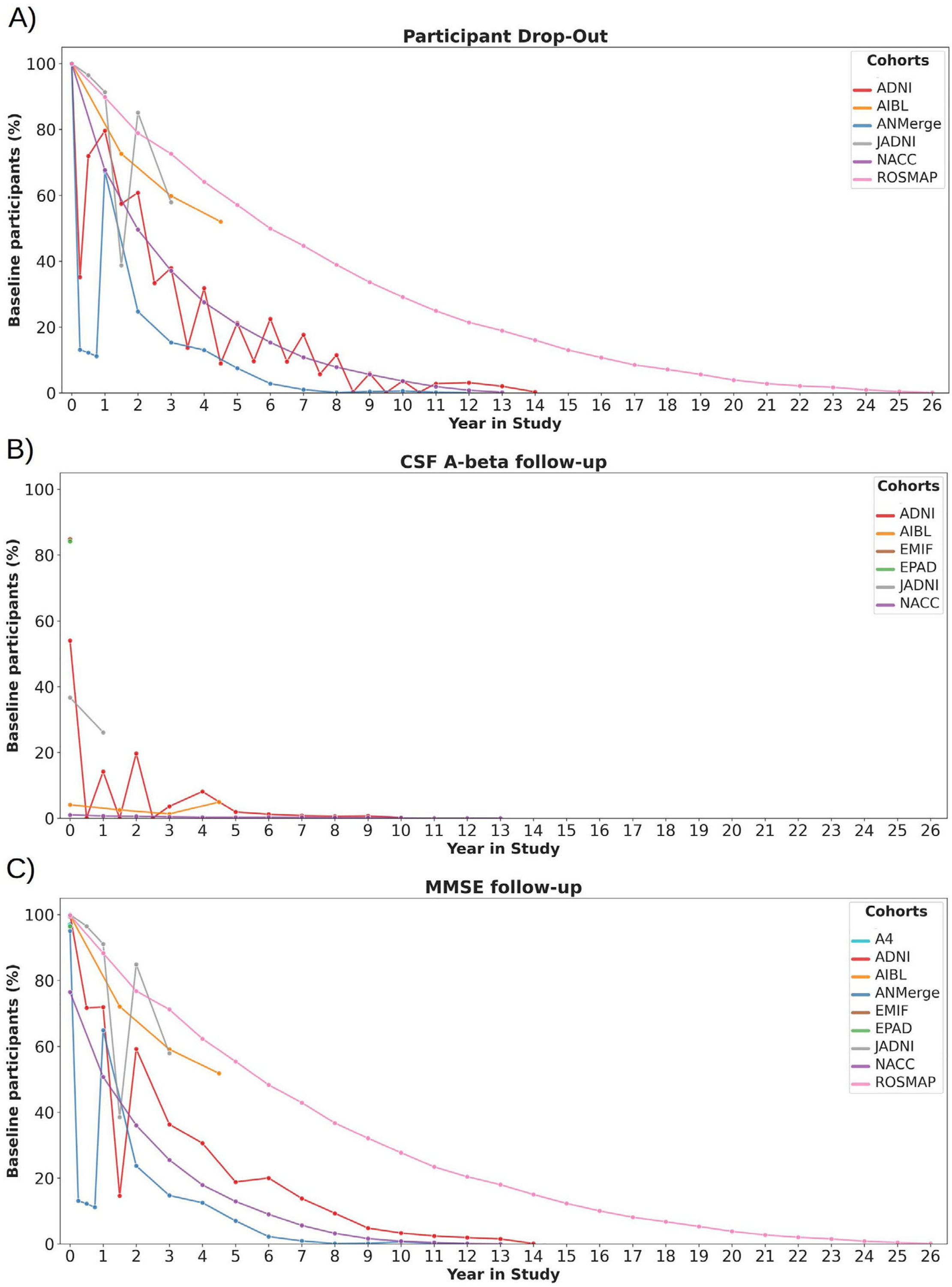


FIGURE 3 Longitudinal follow-up as the proportion of participants at study baseline (ie, participants were aligned based on their first visit). A, At least one variable measured. B, CSF amyloid beta. C, MMSE scores. CSF = cerebrospinal fluid. MMSE = Mini Mental State Examination

there are only limited data from non-White participants available, trained models could fail to learn such ethnoracial-specific signals, which, in turn, would result in poor performance for individuals of non-White background.

As mentioned previously, the abundance of longitudinal CSF data was limited throughout all acquired data sets. One possible reason explaining participants' reluctance to provide CSF samples, especially repeatedly, is the invasiveness of its sampling procedure.[24] Although cross-sectional CSF biomarkers can support AD diagnosis, longitudinal measurements are fundamental to understand disease progression on a biomarker-level. Given the low CSF sample sizes currently available, it remains questionable whether longitudinal analyses of these data can generate robust insights on conversions between normal and abnormal values of CSF biomarkers.

4.3 | Actionable knowledge through data-driven landscapes

The evident contradictions found between our data-driven investigation and the metadata-based approaches (Section 3.4) can be divided into two types. Type 1 describes that we found variables in the data sets that were reported as missing according to metadata resources. From this type of contradiction, we can conclude that approaches relying solely on metadata and literature potentially suffer in accuracy when estimating the real content available in cohort data sets. Contradiction type 2, on the other hand, resembles cases in which metadata sources reported a variable to be present, while we were not able to find it in the underlying data. Type 2 contradictions do not lead to the same conclusion as type 1, since it may be possible that the respective variables have simply not been shared with us. However, it is arguable how practical correct metadata is if the data it describes are not themselves available. We believe that our presented comparison highlights that, despite their significantly higher demand for time and effort, data-driven investigations should be preferred when assessing a data landscape.

4.4 | Future perspectives

The observed differences in demographic characteristics and disease risk factors across studies could severely hamper the comparison and validation of findings across disparate cohorts, since they can significantly influence the patterns and trends exhibited in the data.[2] Until now, only limited insight is available on how much the heterogeneous data landscape limits comparative approaches and cross-cohort disease modeling on AD data. Further systematic investigations are required to ensure that results generated on AD data sets are robust and reproducible across multiple cohorts. To support such endeavors, we aim to improve the ADataViewer to include more data sets, variable mappings, and the results of systematic data set comparisons in the future.

ACKNOWLEDGEMENTS

The authors want to thank Liu Shi and Alejo Nevado-Holgado for her help in accessing data sets.

The authors would also like to thank Lauren DeLong for her helpful comments.

We thank the study participants and staff of the Rush Alzheimer's Disease Center. ROSMAP was supported by NIA grants P30AG010161, R01AG015819, and R01AG017917.

The A4 Study is a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 Study is funded by a public-private-philanthropic partnership, including funding from the National Institutes of Health/National Institute on Aging, Eli Lilly and Company, the Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation, and additional private donors, with in-kind support from Avid and Cogstate. The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) study is funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN studies are led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School and Dr. Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI), University of Southern California. The A4 and LEARN Studies are coordinated by ATRI at the University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University of Southern California. The participants screening for the A4 Study provided permission to share their de-identified data to advance the quest to find a successful treatment for Alzheimer's disease. We would like to acknowledge the dedication of all the participants, the site personnel, and all of the partnership team members who continue to make the A4 and LEARN Studies possible. The complete A4 Study Team list is available at: a4study.org/a4-study-team.

J-ADNI was supported by the following grants: Translational Research Promotion Project from the New Energy and Industrial Technology Development Organization of Japan; Research on Dementia, Health Labor Sciences Research Grant; Life Science Database Integration Project of Japan Science and Technology Agency; Research Association of Biotechnology (contributed by Astellas Pharma Inc., Bristol-Myers Squibb, Daiichi-Sankyo, Eisai, Eli Lilly and Company, Merck-Banyu, Mitsubishi Tanabe Pharma, Pfizer Inc., Shionogi & Co., Ltd., Sumitomo Dainippon, and Takeda Pharmaceutical Company), Japan, and a grant from an anonymous foundation.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of

Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The NACC database is funded by the NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EPAD grant agreement no. 117536, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826421, "TheVirtualBrain-Cloud."

Open access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors have nothing to declare.

REFERENCES

1. Kalra D. The importance of real-world data to precision medicine. *Per Med*. 2019;16(2):79-82.
2. Birkenbihl C, Emon MA, Vrooman H, et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA J*. 2020;11(3):367-376.
3. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16(1):150.
4. Whitwell JL, Wiste HJ, Weigand SD, et al. Comparison of imaging biomarkers in the Alzheimer disease neuroimaging initiative and the Mayo Clinic Study of Aging. *Arch Neurol*. 2012;69(5):614-622.
5. Ferreira D, Hansson O, Barroso J, et al. The interactive effect of demographic and clinical factors on hippocampal volume: a multicohort study on 1958 cognitively normal individuals. *Hippocampus*. 2017;27(6):653-667.
6. Oliveira JL, Trifan A, Silva LAB. EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data. *Int J Med Inf*. 2019;126:35-45.
7. Janssen O, Vos SJ, García-Negredo G, et al. Real-world evidence in Alzheimer's disease: the ROADMAP Data Cube. *Alzheimers Dement (N Y)*. 2020;16(3):461-471.
8. Lawrence E, Vegvari C, Ower A, Hadjichrysanthou C, De Wolf F, Anderson RM. A Systematic review of longitudinal studies which measure alzheimer's disease biomarkers. *J Alzheimers Dis*. 2017;59(4):1359-1379.
9. Sperling RA, Rentz DM, Johnson KA, et al. The A4 study: stopping AD before symptoms begin?. *Sci Transl Med*. 2014;6(228):228fs13-228fs13.
10. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement (N Y)*. 2005;1(1):55-66.
11. Birkenbihl C, Westwood S, Shi L, et al. ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *medRxiv*:2020.
12. Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr*. 2009;21(4):672-687.
13. Bos I, Vos S, Vandenberghe R, et al. The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. *Alzheimers Res Ther*. 2018;10(1):64.
14. Solomon A, Kivipelto M, Molinuevo JL, Tom B, Ritchie CW. European prevention of Alzheimer's dementia longitudinal cohort study (EPAD LCS): study protocol. *BMJ Open*. 2018;8(12):e021017.
15. Iwatsubo T, Iwata A, Suzuki K, et al. Japanese and North American Alzheimer's Disease Neuroimaging Initiative studies: harmonization for international trials. *Alzheimers Dement (N Y)*. 2018;14(8):1077-1087.
16. Besser L, Kukull W, Knopman DS, et al. Version 3 of the National Alzheimer's coordinating center's uniform data set. *Alzheimer Dis Assoc Disord*. 2018;32(4):351.
17. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimers Dis*. 2018;64(s1):S161-S189.
18. Hye A, Lynham S, Thambisetty M, et al. Proteome-based plasma biomarkers for Alzheimer's disease. *Brain*. 2006;129(11):3042-3050.
19. Babulal GM, Quiroz YT, Albensi BC, et al. Perspectives on ethnic and racial disparities in Alzheimer's disease and related dementias: update and areas of immediate need. *Alzheimers Dement (N Y)*. 2019;15(2):292-312.
20. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group: under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*. 1984;34(7):939-939.
21. O'Bryant SE, Gupta V, Henriksen K, et al. Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. *Alzheimers Dement (N Y)*. 2015;11(5):549-560.
22. Weiner MW, Veitch DP, Aisen PS, et al. Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimers Dement (N Y)*. 2015;11(7):865-884.
23. Neville J, Kopko S, Romero K, et al. Accelerating drug development for Alzheimer's disease through the use of data standards. *Alzheimers Dement*. 2017;3(2):273-283.
24. Sand T, Stovner LJ, Dale L, Salvesen R. Side effects after diagnostic lumbar puncture and lumbar iohexol myelography. *Neuroradiology*. 1987;29(4):385-388.
25. Misiura MB, Howell JC, Wu J, et al. Race modifies default mode connectivity in Alzheimer's disease. *Transl Neurodegener*. 2020;9:8.
26. Howell JC, Watts KD, Parker MW, et al. Race modifies the relationship between cognition and Alzheimer's disease cerebrospinal

fluid biomarkers. *Alz Res Therapy*. 2017;9:88. https://doi.org/10.1186/s13195-017-0315-1.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Birkenbihl C, Salimi Y, Domingo-Fernándéz D, et al. Evaluating the Alzheimer's disease data landscape. *Alzheimer's Dement*. 2020;6:e12102. https://doi.org/10.1002/trc2.12102

A.2 ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets

Reprinted with permission from "Salimi, Y., Domingo-Fernández, D., Bobis-Álvarez, C., Hofmann-Apitius, M., and **Birkenbihl, C.**, for the Alzheimer's Disease Neuroimaging Initiative, the Japanese Alzheimer's Disease Neuroimaging Initiative, for the Aging Brain: Vasculature, Ischemia, and Behavior Study, the Alzheimer's Disease Repository Without Borders Investigators, for the European Prevention of Alzheimer's Disease (EPAD) Consortium. (2022). ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets. *Alzheimer's Research & Therapy*, 14(1), 69.".

Alzheimer's Research & Therapy

RESEARCH Open Access

ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets

Yasamin Salimi[1,2*], Daniel Domingo-Fernández[1], Carlos Bobis-Álvarez[3], Martin Hofmann-Apitius[1,2], Colin Birkenbihl[1,2] and for the Alzheimer's Disease Neuroimaging Initiative, the Japanese Alzheimer's Disease Neuroimaging Initiative, for the Aging Brain: Vasculature, Ischemia, and Behavior Study, the Alzheimer's Disease Repository Without Borders Investigators, for the European Prevention of Alzheimer's Disease (EPAD) Consortium

Abstract

Background: Currently, Alzheimer's disease (AD) cohort datasets are difficult to find and lack across-cohort interoperability, and the actual content of publicly available datasets often only becomes clear to third-party researchers once data access has been granted. These aspects severely hinder the advancement of AD research through emerging data-driven approaches such as machine learning and artificial intelligence and bias current data-driven findings towards the few commonly used, well-explored AD cohorts. To achieve robust and generalizable results, validation across multiple datasets is crucial.

Methods: We accessed and systematically investigated the content of 20 major AD cohort datasets at the data level. Both, a medical professional and a data specialist, manually curated and semantically harmonized the acquired datasets. Finally, we developed a platform that displays vital information about the available datasets.

Results: Here, we present ADataViewer, an interactive platform that facilitates the exploration of 20 cohort datasets with respect to longitudinal follow-up, demographics, ethnoracial diversity, measured modalities, and statistical properties of individual variables. It allows researchers to quickly identify AD cohorts that meet user-specified requirements for discovery and validation studies regarding available variables, sample sizes, and longitudinal follow-up. Additionally, we publish the underlying variable mapping catalog that harmonizes 1196 unique variables across the 20 cohorts and paves the way for interoperable AD datasets.

Conclusions: In conclusion, ADataViewer facilitates fast, robust data-driven research by transparently displaying cohort dataset content and supporting researchers in selecting datasets that are suited for their envisioned study. The platform is available at https://adata.scai.fraunhofer.de/.

Keywords: Alzheimer's disease, Dementia, Data harmonization, Semantic mapping, MRI, Variable catalog, Interoperability, Data curation, Cohort study

Background

Alzheimer's disease (AD) and dementia research has progressed considerably thanks to the increased availability of patient-level cohort datasets [1]. Cohort data have, among others, laid the foundation to discover novel biomarkers [2], investigate disease progression [3], and identify disease subtypes [4]. To ensure the robustness and

*Correspondence: yasamin.salimi@scai.fraunhofer.de

[2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany
Full list of author information is available at the end of the article

reproducibility of results achieved in such data-driven analyses, they must be externally validated in independent cohort datasets [5]. Working across multiple cohort datasets is, however, impeded by several profound challenges. The first challenge manifests in the access to further validation cohort datasets, as third-party researchers have to go through time-intensive application processes that often span several weeks before they can actually start getting familiar with the acquired data. Secondly, once access is granted, the validation datasets have to be comparable to the original discovery dataset concerning their assessed variables [6]. This means that (1) a largely overlapping set of variables should have been measured in both cohorts and (2) these variables need to be harmonized across the independent cohort datasets, which is rarely the case by default. Identifying and semantically harmonizing equivalent variables in distinct datasets is an arduous task given that datasets typically employ their own variable naming system [7]. While theoretical guidelines for AD data harmonization have been previously proposed [8], as of now and to the best of our knowledge, no comprehensive mapping catalog is available to the AD research community that would help to unify the variable names across existing cohorts.

Across-cohort interoperability, however, goes beyond the semantic layer as statistical distributions of equivalent variables might differ among cohorts [9]. Our recent study revealed that such systematic statistical differences can bias results of data-driven analyses based on cohort data [10]. However, in practice, researchers only see the factual content of a shared dataset after data download occurred and data investigation started. At this stage, the realization of, for example, incompatible discovery and validation datasets can render the process of data access and exploration a waste of time as the lacking data interoperability would render the envisioned analysis infeasible.

Several funding bodies, for example, the Innovative Medicine Initiative (IMI) or the Alzheimer's Disease Data Initiative (ADDI), have launched large projects to address data problems in the AD domain, for example, the European Medical Information Framework (EMIF) [11], ROADMAP [12], or the ADDI Workbench, and new calls were issued in this direction. In fact, both EMIF and ROADMAP have built information sources on cohort datasets that were assembled from the respective cohorts' self-reported metadata [13, 14]. However, in a recent study, we observed that the information gained through such metadata-driven cohort assessments differs from the content that is factually shared with researchers after successful access applications [15].

In this work, we present ADataViewer, an interactive tool that enables the scientific community to explore 20 AD cohort datasets, both from a semantic and statistical perspective. To establish semantic interoperability across these datasets, we created a variable mapping catalog that harmonizes 1196 unique variables encountered in the datasets, spanning nine data modalities. Leveraging these semantically harmonized versions of the datasets, we developed tools and interfaces that facilitate the exploration of the cohort datasets with respect to longitudinal follow-up, demographics, ethnoracial diversity, measured modalities, and individual variables. Finally, we present ADataViewers' "StudyPicker," a tool that assists researchers in identifying cohort datasets suited for their envisioned analysis.

Methods

Harmonizing variables across cohorts

Semantic harmonization of the datasets was achieved through meticulous manual curation. Two curators systematically investigated variable names, metadata describing the variable content, and the values stored in the respective data tables across each dataset to gain robust mappings between equivalent variables. We opted for a multidisciplinary curation team to combine the complementary strengths of a curator from a medical background with those of a second curator leveraging a data-driven perspective. In the first step, the curators categorized the variables of each dataset according to a set of modalities (e.g., magnetic resonance imaging (MRI), demographics, and genotyping). To facilitate the curation process, mappings were proposed to the curators based on variable name similarity in modalities where the number of features was abundant. For the majority of modalities, we mapped approximately between 10 to 30 variables, with the exception being the MRI modality which comprised more than 1000 variables, as it contained a vast selection of brain region-specific measures derived from the raw images (e.g., volumes or thickness). No specific data model (e.g., FHIR or OMOP) was used. For more detailed curation guidelines, we refer to the Supplementary Material. Whenever possible, variables found in the investigated AD datasets were additionally mapped to ontologies that provided respective semantic context. Further details on the used ontologies and the process of mapping variable names to ontologies are described in the Supplementary Material.

Data access and data privacy

ADataViewer does not store or enable the download of any cohort data itself. All displayed plots and provided exploration tools are fully anonymized and no participant identifying information is disclosed nor stored in the underlying database, not even the original study internal patient identifiers. Shown statistical plots are solely based on summary statistics or univariate analyses that cannot be linked to other variables or personal information. To facilitate access to the datasets, we provide links that lead researchers to the original data portals through which the respective cohorts are distributed.

Results

ADataViewer is an interactive platform that enables the detailed exploration of, at the time of publication, 20 major cohort datasets from the AD domain. Its goal is to provide an overview across their content from a predominantly data-driven perspective. Each section of ADataViewer focuses on distinct aspects of the investigated datasets. The "Modality" section provides an overview of the data modalities collected in each cohort (e.g., magnetic resonance imaging (MRI), autopsy, and genotype data). The "Ethnicity" page displays the ethnoracial diversity in each cohort study as well as aggregated plots over specific geographic regions. In the "Longitudinal" section, the frequency and abundance of follow-up assessments are presented both per cohort and variable. The "Biomarkers" section allows the visualization of variable distributions and their comparison across cohorts. The semantic mappings between cohort name spaces are covered in the "Mappings" section. Finally, the "StudyPicker" leverages on all of these sections to guide researchers to the cohort datasets which provide the best basis for their planned analyses.

Instead of relying solely on study protocols and reported metadata, we based all our investigations on the data that were factually shared by the respective data owners. To transparently mirror the state of the dataset to which researchers will gain access after successful application, we refrained from any extensive data processing (e.g., transforming numerical ranges and value representations). As such, any inconsistencies in the datasets (e.g., extreme outliers) will be accordingly displayed in ADataViewers' tools and visualizations. Consequently, this allows researchers to comprehensively evaluate the data that will actually be available for analysis.

Accessed AD cohort datasets

To enable a comprehensive exploration of the available AD data, it was vital to identify, access, and curate as many cohort-level datasets as possible. Therefore, we systematically scanned data repositories and scientific publications, leading to the identification of 24 cohorts of which most claimed to follow the open science paradigm and share their data with third-party researchers. After applying for access to the corresponding data owners, we acquired 20 of those datasets over the course of 3 years (information on why the four remaining datasets were not accessed is provided in the Supplementary Material). These datasets originated from a heterogeneous pool of studies that followed a variety of different goals ranging from purely observational cohort studies over memory clinic data collections to dedicated clinical trials. Concordantly, the employed participant recruitment procedures, inclusion and exclusion criteria, and measured data modalities varied among them. More information about the collected datasets, their content, and original study aim is given in Table 1; for further study-specific details, we refer to the original publications.

Semantic harmonization of the accessed cohort datasets

To build ADataViewer, we mapped 1196 unique terms across the investigated datasets corresponding to variables from nine different data modalities (Fig. 1). Table 2 shows the total number of mapped terms per modality and cohort. Furthermore, to connect the variables of the cohort datasets to clearly defined semantic concepts, we additionally mapped them to standardized ontologies. In total, 241 concepts from seven distinct referential ontologies were used in this process (more details in the Supplements). All mappings can be explored through interactive visualizations and tables at https://adata.scai.fraunhofer.de/mappings. The genotype and omics modalities of datasets were not mapped as they are already precisely defined by genetic database identifiers (e.g., rsID's or UniProt identifiers) and their corresponding reference genome. A prerequisite for mapping the variables was that they were at least present in two independent cohorts.

The StudyPicker: variable-based selection of cohort datasets

The StudyPicker is a tool that supports researchers in finding datasets based on the requirements of their envisioned analysis (https://adata.scai.fraunhofer.de/study_picker). It takes a collection of variable names as input and ranks the cohorts in ADataViewer based on the availability of these specified variables (Fig. 4A). The generated ranking shows the availability of the variables and the number of participants per cohort for whom these variables have been assessed at the study baseline, as well as their longitudinal coverage (i.e., assessment frequency and the number of participants assessed per visit) (Fig. 4B). Additionally, links are provided that guide interested researchers directly to the data access applications of the respective datasets. The StudyPicker is particularly helpful for hypothesis-driven research or validation studies in which the variables that are elementary to conduct the planned analysis are often known in advance.

Table 1 AD cohorts available for exploration using ADataViewer

Cohort	Consortium	Patients at baseline	Modalities	Longitudinal (yes/no)	Study type
A4 [16]	Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease	6945	7	No[a]	Clinical trial
ABVIB [17]	Aging Brain: Vasculature, Ischemia, and Behavior	280	2	Yes	Observational study
ADNI [18]	The Alzheimer's Disease Neuroimaging Initiative	2249	12	Yes	Observational study
AIBL [19]	The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing	1378	9	Yes	Observational study
ANMerge [20]	AddNeuroMed	1703	10	Yes	Observational study
ARWIBO [21]	Alzheimer's Disease Repository Without Borders	2617	10	Yes	Observational study
DOD-ADNI [22]	Effects of TBI & PTSD on Alzheimer's Disease in Vietnam Vets	458	11	Yes	Observational study
EDSD [23]	The European DTI Study on Dementia	474	7	No	Observational study
EMIF-1000 [24]	European Medical Information Framework	1199	10	No	Meta-cohort
EPAD V.IMI [25]	European Prevention of Alzheimer's Dementia	2096	9	Yes	Observational study
I-ADNI [26]	The Italian Alzheimer's Disease Neuroimaging Initiative	262	5	No	Observational study
JADNI [27]	Japanese Alzheimer's Disease Neuroimaging Initiative	567	9	Yes	Observational study
NACC [28]	The National Alzheimer's Coordinating Center	40,948	11	Yes	Memory clinic database
OASIS-1 [29] and OASIS-2 [30]	Open Access Series of Imaging Studies	564	3	Yes	Observational study
PREVENT-AD [31]	Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease	348	8	Yes	Clinical trial
PharmaCog [32]	Prediction of Cognitive Properties of New Drug Candidates for Neurodegenerative Diseases in Early Clinical Development	147	6	Yes	Observational study
ROSMAP [33]	The Religious Orders Study and Memory and Aging Project	3626	7	Yes	Observational study
VASCULAR [34]	The Vascular Contributors to Prodromal Alzheimer's disease	250	8	No	Non-interventional cohort study
VITA [35]	Vienna Transdanube Aging	606	5	Yes	Observational study
WMH-AD [36]	White Matter Hyperintensities in Alzheimer's Disease	90	5	No	Observational study

A complete overview about the collected data modalities can be found under https://adata.scai.fraunhofer.de/modality

[a] Follow-up assessments were planned for A4 but no according data was released at the time of this publication

Detailed exploration of dataset content through interactive visualizations

Next to the semantic perspective, ADataViewer also allows for a detailed exploration of the integrated datasets based on descriptive statistics. Statistical distributions of numerical and categorical variables of interest can be visualized and compared across the available cohorts (https://adata.scai.fraunhofer.de/biomarkers). This functionality enables comparisons between individual diagnosis groups (i.e., cognitively unimpaired (CU), mild cognitive impairment (MCI), AD) as well as the complete cohorts. Using these visualizations, researchers can investigate distributions and value representations encountered in the datasets and identify possible differences among them before starting their analysis.

A longitudinal view of the data can be generated in the "Longitudinal" section. Dedicated visualizations display the follow-up per cohort on a variable level (Fig. 2).

Meta-analysis of cohort study content, assessed variables, and common modalities

Besides the exploration and comparison of specific cohorts, ADataViewer helps to get a comprehensive

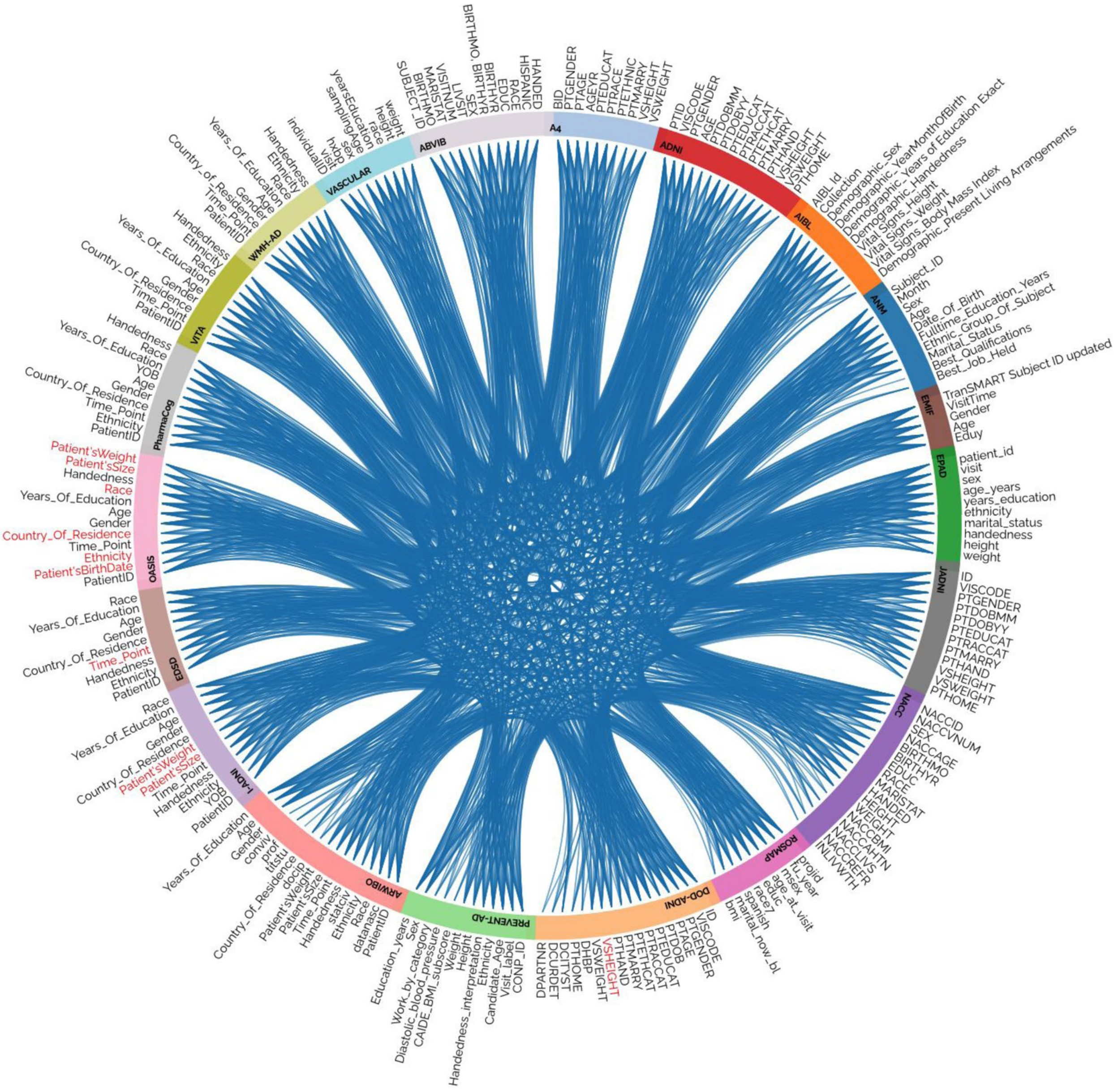


Fig. 1 Mapping of demographic variables across the 20 cohorts. Red labels indicate variables mentioned in the metadata which consisted purely of missing data in the shared dataset. The corresponding plot for each modality as well as the underlying mapping tables for data harmonization are available at https://adata.scai.fraunhofer.de/mappings.

overview of the state of the data landscape formed by the underlying cohorts. Here, the modality map (https://adata.scai.fraunhofer.de/modality) displays how commonly specific data modalities were included in cohort studies and, simultaneously, highlights areas that currently remain underexplored. Along the same line, Fig. 3 shows an excerpt from an interactive visualization that depicts how many studies measured each individual variable. Furthermore, the plots displaying the ethnoracial diversity encountered in each individual cohort, and across cohorts grouped by geographic location, reveal over- and underrepresentation of ethnoracial groups in data-driven AD research. All of this information can be vital when designing a novel cohort study aiming either for compatibility to other studies or at illuminating blind spots previously underrepresented in the AD data landscape.

Table 2 Number of mapped unique variables per cohort and modality

Dataset	Demographics	Clinical	MRI	PET	CSF	Plasma	Comorbidities	Family	Lifestyle
A4	13	5	44	1	0	0	2	6	4
ABVIB	12	0	0	0	0	0	0	0	0
ADNI	17	23	247	3	10	11	14	8	5
AIBL	15	16	3	2	3	0	12	2	5
ANMerge	14	11	136	0	0	0	1	3	1
ARWIBO	21	14	1026	21	3	6	13	3	2
DOD-ADNI	21	20	249	1	3	0	18	6	6
EDSD	12	8	1026	8	3	2	4	2	0
EMIF-1000	8	4	3	1	6	0	3	0	4
EPAD V.IMI	14	11	80	0	3	0	17	5	4
I-ADNI	15	10	1026	8	3	1	1	2	0
JADNI	15	21	871	2	3	0	14	6	4
NACC	20	17	123	2	3	0	14	3	6
OASIS	16	3	1026	8	3	2	0	2	0
PREVENT-AD	15	4	0	0	7	0	5	5	0
PharmaCog	13	16	1026	8	3	2	0	2	0
ROSMAP	12	9	0	0	0	0	8	0	1
VASCULAR	9	8	31	0	0	0	3	0	2
VITA	12	3	1026	8	3	2	0	2	0
WMH-AD	12	4	1025	8	3	2	0	2	0
Total unique terms	**23**	**34**	**1050**	**24**	**14**	**15**	**20**	**9**	**7**

Exemplary application scenarios employing ADataViewer

While there are multiple scenarios in which ADataViewer can support AD research, we focus on two scenarios below. Another application scenario not explained here, however, one that would follow similar routes as the ones outlined below, would be the writing of grant applications and identifying datasets to include into the proposal.

Scenario 1

A researcher is searching for a discovery and validation cohort to model cognitive decline in the light of hippocampus atrophy, amyloid PET, and depression. The variables of interest are the Mini-Mental State Examination (MMSE), Clinical Dementia Rating Sum of Boxes (CDRSB), hippocampus volume, Amyvid Positron Emission Tomography (AV PET), Geriatric Depression Scale

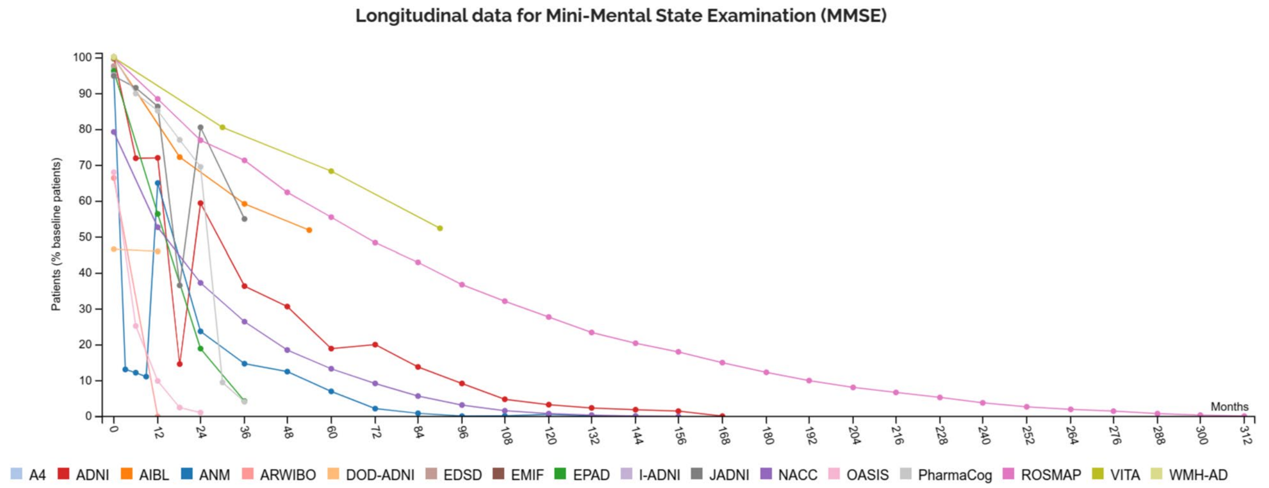


Fig. 2 Exemplary longitudinal plot of MMSE assessments generated using ADataViewer. Displayed are cohorts and their respective number of assessed participants for the selected variable

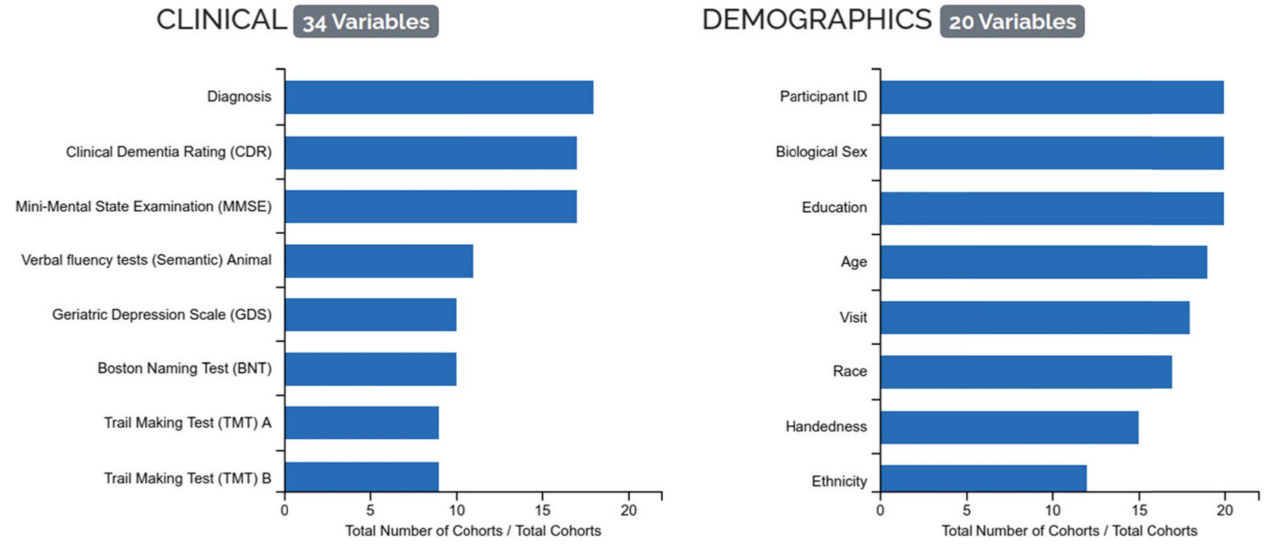


Fig. 3 Assessment frequency of exemplary variables across cohorts. Interactive figure displaying the number of studies in which each specific variable was encountered (https://adata.scai.fraunhofer.de/biomarkers)

(GDS), and variables to correct for possible confounding (age, biological sex, education, and APOE ε4 allele presence).

Given such a set of variables of interest, the StudyPicker of ADataViewer is the appropriate starting point to identify relevant cohorts. After submitting the variable query, we can directly observe that NACC, A4, ADNI, and DOD-ADNI contain all specified variables of interest (Fig. 4A). However, after inspecting the follow-up plots, it is revealed that only NACC and ADNI hold sufficient longitudinal data to detect time-dependent relationships (here, 463 and 557 patients over 24 months of study runtime, respectively) (Fig. 4B and Fig. S1). Besides these two cohorts, EPAD, including 1845 participants, could also provide a rich basis for the planned analysis if AV PET would be omitted (Fig. 4A).

For a final evaluation on whether NACC and ADNI would suit the study needs, the "Biomarkers" section can be used to compare cohort demographics and variable distributions. For example, comparing the age of participants in NACC and ADNI reveals a higher variance in the NACC data and the presence of younger participants who would have been excluded from the ADNI study (Fig. 4C). Furthermore, investigating the hippocampal volumes exposes a difference in value representation between the cohorts, as NACC values have been reported as normalized values (Fig. S2). Consequently, it could be concluded that both datasets could be viable options for the discovery and replication process of a data-driven study, given that the representations of the hippocampal volume can be unified. Finally, the application process for data access can be initiated directly through the StudyPicker.

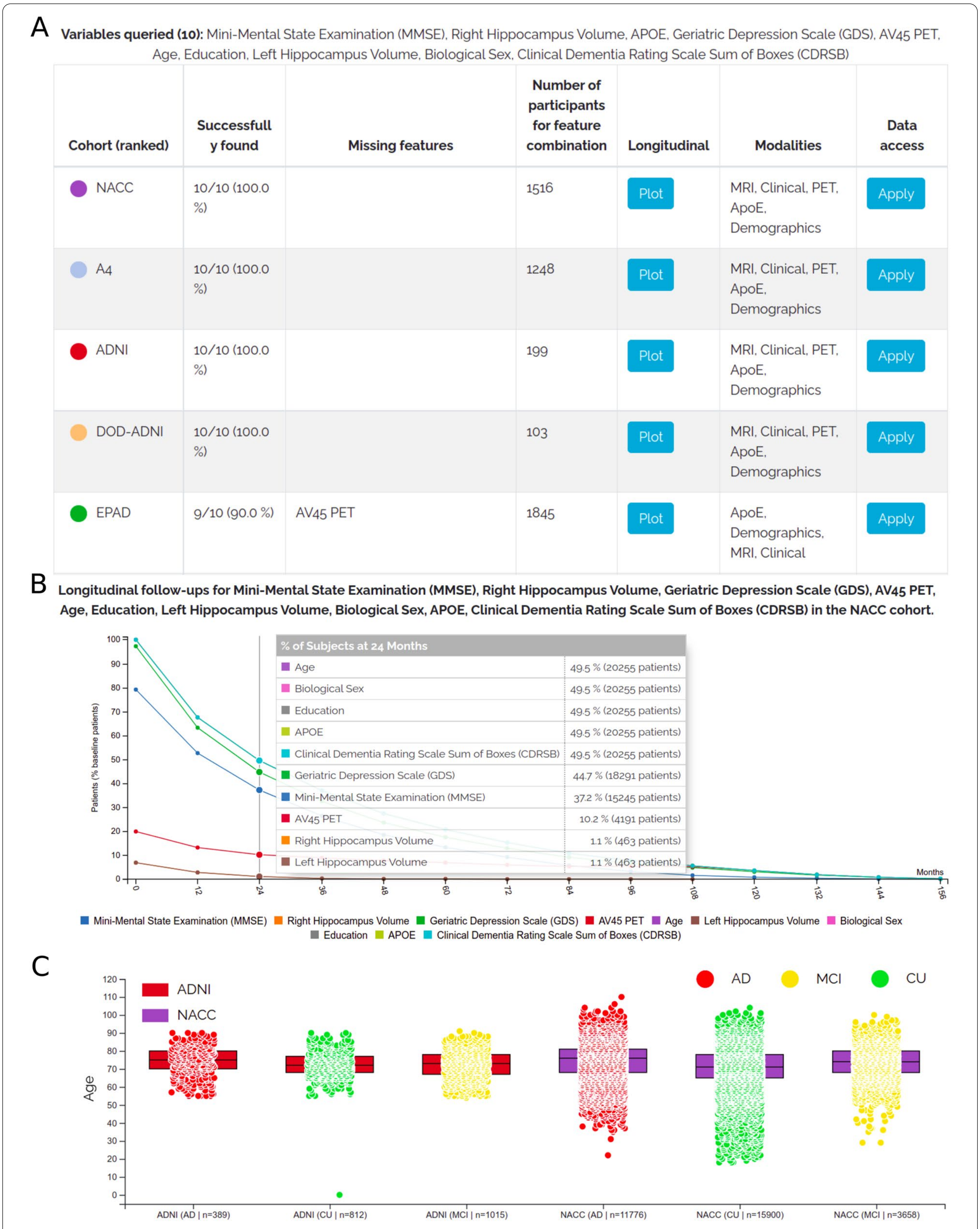
Scenario 2

A consortium is planning to conduct a longitudinal cohort study that aims at investigating AD in previously underrepresented ethnoracial groups. The assessed variables, however, should be compatible with other landmark AD cohorts to allow for a comparison of achieved results.

First, the ethnoracial diversity encountered across previous AD cohorts can be explored in the "Ethnicity" section of ADataViewer. Their investigation demonstrates that 19 of the 20 cohorts enrolled predominantly caucasian/white participants. Keeping our proposed study goals in mind, it would therefore make sense to exclude caucasian/white participants from the recruitment of the

(See figure on next page.)

Fig. 4 Using ADataViewer to identify suitable cohort datasets in a use case scenario. Selection of this case scenario was with the aim to evaluate cognitive decline in the light of depression, AV PET, and hippocampal atrophy. All graphs were created using the tools of ADataViewer. **A** Excerpt of the ranking received by entering the variables of interest specified in application scenario 1 into the StudyPicker. **B** Longitudinal coverage of the specified variables in the NACC cohort. See Fig. S1 for the other cohorts' plots. **C** Comparison of the age distributions encountered across diagnostic groups of ADNI and NACC



A

Variables queried (10): Mini-Mental State Examination (MMSE), Right Hippocampus Volume, APOE, Geriatric Depression Scale (GDS), AV45 PET, Age, Education, Left Hippocampus Volume, Biological Sex, Clinical Dementia Rating Scale Sum of Boxes (CDRSB)

Cohort (ranked)	Successfully found	Missing features	Number of participants for feature combination	Longitudinal	Modalities	Data access
● NACC	10/10 (100.0 %)		1516	Plot	MRI, Clinical, PET, ApoE, Demographics	Apply
● A4	10/10 (100.0 %)		1248	Plot	MRI, Clinical, PET, ApoE, Demographics	Apply
● ADNI	10/10 (100.0 %)		199	Plot	MRI, Clinical, PET, ApoE, Demographics	Apply
● DOD-ADNI	10/10 (100.0 %)		103	Plot	MRI, Clinical, PET, ApoE, Demographics	Apply
● EPAD	9/10 (90.0 %)	AV45 PET	1845	Plot	ApoE, Demographics, MRI, Clinical	Apply

B

Longitudinal follow-ups for Mini-Mental State Examination (MMSE), Right Hippocampus Volume, Geriatric Depression Scale (GDS), AV45 PET, Age, Education, Left Hippocampus Volume, Biological Sex, APOE, Clinical Dementia Rating Scale Sum of Boxes (CDRSB) in the NACC cohort.

% of Subjects at 24 Months	
Age	49.5 % (20255 patients)
Biological Sex	49.5 % (20255 patients)
Education	49.5 % (20255 patients)
APOE	49.5 % (20255 patients)
Clinical Dementia Rating Scale Sum of Boxes (CDRSB)	49.5 % (20255 patients)
Geriatric Depression Scale (GDS)	44.7 % (18291 patients)
Mini-Mental State Examination (MMSE)	37.2 % (15245 patients)
AV45 PET	10.2 % (4191 patients)
Right Hippocampus Volume	1.1 % (463 patients)
Left Hippocampus Volume	1.1 % (463 patients)

Fig. 4 (See legend on previous page.)

envisioned study to focus on the currently underrepresented groups.

To achieve high compatibility with previous AD studies, the planned study should align its follow-up intervals and the assessed variables/data modalities to them. Here, the data modality map indicates that we should include demographics, clinical assessments, MRI, cerebrospinal fluid (CSF) biomarkers, at least APOE genotyping, administered medication, comorbidities, and the family history of participants to achieve a strong overlap in data modalities (Fig. S3). More specifically, the most prominently assessed variables per modality can be explored in the "Biomarkers" section (Fig. 3). For example, we can observe that Clinical Dementia Rating (CDR) and MMSE are the most conducted cognitive assessments; demographics most commonly cover the biological sex, age, years of education, and ethnoracial group of participants; and phosphorylated tau, total tau, and beta-amyloid were abundantly measured as CSF markers. By leveraging this information, we can make an informed decision on the variables we want to measure in the envisioned cohort study, such that an exploration of AD progression is feasible and that possible differences to cohorts of other ethnoracial compositions can be systematically evaluated. Additionally, the value ranges commonly encountered per variable can be explored using the biomarker boxplots (Fig. 4C). Once the cohort study was conducted, we can use the provided variable mapping catalog to harmonize the new cohort dataset to all 20 datasets currently present in ADataViewer.

Discussion

ADataViewer aims at advancing patient data-driven AD research by increasing the findability and interoperability of cohort datasets and providing a deeper understanding of their content, both from a semantic and statistical perspective. The platform supports the variable-level exploration of 20 AD cohort datasets and enables researchers to identify datasets suited for their envisioned studies before spending time on data access applications. In this context, we created, to the best of our knowledge, the most comprehensive variable mapping catalog in the AD domain that semantically harmonizes 1196 unique variables across all investigated cohorts.

Aspiring to contribute to a FAIR data paradigm (findable, accessible, interoperable, reusable) in AD research [37], ADataViewer increases the findability of AD cohort datasets by displaying and suggesting possible data resources to researchers, enables better accessibility through direct links to the respective data access points, provides the variable mapping catalog to establish data interoperability, and facilitates the reuse of data for validation purposes. We believe that the presented platform can elevate data-driven AD research to be faster and more robust, because it becomes significantly easier to access the right datasets and validate results across multiple independent cohorts. In turn, this will help to better understand the heterogeneity across AD patients [38] and help to reveal possible cohort-specific findings [10].

Collecting patient-level data is a vastly expensive process. Therefore, studies are often limited concerning their sample size, follow-up time, and variety of assessed data modalities. ADataViewer transparently provides researchers with information about what they can expect from specific datasets and whether it makes sense for them to spend a substantial amount of time on the acquisition of the individual data resource. Limiting the time spent on unfruitful dataset acquisitions will accelerate and benefit the actual analysis of the data. On this note, we would like to emphasize that ADataViewer is not meant to promote only the largest, most complete cohorts, but to show all available datasets that contain the information of interest for a conceived project. While larger cohorts often fare better as discovery cohorts, any cohort with equivalent information, regardless of the size, could present a valuable resource for the subsequent validation of results and should therefore be considered.

Given the restrictions of sensible personal data, there are multiple initiatives testing and establishing federated learning concepts that aim to facilitate secure remote access to multiple sensible datasets [39]. These concepts rely on interoperable data and our mappings and data descriptions could provide a starting point to establish such comprehensive interoperability by extending them into a complete data model following, for example, the OMOP or FHIR standard.

We plan to update ADataViewer as well as its underlying information (e.g., the mappings) whenever we get access to new datasets. However, an automatic periodic updating is infeasible, as the data is usually not shared via programmatic interfaces but through personal contacts and access-restricted data portals.

Limitations

One strength and simultaneous limitation of this work was its overarching premise that the data investigation was not based purely on descriptive metadata but on the dataset that was factually shared with us. Therefore, all results are based on the status of the distributed data and could vary from the content mentioned in official study reports or other versions of the same dataset. Ultimately, however, what drives the advancement of AD research is the factually shared, analyzable data and not what could potentially be available in theory.

The decision on how strict equivalence of variables is defined inevitably remains arbitrary to some degree. Here, we define two variables as semantically equivalent if the same information is presented in principle (i.e., the content of both variables can at least be broken down into the same information, see Supplementary Material for examples). Therefore, the acquisition method (e.g., type of MRI scanner) between two variables that were declared to be semantically equivalent may still differ and subsequent pre-processing of the raw data might be necessary to account for resulting statistical differences (e.g., elimination of batch effects). Sharing statistically harmonized data via ADataViewer is infeasible due to legal data sharing restrictions. However, the presented semantic mapping catalog presents a starting point to directly identify equivalent variables of interest and initiate the following pre-processing steps.

Conclusion

With ADataViewer, we aim to contribute to a robust, data-driven research culture that carefully reproduces and validates scientific results across multiple comparable datasets. As such, instead of pointing towards a single data resource, ADataViewer transparently displays the content of all integrated AD cohort datasets and the StudyPicker proposes all of these resources that match the researcher's requirements. Our provided variable mappings build the basis for in-depth dataset comparisons and can act as a starting point to select and harmonize suited discovery and validation datasets.

Abbreviations

A4: Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease; ABVIB: Aging Brain Vasculature, Ischemia, and Behavior; AD: Alzheimer's Disease; ADDI: Alzheimer's Disease Data Initiative; ADNI: Alzheimer's Disease Neuroimaging Initiative; AIBL: Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge: AddNeuroMed; ARWIBO: Alzheimer's Disease Repository Without Borders; AV PET: Amyvid Positron Emission Tomography; CDR: Clinical Dementia Rating; CDRSB: Clinical Dementia Rating Sum of Boxes; CSF: Cerebrospinal Fluid; CU: Cognitively Unimpaired; DOD-ADNI: Effects of TBI & PTSD on Alzheimer's Disease in Vietnam Vets; EDSD: European DTI Study on Dementia; EMIF: European Medical Information Framework; EPAD: European Prevention of Alzheimer's Dementia; GDS: Geriatric Depression Scale; I-ADNI: Italian Alzheimer's Disease Neuroimaging Initiative; IMI: Innovative Medicine Initiative; JADNI: Japanese Alzheimer's Disease Neuroimaging Initiative; MCI: Mild Cognitive Impairment; MMSE: Mini-Mental State Examination; MRI: Magnetic Resonance Imaging; NACC : National Alzheimer's Coordinating Center; OASIS: Open Access Series of Imaging Studies; PREVENT-AD: Pre-symptomatic Evaluation of Experimental or Novel Treatments for Alzheimer's Disease; PharmaCog: Prediction of Cognitive Properties of New Drug Candidates for Neurodegenerative Diseases in Early Clinical Development; ROSMAP: Religious Orders Study and Memory and Aging Project; VASCULAR: Vascular Contributors to Prodromal Alzheimer's Disease; VITA: Vienna Transdanube Aging; WMH-AD: White Matter Hyperintensities in Alzheimer's Disease.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13195-022-01009-4.

Additional file 1: Figure S1. Longitudinal follow-up plots the specified variables in a case scenario. **Figure S2.** Distribution of hippocampus volume displayed with boxplots using the "Biomarkers" tool of the ADataViewer. **Figure S3.** The modality map, describing which data modalities have been assessed per cohort.

Acknowledgements

We want to commend all data owners on their adherence to open science principles by sharing their data. We believe that their commitment is invaluable for any scientific research.

We thank the study participants and staff of the Rush Alzheimer's Disease Center. ROSMAP was supported by NIA grants P30AG010161, R01AG015819, and R01AG017917.

The A4 Study is a secondary prevention trial in preclinical Alzheimer"s disease, aiming to slow the cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 Study is funded by a public-private-philanthropic partnership, including funding from the National Institutes of Health-National Institute on Aging, Eli Lilly and Company, Alzheimer"s Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation and additional private donors, with in-kind support from Avid and Cogstate. The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study is funded by the Alzheimer"s Association and GHR Foundation. The A4 and LEARN Studies are led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School and Dr. Paul Aisen at the Alzheimer"s Therapeutic Research Institute (ATRI), University of Southern California. The A4 and LEARN Studies are coordinated by ATRI at the University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University of Southern California. The participants screening for the A4 Study provided permission to share their de-identified data in order to advance the quest to find a successful treatment for Alzheimer's disease. We would like to acknowledge the dedication of all the participants, the site personnel, and all of the partnership team members who continue to make the A4 and LEARN Studies possible. The complete A4 Study Team list is available on a4study.org/a4-study-team.

Data collection and sharing of ABVIB was funded by the National Institutes on Aging (NIA) P01 AG12435.

Data collection and sharing of ARWIBO was supported by the Italian Ministry of Health, under the following grant agreements: Ricerca Corrente IRCCS Fatebenefratelli, Linea di Ricerca 2; Progetto Finalizzato Strategico 2000-2001 "Archivio normativo italiano di morfometria cerebrale con risonanza magnetica (età 40+)"; Progetto Finalizzato Strategico 2000-2001 "Decadimento cognitivo lieve non dementigeno: stadio preclinico di malattia di Alzheimer e demenza vascolare. Caratterizzazione clinica, strumentale, genetica e neurobiologica e sviluppo di criteri diagnostici utilizzabili nella realtà nazionale,"; Progetto Finalizzata 2002 "Sviluppo di indicatori di danno cerebrovascolare clinicamente significativo alla risonanza magnetica strutturale"; Progetto Fondazione CARIPLO 2005-2007 "Geni di suscettibilità per gli endofenotipi associati a malattie psichiatriche e dementigene"; "Fitness and Solidarietà"; and anonymous donors.

EPAD LCS is registered at www.clinicaltrials.gov Identifier: NCT02804789. Data used in the preparation of this article were obtained from the EPAD LCS data set V.IMI, doi:10.34688/epadlcs_v.imi_20.10.30. The EPAD LCS was launched in 2015 as a public-private partnership, led by Chief Investigator Professor Craig Ritchie MB BS. The primary research goal of the EPAD LCS is to provide a well-phenotyped probability-spectrum population for developing and continuously improving disease models for Alzheimer's disease in individuals without dementia. This work used data and/or samples from the EPAD project which received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking EPAD grant agreement n° 115736 and an Alzheimer's Association Grant (SG21-818099-EPAD).

PharmaCog was funded through the European Community's 'Seventh Framework' Programme (FP7/2007-2013) for an innovative scheme, the Innovative Medicines Initiative (IMI). IMI is a young and unique public-private partnership, founded in 2008 by the pharmaceutical industry (represented by the European Federation of Pharmaceutical Industries and Associations), EFPIA and the European Communities (represented by the European Commission).

J-ADNI was supported by the following grants: Translational Research Promotion Project from the New Energy and Industrial Technology Development Organization of Japan; Research on Dementia, Health Labor Sciences Research Grant; Life Science Database Integration Project of Japan Science and Technology Agency; Research Association of Biotechnology (contributed by Astellas Pharma Inc., Bristol-Myers Squibb, Daiichi-Sankyo, Eisai, Eli Lilly and Company, Merck-Banyu, Mitsubishi Tanabe Pharma, Pfizer Inc., Shionogi & Co., Ltd., Sumitomo Dainippon, and Takeda Pharmaceutical Company), Japan, and a grant from an anonymous foundation.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), and P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). These data are from the Vascular Contributors to Prodromal Alzheimer's disease (VASCULAR) study. The study is part of the M2OVE-AD supported studies and is supported by NIH/NIA (PI: Ihab Hajjar, MD, MS grant numbers: AG051633 and AG057470. The investigators acknowledge the contribution of the study participants who donated their time and effort for this study.

Alzheimer's Disease Neuroimaging Initiative: Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Japanese Alzheimer's Disease Neuroimaging Initiative: Data used in the preparation of this article were obtained from the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) database deposited in the National Bioscience Database Center Human Database, Japan (Research ID: hum0043.v1, 2016). As such, the investigators within J-ADNI contributed to the design and implementation of J-ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of J-ADNI investigators can be found at https://humandbs.biosciencedbc.jp/ en/hum0043-j-adni-authors.

Data used in the preparation of this article were obtained from the Aging Brain: Vasculature, Ischemia, and Behavior Study (ABVIB). As such, the key investigators within the ABVIB contributed to the design and implementation of ABVIB and/or provided data but did not participate in the analysis or writing of this report: Helena C. Chui M.D. (Principal Investigator), Charles C. DeCarli, M.D., William G. Ellis, M.D., William J. Jagust, M.D., Joel H. Kramer, Ph.D., Meng Law, M.D., Dan Mungas Ph,D. Bruce R. Reed, Ph.D., Nerses Sanossian, M.D., Michael W. Weiner, M.D. Wendy J. Mack, Ph.D., Harry V. Vinters, M.D., Chris Zarow, Ph.D., Ling Zheng, Ph.D.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Repository Without Borders (ARWiBo) database (www.arwibo.it). As such, the researchers within the ARWiBo contributed to the design and implementation of ARWiBo and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ARWiBo researchers can be found at www.arwibo.it/acknowledgement.it.

Data used in the preparation of this article were obtained from the Longitudinal Cohort Study (LCS), delivered by the European Prevention of Alzheimer's Disease (EPAD) Consortium. As such, investigators within the EPAD LCS and EPAD Consortium contributed to the design and implementation of EPAD and/or provided data but did not participate in the analysis or writing of this report. A complete list of EPAD Investigators can be found at http://ep-ad.org/wp-content/uploads/2020/12/202010_List-of-epadistas.pdf.

Authors' contributions

CB conceived and supervised the project. YS and CB collected the datasets. YS prepared the data for ADataViewer. DDF implemented the platform. YS and CBA curated the variable mappings. CB drafted the manuscript. DDF, YS, and MHA revised the manuscript. MHA acquired the funding. The author(s) read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EPAD grant agreement n°117536, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 826421, "TheVirtualBrain-Cloud".

Availability of data and materials

All investigated datasets used in this study can be obtained from the respective data owners. Links are provided at https://adata.scai.fraunhofer.de/cohorts.

Declarations

Ethics approval and consent to participate

All investigated studies acquired informed consent for data collection and sharing from their participants. All cohort studies got ethical approval. For more details, we refer to their individual references.

Consent for publication

The publication guidelines of each individual cohort study were followed and the manuscript was submitted and subsequently approved by all data owners that requested manuscript clearing.

Competing interests

DDF received a salary from Enveda Biosciences, and the company has no competing interests with the published results. The rest of the authors declare that they have no competing interests.

Author details
[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany. [2]Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany. [3]University Hospital Ntra. Sra. de Candelaria, Santa Cruz de Tenerife 38010, Spain.

Received: 26 November 2021 Accepted: 27 April 2022
Published online: 21 May 2022

References

1. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Cedarbaum J, et al. Impact of the Alzheimer's Disease Neuroimaging Initiative, 2004 to 2014. Alzheimers Dement. 2015;11(7):865–84.
2. Shi L, Westwood S, Baird AL, Winchester L, Dobricic V, Kilpert F, et al. Discovery and validation of plasma proteomic biomarkers relating to brain amyloid burden by SOMAscan assay. Alzheimers Dement. 2019;15(11):1478–88.
3. Koval I, Bône A, Louis M, Lartigue T, Bottani S, Marcoux A, et al. AD Course Map charts Alzheimer's disease progression. Sci Rep. 2021;11(1):8020.
4. Vogel JW, Young AL, Oxtoby NP, Smith R, Ossenkoppele R, Strandberg OT, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. Nat Med. 2021;27(5):871–81.
5. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. BMC Med. 2018;16(1):150.
6. Golriz Khatami S, Robinson C, Birkenbihl C, Domingo-Fernández D, Hoyt CT, Hofmann-Apitius M. Challenges of integrative disease modeling in Alzheimer's disease. Front Mol Biosci. 2020;6:158.
7. Cunningham JA, Van Speybroeck M, Kalra D, Verbeeck R. Nine principles of semantic harmonization. AMIA Annu Symp Proc. 2017;2016:451–9.
8. Neville J, Kopko S, Romero K, Corrigan B, Stafford B, LeRoy E, et al. Accelerating drug development for Alzheimer's disease through the use of data standards. Alzheimers Dement (N Y). 2017;3(2):273–83.
9. Birkenbihl C, Emon MA, Vrooman H, Westwood S, Lovestone S, AddNeuroMed Consortium, et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia- lessons for translation into clinical practice. EPMA J. 2020;11(3):367–76.
10. Birkenbihl C, Salimi Y, Fröhlich H. Japanese Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Neuroimaging Initiative. Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. Alzheimers Dement. 2021. https://doi.org/10.1002/alz.12387.
11. Lovestone S, EMIF Consortium. The European medical information framework: a novel ecosystem for sharing healthcare data across Europe. Learn Health Syst. 2019;4(2):e10214.
12. Gallacher J, de Reydet de Vulpillieres F, Amzal B, Angehrn Z, Bexelius C, Bintener C, et al. Challenges for optimizing real-world evidence in Alzheimer's disease: the ROADMAP project. J Alzheimers Dis. 2019;67(2):495–501.
13. Oliveira JL, Trifan A, Bastião Silva LA. EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data. Int J Med Inform. 2019;126:35–45.
14. Janssen O, Vos SJB, García-Negredo G, Tochel C, Gustavsson A, Smith M, et al. Real-world evidence in Alzheimer's disease: the ROADMAP Data Cube. Alzheimers Dement. 2020;16(3):461–71.
15. Birkenbihl C, Salimi Y, Domingo-Fernándéz D, Lovestone S, AddNeuroMed Consortium, Fröhlich H, et al. Evaluating the Alzheimer's disease data landscape. Alzheimers Dement (N Y). 2020;6(1):e12102.
16. Sperling RA, Rentz DM, Johnson KA, Karlawish J, Donohue M, Salmon DP, et al. The A4 study: stopping AD before symptoms begin? Sci Transl Med. 2014;6(228):228fs13.
17. Rodriguez FS, Zheng L, Chui HC. Aging Brain: Vasculature, Ischemia, and Behavior Study. Psychometric characteristics of cognitive reserve: how high education might improve certain cognitive abilities in aging. Dement Geriatr Cogn Disord. 2019;47(4-6):335–44.
18. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement. 2005;1(1):55–66.
19. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr. 2009;21(4):672–87.
20. Birkenbihl C, Westwood S, Shi L, Nevado-Holgado A, Westman E, Lovestone S, et al. ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. J Alzheimers Dis. 2021;79(1):423–31.
21. Frisoni GB, Prestia A, Zanetti O, Galluzzi S, Romano M, Cotelli M, et al. Markers of Alzheimer's disease in a population attending a memory clinic. Alzheimers Dement. 2009;5(4):307–17.
22. Weiner MW, Veitch DP, Hayes J, Neylan T, Grafman J, Aisen PS, et al. Effects of traumatic brain injury and posttraumatic stress disorder on Alzheimer's disease in veterans, using the Alzheimer's Disease Neuroimaging Initiative. Alzheimers Dement. 2014;10(3 Suppl):S226–35.
23. Brueggen K, Grothe MJ, Dyrba M, Fellgiebel A, Fischer F, Filippi M, et al. The European DTI Study on Dementia - a multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment. Neuroimage. 2017;144(Pt B):305–8.
24. Bos I, Vos S, Vandenberghe R, Scheltens P, Engelborghs S, Frisoni G, et al. The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. Alzheimers Res Ther. 2018;10(1):64.
25. Solomon A, Kivipelto M, Molinuevo JL, Tom B, Ritchie CW, EPAD Consortium. European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): study protocol. BMJ Open. 2019;8(12):e021017.
26. Cavedo E, Redolfi A, Angeloni F, Babiloni C, Lizio R, Chiapparini L, et al. The Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI): validation of structural MR imaging. J Alzheimers Dis. 2014;40(4):941–52.
27. Iwatsubo T. Japanese Alzheimer's Disease Neuroimaging Initiative: present status and future. Alzheimers Dement. 2010;6(3):297–9.
28. Besser L, Kukull W, Knopman DS, Chui H, Galasko D, Weintraub S, et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. Alzheimer Dis Assoc Disord. 2018;32(4):351–8.
29. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. J Cogn Neurosci. 2010;22(12):2677–84.
30. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci. 2007;19(9):1498–507.
31. Breitner JCS, Poirier J, Etienne PE, Leoutsakos JM. Rationale and Structure for a New Center for Studies on Prevention of Alzheimer's Disease (StoP-AD). J Prev Alzheimers Dis. 2016;3(4):236–42.
32. Galluzzi S, Marizzoni M, Babiloni C, Albani D, Antelmi L, Bagnoli C, et al. Clinical and biomarker profiling of prodromal Alzheimer's disease in workpackage 5 of the Innovative Medicines Initiative PharmaCog project: a 'European ADNI study'. J Intern Med. 2016;279(6):576–91.
33. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012;9(6):628–45.
34. Emory University School of Medicine (2021, July). VASCULAR (VAScular ContribUtors to prodromaL AlzheimeR's disease). https://med.emory.edu/departments/medicine/divisions/geriatrics-gerontology/research/labs/bsharp/studies.html
35. Fischer P, Jungwirth S, Krampla W, Weissgram S, Kirchmeyr W, Schreiber W, et al. Vienna Transdanube Aging "VITA": study design, recruitment strategies and level of participation. J Neural Transm Suppl. 2002;62:105–16.
36. Damulina A, Pirpamer L, Seiler S, Benke T, Dal-Bianco P, Ransmayr G, et al. White matter hyperintensities in Alzheimer's disease: a lesion probability mapping study. J Alzheimers Dis. 2019;68(2):789–96.
37. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
38. Verdi S, Marquand AF, Schott JM, Cole JH. Beyond the average patient: how neuroimaging models can address heterogeneity in dementia. Brain. 2021;144(10):2946-53.
39. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.

Publisher's Note

A.3 ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset

Reprinted with permission from "Birkenbihl, C., Westwood, S., Shi, L., Nevado-Holgado, A., Westman, E., Lovestone, S., Hofmann-Apitius, M., and AddNeuroMed Consortium. (2021). ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *Journal of Alzheimer's Disease*, 79(1), 423-431.".

Journal of Alzheimer's Disease 79 (2021) 423–431
DOI 10.3233/JAD-200948
IOS Press

ANMerge: A Comprehensive and Accessible Alzheimer's Disease Patient-Level Dataset

Colin Birkenbihl[a,b,*], Sarah Westwood[c], Liu Shi[c], Alejo Nevado-Holgado[c], Eric Westman[d], Simon Lovestone[c] on behalf of the AddNeuroMed Consortium and Martin Hofmann-Apitius[a,b]

[a]*Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany*

[b]*Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany*

[c]*Department of Psychiatry, University of Oxford, Oxford, UK*

[d]*Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden*

Accepted 22 October 2020

Abstract.

Background: Accessible datasets are of fundamental importance to the advancement of Alzheimer's disease (AD) research. The AddNeuroMed consortium conducted a longitudinal observational cohort study with the aim to discover AD biomarkers. During this study, a broad selection of data modalities was measured including clinical assessments, magnetic resonance imaging, genotyping, transcriptomic profiling, and blood plasma proteomics. Some of the collected data were shared with third-party researchers. However, this data was incomplete, erroneous, and lacking in interoperability.

Objective: To provide the research community with an accessible, multimodal, patient-level AD cohort dataset.

Methods: We systematically addressed several limitations of the originally shared resources and provided additional unreleased data to enhance the dataset.

Results: In this work, we publish and describe ANMerge, a new version of the AddNeuroMed dataset. ANMerge includes multimodal data from 1,702 study participants and is accessible to the research community via a centralized portal.

Conclusion: ANMerge is an information rich patient-level data resource that can serve as a discovery and validation cohort for data-driven AD research, such as, for example, machine learning and artificial intelligence approaches.

Keywords: AddNeuroMed, Alzheimer's disease, biomarkers, cohort analysis, cohort studies, data-driven science, dataset, dementia, genome wide association studies, magnetic resonance imaging, multimodal

INTRODUCTION

Alzheimer's disease (AD) is a progressive disease whose pathology develops years before cognitive symptoms arise and a diagnosis is made by a clinician [1]. Early intervention in non-cognitively impaired, pre-symptomatic disease stages is instrumental to any future disease modifying therapy. Enabling such an early intervention poses the problem of diagnosing a patient with AD before cognitive symptoms indicate disease presence. One approach to establish whether a specific individual is in the pre-symptomatic stages of the disease is a diagnosis based on informative disease biomarkers. The critical prerequisite for discovery and validation of such biomarkers are resourceful patient-level datasets [2]. However, findable AD cohort datasets which are accessible to the research community are scarce.

*Correspondence to: Colin Birkenbihl, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53754 Sankt Augustin, Germany. Tel.: +49 2241 14 2420; E-mail: colin.birkenbihl@scai.fraunhofer.de.

Open science is a paradigm aimed at increasing societal benefit of research through dissemination and sharing of scientific data. This enables usage and analysis of collected data by the whole research community which subsequently will increase the achieved knowledge gain. Currently, the prime example of following the open science paradigm in the AD field is the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3]. ADNI is an information rich, comprehensive clinical AD cohort dataset that enables secure, yet easy access to its patient level data for researchers with reasonable study interest. In only a few days, raw data as well as a preprocessed version of ADNI (ADNIMERGE) are accessible via the Laboratory of Neuro Imaging (LONI) service (https://loni.usc.edu/). With regard to clinical data, initial preprocessing, arranging, and cleaning of data is often the most time-consuming step in data analysis. Due to that, a major cumulative time save is possible by sharing an already preprocessed, easy-to-analyze dataset instead of a raw data collection. Here, researchers can simply use the provided ADNIMERGE and thereby avoid investing additional time into data preprocessing and cleaning.

While ADNI is a tremendously important resource, as every cohort dataset, it comes with its own limitations and biases [4]. To ensure reliability of observations made in one cohort, validation in data from independent cohorts is necessary [5]. Still, apart from ADNI there are not many AD cohort studies which 1) share their data in a similarly comprehensive version and 2) keep the bureaucracy during an access application as straightforward as ADNI does. From our experience, access applications are often time consuming and if access is granted, shared data is sometimes lacking important information. Therefore, other easily accessible and information rich alternatives besides ADNI are crucial.

In 2005, Lovestone et al. started AddNeuroMed, a project funded by InnoMed, a precursor of the Innovative Medicine Initiative (IMI) [6]. It aimed at collecting longitudinal patient data at multiple sites across Europe to identify urgently needed progression biomarkers for AD. For this purpose, a broad spectrum of variables was measured including demographics, neuropsychological assessments, genetic variations and transcriptomics, blood plasma proteomics, and structural magnetic resonance imaging (MRI) of the brain. In 2015, a subset of the collected data was uploaded on Synapse (https://www.synapse.org/). Next to the original AddNeuroMed data, some data from participants of the Maudsley BRC Dementia Case Registry at King's Health Partners cohort (DCR) and the Alzheimer's Research Trust UK cohort (ART) was included [7]. Although the shared AddNeuroMed collection is a large dataset, involving more than 1,700 participants, it has only been cited about 65 times. In contrast, ADNI, which involves roughly 2,400 individuals, was cited more than 1,300 times. Compared to the impact ADNI has had on recent research activities, it seems AddNeuroMed has not reached its full potential. One probable reason for the comparably lower data usage might be the findability and the state of the data published on Synapse. The dataset 1) has never been officially published, 2) is not easy to work with due to missing organization, and 3) is not complete with several entries being erroneous or lacking information. To enable the research community to leverage the full potential of this dataset, a lot of data preprocessing efforts are needed and it is vital to point the community toward this unsalvaged resource.

In this work, we present and publish a new, improved, and updated version of AddNeuroMed called ANMerge. ANMerge is a comprehensive, preprocessed AD cohort dataset which is again accessible via Synapse (https://doi.org/10.7303/syn22252881). It is fully interoperable in between its modalities, and rigorous data curation was performed to ensure higher information density and usability. Furthermore, we present a detailed overview on which and how much data is available in the dataset. Finally, we highlight the increased preprocessing efforts involved in creating such a dataset. By making ANMerge accessible, we aim to provide the AD research community with an information rich alternative to previously published cohort datasets, and thereby support the discovery and robust validation of scientific insights.

METHODS

Data collection

AddNeuroMed data collection was performed at six different centers across Europe: University of Kuopio, Finland; Aristotle University of Thessaloniki, Greece; King's College London, United Kingdom; University of Lodz, Poland; University of Perugia, Italy; and University of Toulouse, France [6]. The participation of those centers highlights AddNeuroMed as a major cross-European effort in AD related data collection. At each site, all protocols and procedures were approved by Institutional Review Boards and informed consent was obtained

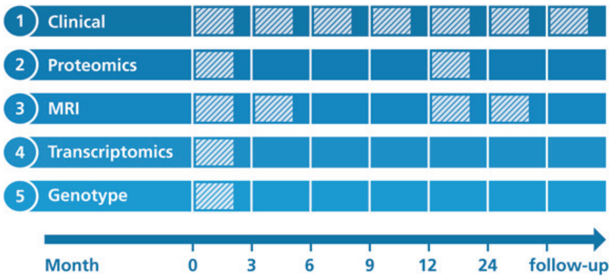


Fig. 1. Overview on longitudinal data collection per modality. Proteomics, Proteomic data from blood plasma. Transcriptomics, Transcriptomic data from blood plasma. MRI, Structural magnetic resonance imaging.

for all patients according to the Declaration of Helsinki (1991) [8]. In cases where dementia compromised capacity assent from the patient and consent from a relative, according to local law, was obtained.

Exclusion criteria included other neurological or psychiatric diseases, significant unstable systemic illness or organ failure, and alcohol or substance misuse. AD diagnosis followed the Diagnostic and Statistical Manual for Mental Diagnosis, fourth edition and National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association criteria [9]. AD patients were included if they exhibited a Mini-Mental State Examination (MMSE) score in the range of 12–28, a Clinical Dementia Rating (CDR) scale score of above 0.5, and were aged 65 years or above. Individuals were considered as mild cognitive impairment (MCI) according to the Petersen criteria [10]. For inclusion, MCI patients aged 65 or above, the MMSE score ranged between 24 and 30, and they scored 0.5 on the CDR. Participants were considered to be cognitively healthy if they showed normal performance on cognitive tests (within 1.5 SD of average for age, gender and education) and scored 0 on the CDR [11].

AddNeuroMed's study protocols were designed to be at least partially compatible with ADNI [6]. Figure 1 illustrates when data collection was performed for each modality.

Clinical assessments

At each participant's visit throughout the study, a broad collection of neurocognitive and psychological assessments were performed, including the MMSE, CDR, GDS (Geriatric Depression Scale), NPI (Neuropsychiatric Inventory), ADAS-Cog (Alzheimer's Disease Assessment Scale-Cognitive Subscale), ADCS-ADL (Alzheimer's Disease Cooperative Study Activities of Daily Living Scale), the full CERAD battery [12], the Hachinski Ischemic Score, and the Webster Rating Scale. The frequency with which assessments were made varied between diagnostic groups. During the first year, AD cases completed assessments every three months and annual follow-up visits afterwards. MCI patients and healthy individuals from AddNeuroMed, as well as all participants from the ART and DCR cohorts, were assessed regularly every twelve months.

Proteomics

Proteomic data were measured in blood plasma using a Slow Off-rate Modified Aptamer (SOM Amer)-based array called 'SOMAscan' (SomaLogic, Inc, Boulder, Colorado). Data collection was performed at baseline and again one year into the study. Details on data acquisition are presented in Kiddle et al. [13] and Sattlecker et al. [14]. In brief, using chemically altered nucleotides the protein signal is turned into a nucleotide signal that can be measured using microarrays. Per sample 8 μL plasma were required and levels of 1,001 distinct proteins were measured. An in-depth description of the array technology can be found in Gold et al. [15].

Genotyping

AddNeuroMed participants were genotyped in three batches. For batch one, the Illumina Human Hap610-Quad Beadchip was used, while batches two and three were processed using the Illumina HumanOmniExpress-12 v1.0. More information can be found in the method section of Loudursamy et al. [16] and Proitsi et al. [17]. All genotyping was performed at the Centre National de Génotypage in France.

Transcriptomics

Blood samples for the collection of gene expression data were taken at study baseline. Transcriptional profiling was performed in two batches using the Illumina HumanHT-12 v3 (batch one) and v4 (batch two) Expression BeadChip kits. Original raw data can be found in GEO[1]. Preprocessed raw data files, as well as post quality control, batch corrected expression values, are distributed via Synapse. The processed data underwent background correction, log base two transformation and all values were robust spline normalized [18]. Outlying samples were excluded. Batch correction was performed using ComBat [19]. All data were subset to probes that could reliably be detected in at least 80% of samples in at least one diagnostic group. More details on the processing of the data is explained in Voyle et al. [18].

Magnetic resonance imaging

1.5 Tesla T1-weighted MRI images were taken at three different timepoints throughout the study (Month 0, 3, 12). The first 3-month interval was explicitly chosen to contrast the 6-month MRI follow-up of ADNI and thereby evaluate if 3 months could potentially be enough to observe substantial changes in brain structure. Protocols for imaging were aligned to the ADNI study. Details on the AddNeuroMed MRI data acquisition have been described in Simmons et al. [8, 20]. ANMerge provides access to collected raw images as well as processed brain volumes and cortical thickness calculated using FreeSurfer 5.3 and 6.0.

Data preprocessing

As a first step, manual investigation of all raw AddNeuroMed data files was inevitable to assess the availability and state of each data type. To avoid irreproducible changes to the data, we did not alter any entry manually but relied on programming for each data changing step.

We tried to build the most informative and complete, yet minimally complex, version of AddNeuroMed possible. Therefore, we carefully selected variables from the raw data for inclusion into ANMerge. To limit the number of variables in ANMerge, we only included total scores of clinical assessments in the new ANMerge files instead of listing all sub-scores and individual answers. Variables not considered for inclusion into ANMerge, such as the test subscores, are accessible through the additionally provided raw data.

Not all participants from the DCR and ART cohorts underwent data collection in the course of AddNeuroMed. However, since clinical assessments between the original AddNeuroMed study and DCR were largely overlapping, we decided to include all DCR participants into ANMerge, even if they lacked other modalities apart from clinical data. From the ART cohort, only those individuals who had been assessed in at least one modality next to the clinical data were included in order to reduce sparsity in the resulting tables.

In the original AddNeuroMed data, modality specific data tables lacked interoperability because distinct patient identifiers were used for many of them. Additionally, only the visit numbers were reported instead of the actual months in study. This was misleading due to differences in assessment intervals between diagnostic groups (e.g., visit 2 for healthy and MCI participants corresponds to visit 5 of AD patients). Information which is not subject to change (e.g., *APOE* genotype) was only reported at baseline which led to sparsity in follow-up visit entries. Furthermore, to increase interoperability not only within AddNeuroMed itself but also to other data resources, we mapped variable names to public database identifiers wherever possible. Finally, we enriched ANMerge with data previously not available in the Synapse version. Among others, we added missing diagnoses and clinical assessment scores as well as months in study as an unambiguous time scale.

RESULTS

Overview on data

The resulting ANMerge dataset comprises four data modality specific subtables, genotype data in PLINK format and one combined table providing all preprocessed information as one. Respectively, one subtable was created for clinical data, proteomics, FreeSurfer calculated MRI features, and gene expression values. Next to diagnosis and clinical assessments, the clinical subtable also provides participants demographics, family history, and medication data.

In total, the dataset comprises information on 1,702 patients, out of which 773, 665, and 264 originated from the AddNeuroMed, DCR, and ART cohorts, respectively (Table 1). Data on 4,585 individual participant visits are reported. At study baseline, 512 participants had been diagnosed with AD, 397 with MCI, and 793 were non-cognitively impaired individuals. Table 1 describes the average characteristics of each diagnosis group at baseline. On average, cognitively affected individuals (i.e., MCI and AD) in ANMerge were 77 years old at baseline, completed 9.7 years of full-time education and 59% of them were female. Healthy individuals averaged to an age of 74.5 years, underwent 12.3 years of education and 59% are female. During study runtime 48 and 11 healthy participants converted to MCI and AD respectively. Out of all patients diagnosed with MCI at baseline 70 converted to AD.

Table 1
Summary statistics describing the ANMerge dataset at baseline

Diagnosis	N	ANM	DCR	ART	Age (SD)	Female %	Education (SD)	*APOE* ϵ4 positive %
CTL	793	266	423	104	74.5 (6.4)	59	12.3 (4.3)	25
MCI	397	247	89	61	76.0 (6.5)	55	10.0 (4.3)	40
AD	512	260	153	99	78.6 (7.2)	63	9.4 (4.3)	54
Total	1702	773	665	264	76.4 (6.9)	59	10.9 (4.5)	39

N, Number of participants with the corresponding diagnosis; ANM, Number of participants originally from the AddNeuroMed study; DCR, Number of participants originally from the DCR study; ART, Number of participants originally from ART study; CTL, Healthy control participants; SD, Standard deviation.

Table 2
Number of assessed variables and participants per modality subtables

Modality	Participants	Variables
Clinical	1,702	40
Proteomics	680	1,016
MRI	453	136
Gene expression	709	56,701
Genotype	1,014	789,470

Not every study participant took part in data collection of all modalities. For our evaluation, we considered participants as represented in a modality if at least one modality specific variable was measured. This implies that not necessarily all variables of that modality were available for a given participant (e.g., an individual listed in the clinical table might have MMSE scores but no ADAS-Cog). We found that clinical data is reported for all 1,702 participants, while MRI, proteomic, gene expression, and genotype data were collected for subsets of several hundred participants each (Table 2 ‘Participants’). Figure 2 demonstrates the number of patients assessed across multiple modalities. In total, 239 participants have been assessed with regard to all five data modalities. By reducing the number of modalities included into an analysis, subsequently the number of available participants rises. For example, when conducting a multimodal study using transcriptomic, genotype and clinical variables data from 614 participants would be available. Focusing only on genotype and clinical data yields 1,010 analyzable subjects.

All in all, data on more than 800,000 variables are reported in ANMerge. 40 of them correspond to the clinical modality, 56,701 originate from gene expression analysis, 136 are MRI variables, and 1,016 were assessed in blood proteomics (Table 2 ‘Variables’).

As with most clinical studies, AddNeuroMed exhibits a declining number of participants over study runtime (Fig. 3). For most patients ($n = 1{,}136$) at least one additional visit 12 months after baseline is available in the data. The drop of AD patients at month 3 to 9 is explained by the fact that only AD cases recruited in the original AddNeuroMed study had three monthly visits during the first year, while

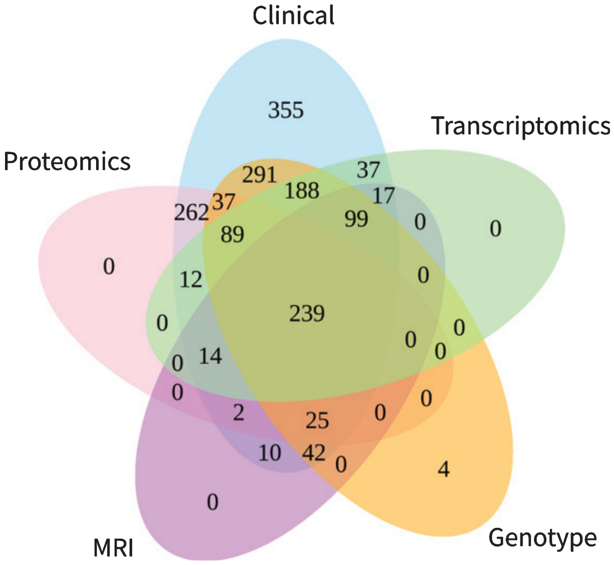


Fig. 2. Participant overlap across modalities. The numbers illustrate the number of participants with available information for the intersection of the respective modalities.

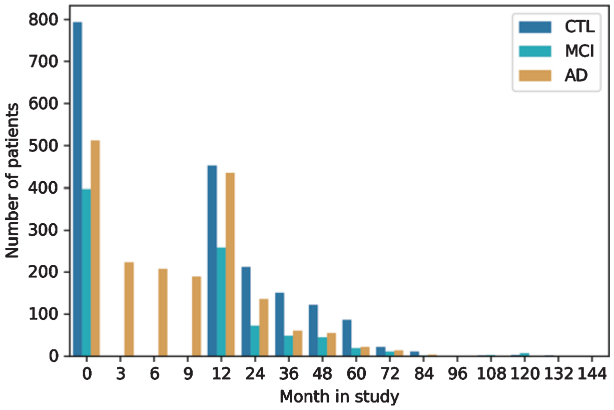


Fig. 3. Longitudinal follow-up and patient drop-out throughout study runtime per diagnosis group. CTL, healthy controls; MCI, mild cognitive impaired participants; AD, Alzheimer’s disease patients.

ART and DCR assessed all patients annually. The longest follow-up exhibited in the data spanned 12 years.

Data after preprocessing

The new ANMerge dataset is divided into modality specific subtables which makes unimodal analysis straightforward. During the preprocessing of AddNeuroMed we addressed multiple issues detected in the original data. The previous version of Add NeuroMed was indexed using distinct patient identifiers across its modalities, thereby impeding multimodal analysis due to missing internal interoperability. Standard data integration techniques like table joins were impossible. By mapping all present identifiers to a unique one, we enabled inter-modality interoperability such that tables can now easily be analyzed together. Additionally, we provide a new identifier mapping file which helps to map the unified identifiers to the raw data for backwards compatibility. To increase interoperability also beyond ANMerge itself, we mapped variable names to public database identifiers. For example, proteomic variables are now also given as UniProt identifiers, genotype data is encoded as rs-numbers, and gene expression probes as Illumina IDs [21]. All of these identifiers can be easily mapped to other resources and be enriched with information from public databases. Instead of relying on the misleading reported visit numbers, in ANMerge we added an unambiguous time scale (months in study) to patient entries to make longitudinal follow-up easier to understand. Information that will stay permanent (e.g., *APOE* ε4 status) throughout study runtime is now reported at every visit for that respective patient, not only at baseline. Multiple issues found in the data (e.g., typos and erroneous entries) have been corrected.

Although proteomic and transcriptomic data, for example, were presented for some DCR and ART participants in the previous AddNeuroMed version, no corresponding clinical data was available, including important information like participant diagnosis. ANMerge now has all available clinical data for the two associated cohorts, which critically increases the amount of actionable information in the dataset.

Accessing ANMerge

ANMerge and the underlying data are available under https://doi.org/10.7303/syn22252881. To ensure data privacy, a straight-forward data access application has to be completed. During this access application, researchers are asked to 1) register a Synapse account, 2) have all collaborators who will access the data sign a data use certificate (DUC), 3) provide a brief research proposal (1–3 paragraphs), and 4) agree that the appropriate citation of ANMerge will be used. By signing the DUC, applicants confirm that the planned study underwent ethical review. If successful, access approval is granted within approximately 14 days.

DISCUSSION

In this work, we presented ANMerge, a longitudinal multimodal AD cohort dataset that we made accessible to the research community. Since the most time-consuming part about data analysis is often the preprocessing of data, we believe that the cumulative time save, achieved by sharing readily preprocessed datasets, can lead to faster global scientific advancement. Additionally, by describing the characteristics of the dataset in detail, we aim to enable researchers to evaluate on first sight if ANMerge is suited for their analysis.

Establishing reliable results through external validation on independent cohorts is of utmost importance, especially when dealing with high complex diseases like AD. Up to date, and to the best of our knowledge, the vast majority of data-driven approaches in AD relied solely on ADNI data. To validate discoveries made in ADNI on other datasets, a high overlap in measured variables is a prerequisite. Previously, we could demonstrate that despite evident differences to ADNI, ANMerge is a viable validation dataset [22].

Providing clean, preprocessed datasets is a key prerequisite to enable any data-driven AD research. However, small cohort studies, for example conducted in single hospitals, often lack the resources to provide such readily preprocessed data. In an era where data re-use beyond the initial study itself becomes increasingly important, we believe that adequate data preprocessing and sharing should resemble a planned position in the initial funding proposal for all cohort studies.

Limitations

While AddNeuroMed collected a valuable dataset, it still has some noteworthy limitations. The main limitation of the data is that the amyloid status of participants is unknown. No positron emission tomography (PET) imaging was performed and cerebrospinal fluid markers were not assessed. This difference to the ADNI data could partially explain the comparably lower number of citations of the original AddNeuroMed data.

As in many clinical cohort datasets, missing data is a considerable issue in AddNeuroMed. Not every patient was involved in the assessment of every data modality and within a modality not necessarily all variables were measured for each patient.

Compared to ADNI, AddNeuroMed lacks comprehensive documentation. Retrospectively searching for study procedures and protocols of an already concluded, older cohort study proved to be very difficult. The original study website is not available anymore and exhaustive study protocols were not findable. However, we tried to address this limitation by collecting and assembling all available information and links in this publication. While the original AddNeuroMed dataset provided descriptive data dictionaries for most clinical variables, we extent the documentation by meaningful connections of other modalities to public databases (e.g., UniProt or dbSNP) by mapping their variable names to appropriate identifiers wherever possible.

The genotype and transcriptomic data presented in ANMerge was acquired in two separate batches of participants. This implies that the data can be subject to systematic batch effects and appropriate adjustments should be made [23].

Conclusion

Over the last years, the AD field witnessed a fortunate shift to a more accessible and comprehensible data culture. New studies such as PREVENT-AD [24] and EPAD [25] recently joined the ranks of ADNI, DIAN [26], and others by making their data accessible to third party researchers. Currently running studies, for example the Deep Frequent Phenotype Study [27], already emphasized that the collected data will be published. On the metadata-level, projects such as EMIF [28] and ROADMAP [29] aimed at aiding researchers to understand the datasets in our field by providing comprehensive metadata resources. This shift in the AD data landscape toward increasingly accessible and understandable datasets marks an important development to facilitate data-driven research in the dementia domain.

By publishing ANMerge, we want to contribute to a culture of data sharing in AD research and follow the open science paradigm. Participation in observational clinical cohort studies represents an immense investment by volunteering patients and healthy individuals. They undergo extensive and sometimes intrusive repeated measurements, most of the time without any direct benefit for the individuals themselves, with the ultimate aim to contribute to disease research. We believe that it is an ethical imperative to honor their investment by enabling their data to be used for generating the most societal benefit possible.

ACKNOWLEDGMENTS

The authors of this article wanted to thank every participant and professional involved in AddNeuroMed, DCR, and ART data collection.

This project has received partial support from the Innovative Medicines Initiative Joint Undertaking "AETIONOMY" under grant agreement #115568, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution.

This project was partially funded by an MRC Mental Health Data Pathfinder Award to the University of Oxford (MC_PC_17215), and supported by the NIHR Oxford Health Biomedical Research Centre.

The original AddNeuroMed study was supported by InnoMed (Innovative Medicines in Europe), an Integrated Project funded by the European Union of the Sixth Framework program priority FP6-2004-LIFESCIHEALTH-5, Life Sciences, Genomics, and Biotechnology for Health.

Additional support for data collection was granted by the Alzheimer's Research Trust and from the NIHR Specialist Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and King's College London, Institute of Psychiatry, London, United Kingdom.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-0948r2).

AVAILABILITY OF DATA AND MATERIALS

All data are available under: https://doi.org/10.7303/syn22252881

REFERENCES

[1] Sperling RA, Jack CR Jr, Aisen PS (2011) Testing the right target and right drug at the right stage. *Sci Transl Med* **3**, 111cm33.

[2] Morgan AR, Touchard S, Leckey C, O'Hagan C, Nevado-Holgado AJ; NIMA Consortium, Barkhof F, Bertram L, Blin O, Bos I, Dobricic V, Engelborghs S, Frisoni G, Frölich L, Gabel S, Johannsen P, Kettunen P, Kłoszewska I, Legido-Quigley C, Lleó A, Martinez-Lage P, Mecocci P, Meersmans K, Molinuevo JL, Peyratout G, Popp J, Richardson J, Sala I, Scheltens P, Streffer J, Soininen H, Tainta-Cuezva M, Teunissen C, Tsolaki M, Vandenberghe R, Visser PJ, Vos S, Wahlund LO, Wallin A, Westwood S, Zetterberg H, Lovestone S, Morgan BP; Annex: NIMA–Wellcome Trust Consortium for Neuroimmunology of Mood Disorders and Alzheimer's Disease (2019) Inflammatory biomarkers in Alzheimer's disease plasma. *Alzheimers Dement* **15**, 776-787.

[3] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005) Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement* **1**, 55-66.

[4] Whitwell JL, Wiste HJ, Weigand SD, Rocca WA, Knopman DS, Roberts RO, Boeve BF, Petersen RC, Jack CR Jr; Alzheimer Disease Neuroimaging Initiative (2012) Comparison of imaging biomarkers in the Alzheimer Disease Neuroimaging Initiative and the Mayo Clinic Study of Aging. *Arch Neurol* **69**, 614-622.

[5] Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, Maathuis MH, Moreau Y, Murphy SA, Przytycka TM, Rebhan M, Röst H, Schuppert A, Schwab M, Spang R, Stekhoven D, Sun J, Weber A, Ziemek D, Zupan B (2018) From hype to reality: Data science enabling personalized medicine. *BMC Med* **16**, 150.

[6] Lovestone S, Francis P, Strandgaard K (2007) Biomarkers for disease modification trials–the innovative medicines initiative and AddNeuroMed. *J Nutr Health Aging* **11**, 359-361.

[7] Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers HL, Hooper C, Rijsdijk F, Tabrizi SJ, Banner S, Shaw CE, Foy C, Poppe M, Archer N, Hamilton G, Powell J, Brown RG, Sham P, Ward M, Lovestone S (2006) Proteome-based plasma biomarkers for Alzheimer's disease. *Brain* **129**, 3042-3050.

[8] Simmons A, Westman E, Muehlboeck S, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Wahlund LO, Soininen H, Lovestone S, Evans A, Spenger C; AddNeuroMed Consortium (2009) MRI measures of Alzheimer's disease and the AddNeuroMed study. *Ann N Y Acad Sci* **1180**, 47-55.

[9] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-944.

[10] Petersen RC (2004) Mild cognitive impairment as a diagnostic entity. *J Intern Med* **256**, 183-194.

[11] Hye A, Riddoch-Contreras J, Baird AL, Ashton NJ, Bazenet C, Leung R, Westman E, Simmons A, Dobson R, Sattlecker M, Lupton M, Lunnon K, Keohane A, Ward M, Pike I, Zucht HD, Pepin D, Zheng W, Tunnicliffe A, Richardson J, Gauthier S, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, Vellas B, Lovestone S (2014) Plasma proteins predict conversion to dementia from prodromal disease. *Alzheimers Dement* **10**, 799-807.

[12] Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C (1989) The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* **39**, 1159-1165.

[13] Kiddle SJ, Sattlecker M, Proitsi P, Simmons A, Westman E, Bazenet C, Nelson SK, Williams S, Hodges A, Johnston C, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, Vellas B, Newhouse S, Lovestone S, Dobson RJ (2014) Candidate blood proteome markers of Alzheimer's disease onset and progression: A systematic review and replication study. *J Alzheimers Dis* **38**, 515-531.

[14] Sattlecker M, Kiddle SJ, Newhouse S, Proitsi P, Nelson S, Williams S, Johnston C, Killick R, Simmons A, Westman E, Hodges A, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, Vellas B, Lovestone S; AddNeuroMed Consortium, Dobson RJ (2014) Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimers Dement* **10**, 724-734.

[15] Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, Carter J, Dalby AB, Eaton BE, Fitzwater T, Flather D, Forbes A, Foreman T, Fowler C, Gawande B, Goss M, Gunn M, Gupta S, Halladay D, Heil J, Heilig J, Hicke B, Husar G, Janjic N, Jarvis T, Jennings S, Katilius E, Keeney TR, Kim N, Koch TH, Kraemer S, Kroiss L, Le N, Levine D, Lindsey W, Lollo B, Mayfield W, Mehan M, Mehler R, Nelson SK, Nelson M, Nieuwlandt D, Nikrad M, Ochsner U, Ostroff RM, Otis M, Parker T, Pietrasiewicz S, Resnicow DI, Rohloff J, Sanders G, Sattin S, Schneider D, Singer B, Stanton M, Sterkel A, Stewart A, Stratford S, Vaught JD, Vrkljan M, Walker JJ, Watrobka M, Waugh S, Weiss A, Wilcox SK, Wolfson A, Wolk SK, Zhang C, Zichi D (2010) Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004.

[16] Lourdusamy A, Newhouse S, Lunnon K, Proitsi P, Powell J, Hodges A, Nelson SK, Stewart A, Williams S, Kloszewska I, Mecocci P, Soininen H, Tsolaki M, Vellas B, Lovestone S; AddNeuroMed Consortium, Dobson R, Alzheimer's Disease Neuroimaging Initiative (2012) Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum Mol Genet* **21**, 3719-3726.

[17] Proitsi P, Lupton MK, Velayudhan L, Newhouse S, Fogh I, Tsolaki M, Daniilidou M, Pritchard M, Kloszewska I, Soininen H, Mecocci P, Vellas B; Alzheimer's Disease Neuroimaging Initiative, Williams J; GERAD1 Consortium, Stewart R, Sham P, Lovestone S, Powell JF (2014) Genetic predisposition to increased blood cholesterol and triglyceride lipid levels and risk of Alzheimer disease: A Mendelian randomization analysis. *PLoS Med* **11**, e1001713.

[18] Voyle N, Keohane A, Newhouse S, Lunnon K, Johnston C, Soininen H, Kloszewska I, Mecocci P, Tsolaki M, Vellas B, Lovestone S, Hodges A, Kiddle S, Dobson RJ (2016) A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis. *J Alzheimers Dis* **49**, 659-669.

[19] Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127.

[20] Simmons A, Westman E, Muehlboeck S, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Wahlund LO, Soininen H, Lovestone S, Evans A, Spenger C (2011) The AddNeuroMed framework for multi-centre MRI assessment of Alzheimer's disease: Experience from the first 24 months. *Int J Geriatr Psychiatry* **26**, 75-82.

[21] Du P, Kibbe WA, Lin SM (2008) lumi: A pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547-1548.

[22] Birkenbihl C, Emon MA, Vrooman H, Westwood S, Lovestone S; AddNeuroMed Consortium, Hofmann-Apitius M, Fröhlich H, Alzheimer's Disease Neuroimaging Initiative (2020) Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia - lessons for translation into clinical practice. *EPMA J* **11**, 367-376.

[23] Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS (2004) Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105-114.

[24] Tremblay-Mercier J, Madjar C, Das S, Dyke SO, Étienne P, Lafaille-Magnan M, Bellec P, Collins DL, Rajah MN, Bohbot VD, Leoutsakos J, Iturria-Medina Y, Kat J, Hoge RD, Gauthier S, Chakravarty MM, Poline J, Rosa-Neto P, Villeneuve S, Evans AC, Poirier J, Breitner JCS, the PREVENT-AD Research Group (2020) Creation of an open science dataset from PREVENT-AD, a longitudinal cohort study of pre-symptomatic Alzheimer's disease. *bioRxiv*; doi: https://doi.org/10.1101/2020.03.04.976670

[25] Solomon A, Kivipelto M, Molinuevo JL, Tom B, Ritchie CW EPAD Consortium (2019) European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): Study protocol. *BMJ Open* **8**, e021017.

[26] Morris JC, Aisen PS, Bateman RJ, Benzinger TL, Cairns NJ, Fagan AM, Ghetti B, Goate AM, Holtzman DM, Klunk WE, McDade E, Marcus DS, Martins RN, Masters CL, Mayeux R, Oliver A, Quaid K, Ringman JM, Rossor MN, Salloway S, Schofield PR, Selsor NJ, Sperling RA, Weiner MW, Xiong C, Moulder KL, Buckles VD (2012) Developing an international network for Alzheimer research: The Dominantly Inherited Alzheimer Network. *Clin Investig (Lond)* **2**, 975-984.

[27] Koychev I, Lawson J, Chessell T, Mackay C, Gunn R, Sahakian B, Rowe JB, Thomas AJ, Rochester L, Chan D, Tom B, Malhotra P, Ballard C, Chessell I, Ritchie CW, Raymont V, Leroi I, Lengyel I, Murray M, Thomas DL, Gallacher J, Lovestone S (2019) Deep and Frequent Phenotyping study protocol: An observational study in prodromal Alzheimer's disease. *BMJ Open* **9**, e024498.

[28] Oliveira JL, Trifan A, Bastião Silva LA (2019) EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data. *Int J Med Inform* **126**, 35-45.

[29] Gallacher J, de Reydet de Vulpillieres F, Amzal B, Angehrn Z, Bexelius C, Bintener C, Bouvy JC, Campo L, Diaz C, Georges J, Gray A, Hottgenroth A, Jonsson P, Mittelstadt B, Potashman MH, Reed C, Sudlow C, Thompson R, Tockhorn-Heidenreich A, Turner A, van der Lei J, Visser PJ, ROADMAP Consortium (2019) Challenges for optimizing real-world evidence in Alzheimer's disease: The ROADMAP Project. *J Alzheimers Dis* **67**, 495-501.

A.4 Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling

Reprinted with permission from "Birkenbihl, C.[†], Salimi, Y.[†], Fröhlich, H., Japanese Alzheimer's Disease Neuroimaging Initiative, and Alzheimer's Disease Neuroimaging Initiative. (2022). Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimer's & Dementia*, 18(2), 251-261.".

Received: 16 December 2020 | Revised: 19 April 2021 | Accepted: 25 April 2021

DOI: 10.1002/alz.12387

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

FEATURED ARTICLE

Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling

Colin Birkenbihl[1,2] | Yasamin Salimi[1,2] | Holger Fröhlich[1,2] | for the Japanese Alzheimer's Disease Neuroimaging Initiative[#] | the Alzheimer's Disease Neuroimaging Initiative[†]

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

[2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Correspondence
Colin Birkenbihl, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53757 Sankt Augustin, Germany.
E-mail: colin.birkenbihl@scai.fraunhofer.de

Colin Birkenbihl and Yasamin Salimi contributed equally to this work.

[#] Data used in preparation of this article were obtained from the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) database deposited in the National Bioscience Database Center Human Database, Japan (Research ID: hum0043.v1, 2016). As such, the investigators within J-ADNI contributed to the design and implementation of J-ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of J-ADNI investigators can be found at: https://humandbs.biosciencedbc.jp/en/hum0043-j-adni-authors.

[†] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Funding information
European Union's Horizon 2020, Grant/Award Number: 826421

Abstract

Introduction: Given study-specific inclusion and exclusion criteria, Alzheimer's disease (AD) cohort studies effectively sample from different statistical distributions. This heterogeneity can propagate into cohort-specific signals and subsequently bias data-driven investigations of disease progression patterns.

Methods: We built multi-state models for six independent AD cohort datasets to statistically compare disease progression patterns across them. Additionally, we propose a novel method for clustering cohorts with regard to their progression signals.

Results: We identified significant differences in progression patterns across cohorts. Models trained on cohort data learned cohort-specific effects that bias their estimations. We demonstrated how six cohorts relate to each other regarding their disease progression.

Discussion: Heterogeneity in cohort datasets impedes the reproducibility of data-driven results and validation of progression models generated on single cohorts. To ensure robust scientific insights, it is advisable to externally validate results in independent cohort datasets. The proposed clustering assesses the comparability of cohorts in an unbiased, data-driven manner.

KEYWORDS
Alzheimer's disease, cohort study, data mining, data-driven, disease modeling, machine learning, sampling bias, statistical learning, translational research

1 | BACKGROUND

In the last decade, understanding the progressive dynamics of Alzheimer's disease (AD) and AD clinical syndrome,[1] proved to be one of the fundamental challenges in our field.[2,3] Comprehensive knowledge on AD progression opens crucial opportunities for medical intervention to counteract or delay impediments to activities of daily living.[4] One path to facilitate this understanding manifests in the extraction of longitudinal progression signals from patient-level datasets collected in cohort studies. In this context, data mining and machine learning methods can be used to build mathematical models that elucidate and predict progression patterns hidden in the data. In the past, such progression models were used, for example, to approximate biomarker trajectories,[5] to identify distinct progression subtypes,[6] and to assess patient risk of progression toward more impaired disease stages.[7] However, to demonstrate that progression patterns identified in one cohort generalize beyond the discovery dataset itself, it is imperative to externally validate them in an independent dataset.[8] External validation data should originate from a separate cohort study independent from the training data used for building the model. Especially in the context of multifactorial and heterogenous diseases such as AD, external validation turns out to be a non-trivial undertaking.

The key limitation encountered in external validation manifests in the characteristics of clinical AD cohort data.[9] By nature of the disease, AD cohorts are very heterogeneous with respect to their exhibited progression,[10] for example, with respect to brain atrophy[11] and age of disease onset.[12] Furthermore, cohort study participants are recruited according to specific inclusion and exclusion criteria defined based on the goals of the study (e.g., selection of specific age ranges or risk factors). These specific sampling procedures shape potentially distinct statistical distributions from which each study's participants are recruited and, in turn, inevitably introduce cohort-specific statistical biases into the collected dataset itself.[13,14] These aspects potentially violate the fundamental assumption behind data mining and machine learning approaches that the participants of a validation dataset constitute a representative sample of the same population from which the original training data were drawn (Figure S1 in supporting information). Consequently, this indicates that training and validation data must be independently and identically distributed (i.i.d.) samples.[15] As such, a well-trained model should show similar performance on a validation dataset that was drawn from the identical statistical distribution as the training data, while an overfitted model would fail such validation. However, on a validation dataset that is violating the assumption of being sampled from the same statistical distribution as the training data even a well-trained model would fail, because the validation data falls outside the domain of the model (Figure S1). In conclusion, data-driven models trained on cohort datasets cannot be expected to generalize appropriately beyond the statistical distribution from which this cohort's participants were sampled.[16,17]

The heterogeneity found in AD cohort datasets, therefore, raises several important questions with respect to data-driven modeling of AD. First, it warrants an evaluation as to whether exhibited trends of disease progression are consistent across cohorts despite possible differences in their underlying populations. Further investigation should also determine whether progression models fitted on such datasets learn potential cohort-specific biases that could impede the generalizability of findings. Finally, as of now, there is no way to measure and express the general comparability between patient-level datasets on the level of disease progression. In the past, researchers mainly relied on comparing baseline study characteristics of their studied datasets.[7,18,19] However, for obvious reasons, evaluating variable distributions at a singular time point is a very limited comparison in the scope of disease progression. Deriving a quantitative measure to compare longitudinal progression patterns across multiple clinical studies could aid researchers to better understand the landscape of existing studies and to identify datasets that might fulfill the i.i.d. assumption. Furthermore, it could be used to investigate whether the cause of a significant drop in prediction performance lies in systematic differences between the training and validation datasets (i.e., a probable violation of the i.i.d. assumption) or simply in an overfitted model.

In this work, we evaluated the heterogeneity of disease progression patterns encountered in six longitudinal clinical AD cohort studies. Relying on multi-state models (MSM),[20] a well-established data mining approach in the AD field,[7,21–24] we performed a systematic comparison of progression patterns extracted from these studies to assess whether discovered signals are robust. Furthermore, we investigated whether cohort-specific biases propagate into trained

RESEARCH IN CONTEXT

1. **Systematic review**: The authors reviewed relevant literature using standard bibliographic search engines. Accessible cohort datasets have been discovered through data portals and citations in literature (primarily https://adata.scai.fraunhofer.de/).
2. **Interpretation**: The presented results illustrate the comparability of Alzheimer's disease (AD) progression across six major AD cohorts. We identified evident differences in progression patterns between cohorts and, furthermore, observed that data-driven approaches learn cohort-specific effects from their training data. These findings can impede the generalization of results generated on single cohorts. We propose a novel clustering approach for cohort data that helps to better understand which cohorts are comparable with respect to their exhibited disease progression.
3. **Future directions**: This work emphasizes the need for thorough validation of data-driven results. To eventually support clinical decision-making using data-driven approaches, it might be more promising to build models specific for disease subtypes or use domain adaptation techniques to address the encountered heterogeneity in cohort datasets.

progression models. Finally, we propose a novel method for clustering cohorts based on their exhibited progression patterns. This approach reveals the similarity of cohort studies in a data-driven and unbiased manner. It allows researchers to adequately understand and characterize performances measured via external validation of statistical and machine learning models developed on another cohort. In conclusion, our approach allows for better understanding of statistical differences that have previously been reported between various AD studies.[13]

2 | METHODS

2.1 | Data selection

Six longitudinal datasets stemming from the Alzheimer's Disease Neuroimaging Initiative (ADNI),[25] AddNeuroMed (ANMerge),[26] Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL),[27] Japanese Alzheimer's Disease Neuroimaging (J-ADNI),[28] National Alzheimer's Coordinating Center (NACC),[29] and the Religious Orders Study and Rush Memory and Aging Project (ROSMAP)[30] were used as training datasets for our progression models. All of these studies obtained ethical approval for human data collection and informed patient consent for data sharing. We excluded participants whose mild cognitive impairment (MCI) diagnoses were not attributed to AD. Information on the cohorts with respect to key variables, as well as the number of participants, can be found in Table S1 in supporting information.

2.2 | Progression models applied for statistical analysis

To extract disease progression patterns from the investigated datasets, we fitted one MSM per cohort using the msm R package.[20] The states in our models represent the three commonly assessed stages for AD progression: cognitively unimpaired (CU), MCI, and AD. Consequently, transitions between states illustrate conversions from one clinical diagnosis stage to another. We modeled AD as an absorbing state, that is, we assumed that patients were not able to recover once deterioration was advanced enough to receive an AD diagnosis. However, because the classification of patients into CU, MCI, and AD in all cohorts had been performed based on clinical assessments, reversions from AD were observed in the data. These reversions were modeled as misclassifications. A graphical representation of the model can be seen in Figure S3 in supporting information. Each transition rate was estimated based on a set of covariates to account for the individual compositions of the cohorts. For determining the most informative combination of covariates, we performed a rigorous model selection using the Akaike's information criterion (AIC). The choice of covariates was mainly limited by their availability across the cohorts (Figure S2 in supporting information). Ultimately, the selected covariates comprised participant's age, biological sex, completed years of education, apolipoprotein E (*APOE*) ε4 status, and the Mini-Mental State Examination (MMSE). Likelihood-ratio tests comparing each MSM to a null model demonstrated that all models extracted progression signals from their training dataset ($P < .05$). To rule out potential overfitting of the models, we built 150 models on repeated bootstrap samples from each respective cohort and observed low variation in model estimates (Table S3 in supporting information). Application of interval censoring allowed for the inclusion of participants with missing intermediate visits while right censoring was used for individuals who did not receive an AD diagnosis during study runtime. More details on the methodology and model selection are presented in the supporting information.

2.3 | Comparison of data mined progression patterns across cohorts

To explore and assess the heterogeneity in disease progression trends across cohorts, we estimated several progression patterns using each cohort's MSM: the state transition probabilities, probability of staying AD diagnosis free over time, and sojourn times (i.e., the expected time a participant spends in a considered state). All patterns were separately investigated for the CU and MCI states. For estimation of a cohort's progression patterns starting in the CU state, we used the covariate values observed at the study baseline of each of the respective cohort's CU participants. Similarly, for estimating transitions from the MCI state, we relied on the covariate values of participants at their first MCI diagnosis. Where appropriate, uncertainty of estimates was quantified using 95% confidence intervals (CI). Differences between cohort-specific distributions of the aforementioned progression estimates were determined using Kruskal-Wallis and pairwise Mann-Whitney *U* tests employing a confidence level of 95%. *P*-values were corrected for multiple testing using the Bonferroni-Holm method.

2.4 | Evaluation of cohort biases in statistical models

The second set of analyses aimed at elucidating whether MSMs fitted to data from a single cohort would learn cohort-specific effects that reduce generalizability to other cohorts. Hazard ratios, for example, are covariate-specific parameters of a model that quantify the influence of covariates onto the transition risk between two states. Comparing these ratios, it becomes apparent whether models learned the same covariate influences from distinct cohorts. Furthermore, we used each cohort's previously trained MSM to estimate the progression patterns for the same, combined set of participants from all cohorts. By fixing the data to be estimated across models, all variability in the progression patterns stems from the cohort-specific effects learned by the model. To evaluate the existence of these cohort-specific biases, we performed Kruskal-Wallis tests and pairwise Mann-Whitney

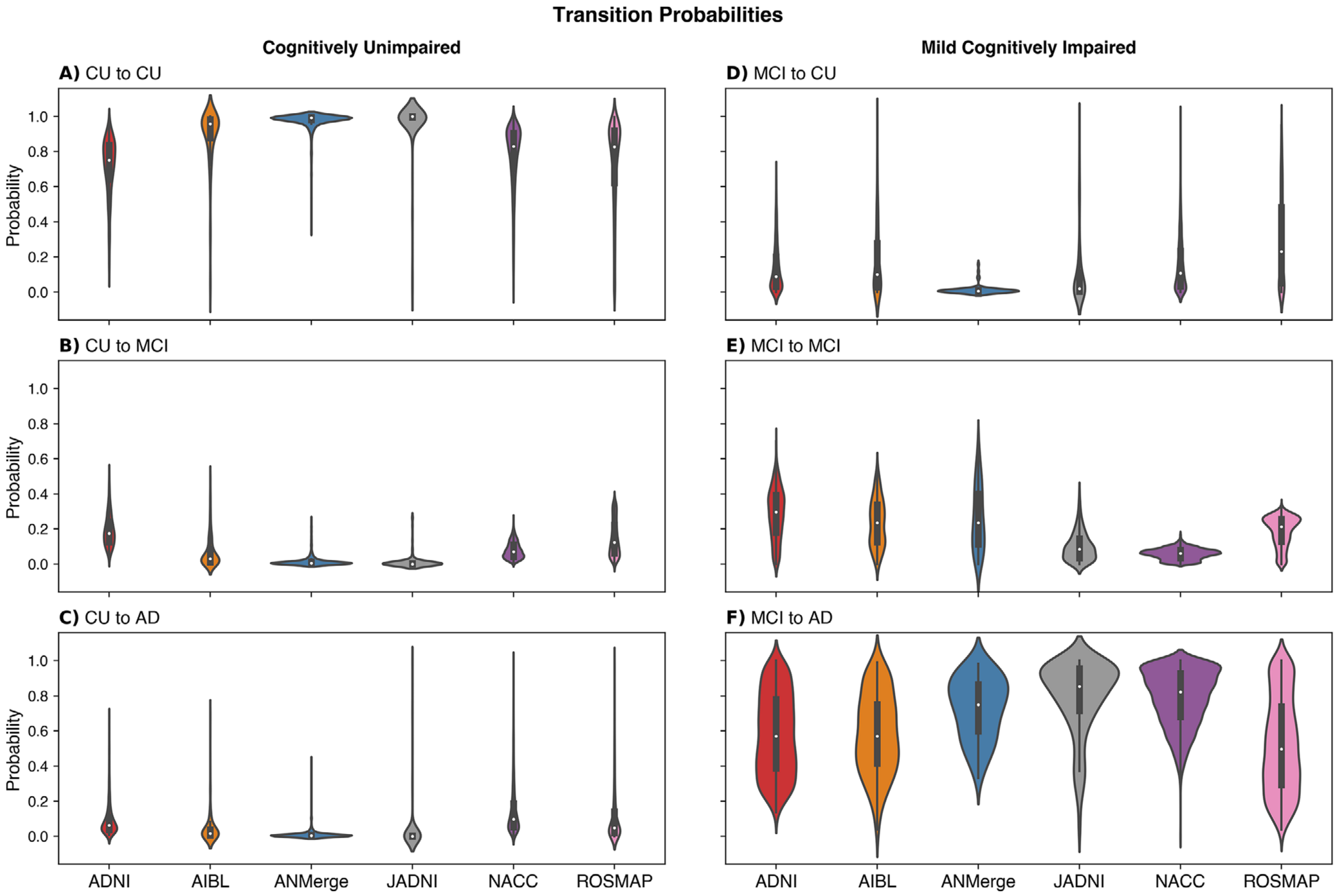


FIGURE 1 Probabilities to transition from one state to another are estimated for a 10-year period. Median probabilities are marked with white points. Statistical distributions are shown as box plots as well as superimposed kernel density estimates, resulting in violin plots. Because most deviations between depicted distributions were significant, we omit indication of significance for brevity. A-C, Transition probabilities starting from the cognitively unimpaired (CU) state. D-F, Transition probabilities starting from the mild cognitive impairment (MCI) state. AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

U tests, again correcting for multiple testing using Bonferroni-Holm and assuming a confidence level of 95%.

2.5 | Cohort similarity clustering

Whereas previous analyses focused on statistical differences between cohorts, we additionally developed an approach to cluster cohorts based on their global similarity across progression patterns. More specifically, each cohort's MSM was used to calculate the log-likelihood of observing the actual transitions of all the participants of each other cohort. These pairwise log-likelihoods were afterward averaged across the number of participants per cohort to eliminate biases toward cohort size. This resulted in a pairwise similarity matrix between cohorts which was subsequently transformed into a symmetric distance matrix. Mathematical details can be found in the supporting information. The resulting distance matrix was then used in an agglomerative hierarchical clustering approach using average linkage.

3 | RESULTS

3.1 | Progression patterns differ across cohorts

Transition probabilities estimated for a 10-year period varied significantly between cohorts (Figure 1). While we observed in all cohorts that participants in the CU state were most likely to remain CU over the next 10 years, the proportions of probabilities showed evident differences (Figure 1A-C). We discovered a range of 25% difference between the maximum and minimum observed median probability to remain CU (J-ADNI, > 99%; ADNI, 75%). All observed differences between pairwise combinations of cohorts were significant ($P < .001$), with the exception of ROSMAP–NACC for remaining in the CU state ($P = .3$).

When investigating the estimated transition probabilities from the MCI state (Figure 1D-F), all cohorts exhibited their most probable transition toward the AD state. J-ADNI showed the highest median probability across cohorts with 85%, while ROSMAP held the lowest median probability with 50%, exposing a difference of 35% between them.

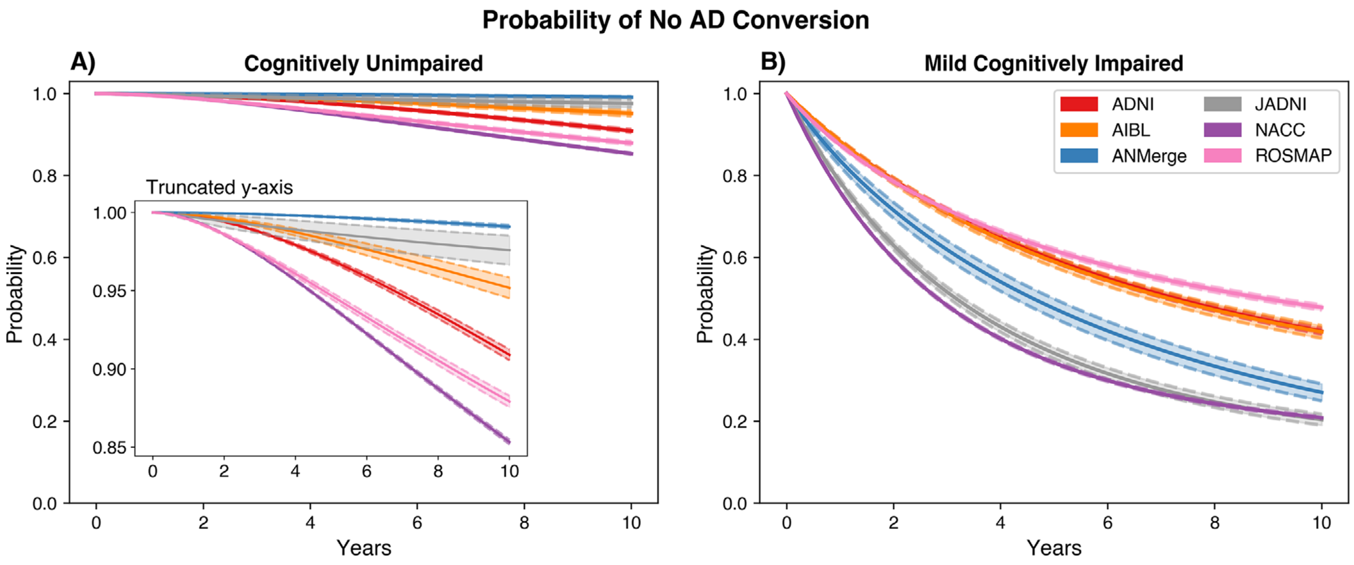


FIGURE 2 Average probability of staying AD diagnosis free over time for each cohort. Dashed lines indicate the standard errors of the estimates. A, Starting from cognitively unimpaired. B, Starting from mild cognitive impairment. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

Additionally, compared to the other cohorts, ROSMAP showed a considerably higher median probability to revert from MCI back to CU of 23%. All pairwise differences across cohorts proved to be significant ($P < .001$). Numerical values for the transition probabilities are presented in Table S4 in supporting information.

In concordance with the transition probabilities, the probability of staying AD diagnosis free over time differed substantially across cohorts. Starting in the CU state (Figure 2A), the trajectories of cohorts deviated significantly after approximately 4 years. NACC and ROSMAP exhibited the steepest decline (respectively, 85% and 87% after 10 years), while the probability for ANMerge stayed relatively stable (99%). Considering the MCI state as a starting point, the probability of remaining AD diagnosis free exhibited a steeper decline (Figure 2B). After 10 years, the most extreme estimates were made for ROSMAP (48%) and J-ADNI (20%), while no significant differences were observed between J-ADNI and NACC (both 20%), as well as between AIBL and ADNI (both 42%). Ultimately, we discovered a maximum deviation of 14% for the CU state and 28% for the MCI state.

All pairwise comparisons between the cohorts' sojourn time estimates turned out to be significant for the CU state ($P < .001$, with exception of ADNI–ROSMAP, $P < .05$; Figure 3A). Given their respective MSMs, ROSMAP displayed the shortest sojourn time with a median of 27.5 years, followed by ADNI (29.7 years), NACC (38.7 years), AIBL, ANMerge, and J-ADNI (all > 100 years). In the MCI state, again, most deviations were found to be significant ($P < .001$; Figure 3B). The only exception to this was ANMerge, which did not differ significantly from ADNI ($P = .9$) and AIBL ($P = .88$). The median sojourn time in the MCI state showed relatively lower values for J-ADNI (3.8 years) and NACC (3.1 years), while ADNI, AIBL, and ANMerge showed relatively higher values (7.7, 6.5, and 6.9 years, respectively). ROSMAP is placed in between with a median of 5 years. Detailed descriptions of the sojourn times distributions can be found in Table S5 in supporting information.

3.2 | Comparison of cohort-specific models

In the second set of analyses, we explored the cohort-specific biases learned by our MSMs from their respective training datasets. We observed that the cohort-specific models learned significantly different relationships between covariate values and the disease progression. Non-overlapping CIs indicated significant differences in hazard ratios for the transition from CU to MCI between ROSMAP (CI: 1.05 to 1.1), NACC (1.0 to 1.04), and ADNI (0.86 to 0.99) regarding education level. With respect to the MMSE, significant differences were found for ROSMAP, NACC, J-ADNI, and ADNI (CIs: 0.60 to 0.67, 0.76 to 0.81, 0.11 to 0.58, and 0.76 to 0.98, respectively; Figure 4A). The influence of education in J-ADNI (CI: 1.15 to 1.92) differed significantly from ADNI (0.93 to 1.12), NACC (0.94 to 1.04), and ROSMAP (0.93 to 1.03) with respect to reverting from MCI to CU (Figure 4B). Regarding the conversion from MCI to AD, significant differences were discovered in the hazard ratios for age between ROSMAP (1.02 to 1.05) and NACC (1.00 to 1.01), for *APOE* ε4 status between NACC (1.10 to 1.31) and ADNI (1.34 to 1.82), and for MMSE between NACC (0.83 to 0.87), ADNI (0.7 to 0.76), and ROSMAP (0.74 to 0.79; Figure 4C). In several cases, large CIs hampered the interpretation of the hazard ratios. The exact estimates of all hazard ratios are presented in Table S6 in supporting information.

When applying each MSM to the same set of data, the difference in the estimated progression patterns across models resembled the consequences of the learned cohort-specific biases (Figure 5). Numerical descriptions of the distributions in Figure 5 can be found in

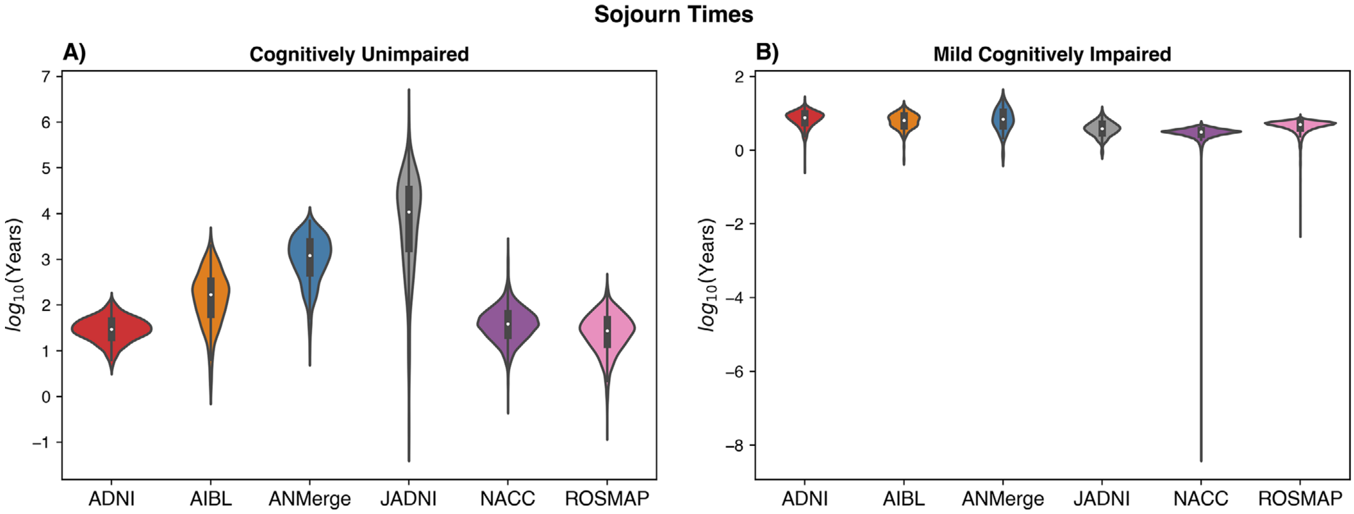


FIGURE 3 Sojourn times of cohort participants on a log10-scale. Because most deviations between depicted distributions were significant, we omit indication of significance for brevity and refer to the text. A, Occupying the cognitively unimpaired state. B, Occupying the mild cognitive impairment state. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

Tables S6 and S7 in supporting information. For all evaluated patterns (i.e. the transition probabilities, Figure 5A; sojourn times, Figure 5B; and estimated probability of staying AD diagnosis free, Figure 5C), significant Kruskal-Wallis tests underlined the presence of cohort-specific effects ($P < .001$). Additional pairwise comparisons using Mann-Whitney *U* tests are presented in the supporting information. We observed that naive pooling of datasets and training models on a combination of multiple, complete cohorts expectedly biases the estimates toward the cohort with the largest sample size (Figure S4 in supporting information).

We also found differences between cohorts when extracting progression patterns for a cohort's representative individual (Figure S5 in supporting information) and even when applying the same exemplary patients to each cohort's specific MSM (Figure S6 in supporting information).

3.3 | Clustering reveals overall similarity of studies

Figure 6 presents the results achieved by clustering the investigated cohorts based on the similarity of their progression patterns. ANMerge, AIBL, and NACC displayed close proximity indicating that their participants exhibited similar disease progression in combination with their trained MSMs. Furthermore, ADNI and J-ADNI formed a cluster that connected with the previously mentioned cluster in relatively high distance. ROSMAP was placed far from all other cohorts, constituting its own cluster.

4 | DISCUSSION

In this work, we explored the heterogeneity in AD progression across multiple, independent cohort datasets and the implications for data-driven approaches for progression modeling. Evident differences in mined progression patterns surfaced between six investigated cohorts. This finding raises concerns regarding the reliability of results discovered in single data resources and underlines the need for external validation. Furthermore, we demonstrated that models learn cohort-specific effects from their training dataset, which can impede model generalization. Last, we proposed a novel approach to identify similar cohort datasets that could help to find datasets that come closer to fulfilling the i.i.d. assumption. We demonstrated this approach by highlighting how six major AD cohorts relate to each other with regard to their exhibited disease progression.

4.1 | Progression trends differ across cohort datasets

Analyzing the characteristic progression trends extracted from the investigated cohorts revealed substantial differences among them. The observation of lower variability in estimates for the CU state compared to the MCI state can be explained by the fact that only a fraction of the CU participants will eventually develop cognitive symptoms. Thus, a substantial amount of CU participants are expected to show no signals of AD progression at all. Overall, the discovered heterogeneity could likely stem from differences in the recruitment processes of cohort studies. Compositional shifts across sampled populations pose a critical confounder comparing cohort datasets and model performance.[13] Here, statistical matching could potentially help to identify comparable subsets.

4.2 | Data-driven models learn systematic biases present in cohort datasets

Using all cohort-specific MSMs to estimate progression patterns for the same set of participants revealed the presence of strong

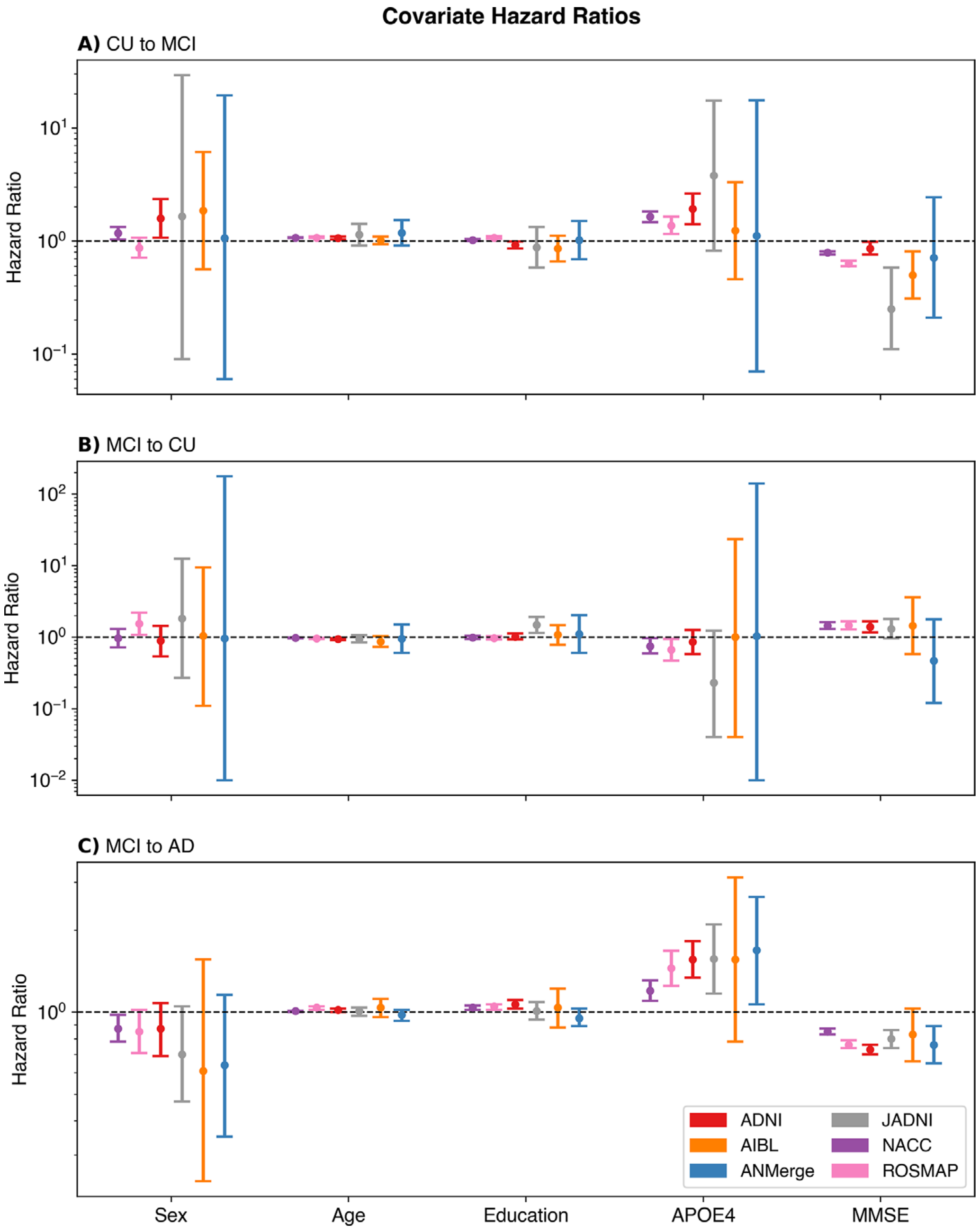
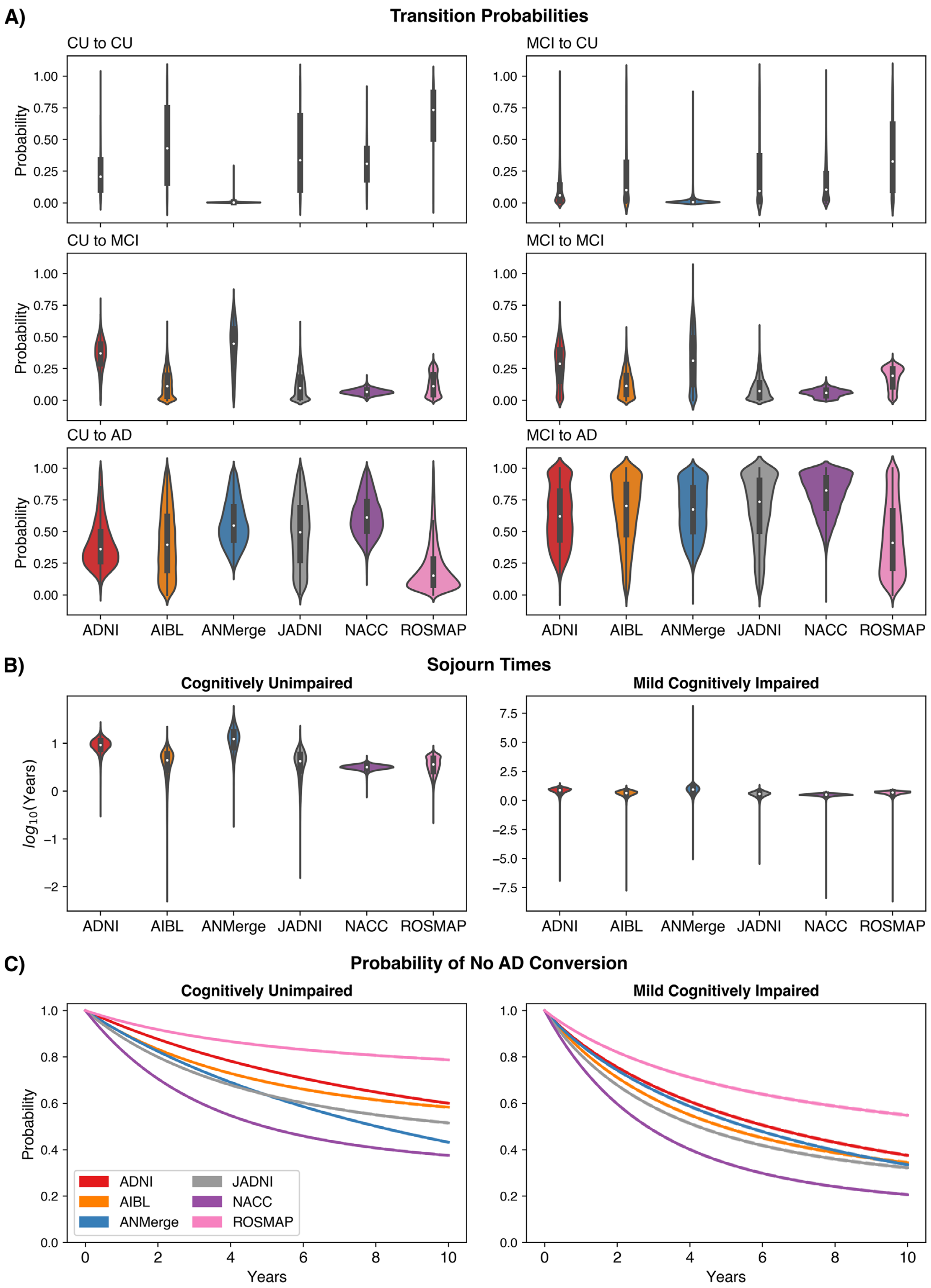


FIGURE 4 Covariate hazard ratios learned per cohort-specific multi-state models. For readability, significant deviations are not indicated visually. Instead, we refer to the text for the corresponding evaluations. A, B, C, Impact on transition from cognitively unimpaired (CU) to mild cognitive impairment (MCI), MCI to CU, and MCI to Alzheimer's disease (AD), respectively. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

cohort-specific effects that the models learned from their training datasets. The estimated covariate hazard ratios are an integral component of the cohort-specific progression signals and while we could observe commonalities in the directional influence of covariates, partially described by previous studies as well,[7,21] the magnitude of these influences exposed several significant differences. With regard to education, even contradicting influences were found. Differences in such fundamental parameters of a model propagate into, and thereby bias, their estimates; this became apparent in the subsequently estimated progression patterns.

Naive pooling of data from several cohorts does not necessarily pose a solution for addressing the biases but leads to an overshadowing of signals in smaller cohorts by larger ones. Instead, more considerate methods must be applied, such as sampling the same number of participants from each cohort, weighting of subjects to favor smaller datasets, or ensemble techniques that combine


A)
Transition Probabilities
CU to CU
MCI to CU
CU to MCI
MCI to MCI
CU to AD
MCI to AD
Probability
1.00
0.75
0.50
0.25
0.00
ADNI
AIBL
ANMerge
JADNI
NACC
ROSMAP
B)
Sojourn Times
Cognitively Unimpaired
Mild Cognitively Impaired
log_{10}(Years)
1
0
−1
−2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
C)
Probability of No AD Conversion
Cognitively Unimpaired
Mild Cognitively Impaired
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Years
ADNI
AIBL
ANMerge
JADNI
NACC
ROSMAP

dataset-specific models. Future work should explore these options in more detail.

4.3 | Clustering allows assessment of cohort similarities

Our proposed approach to measure cohort similarity with regard to their global disease progression trends (informed by neuropsychological tests, biological sex, completed years of education, *APOE* ε4 status) elicited commonalities across cohorts that mirror the design of these studies. Finding ADNI and J-ADNI in one cluster together is reassuring as J-ADNI was designed as a complementary cohort to ADNI, and similar trends have been observed in both cohorts.[28] Their use of equal eligibility criteria for participant recruitment counteracts the risk of sampling from two distinct populations. The distance we observe between them could be explained partially due to differences in ethnoracial composition[31] and lifestyle.[32] ROSMAP, on the other hand, is a special case in the landscape of AD cohorts. Its participants are exclusively recruited from religious orders, are considerably older, and hold a higher proportion of female participants compared to the other cohorts.[13,30]

Our proposed method enables a quantitative description of differences across cohorts and, subsequently, an evaluation of cohort similarity based not only on cross-sectional values of covariates but on their general progression. Consequently, it could help researchers to better understand and characterize performance measures obtained during the external validation of machine learning models. More specifically, our cohort clustering can be used post hoc to indicate whether failed validation was likely caused by overfitting or systematic biases between discovery and validation cohort originating from, for example, sampling of distinct statistical distributions.

4.4 | Limitations

It is unknown how many of the CU participants per cohort would have eventually developed cognitive symptoms during their lifetime. While the models account for this factor using censoring, estimates based on the CU participants could be biased depending on the size of the participant fraction with prodromal AD.

One limitation of MSMs is the assumption that disease progression depends only on the current state of a participant. While this is a necessary and widely accepted assumption in the literature,[7,21–24] there is no universal way to prove that it always holds true for all possible state transitions.

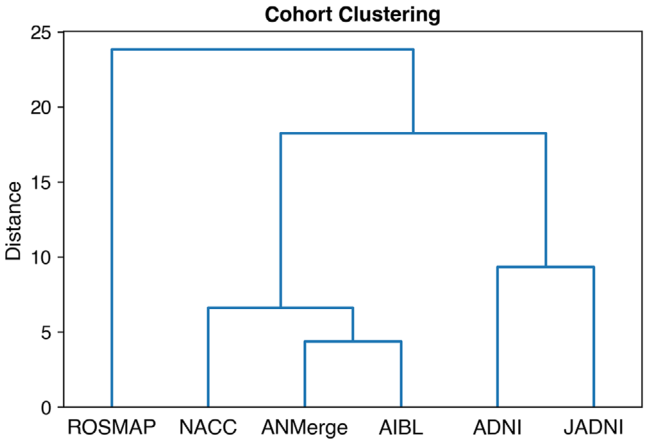


FIGURE 6 Cohort dendrogram resulting from the clustering of pairwise log-likelihoods. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

In recent years, AD is more considered a biological entity[1] and while we aimed to account for as many clinically relevant covariates as possible, we were unable to include emerging biomarkers in our MSMs. Given the limited number of individuals participating in longitudinal biomarker collection, the inclusion of biomarkers would have led to underpowered models and reduced the number of cohorts available for analysis. However, using this limited set of covariates, our model selection showed that all chosen covariates added meaningful information to the models and that progression signals could successfully be learned.

5 | CONCLUSION

Applying machine learning and statistical modeling to single data resources can bias results and might render the generalizability of the models used infeasible. Ideally, it would be imperative that we go beyond single data resources and instead investigate and validate findings across the landscape of AD data we have at our disposal. In practice, however, external validation of data-driven machine learning models is often limited by the availability of semantically and statistically comparable datasets.[13] For some investigations only single cohorts might be suitable. While results originating from such single-cohort investigations hold value as initial indications, they should be (1) regarded as cohort-specific findings pending external validation, and

FIGURE 5 Consequences of learned cohort-specific biases onto estimated progression patterns. The same set of participants was considered under each cohort's trained multi-state models (i.e., variability in estimates stems from the models, not the data). Deviations between estimates illustrate the learned biases. Because most deviations between depicted distributions were significant, we omit indication of significance for brevity. AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; CU, cognitively unimpaired; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; MCI, mild cognitive impairment; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

(2) meticulously validated internally. Here, resampling techniques and cross-validation can help to increase the robustness of single cohort studies.[8]

Dealing with such heterogeneous data as is encountered in our field, building a single model that serves all predictive purposes and is applicable to the general AD population seems inconceivable. Instead, the more promising alternative to support clinical decision-making using data-driven approaches for AD and dementia could be to build subpopulation-specific models that embrace the specifics of their target group. Here, the stratification of the AD population into specific progression subtypes could guide which model is applicable to which patient. Alternatively, artificial intelligence methods from the field of domain adaptation (e.g., transfer learning) might help to manage the heterogeneous signals when applying models across cohorts.

ACKNOWLEDGMENTS

We thank the study participants and staff of the Rush Alzheimer's Disease Center. ROSMAP was supported by NIA grants P30AG010161, R01AG015819, and R01AG017917.

J-ADNI was supported by the following grants: Translational Research Promotion Project from the New Energy and Industrial Technology Development Organization of Japan; Research on Dementia, Health Labor Sciences Research Grant; Life Science Database Integration Project of Japan Science and Technology Agency; Research Association of Biotechnology (contributed by Astellas Pharma Inc., Bristol-Myers Squibb, Daiichi-Sankyo, Eisai, Eli Lilly and Company, Merck-Banyu, Mitsubishi Tanabe Pharma, Pfizer Inc., Shionogi & Co., Ltd., Sumitomo Dainippon, and Takeda Pharmaceutical Company), Japan, and a grant from an anonymous foundation.

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private-sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 826421, "TheVirtualBrain-Cloud."

CONFLICTS OF INTEREST

The authors have nothing to declare.

AUTHOR CONTRIBUTIONS

Colin Birkenbihl and Holger Fröhlich designed the study. Yasamin Salimi and Colin Birkenbihl implemented the methods and ran the experiments. Colin Birkenbihl wrote the manuscript. Holger Fröhlich and Yasamin Salimi revised the manuscript. Holger Fröhlich supervised the project.

REFERENCES

1. Jr Jack RC, Bennett DA, Blennow K, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14(4):535-562.
2. Jr Jack RC, Knopman DS, Jagust WJ. et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol*. 2013;12(2):207-216.
3. Winblad B, Amouyel P, Andrieu S, et al. Defeating Alzheimer's disease and other dementias: a priority for European science and society. *Lancet Neurol*. 2016;15(5):455-532.
4. Sperling RA, Jack CR, Aisen PS. Testing the right target and right drug at the right stage. *Sci Transl Med*. 2011;3(111):111cm33-111cm33.
5. Hadjichrysanthou C, Evans S, Bajaj S, et al. The dynamics of biomarkers across the clinical spectrum of Alzheimer's disease. *Alzheimers Res Ther*. 2020;12(1):1-16.
6. de Jong J, Emon MA, Wu P, et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *Gigascience*. 2019;8(11):giz134.

7. Vermunt L, Sikkes SA, Van Den Hout A, et al. Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype. *Alzheimers Dement*. 2019;15(7):888-898.
8. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16(1):150.
9. Golriz Khatami S, Robinson C, Birkenbihl C, Domingo-Fernández D, Hoyt CT, Hofmann-Apitius M. Challenges of integrative disease modeling in Alzheimer's disease. *Front Mol Biosci*. 2020;6:158.
10. Ryan J, Fransquet P, Wrigglesworth J, Lacaze P. Phenotypic heterogeneity in dementia: a challenge for epidemiology and biomarker studies. *Front Public Health*. 2018;6:181.
11. Habes M, Grothe MJ, Tunc B, McMillan C, Wolk DA, Davatzikos C. Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods. *Biol Psychiatry*. 2020.
12. Jacobs D, Sano M, Marder K, et al. Age at onset of Alzheimer's disease: relation to pattern of cognitive dysfunction and rate of decline. *Neurology*. 1994;44(7):1215-1215.
13. Birkenbihl C, Salimi Y, Domingo-Fernándéz D, et al. Evaluating the Alzheimer's disease data landscape. *Alzheimers Dement*. 2020;6(1):e12102.
14. Birkenbihl C, Emon MA, Vrooman H, et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA J*. 2020;11(3):367-376.
15. Vapnik V. *Statistical Learning Theory*. New York: Wiley; 1998:624.
16. Ben-David S, Blitzer J, Crammer K, Pereira F, (2007). Analysis of representations for domain adaptation. *In advances in neural information processing systems* (pp. 137-144). MIT press.
17. Sun B, Feng J, Saenko K, (2016). Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 2058-2065). AAAI Press.
18. Whitwell JL, Wiste HJ, Weigand SD, et al. Comparison of imaging biomarkers in the Alzheimer disease neuroimaging initiative and the Mayo Clinic Study of Aging. *Arch Neurol*. 2012;69(5):614-622.
19. Ferreira D, Hansson O, Barroso J, et al. The interactive effect of demographic and clinical factors on hippocampal volume: a multicohort study on 1958 cognitively normal individuals. *Hippocampus*. 2017;27(6):653-667.
20. Jackson CH. Multi-state models for panel data: the msm package for R. *J stat softw*. 2011;38(8):1-29.
21. Robitaille A, van den Hout A, Machado RJ, et al. Transitions across cognitive states and death among older adults in relation to education: a multistate survival model using data from six longitudinal studies. *Alzheimers Dement*. 2018;14(4):462-472.
22. Zhang L, Lim CY, Maiti T, et al. Analysis of conversion of Alzheimer's disease using a multi-state markov model. *Stat Methods Med Res*. 2019;28(9):2801-2819.
23. Brookmeyer R, Abdalla N. Estimation of lifetime risks of Alzheimer's disease dementia using biomarkers for preclinical disease. *Alzheimers Dement*. 2018;14(8):981-988.
24. Jr Jack CR, Therneau TM, Wiste HJ, et al. Rates of transition between amyloid and neurodegeneration biomarker states and to dementia among non-demented individuals: a population-based cohort study. *Lancet Neurol*. 2016;15(1):56.
25. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *AlzheimersDement*. 2005;1(1):55-66.
26. Birkenbihl C, Westwood S, Shi L, et al. ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *J Alzheimers Dis*. 2021;79(1):423-431.
27. Ellis KA, Bush AI, Darby D, et al. The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr*. 2009;21(4):672-687.
28. Iwatsubo T, Iwata A, Suzuki K, et al. Japanese and North American Alzheimer's disease neuroimaging initiative studies: harmonization for international trials. *Alzheimers Dement*. 2018;14(8):1077-1087.
29. Besser L, Kukull W, Knopman DS, et al. Version 3 of the national Alzheimer's coordinating center's uniform data set. *Alzheimer Dis Assoc Disord*. 2018;32(4):351.
30. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimers Dis*. 2018;64(s1):S161-S189.
31. Babulal GM, Quiroz YT, Albensi BC, et al. Perspectives on ethnic and racial disparities in Alzheimer's disease and related dementias: update and areas of immediate need. *Alzheimers Dement*. 2019;15(2):292-312.
32. Xu W, Tan L, Wang HF, et al. Meta-analysis of modifiable risk factors for Alzheimer's disease. *J Neurol, Neurosurg Psychiatry*. 2015;86(12):1299-1306.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Birkenbihl C, Salimi Y, Fröhlich H. Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimer's Dement*. 2022;18:251–261. https://doi.org/10.1002/alz.12387

A.5 Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice

Reprinted with permission from “Birkenbihl, C., Emon, M. A., Vrooman, H., Westwood, S., Lovestone, S., AddNeuroMed Consortium, , Hofmann-Apitius M., Fröhlich, H., and Alzheimer’s Disease Neuroimaging Initiative. (2020). Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA Journal*, 11, 367-376.”.

EPMA Journal (2020) 11:367–376
https://doi.org/10.1007/s13167-020-00216-z

RESEARCH

Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia - lessons for translation into clinical practice

Colin Birkenbihl[1,2] · Mohammad Asif Emon[1,2] · Henri Vrooman[3,4] · Sarah Westwood[5] · Simon Lovestone[5] · On behalf of the AddNeuroMed Consortium · Martin Hofmann-Apitius[1,2] · Holger Fröhlich[1,2,6] · Alzheimer's Disease Neuroimaging Initiative

Received: 11 April 2020 / Accepted: 5 June 2020 / Published online: 22 June 2020

Abstract

Artificial intelligence (AI) approaches pose a great opportunity for individualized, pre-symptomatic disease diagnosis which plays a key role in the context of personalized, predictive, and finally preventive medicine (PPPM). However, to translate PPPM into clinical practice, it is of utmost importance that AI-based models are carefully validated. The validation process comprises several steps, one of which is testing the model on patient-level data from an independent clinical cohort study. However, recruitment criteria can bias statistical analysis of cohort study data and impede model application beyond the training data. To evaluate whether and how data from independent clinical cohort studies differ from each other, this study systematically compares the datasets collected from two major dementia cohorts, namely, the Alzheimer's Disease Neuroimaging Initiative (ADNI) and AddNeuroMed. The presented comparison was conducted on individual feature level and revealed significant differences among both cohorts. Such systematic deviations can potentially hamper the generalizability of results which were based on a single cohort dataset. Despite identified differences, validation of a previously published, ADNI trained model for prediction of personalized dementia risk scores on 244 AddNeuroMed subjects was successful: External validation resulted in a high prediction performance of above 80% area under receiver operator characteristic curve up to 6 years before dementia diagnosis. Propensity score matching identified a subset of patients from AddNeuroMed, which showed significantly smaller demographic differences to ADNI. For these patients, an even higher prediction performance was achieved, which demonstrates the influence systematic differences between cohorts can have on validation results. In conclusion, this study exposes challenges in external validation of AI models on cohort study data and is one of the rare cases in the neurology field in which such external validation was performed. The presented model represents a proof of concept that reliable models for personalized predictive diagnostics are feasible, which, in turn, could lead to adequate disease prevention and hereby enable the PPPM paradigm in the dementia field.

Keywords Predictive preventive personalized medicine (3 PM/PPPM) · Disease risk prediction · Cohort data · Model validation · Machine learning · Disease modeling · Artificial intelligence · Individualized patient profiling · Interdisciplinary · Multiprofessional · Risk modeling · Survival analysis · Bioinformatics · Alzheimer's disease · Neurodegeneration · Precision medicine · Cohort comparison · Health data · Medical data · Data science · Translational medicine · Digital clinic · Propensity score matching · Sampling bias · Model performance · Dementia

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13167-020-00216-z) contains supplementary material, which is available to authorized users.

✉ Colin Birkenbihl
colin.birkenbihl@scai.fraunhofer.de

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany

[2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, 53115 Bonn, Germany

[3] Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Center, Rotterdam, Netherlands

[4] Department of Medical Informatics, Erasmus MC University Medical Center, Rotterdam, Netherlands

[5] Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK

[6] UCB Biosciences GmbH, Alfred-Nobel Str. 10, 40789 Monheim am Rhein, Germany

Introduction

Dementia is a disease manifesting in cognitive decline of patients which ultimately leads to an inability to perform activities of daily living. Subsequently, patients are in need of full-time professional care. With an increasingly aging population, it is estimated that in 2050 there will be 1.5 billion dementia cases worldwide [1]. The economic implications are tremendous: as of now, annually $600 billion are spent on dementia care globally, surpassing the costs of cancer and heart disease, and without adequate treatment or prevention, expenses will further increase [2].

Dementia is a progressive disease that likely onsets years before cognitive symptoms arise. Treating patients who are already exhibiting cognitive symptoms shows only limited success [3, 4]. Accordingly, it has been proposed to transition to the paradigm of personalized, predictive, and preventive medicine (PPPM) in order to treat patients in pre-symptomatic dementia stages, when irreversible brain damages have not yet occurred (i.e., when patients are cognitively healthy or mild cognitive impaired, MCI, the prodromal stage of dementia) [5–8]. However, pre-symptomatic dementia diagnosis remains difficult, as reliable prognostic biomarkers have yet to be developed, and therefore, up to date, diagnosis is still mainly based on cognitive function [8].

Artificial intelligence as a powerful instrument to implement PPPM approach

Methods from the field of artificial intelligence (AI), and more specifically machine learning, pose a great opportunity to drive the transition towards the PPPM paradigm [9]. These methods involve the use of biomedical data to build (i.e., "train") models which are capable of addressing a plethora of problems encountered in health research: Given a suitable data, they can assist diagnosis [10], model disease progression [11], identify patient subgroups for stratification [12], analyze survival chances [13], assist disease monitoring, and support appropriate therapies and medication [14].

Often, these approaches conglomerate into one crucial aspect: they model and predict disease-relevant aspects on a personalized level and can incorporate multimodal biomedical signals as predictors. Especially these personalized predictions substantiate why AI strategies are of such relevance to the PPPM paradigm.

Pre-symptomatic personalized dementia risk diagnosis

In the context of pre-symptomatic diagnosis, so-called AI-based disease risk models allow for predicting personalized risk years of patients, before onsetting cognitive symptoms will lead to a dementia diagnosis by a clinician. The potential of these models is an earlier identification and subsequent treatment of patients, which likely increases the chances of preventing or slowing down disease progression [15]. Several factors contributing to dementia risk are known and can be used as predictors. These contain unmodifiable patient characteristics such as biological sex, age, APOEε4 allele status, and dementia-linked single nucleotide polymorphisms (SNPs) [16–18]. Additionally, a variety of modifiable variables are known to affect dementia risk such as education, physical activity, and smoking [18]. Disease risk models can combine these predictive features to estimate the personalized dementia risk of an individual. This leads to highly multivariate models that do not only rely on single biomarkers.

Implications of training models on cohort data: the need for validation studies

The basis for training and validation of such machine learning models are data that usually originate from a particular study (e.g., observational cohort studies). Two landmark studies in the dementia field are the Alzheimer's Disease Neuroimaging Initiative (ADNI) [19] and AddNeuroMed [20]. ADNI is one of the worldwide largest dementia cohorts that displays an unmatched degree of deep multimodal phenotyping and longitudinal follow-up. Among others, it is funded by the National Institutes of Health (NIH) and is the most referential dementia data resource with more than 1300 citations. By sharing their complete dataset, ADNI represents a prime example in the context of open science and has enabled groundbreaking advancements in dementia research. Likewise, AddNeuroMed is up to date the largest European dementia cohort and involved participants coming from six sites all across the European Union [21]. It was the first project funded by the Innovative Medicine Initiative (IMI) and paved the way for the employed joint public-private funding scheme. Like ADNI, AddNeuroMed shares all collected patient-level data with third-party researchers.

In our earlier work [22], we have used data from ADNI to develop a machine learning model that predicts an individual patient's risk to be diagnosed with dementia. In an internal validation, the model showed a strong performance when sequentially leaving out parts of the ADNI data from model training and using them as a test set in a nested cross-validation. However, a grand challenge in biomedicine is that clinical studies are never representative of the entire population [23], since they are inherently biased by their study design. These biases can be caused by multiple reasons, some of which are inclusion and exclusion criteria, types of collected data, or sampling and laboratory procedures. Therefore, an important question is how far an artificial intelligence model trained with data from one study can generalize (i.e., achieve sufficiently high prediction performance) to patients from another study. For this purpose, the model has to be tested on independent data. This process is called external validation. External validation is usually done retrospectively and can be understood as the first step of the long-lasting validation process [24]. The steps would comprise a prospective validation study, approval as a diagnostic tool by a regulatory agency, and finally a utility assessment, which has to carefully compare the economic costs with the achievable benefit for the patient.

To enable the paradigm shift towards AI-supported translational PPPM approaches, an adequate model validation is vital. Here, a core aspect of machine learning theory is that the training and validation data are drawn from the same underlying statistical distribution. If the training data and the validation dataset originated from two significantly different populations, validation can fail because the model is not familiar with the specific values it is presented with, even though it has successfully learned the distribution underlying the training data. Therefore, a critical question is how to quantify and decide whether a patient from an external validation study is sufficiently comparable with the original training data, given the study protocols were similar. This is an essential prerequisite for an artificial intelligence model to make reliable predictions. More broadly, any kind of statistical analysis derived from two independent studies for the same medical research question is confronted with the same issue: Only if a sufficiently similar subset of patients can be identified, statistics can be expected to be directly comparable. For example, if patients differ significantly in their age distribution in two dementia studies, their cognitive impairment scores cannot be directly compared. However, a suitable subset of patients out of both studies may be identifiable that are in the same age range and thus allow for a less biased comparison.

State of the art: cohort comparisons and dementia risk prediction

Few evaluations of the comparability of longitudinal cohort studies in the dementia domain have been made [25, 26]. All of these works focused only on a small subset of dementia-relevant features and were based on a reduced patient subset of their investigated cohorts. In conclusion, there is an unmet need for a systematic in-depth comparison of cohorts in the dementia field.

Since the appearance of our model publication, a number of alternative machine learning algorithms for predicting dementia risk have been suggested [27–30]. Our model differentiates from those, because it is able to predict dementia risk as a function of time. Additionally, to the best of our knowledge, none of the other models were externally validated.

Novelty beyond the state of the art

Our presented work makes two major contributions: first, we statistically analyzed the differences between two important dementia cohort studies, namely, ADNI and AddNeuroMed, in order to understand and characterize their relative sampling biases. We demonstrate that substantial differences between both studies exist in demographic, clinical, and MRI features, raising concerns regarding the generalizability of statistical analysis results and more complex modeling efforts that have solely used one of these datasets. As a second major contribution, we show that, despite the existing differences between both studies, external validation of our earlier developed dementia risk model [22] demonstrated a high prediction performance of disease diagnosis (AUC = 0.81) up to 6 years before made by a clinician. To explore the effect of systematic differences between cohorts on validation performance, we used propensity score matching (PSM) [31] to identify a subset of AddNeuroMed patients which are sufficiently similar to ADNI participants with respect to a subset of key demographic features. For those subjects, an even higher prediction performance of 88% AUC was achieved, which illustrates that systematic sampling biases can significantly influence the prediction performance of AI-based models in PPPM.

We would like to highlight that, to the best of our knowledge, our model is the only artificial intelligence-based dementia risk model that has been externally validated so far (AUC = 0.81). Hence, we see the external validation of our model as an important contribution of this work, which demonstrates that, instead of solely relying on symptomatic diagnosis, a validated PPPM approach in the dementia domain is feasible.

Material and methods

Clinical studies and investigated features

We selected two major dementia cohorts (i.e., ADNI and AddNeuroMed) for comparison and artificial intelligence model validation. Both studies were conducted following the Declaration of Helsinki and informed consent of participants was acquired. In order to compare the selected cohorts, and to be in a position to apply an artificial intelligence model trained on ADNI data to patients from AddNeuroMed, we first had to identify variables which were jointly available in both studies. Because demographic variables are usually well defined and clinical and MRI procedures in AddNeuroMed were aligned to ADNI protocols [20, 21], we focused on demographic, clinical, and MRI variables in our comparison. In addition, we had to ensure that brain volumes were calculated identically for both cohorts. Therefore, we reprocessed raw MRI images from ADNI and AddNeuroMed using the same pipeline and brain parcellation method (see Supplementary Material). In total, 200 variables were measured in both studies and could be compared with each other. Determined by AddNeuroMed, the longest available follow-up we could investigate spanned 84 months.

Propensity score matching

Statistical matching or PSM is a procedure used to identify comparable patients from two cohorts. The goal is to assign patients of one cohort an individual counterpart from another dataset such that the matched pair is comparable with regard to a specified set of matching features. Classically, PSM has been used to study treatment effects outside the framework of randomized controlled trials [32], e.g., in pharma-epidemiology [33].

Matching two dementia cohorts based on sex, age, APOEε4 status, and education level of patients will result in two sub-cohorts that are similar to each other with respect to the distribution of these matching features. PSM starts by fitting a logistic regression model which discriminates between patients of two cohorts. One class represents patients from study 1 (i.e., ADNI) and the other class study 2 (i.e., AddNeuroMed), and predictors or matching features are those clinical variables for which differences between these studies should be eliminated. The logistic regression results in a propensity score per patient in both cohorts (Fig. 1A). The score thereby represents the probability of a patient to belong to study 1. In a second step, this propensity score is then used to find suitable matching partners of ADNI patients in AddNeuroMed.

One way this can be done, which we followed here, uses the concept of a caliper [34]. For a given ADNI patient X, an AddNeuroMed patient Y is accepted as a matching partner, if their propensity score differs by at most a certain fraction of standard deviations of the propensity score. If multiple matching partners are available within the caliper range, one is selected randomly, with resampling being usually not permitted. Participants for whom no partner from the other cohort could be found within the caliper range are discarded.

The caliper can thus potentially significantly affect the matching. Althauser et al. reported that a caliper of 1 standard deviation removes approximately 75% of the initial bias, while a caliper of 0.2 can remove 98% [34]. We tested different calipers for matching: 1.5, 1.3, 1, 0.7, 0.5, 0.3, and 0.1. For each of those calipers, 100 matchings were performed and the matching quality was assessed (Supplementary Fig. 1, 2, and Supplementary Table 1). Based on this evaluation, we here decided on a caliper of 1.

To conduct PSM, we used the R package MatchIt [35]. As matching features, we selected patient age, sex, the number of full-time education years, and APOEε4 allele count. After PSM, the resulting sub-cohorts should show comparable characteristics with respect to these variables.

Statistical cohort comparisons

We performed a comparison of ADNI and AddNeuroMed for each baseline diagnosis group separately (healthy, MCI, dementia), one before and one after PSM. We evaluated whether PSM was able to eliminate differences between ADNI and AddNeuroMed with respect to chosen matching features. Furthermore, we also investigated how PSM influenced the differences in features not matched for. To ensure robust results, we compared features for 100 matchings and set the results against those gained from comparing features in 100 randomly selected patient subgroups of the same sample size.

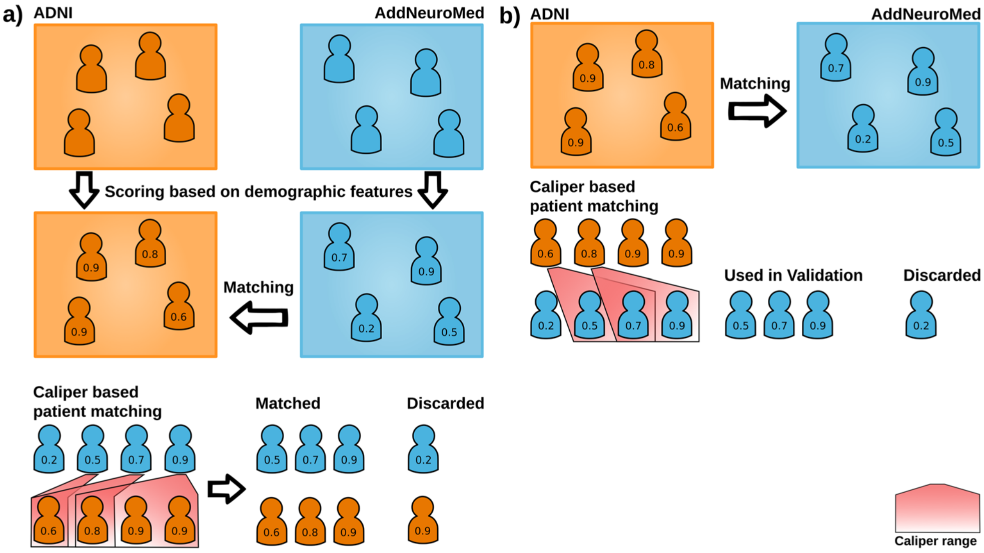


Fig. 1 Caliper-based propensity score matching. (**A**) Procedure of caliper-based nearest neighbor propensity score matching as it was used in the comparison of ADNI and AddNeuroMed. The first step in the matching process is the calculation of a propensity score for each patient, followed by the matching of patients based on a caliper. The results are two cohorts consisting of patients similar with respect to the chosen matching features. Patients without match are discarded. (**B**) Caliper-based PSM as it was used for model validation. Only AddNeuroMed patients that found a match in ADNI were kept and used to validate the dementia risk model

The amount of matched/randomly selected patients from each diagnosis group can be seen in Table 1.

We declared a continuous feature to be significantly different between the two cohorts if the 95% confidence interval of the difference between the population means (after correction for multiple testing via Bonferroni's method) did not cover 0. For categorical variables (such as sex or APOEε4 status), we estimated the 95% confidence interval for the difference in proportions of each variable category (e.g., 0, 1, 2 APOEε4 risk alleles). We assessed the absolute number of significant deviations for each diagnosis cohort separately. Due to the randomness involved in the matching procedure, we repeated the comparisons 100 times, each with newly matched sub-cohorts. To evaluate if the number of found differences in matched subgroups is significantly lower than the number of differences found between random subsamples, we applied a one-tailed Wilcoxon test using an alpha level of 5%.

Since PSM cannot deal with missing data, only cases that were complete with regard to the chosen matching features were considered. After excluding incomplete cases and conducting the matching, the ADNI and AddNeuroMed subcohorts consisted of 199 healthy controls, 147 MCI patients, and 150 dementia cases each (Table 1 "Match").

Validation of an artificial intelligence-based model to predict dementia diagnosis

In our previous work [22], we proposed an artificial intelligence model based on stochastic gradient boosted decision trees (GBM) [36] for predicting the time-dependent risk of a patient to convert from a healthy or MCI state to diagnosed dementia. The model was originally trained on data from 315 cognitively normal and 609 MCI ADNI participants. Fourteen (4.4%) of the normal and 238 (39%) of the MCI patients developed dementia during the 96 months in the study. GBMs inherently perform a feature selection in the training process, which ultimately leads to sparse models. The final predictors used in the model included clinical baseline information (e.g., diagnosis, age, sex, education, and cognition scores), glucose uptake (FDG), amyloid β deposit (AV45), brain volumes (36 variables),s and genotype (APOEε4 status, 100 dementia associated SNPs, 116 polygenic pathway impact scores, and 32 principal components describing genetic variability based on 53014 SNPs within each individual). Prediction performance was assessed via 10 times repeated 10-fold cross-validation, resulting in a Harrell's C-index of ~ 0.86. Briefly, Harrell's C-index is a generalization of the area under the ROC curve for classification and ranges from 0 to 1, where 0.5 indicates chance level [37]. More details regarding our published model, including a comparison against several competing AI models, can be found in [22].

Table 1 Sample size reduction when applying PSM to ADNI and AddNeuroMed

	Healthy			MCI			Dementia		
Cohort	*n*	CC	Matched	*n*	CC	Matched	*n*	CC	Matched
ADNI	417	415	199	872	866	147	342	338	150
ANM	793	266	199	397	238	147	512	262	150

n number of cases before PSM, *CC* number of complete cases with regard to the matching features, *Matched* number of matched patients following the approach depicted in Fig. 1A, *MCI* mild cognitive impaired

Since not all features used in the original model were present in AddNeuroMed, we had to restrict ourselves to the CDRSB (clinical dementia rating scale sum of boxes score) and MMSE (Mini-Mental State Examination) total scores as cognition assessments. In consequence, a revised AI model (stochastic gradient boosted decision trees—GBM) had to be trained on ADNI data. The training and subsequent evaluation procedure was identical to the one published in [22] and is described in the Supplementary Material in more detail.

In our case, the revised GBM model achieved a lower cross-validated C-index than our original one of ~ 0.83 (Supplementary Fig. 3 and Supplementary Table 2). Due to the restriction on features available in both cohorts, the revised model contained fewer features (n = 32) than the original one. It included 24 MRI-derived volumes of different brain regions, age, CDRSB, MMSE, baseline diagnosis (i.e., MCI or cognitively normal), 3 principal components describing genetic variance within each individual (computed from the same set of SNPs as in our original model), APOEε4 status, and 1 dementia-associated SNP (rs7364180) in the coiled-coil domain containing 134 gene (CCDC134). This revised model was subsequently evaluated on cognitively normal and MCI AddNeuroMed patients.

In addition, we investigated whether the AI model would yield better prediction performance on a subset of AddNeuroMed subjects that were more similar to ADNI patients with regard to their demographics. For that purpose, we performed PSM as shown in Fig. 1B. Based on ADNI, we scored AddNeuroMed patients and included those participants into a validation dataset who received an ADNI matching partner based on our matching variables. Additionally, baseline MMSE was included to correct for differences in cognitive impairment. No a priori stratification by baseline diagnosis was performed before PSM to avoid overoptimism. After matching, we further only included patients for whom MRI images were available. This limited the highest achievable number of validation participants to 244. The resulting average-matched validation cohort contained 164 AddNeuroMed patients of which 20 converted to dementia during the runtime of the study (Supplementary Fig. 2). To ensure that our results were robust, we repeated the validation process for 100 matchings.

Results

ADNI and AddNeuroMed differ significantly in key features

The presence of fundamental differences between ADNI and AddNeuroMed became evident by performing a comparison of the unmatched, full diagnosis groups. Table 2 shows an overview of the demographic characteristics of the two cohorts. With the control group as an exception, AddNeuroMed patients are on average roughly 3 years older than the ADNI population. In AddNeuroMed, the proportion of women is higher and in general there are fewer APOEε4 carriers. The most prominent difference could be observed in the education of study participants. On average, healthy ADNI participants underwent at least 4 years more education, and the cognitively affected cases showed a difference of almost 6 years compared with AddNeuroMed participants.

We could identify 200 features from the clinical, imaging, and demographic modalities that were common between ADNI and AddNeuroMed. In total, 48, 136, and 138 out of the 200 common features differed substantially between the controls, MCI, and dementia patients, respectively (Table 3 "Unmatched"). These results underline the presence of significant differences between both cohorts.

Propensity score matching allows for identifying comparable subjects

PSM resulted in ~ 363 patients from AddNeuroMed that could principally be matched to ADNI following the PSM protocol in Fig. 1B. Keeping only patients for which MRI data was available led to a dataset comprising on average 164 patients. In Fig. 2, we show the distribution of propensity scores before and after PSM. The shift to more similar distributions after PSM highlights that differences in age, sex, MMSE, education, and APOEε4 status between matched patients from both studies are evidently reduced. Evaluation of individual confidence intervals of those features showed similar results, since significant differences observed in the matching features before PSM vanished after (Supplementary Fig. 4), the education of participants being the only exception. Hence, PSM allows for identifying more comparable subjects from AddNeuroMed with respect to key features.

Additionally, we explored whether PSM would reduce the number of significantly different features that were not used as matching variables. This was done by running 100 PSMs, each selecting the amount of matched patients previously shown in Table 1. We then compared the selected subsamples of ADNI and AddNeuroMed to identify significant differences in non-matching variables. This was done (a) in the matched subsamples and (b) as a control in 100 randomly selected patient subsets from both cohorts, which included the same number of patients as selected by PSM.

We found that the number of significantly different variables was, on average, reduced to 22 (± 7 std. dev.; i.e., reduction by 15%), 23 (± 10 std. dev.; i.e., reduction by 67%), and 17 (± 4 std. dev.; i.e., reduction by 66%) for controls, MCI, and dementia patients compared with the random samples (Table 3). Comparing the number of significant differences found in random samples and matched samples using a Wilcoxon test showed that the reduction was significant in all cases (healthy, $p = 0.001$; dementia and MCI, $p < 0.001$). This finding can be explained by the fact that matching variables are correlated with further variables.

Table 2 Demographic composition of ADNI and AddNeuroMed per diagnosis

	Age	Females (%)	Education	0 *APOEε4* (%)	1 *APOEε4* (%)	2 *APOEε4* (%)
Cognitively normal controls						
ADNI	74.8	49.9	16.3	74.9	24.8	2.7
ANM	74.5	59.4	12.3	74.6	23.2	2.2
CI	[-1.33, 0.82]	**[0.02, 0.17]**	**[-4.56, -3.37]**	[-0.05, 0.09]	[-0.09, 0.05]	[-0.03, 0.02]
Mild cognitive impairment						
ADNI	73.0	40.9	15.9	49.7	39.4	10.9
ANM	76.0	54.7	10.0	60.4	35.8	3.8
CI	**[1.81, 4.25]**	**[0.06, 0.21]**	**[-6.39, -5.32]**	**[0.02, 0.19]**	[-0.12, 0.05]	**[-0.11, -0.03]**
Dementia						
ADNI	75.0	44.7	15.2	33.5	47.3	19.2
ANM	78.6	62.9	9.4	45.7	41.3	13.0
CI	**[2.17, 4.97]**	**[0.1, 0.27]**	**[-6.48, -5.1]**	**[0.03, 0.21]**	[-0.15, 0.03]	[-0.13, 0.0]

Average age and education are reported in years. *CI* multiple testing adjusted 95% confidence interval of the difference in means for education and age, and for the difference in proportions for Female and APOEε4 status. Significant intervals are emboldened. *0, 1, 2 APOEε4* fraction of individuals with 0, 1, or 2 APOEε4 alleles. *Females* proportion of female study participants. *ANM* AddNeuroMed

Table 3 Number of significant differences between ADNI and AddNeuroMed

Diagnosis	Unmatched	Random	Matched (mean, SD)	% rel. change (mean)	*p* value	Min	Max
Controls	48	26 (10.4)	22 (6.8)	-15	0.001	11	40
Mild cognitive impaired	136	67 (22.4)	23 (10.0)	-67	< 0.001	4	47
Dementia	138	66 (22.4)	17 (4.3)	-66	< 0.001	8	30

Unmatched number of features found significant by comparing the complete unmatched diagnosis groups. *Random* number of features found significant by comparing random subsamples with the same sample size as the matched subgroups. *Matched* mean number of significant differences found across all 100 matching and comparison runs. Standard deviation in brackets. *% rel. change* relative change in the number of significant features with and without PSM. *Min* minimal number of significant differences found in a single run. *p value p* value indicating if the amount of significant differences in matched subgroups is significantly lower compared with the random sample. *Max* maximal number of significant differences found in a single run

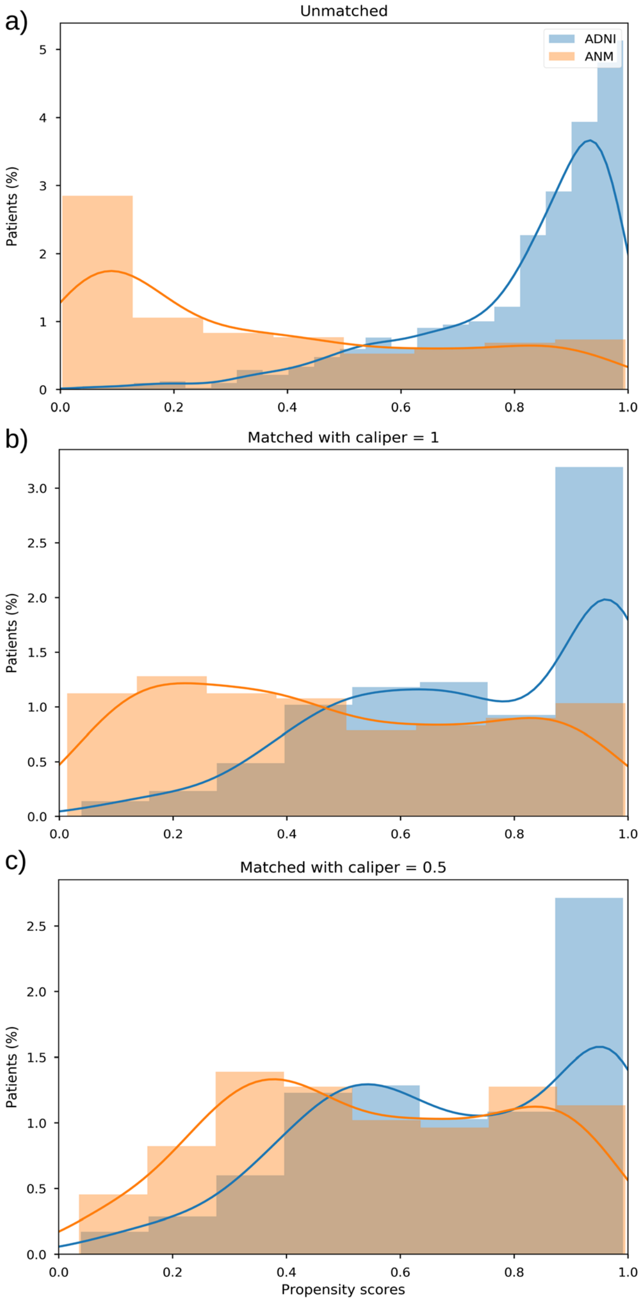


Fig. 2 Distribution of propensity scores before and after PSM. (**A**) Scores for the full unmatched cohorts. (**B**) Scores for matched patients using a caliper of 1. (**C**) Scores for matched patients using a caliper of 0.5. *ANM* AddNeuroMed

Supplementary Fig. 4 shows which features differed consistently between AddNeuroMed and ADNI.

Artificial intelligence model shows high prediction performance in external validation

We initially applied our artificial intelligence-based dementia risk model to all cognitively normal and mild cognitively impaired AddNeuroMed participants with available MRI data ($n = 244$, 30 (12%) received the diagnosis “Alzheimer’s disease” during the course of the study). Due to the highlighted differences between ADNI (our training cohort) and AddNeuroMed (our validation set), prediction performance of the model dropped from 0.83 C-index in ADNI to 0.81 C-index in AddNeuroMed (Fig. 3A). In Fig. 3B, we present the prediction performance as the area under receiver operator characteristic curve over time (AUC(t)) to show that our algorithm can predict dementia diagnosis up to 6 years prior to diagnosis with an AUC of ~ 0.8. The observed low prediction performance at month 0 is an artifact, because no conversions can take place at baseline. Likewise, after 6 years, prediction performance drops, because only few observations were available.

For comparison and motivated by the findings in the last section, we next investigated the prediction performance for AddNeuroMed subjects that were putatively similar to ADNI according to PSM. Our model made a prediction for each of the matched AddNeuroMed patients, and we repeated this procedure for 100 different matchings and averaged the performance. This resulted in a significantly higher C-Index of ~ 0.88, which is comparable with the result reported in our earlier publication using cross-validation (Fig. 3A). Similarly, the AUC at 6 years prior to diagnosis increased to ~ 0.88 as well (Fig. 3B). In conclusion, PSM successfully eradicated differences between cohorts by identifying AddNeuroMed subjects that were more similar to those in ADNI.

Discussion

In order to take dementia treatment to the era of PPPM, presymptomatic diagnosis is vital. AI and machine learning

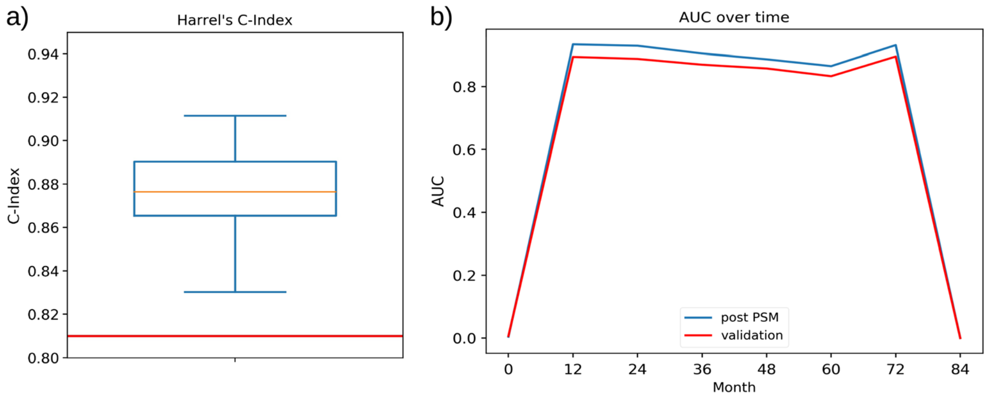


Fig. 3 Performance of the dementia risk model on external validation and matched AddNeuroMed data calculated for 100 different matchings. (**A**) Harrell's C-index of the model. The red line is indicates the model performance on the full unmatched AddNeuroMed cohort. (**B**) Area under the ROC over time (AUC(t)) showing the predictive performance over time before diagnosis. The standard error is plotted around the mean trajectory

methods trained on clinical cohort study data can build a foundation to enable this translation, because they can work with the highly multifactorial nature of dementia and succeed, where single biomarkers are not able to provide a reliably prediction. However, translation of AI models into clinical practice requires a sufficient multi-step validation: (i) an internal validation on the discovery cohort (done in our previous work); (ii) an external validation on a further cohort (done here); (iii) a validation via a prospective clinical study; (iv) an assessment as a diagnostic tool by a regulatory agency; and (v) a careful utility analysis, which includes health economic considerations.

Cohort differences affect model generalizability; predictive dementia diagnosis is possible

This work demonstrated the presence of substantial differences between ADNI and AddNeuroMed, two major dementia cohort studies. Nonetheless, we were able to externally validate our model for personalized dementia risk prediction on the complete AddNeuroMed data, achieving an AUC of ~ 0.81 to predict dementia diagnosis 6 years before diagnosed by a clinician. Due to the identified differences, it is not surprising to observe a lower performance compared with the ~ 0.86 AUC we previously reported on ADNI [22]. Notably, with the help of PSM, we were able to identify a subset of AddNeuroMed subjects that were more comparable with those in ADNI with regard to demographic features. For these matched patients, a significantly higher prediction performance of ~ 0.88 AUC was observed. This again highlights the influence which systematic biases across cohorts can potentially have on the performance of AI-based approaches. We would like to emphasize that this work is one of the very rare cases in the neurology field, in which an AI model was properly validated based on a separate study. As pointed out above, such external validation is crucial to enable a paradigm shift towards an AI-based PPPM paradigm.

In general, the observable differences between ADNI and AddNeuroMed question the generalizability of published statistical analyses that in the past have only used a single dataset. Our concerns are further supported by results of Whitwell et al. [26] as well as Ferreira et al. [25], who also reported significant differences between dementia cohorts. Because there is such a strong bias in cohort data from dementia patients, from our point of view, it is extremely important that scientific findings are tested in independent cohort studies.

Limitations and outlook

For this work, there was only a relatively small number of initially cognitively normal and MCI patients in AddNeuroMed, which later on received the dementia diagnosis (30 out of 244). Hence, additional cohort studies should be employed to further validate the presented AI model. Since each of these studies will have their own biases compared with ADNI as well, such a validation would additionally strengthen the confidence into the model. The next step in order to allow for an implementation of the presented model into a clinical context would be a dedicated prospective study.

Expert recommendations: AI-supported personalized treatments

The crucial role that AI models can play in the process of shifting the diagnosis and treatment of dementia towards the PPPM paradigm stems mainly from their capability of performing personalized predictions. By incorporating patient-specific multivariate information, they provide disease risk assessments for individuals which can potentially impact the time as well as the type of treatment that patients receive. Thereby, reliable AI models can constitute personalized treatment algorithms that open opportunities for critical medical interventions which delay the progression of diseases or even to prevent disease onset at all. Furthermore, AI methods could even suggest the appropriate personalized treatment given the patient specific biomarker signatures. The accompanied reduction in economic costs as well as emotional burden suffered by patients and caregivers would be significant.

Conclusion

Altogether, our work highlighted the inevitable necessity to validate AI models on separate cohort datasets to, at some point, make the translation of AI-based PPPM approaches into clinical routine [6, 24, 38]. Moreover, our work showed the non-trivial challenges that are associated with conducting such efforts. Additional real-world evidence data from clinical practice (e.g., electronic health care records) are now starting to play an increasing role in this context and could potentially help to reduce the cohort selection biases outlined here.

Acknowledgments The authors thank Charles Tapley Hoyt for critical reading of the manuscript.

Code availability All used packages are published and are cited accordingly.

Author contributions Designed the project: CB, MHA, HF. Supervised the project: HF. Retrieved the data: CB, SW, SL. Analyzed the data: CB, MAE, HV, HF. Drafted the manuscript: CB, HF. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Funding information Open Access funding provided by Projekt DEAL. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826421. The research leading to these results has received partial support from the Innovative Medicines Initiative Joint Undertaking under grant agreement #115568, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California.

Data availability All used datasets are findable in official repositories and are cited accordingly. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Compliance with ethical standards

All cohort studies used were conducted following the Declaration of Helsinki and informed consent of participants was acquired.

Conflicts of interests The authors declare that they have no conflict of interest.

Abbreviations ADNI, Alzheimer's Disease Neuroimaging Initiative; AI, artificial intelligence; AUC, area under receiver operator characteristic curve; CDRSB, clinical dementia rating sum of boxes; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; PPPM, predictive preventive personalized medicine; PSM, propensity score matching; SNP, single nucleotide polymorphism

References

1. Prince MJ, Guerchet M, Prina M. The Global Impact of Dementia 2013-2050: Policy Brief for Heads of Government. Alzheimer's Dis Int. 2013.
2. Wimo A, Jönsson L, Bond J, Prince M, Winblad B, International AD. The worldwide economic impact of dementia 2010. Alzheimers Dement. 2013;9(1):1–11.
3. Folch J, Busquets O, Ettcheto M, Sánchez-López E, Castro-Torres RD, Verdaguer E, et al. Memantine for the treatment of dementia: a review on its current and future applications. J Alzheimers Dis. 2018;62(3):1223–40.
4. Mehta D, Jackson R, Paul G, Shi J, Sabbagh M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. Expert Opin Investig Drugs. 2017;26(6):735–9.

5. Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley J, Ames D, et al. Dementia prevention, intervention, and care. Lancet. 2017;390(10113):2673–734.
6. Golubnitschaja O. Neurodegeneration: accelerated ageing or inadequate healthcare? EPMA J. 2010;**1**:211–5. https://doi.org/10.1007/s13167-010-0030-5.
7. Sperling RA, Jack CR, Aisen PS. Testing the right target and right drug at the right stage. Sci Transl Med. 2011;3(111):111 cm33-111 cm33.
8. Mandel, S. (Ed.). Neurodegenerative Diseases: Integrative PPPM Approach as the Medicine of the Future: Springer Science & Business Media; 2013.
9. Barrett M, Boyne J, Brandts J, Brunner-La Rocca HP, De Maesschalck L, De Wit K, et al. Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. EPMA J. 2019:1–20.
10. Zellweger MJ, Tsirkin A, Vasilchenko V, Failer M, Dressel A, Kleber ME, et al. A new non-invasive diagnostic tool in coronary artery disease: artificial intelligence as an essential element of predictive, preventive, and personalized medicine. EPMA J. 2018;9(3):235–47.
11. Fisher CK, Smith AM, Walsh JR. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. Sci Rep. 2019;9(1):1–14.
12. de Jong J, Emon MA, Wu P, Karki R, Sood M, Godard P, et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. GigaScience. 2019;8(11):giz134.
13. Obrzut B, Kusy M, Semczuk A, Obrzut M, Kluska J. Prediction of 5–year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods. BMC Cancer. 2017;17(1):840.
14. Castaneda C, Nalley K, Mannion C, Bhattacharyya P, Blake P, Pecora A, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinf. 2015;5(1):4.
15. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer''s disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984;34(7):939.
16. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31(12):1102.
17. Perera G, Pedersen L, Ansel D, Alexander M, Arrighi HM, Avillach P, et al. Dementia prevalence and incidence in a federation of European Electronic Health Record databases: the European Medical Informatics Framework resource. Alzheimers Dement. 2018;14(2):130–9.
18. Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. Lancet Neurol. 2014;13(8):788–94.
19. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers Dement. 2005;1(1):55–66.
20. Lovestone S, Francis P, Strandgaard K. Biomarkers for disease modification trials-The innovative medicines initiative and AddNeuroMed. J Nutrition Health Aging. 2007;11(4):359.
21. Lovestone S, Francis P, Kloszewska I, Mecocci P, Simmons A, Soininen H, et al. AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Ann N Y Acad Sci. 2009;1180(1):36–46.
22. Khanna S, Domingo-Fernández D, Iyappan A, Emon MA, Hofmann-Apitius M, Fröhlich H. Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms. Sci Rep. 2018;8(1):11173.
23. Lawrence E, Vegvari C, Ower A, Hadjichrysanthou C, De Wolf F, Anderson RM. A systematic review of longitudinal studies which measure Alzheimer's disease biomarkers. J Alzheimers Dis. 2017;59(4):1359–79.
24. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. BMC Med. 2018;16(1):150.
25. Ferreira D, Hansson O, Barroso J, Molina Y, Machado A, Hernández-Cabrera JA, et al. The interactive effect of demographic and clinical factors on hippocampal volume: A multicohort study on 1958 cognitively normal individuals. Hippocampus. 2017;27(6):653–67.
26. Whitwell JL, Wiste HJ, Weigand SD, Rocca WA, Knopman DS, Roberts RO, et al. Comparison of imaging biomarkers in the Alzheimer disease neuroimaging initiative and the Mayo Clinic Study of Aging. Arch Neurol. 2012;69(5):614–22.
27. Grassi M, Loewenstein DA, Caldirola D, Schruers K, Duara R, Perna G. A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach. Int Psychogeriatrics. 2018:1–9.
28. Lee G, Nho K, Kang B, Sohn KA, Kim D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep. 2019;9(1):1952.
29. Park JH, Cho HE, Kim JH, Wall M, Stern Y, Lim H, et al. Electronic health records based prediction of future incidence of Alzheimer's disease using machine learning; 2019. https://doi.org/10.1101/625582.
30. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology. 2018;290(2):456–64.
31. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
32. Kuss O, Blettner M, Börgermann J. Propensity score: an alternative method of analyzing treatment effects. Deutsches Arzteblatt Int. 2016;113(35-36):597–603.
33. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. Epidemiology. 2013;24:401–9.
34. Althauser RP, Rubin D. The computerized construction of a matched sample. Am J Sociol. 1970;76(2):325–46.
35. King G, Ho D, Stuart EA, Imai K. J Stat Software. 2011. MatchIt: nonparametric preprocessing for parametric causal inference. https://doi.org/10.18637/jss.v042.i08.
36. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–78.
37. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543–6.
38. Golubnitschaja O, Baban B, Boniolo G, Wang W, Bubnov R, Kapalla M, et al. Medicine in the early twenty-first century: paradigm and anticipation - EPMA position paper 2016. EPMA J. 2016;**7**:23. https://doi.org/10.1186/s13167-016-0072-4.

A.6 Artificial intelligence-based clustering and characterization of Parkinson’s disease trajectories

Reprinted with permission from “Birkenbihl, C., Ahmad, A., Massat, N. J., Raschka, T., Avbersek, A., Downey, P., Armstrong, M., and Fröhlich, H. (2023). Artificial intelligence-based clustering and characterization of Parkinson’s disease trajectories. *Scientific Reports*, 13(1), 2897.”.

scientific reports

OPEN

Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories

Colin Birkenbihl[3,4]✉, Ashar Ahmad[1,6,7], Nathalie J. Massat[1,2,7], Tamara Raschka[3,4], Andreja Avbersek[1,5], Patrick Downey[1], Martin Armstrong[1] & Holger Fröhlich[3,4]

Parkinson's disease (PD) is a highly heterogeneous disease both with respect to arising symptoms and its progression over time. This hampers the design of disease modifying trials for PD as treatments which would potentially show efficacy in specific patient subgroups could be considered ineffective in a heterogeneous trial cohort. Establishing clusters of PD patients based on their progression patterns could help to disentangle the exhibited heterogeneity, highlight clinical differences among patient subgroups, and identify the biological pathways and molecular players which underlie the evident differences. Further, stratification of patients into clusters with distinct progression patterns could help to recruit more homogeneous trial cohorts. In the present work, we applied an artificial intelligence-based algorithm to model and cluster longitudinal PD progression trajectories from the Parkinson's Progression Markers Initiative. Using a combination of six clinical outcome scores covering both motor and non-motor symptoms, we were able to identify specific clusters of PD that showed significantly different patterns of PD progression. The inclusion of genetic variants and biomarker data allowed us to associate the established progression clusters with distinct biological mechanisms, such as perturbations in vesicle transport or neuroprotection. Furthermore, we found that patients of identified progression clusters showed significant differences in their responsiveness to symptomatic treatment. Taken together, our work contributes to a better understanding of the heterogeneity encountered when examining and treating patients with PD, and points towards potential biological pathways and genes that could underlie those differences.

Parkinson's disease (PD) is an age-associated neurodegenerative disorder that affects approximately seven million people worldwide. Alongside the cardinal motor symptoms of bradykinesia, rigidity, resting tremor, and postural instability in later stages[1], PD patients suffer from a wide range of non-motor symptoms such as sleep disturbances, psychosis, cognitive impairment, and mood disorders[2]. Currently there are no disease modifying treatments available for PD and present medications (e.g., L-DOPA) only offer symptomatic benefits. Designing and conducting clinical trials to test putative disease-modifying treatments is complicated due to the high inter-individual variability of disease progression rates[3–5]. Therefore, understanding the different biological mechanisms that drive differential disease progression is vital to ultimately pave the way for personalised therapies and can help to identify novel target candidates for therapeutic intervention.

Previous attempts to identify PD subtypes focused on ad-hoc classification of the motor characteristics of tremor (tremor dominant sub-type) and postural instability (postural instability and gait dominant sub-type)[6]. Similarly, age at disease diagnosis has been used to classify PD patients into Late Onset PD and Young Onset PD[3]. However, given the broad and complex range of PD symptoms, single-variable subtyping approaches are unlikely to capture the complexity of patients' progression. Here, data-driven multivariate approaches using, for example, cluster analysis[5] offer a promising opportunity to overcome these limitations.

The foundation for such multivariate subtyping approaches is built through multi-modal longitudinal data provided by observational cohort studies such as the Parkinson's Progression Markers Initiative (PPMI)[7]. PPMI

[1]UCB Pharma, Chemin du Foriest 1, 1420 Braine-L'Alleud, Belgium. [2]Veramed Limited, 5th Floor Regal House, 70 London Road, Twickenham TW1 3QS, UK. [3]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany. [4]Bonn-, Aachen International Center for IT, University of Bonn, Friedrich Hirzebruch-Allee 6, 53115 Bonn, Germany. [5]Present address: Regeneron Inc., 777 Old Saw Mill River Road, Tarrytown, NY 10591, USA. [6]Present address: Grünenthal GmbH, 52078 Aachen, Germany. [7]These authors contributed equally: Ashar Ahmad and Nathalie J. Massat. ✉email: colin.birkenbihl@scai.fraunhofer.de

data has been previously used to identify patient subtypes based on cross-sectional imaging data and cerebrospinal fluid biomarkers at study baseline[2,8]. Only a few studies have focused on disease progression which requires the use of longitudinal follow-up data. This aspect was partially addressed by Faghri et al.[9] using PPMI data at 48 months follow-up. The authors identified three PD subtypes using non-negative matrix factorisation. Still, their approach was unable to discern these subtypes with respect to the slope of progression. In this context, recently published neural network-based approaches make it possible to cluster entire longitudinal patient trajectories[10,11]. However, these studies did not explore the biological underpinning of the subtypes nor did they consider how their patients differed in their clinical presentation or in their response to treatment.

The aim of this work was to uncover PD progression clusters by applying an artificial intelligence-based, purely data-driven approach based on multivariate longitudinal trajectories comprised of motor and non-motor scores obtained from *de-novo* patients. Furthermore, using machine learning, we sought to identify associations linking discovered progression clusters to potentially disparate biological pathways, genetic variations, and clinical symptoms. Finally, we aimed to assess any difference in the loss of dopaminergic neurons across clusters and whether patients of distinct progression clusters would respond differently to symptomatic treatment. Such insights could contribute to a deeper understanding and characterisation of the heterogeneous mechanisms at play within PD and offer the opportunity to define novel drug targets.

Results

Multivariate time series analysis identifies three patient clusters with distinct progression profiles.

By clustering the time series data of 407 de novo PD patients from PPMI (267 male, 140 female) using our previously published artificial intelligence-based VaDER approach[11], we identified three groups of PD patients with distinct progression profiles (Supplementary Section S1, Fig. S1). The clustering was conducted based on the multivariate progression of six key clinical assessments of PD symptoms over the course of up to 60 months: the MDS-UPDRS 1, 2, and 3 (off treatment)[12], tremor dominant score (TD), postural instability and gait disorder score (PIGD), and the Epworth sleepiness scale (ESS).

The three resulting clusters contained ’moderate’-progressors ($n = 230$), ‘fast’-progressors ($n = 53$), and ‘slow’-progressors ($n = 124$). Table 1 provides summary statistics of patients from each cluster at study baseline. We found significant differences between the average age at study baseline of slow progressors and the two other respective subtypes (t-test ‘slow’ versus ‘fast’, $p < 0.013$; ‘slow’ versus ’moderate’, $p < 0.019$; ’moderate’ versus ‘fast’, $p > 0.32$). In contrast, no significant difference was observed in the elapsed time from initial diagnosis to study baseline (pairwise U-tests between all three clusters, $p > 0.3$), or distribution of Hoehn and Yahr stages (χ^2-test, $p > 0.15$). With respect to MDS-UPDRS scores at study baseline, we found a significant difference in MDS-UPDRS 1 between the ‘moderate’ cluster and the other two clusters, respectively (U-test, ‘slow’ versus ’fast’, $p < 0.01$; ’moderate’ versus ‘fast’, $p < 0.001$; ‘slow’ versus ’moderate’, $p > 0.59$). For MDS-UPDRS 2, the only significant deviation was observed comparing the ‘moderate’ against ‘fast’-progressors (U-test, ’moderate’ versus ‘fast’, $p < 0.025$; ‘slow’ versus ’fast’, $p > 0.14$; ‘slow’ versus ’moderate’, $p > 0.34$). We identified no significant difference in MDS-UPDRS 3 scores (pairwise U-test for all clusters, $p > 0.69$). Furthermore, we detected no significant differences in the distribution of biological sex (χ^2-test, $p > 0.15$) and the start of symptomatic therapy (Fig. S2).

The mean univariate progression trajectories of these clusters along with their 95% confidence intervals are depicted in Fig. 1. Although the clustering was conducted on multiple outcome measures, we observed a clear separation of clusters across all selected variables except for the TD score between ‘fast’ and ’moderate’ progressors. While ‘fast’ and ‘moderately’ progressing subtypes displayed a clear increase of symptoms over the covered 60 month interval already starting from baseline, ‘slow’-progressors experienced almost no significant symptom worsening across scores until month 24.

Characterisation of PD clusters suggests longitudinal differences in dopaminergic deficiency.

The differences in motor symptom progression rates across subtypes (Fig. 1) were mirrored by significant differences in the age-adjusted trajectories of DaTSCAN measurements, which were available until month 48: the rate in loss of specific-binding ratio (SBR) signal in the caudate region was significantly lower for the cluster exhibiting ‘slow’ progression than for both the ‘fast’ and ’moderate’ progressing clusters, respectively (signal loss of −0.0033 SBR unit/month, 95% CI [−0.0055, −0.0011], $p = 0.004$ compared to the ‘fast’ group, and of −0.0019 SBR unit/month, 95% CI [−0.0032, −0.0003], $p = 0.01$ compared to the ’moderate’ group). No significant difference in SBR was observed between the ‘fast’ and ‘moderate’ progressing groups (details in Supplementary Section S3). The difference in rate of dopaminergic loss between the ‘fast’ and the ‘slow’ progressing clusters was seen equally in the ipsilateral (signal loss of −0.0034 SBR unit/month, 95% CI [−0.0056, −0.0008], $p = 0.008$) and the contralateral (signal loss of −0.0032 SBR unit/month, 95% CI [−0.0057, −0.0008], $p = 0.007$)

Cluster	N	Age (Years) *	Number of Females	Years since diagnosis	UPDRS 1*	UPDRS 2*	UPDRS 3	Hoehn and Yahr stage I	Hoehn and Yahr stage II	Hoehn and Yahr stage III
Slow	124	60.2 ± 9.3	49 (40%)	0.5 ± 0.5	5.4 ± 4.2	5.9 ± 4.2	21.0 ± 9.1	57 (46%)	65 (52%)	2 (0.2%)
Moderate	230	62.7 ± 9.7	78 (34%)	0.6 ± 0.5	5.1 ± 3.5	5.4 ± 3.9	20.6 ± 8.7	95 (41%)	135 (59%)	0 (0%)
Fast	53	64.2 ± 10.8	13 (26%)	0.7 ± 0.8	7.3 ± 4.7	7.2 ± 4.9	21.0 ± 8.8	27 (51%)	26 (49%)	0 (0%)

Table 1. Summary statistics of patients per subtype at study baseline. UPDRS refers to the MDS-UPDRS scale. Presented is the mean and standard deviation of variables as well as the percentage of females per subtype. N, Number of patients per subtype. *Differences were statistically significant; *p*-values are provided in the Result section.

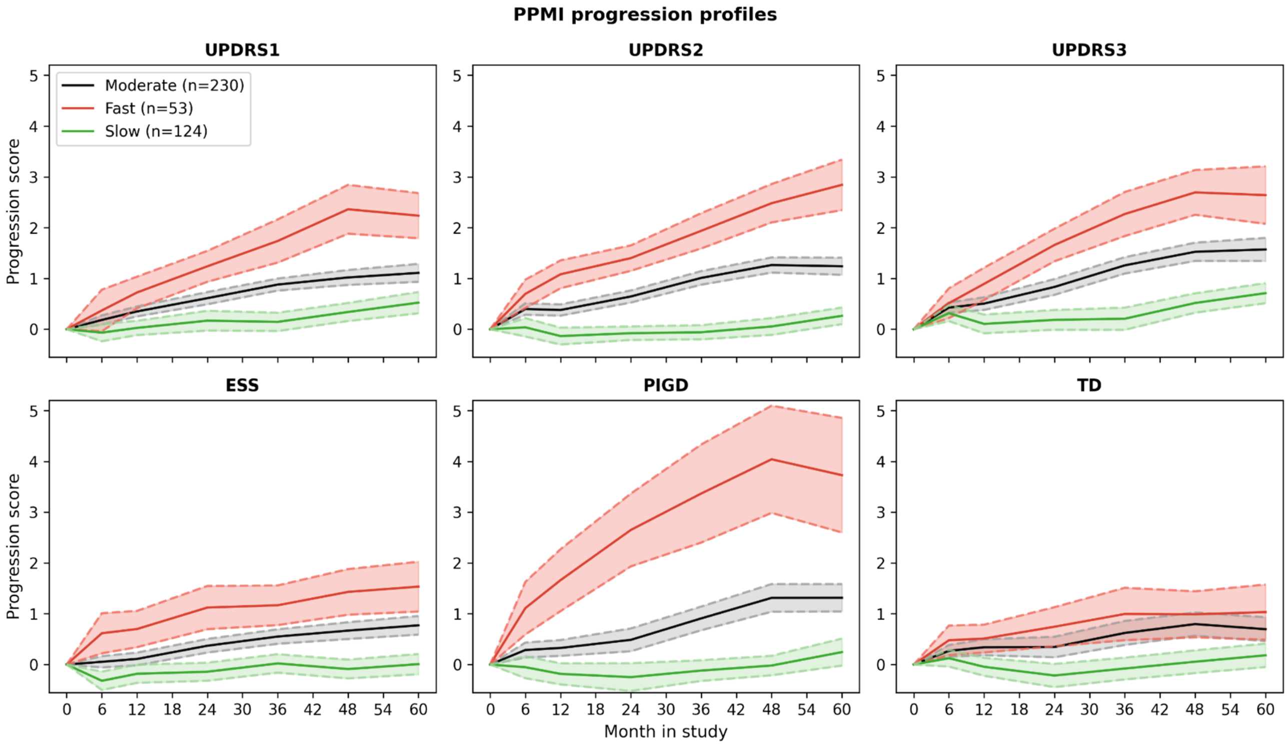


Figure 1. Mean trajectories of the three different progression clusters. Dashed lines depict the 95% confidence interval of the respective trajectory. Confidence intervals grow larger with time as more patients drop-out of the study. The progression score depicted on the y-axis represents the relative change to study baseline normalised by the standard deviation of the respective variable. UPDRS refers to the MDS-UPDRS testing battery, ESS to the Epworth Sleepiness Scale, PIGD to the Postural Instability Gait Disorder, and TD to the Tremor Dominant Score.

sides of the caudate region. In contrast, the difference in rate of progression between the 'moderate' and the 'slow' progressing subtypes was stronger in the contralateral side (signal loss of -0.0022 SBR unit/month, 95% CI $[-0.0038, -0.0006]$, $p=0.006$) as compared to the ipsilateral (signal loss of -0.0016 SBR unit/month, 95% CI $[-0.0030, +0.0002]$, $p=0.07$) sides of the caudate region. No significant difference in SBR rates were observed in the putamen, and changes in the striatum were intermediary between those observed in the caudate and the putamen.

Machine learning revealed associations between clusters and underlying biology. To discover further associations between the identified progression clusters and clinical as well as biomarker and genetic variables, we developed machine learning models based on patients' baseline visit data. Additionally, we built a second version of these models that included the 3-month follow-up data, both in the form of raw values and of change relative to baseline values. The variables included into the models comprised demographic and clinical data, including MDS-UPDRS item-level data (86 variables at baseline; 217 including 3 month follow-up), CSF biomarkers (amyloid beta, phosphorylated tau, total tau), blood serum transcriptomic data (7 variables), 3472 SNPs gained through a linkage disequilibrium analysis of an initial set of 145 PD associated SNPs obtained from DisGeNET[13], and brain region specific DaTSCAN (5 variables). We also calculated burden-scores for biological pathways stemming from Kegg[14], Reactome[15], and NeuroMMSig[16] (36, 10, and 12 pathways, respectively). These scores were based on the SNP data of each respective patient and described the amount of genetic variation affecting a pathway (see Method section for details). A full list of all variables is presented in the Supplementary Spreadsheet.

The machine learning algorithm of choice was a sparse group LASSO (SGL)[17]. We developed three distinct models, each discriminating one of the clusters from the respective other two (i.e., one versus rest approach). The significance of the most strongly associated variables was then determined by bootstrapping each model 200 times and investigating whether the resulting confidence intervals (CI) of standardised coefficients contained zero. CIs were Bonferroni-corrected to account for multiple testing. Further methodological details are described in Supplementary Section S4.

The built models revealed several significant associations between measured variables and progression clusters, which were interpretable from a clinical as well as a biological point of view.

Progression clusters are associated with distinct symptoms and genetic loci. The coefficients of each machine learning model highlight how specific variables influence the probability that a patient belongs

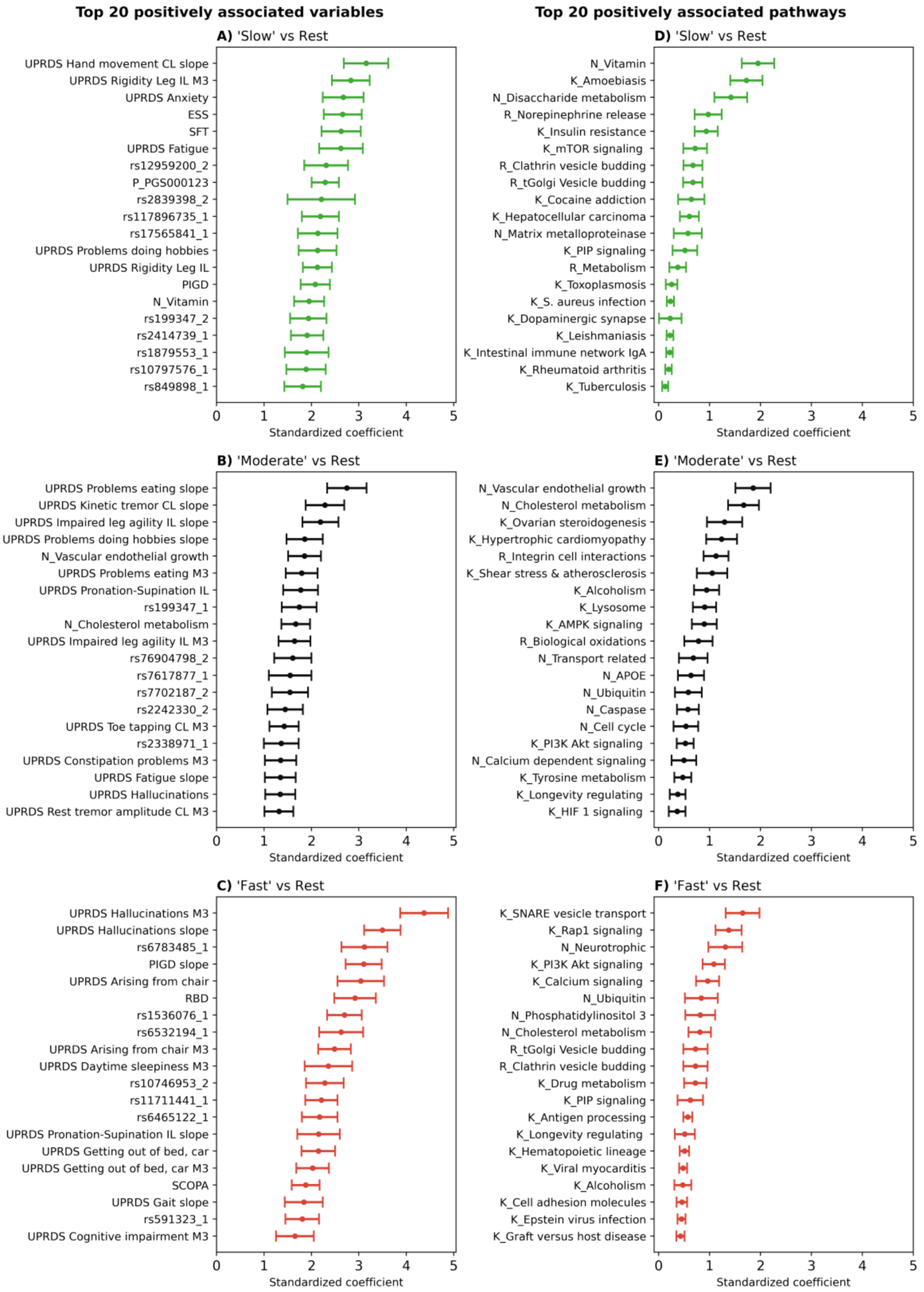


Figure 2. Top 20 variables associated with the respective progression cluster (sparse group LASSO using baseline data + 3-month follow-up). The plots show the standardised coefficient together with their Bonferroni-corrected 95% confidence intervals for each variable. A stronger positive coefficient value in the plot indicates a higher likelihood of a patient belonging to the respective cluster. A corresponding plot for baseline data only is shown in Fig. S7. (**A**–**C**, most associated variables for 'slow', 'moderate' and 'fast' progression. The number after SNP IDs indicates the number of non-reference alleles. 'M3' denotes variables measured at the 3 month visit. 'slope' indicates the calculated slope of the corresponding score measured 3 months after baseline. PGS denotes polygenic risk scores. 'CL' means contralateral, while 'IL' refers to ipsilateral. (**D**–**F**), most associated biological pathways. Pathways starting with 'K_', 'R_', or 'N_' originate from Kegg, Reactome, and NeuroMMSig, respectively.

to a particular cluster. For interpretability, we focused on significant positive interactions (i.e., variables that increase the chance of belonging to the respective cluster; Fig. 2A–C).

The variable most strongly associated with 'fast' PD progression was the presence and severity of hallucinations at the 3 month follow-up visit (NP1HALL m3, 95%CI [3.91, 5.0]), with the increase in experienced hallucinations following in third position (NP1HALL slope, 95%CI [3.07, 3.9]). In fourth position, the increase in postural instability and gait disorder severity over the first 3 months was found (PIGD slope, 95% CI [2.73, 3.55]). Additionally, 'fast' progressing patients experienced more difficulties when rising from a lying or sitting position compared to the other two subtypes (95% CI: NP3RISNG [2.56, 3.63], NP3RISNG m3 [2.16, 2.98], NP2RISE m3 [1.9, 2.65], NP2RISE [1.8, 2.64]). REM sleep behaviour disorder (RBD) proved to be another association for 'fast' progression (95% CI [2.33, 3.24]). Furthermore, several SNPs (rs6783485-LOC105377110, rs1536076-SH3GL2, rs6532194-chromosome 4:89859751, rs11711441–chromosome 3:183103487, and rs591323-LOC105379297) were found to be among the top 20 associated variables for 'fast' progression. Notably, all these SNPs were taken from DisGeNET, because of their known association to PD according to GWAS studies. In all cases, the non-reference-allele increased the risk of 'faster' PD progression.

'Slow' PD progression was associated with increasing difficulties when performing the hand movement task of the MDS-UPDRS (NP3HMOV slope 95% CI [2.93, 3.38]). Furthermore, a series of highly associated variables were connected to daytime sleepiness (ESS 95% CI [2.27, 3.06]) and general fatigue (NP1FATG 95% CI [2.16, 2.97]). Patients of the 'slow' cluster also suffered more often from anxiety (95% CI: NP1ANXS [2.15, 2.93]; NP1ANXS m3 [0.89, 1.53]) and were the only subtype which showed a significant positive association with depression, albeit the coefficient remained rather small (geriatric depression scale 95% CI [0.1, 0.65]). Additionally, better semantic fluency was also connected to 'slower' disease progression (SFT 95% CI [2.06, 2.84]). With regard to motor symptoms, 'slow' progression was associated with rigidity of the ipsilateral extremities at baseline, month 3, and their relative increase in severity (95% CI: NP3RIGL_IL m3 [2.23, 3.09]; NP3RIGL_IL [1.74, 2.54]; NP3RIGU_IL [1.0, 1.61]). Further, we found a significant positive association of the polygenic risk score PGS000123[18] and multiple genetic loci with the probability to belong to the 'slow'-progressors. SNPs rs17565841 (OCA2), and rs12959200 (chromosome 18:73599819) placed among the top 10 associations (95% CI: [2.11, 2.71], [1.95, 3.05], [1.91, 2.77], respectively). Once again, these SNPs were taken from DisGeNET because of their known association to PD according to GWAS studies.

For 'moderate' disease progression, the strongest association was the worsening of performing the eating task of the MDS-UPDRS over the first 3 months (NP2EAT slope 95% CI [2.3, 3.08]). Further, reduced agility in the ipsilateral leg was associated with 'moderate' progression (95% CI: NP3LGAG_IL slope [1.79, 2.55]; NP3LGAG_IL m3 [1.36, 2.06]). With rs76904798 (chromosome 12:40220632), rs199347 (GPNMB), rs7702187 (SEMA5A), and rs7617877 (LINC00693), we identified several PD associated SNPs which raised the probability for patients to belong to the 'moderate' subtype.

A comprehensive view on all variables and their coefficients can be found in the Supplementary Spreadsheet.

While the SGLs were designed to identify variable associations and not to make reliable forecasts, we additionally evaluated their predictive performance. With a cross-validated area under the receiver operating characteristic curve of 0.62, 0.60, and 0.63 for 'slow', 'moderate', and 'fast' progression, respectively, their performance remained limited.

Genetic burden scores connect the heterogeneity in PD progression to biological pathways. Several biological pathways and genes could be associated with the respective clusters (Fig. 2 D–F). The 'fast' cluster was highly associated with higher genetic burden in the Kegg 'SNARE vesicle transport' pathway (95% CI [1.25, 1.92]), the 'Rap1 signalling' pathway (95% CI [1.1, 1.71]), and NeuroMMSig's 'neurotrophic' subgraph (95% CI [1.25, 1.92]). The patients of the 'moderate' cluster were linked to the 'cholesterol metabolism' subgraph (95% CI [1.56, 2.25]) and 'vascular endothelial growth factor' subgraph (95% CI [1.42, 2.12]) originating from NeuroMMSig. The 'vitamin' and 'disaccharide metabolism' subgraphs from NeuroMMSig, and Kegg's 'amoebiasis pathway' were discovered as strongly associated with the 'slow' progressing clusters (95% CI: [1.6, 2.22], [1.04, 1.66], and [1.14, 1.86], respectively). A list of all mappings between pathways, genes and SNPs can be found in the Supplementary Spreadsheet.

Identified clusters show differences in response to motor symptom therapy. After observing that potentially different biological pathways were involved in the PD pathology of each cluster, we investigated whether the clusters also differed in their response to symptomatic treatment for motor symptoms. To this aim, we selected participants who had initiated Levodopa or Dopamine agonist symptomatic treatment between month 6 and month 9 after baseline and assessed whether progression as measured by MDS-UPDRS 3 score differed by PD cluster. We separately analysed the 'ON'-state MDS-UPDRS 3 score data, in which patients are examined approximately one hour after taking medication (Fig. 3), and the 'OFF'-state MDS-UPDRS 3 score data (Fig. S11). As per PPMI protocol, patients were considered to be in the 'OFF'-state when the last treatment dose was taken at least 6 h before symptoms were assessed[19]. Methodological details can be found in Supplementary Section S6.

Although initially all three PD clusters responded similarly to symptomatic treatment by stabilising their motor scores in the first 9 months after treatment initiation (i.e. 9–18 months post-baseline, Fig. 3, Fig.S11), we observed that patients in the 'fast' progressing cluster continued to progress fastest and all three clusters had significantly different MDS-UPDRS 3 scores in 'ON' and 'OFF'-states at 30 months after baseline (i.e. 21 months post-symptomatic treatment initiation) from each others, i.e. the 95% CIs did not overlap. PD subtypes did not differ according to whether they were prescribed Levodopa (alone or in combination with Dopamine agonist), or Dopamine agonist alone as a first line of PD symptomatic treatment (Table S1). The levodopa equivalent daily

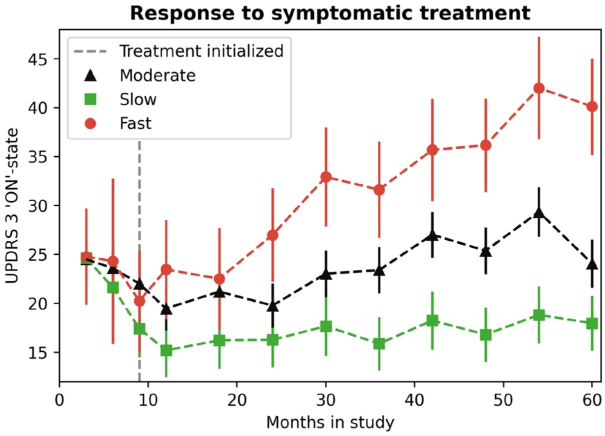


Figure 3. Differential response to symptomatic treatment. Effect plot of modelled MDS-UPDRS 3 'ON'-state score progression prior to and after the initiation of Levodopa or Dopamine agonist in patients who initiated therapy between 6 and 9 months post-baseline using a longitudinal LMEM with time fitted as a categorical variable and baseline score fitted as a covariate. The error bars represent the 95% confidence intervals, based on standard errors computed from the covariance matrix of the fitted regression coefficients.

dose (LEDD) was obtained for the PPMI participants included in this analysis (Table S2). Only beyond 42 months post-baseline, patients in the 'fast' cluster appeared to have taken higher LEDD compared to the patients in the 'moderate' cluster (mean difference at month 54: 186.8, 95%CI [76.2, 267.6], $p < 0.01$), while no significant difference was found for 'fast' versus 'slow', and 'slow' versus 'moderate' progressors, respectively (Figure S12).

Discussion

In this work, we identified three distinct PD progression clusters dividing patients into 'slow', 'moderate', and 'fast'-progressors. This clustering built on the multivariate trajectory of six clinical variables rather than a single univariate outcome. Investigation of potential confounders that could have biased the clustering showed no significant differences of biological sex, disease duration, and Hoehn & Yahr stages across clusters. Also with respect to the type of symptomatic treatment and LEDD, no bias was identified in our clustering. A machine learning model further identified significant associations between clinical measurements taken at study baseline (optionally including 3 months follow-up data), genetic features, biological pathways, and the different progression clusters of patients. Several distinct SNPs and biological mechanisms could be associated with each cluster. Analysis of the observed associations provided insights into the heterogeneity of PD progression and the distinct biological pathways potentially promoting it. Further analysis revealed that patients in different clusters responded differently to symptomatic treatment and displayed significant differences in dopaminergic cell loss. Altogether this makes it improbable that our clustering is just a consequence of patients being in different disease stages at study baseline.

Our clustering differentiates itself from previous clustering approaches in various ways: 1) instead of relying on snapshot, cross-sectional data at any arbitrary point in time, we focus on the progression of key clinical variables over time, 2) this progression is modelled multivariate to better represent the natural progression of PD which occurs across multiple scales, 3) through the inclusion of pathway-specific genetic perturbation scores, we can generate hypotheses connected to possible differences in PD pathology across the identified clusters, 4) we analysed the difference in symptomatic treatment response across clusters, which was seldomly done before[20].

Interpretation of significant associations between variables and PD progression clusters.

Our machine learning models identified that measurements taken early in the disease course already show significant associations with the longitudinal progression of PD's motor and non-motor symptoms. Such significant associations, however, do not imply that the majority of patients in a respective cluster experienced a strongly associated symptom, instead, it indicates that patients suffering from that specific symptom are statistically more likely to belong to the associated cluster. Further, while we found statistically significant differences in MDS-UPDRS 1 & 2 total scores and items between 'faster' progressing patients and the other two clusters, we identified significant associations of individual non-motor symptoms measured via the MDS-UPDRS items with every cluster. This highlights the importance of going beyond high-level clinical assessments when investigating symptom manifestation across PD subgroups.

In the 'fast'-progressing clusters, the presence of psychotic symptoms in the form of hallucinations or delusions was found as the strongest association. Indeed, hallucinations can already be observed in newly diagnosed patients[21] and experiencing such visual or auditory hallucinations was established to be one of the most notable risk factors for increased mortality[22] and earlier placement in care homes[23]. These findings could, on the one hand, be explained by the difficulties of living with psychosis but, on the other, also point towards a faster disease progression in general. In this context, the association between RBD and our 'fast' progressing cluster is noteworthy, as RBD is one of the major risk factors for hallucinations[24] and was also hypothesised to be an early sign of faster disease progression[25]. Furthermore, RBD has been connected to reduced striatal dopaminergic activity[26], which is in line with our observations for the 'fast' progressing cluster. In concordance, Wang et al. discovered slower and faster progressing subtypes based on brain pathology with the faster subtype showing increased RBD and decreased dopaminergic brain efficiency in the caudate and putamen at study baseline[27]. In another subtyping effort by Fereshtehnejad et al. a 'diffuse malignant' PD cluster was described that showed faster disease progression and was characterised by lower CSF amyloid beta values[28]. Indeed, our 'fast' progressing cluster was also associated with lower amyloid beta in CSF, however, considerably older and more affected by hallucinations than the presented 'diffuse malignant' subtype. Since the investigated PPMI patients were de novo PD patients, the significant difference in age across clusters at baseline added further evidence to a previously discovered trend that patients with later disease onset often experience faster progression[29,30].

The 'slow' cluster showed strong associations with non-motor symptoms such as fatigue, sleepiness, and anxiety. While these symptoms have received increasing recognition in recent years, they remain poorly understood aspects of PD[2] and little is known about disease progression in patients that suffer from them. Previously, a more benign PD progression was noticed among patients with resting tremor[31], a finding that was in concordance with our analysis that linked 'slow' progression to resting tremor as measured through MDS-UPDRS item 3.17.

Previous case series reported on several associations between slower disease progression and attributes we found to be significant associations with what we called 'moderate' progression[32]. Here, it was described that patients with predominantly worsening tremors, younger age, and no indication of PGID showed reduced disease progression.

Only slight differences in global cognitive performance as measured by the Montreal Cognitive Assessment (MoCA) could be found among the clusters. This could be due to the comparably early time point of assessment (approximately one month after PD diagnosis for most patients), since only subtle cognitive changes are observable in the PPMI cohort over the first 5 years[2]. However, semantic fluency was among the strongest associated variables with 'slow' progression, indicating that this cluster could be more stable with respect to cognitive performance. Patients who suffered from cognitive symptoms measured by the MDS-UPDRS were most often encountered in the 'fast' progressing cluster.

The limited predictive performance of the SGLs can be explained by the relatively small sample sizes of the identified clusters, the modelling strategy which was primarily chosen to identify significant associations rather than to provide predictions, as well as the difficulty of predicting PD progression from baseline measures. Previous attempts on predicting future PD progression based on baseline variables also reported limited performance in external validation[30].

PD progression clusters are associated with distinct biological pathways and gene mutation load. With the inclusion of available genetic data into the models, we were able to identify distinct biological pathways that were associated with the different clusters. This opens up the opportunity of not only identifying new therapeutic targets, but targets that may be positioned more effectively within certain subgroups of patients.

The pathway most predominantly associated with 'fast'-progression was the Kegg 'SNARE interactions in vesicular transport' pathway. Vesicle dysfunction is a known phenomenon in the pathogenesis of PD, the targeting of related proteins (including SCNA and LRRK2) has been discussed for several years now[33] and there are multiple lines of supporting evidence for the role of this pathway in PD. In this pathway, the retrieved SNPs predominantly mapped to genes encoding for vesicle associated membrane proteins (VAMP2, VAMP4) and syntaxins (SXT4, and SXT1B). VAMP2 interacts with SXT1 in the neuronal synapse and is important for vesicle fusion and neurotransmitter translocation[34,35]. VAMP4 and syntaxins interact with LRRK2[36], a major PD risk factor and potential drug target in which mutations promote a PD phenotype[37], with respect to retrograde and post-Golgi signalling. Both VAMP2 and SXT1 showed diagnostic potential in blood-based biomarker studies for PD[38].

The second strongest association found for fast progressors was the 'Rap1 signaling' pathway which is involved in the nigrostriatal dopaminergic pathway in medium spiny neurons[39]. Again, ample evidence lends biological support to the role of this pathway, including the position of the vascular endothelial growth factor (VEGFA) gene in the pathway, that has been shown to protect dopaminergic neurons from cell death. VEGFA has been discussed as a potential target for treating PD[40] and a recent study suggests blocking of VEGFA to prevent blood–brain-barrier disruption, which has been implicated in several neurodegenerative diseases, including PD[41].

Furthermore, this pathway involves several fibroblast growth factors (FGF5, 10, and 20), with FGF20 also being a prominent entity in the 'Neurotrophin' mechanism listed in NeuroMMSig (the third most associated pathway for 'fast' progression). The FGF gene family has also been associated with neuroprotection and neurogenesis, partially by triggering PI3K-AKT signalling which also occurred among our highly associated pathways with respect to 'fast' PD progression[42].

Taken together, it can be postulated that severe perturbations in Golgi vesicle transport that eventually cause apoptosis, in combination with a reduced neuroprotection and neurogenesis to replace damaged cells might promote a 'fast' progressing form of PD.

The 'moderately' progressing cluster was mainly associated with NeuroMMSig's 'Vascular endothelial growth factor' and 'Cholesterol metabolism' pathways. The former was largely defined by VEGFA which was discussed above and might indicate a common mechanism between 'fast' and 'moderate' progressors. The squalene synthase (FDFT1) was the major gene in the 'cholesterol metabolism' pathway to which we could map SNPs. Squalene is an antioxidant and precursor of cholesterol which is essential for synaptic functioning and has been linked to PD and α-synuclein aggregation[43]. This, along with additional supporting evidence for this pathway[44–49], could indicate that oxidative stress might play a more pronounced role in 'moderately' progressing PD compared to the other two subtypes.

The strongest associated pathway for the 'slow' progressing cluster was the 'Vitamin subgraph' which evolved around the solute carrier family 41 member 1 (SLC41A1). This gene is part of the PD related PARK16 locus and is associated with magnesium efflux and homeostasis which is believed to contribute to PD[50]. Furthermore, the 'amoebiasis' pathway was identified as the second highest associated and the connection of the underlying genes to PD has been observed previously[51]. Interestingly, we also found an association of 'slow' progressing PD to the 'disaccharide metabolism' pathway, in which GBA was a key agent. Whilst GBA mutation carriers were not included in the analysed sporadic PD PPMI cohort, three SNPs in our analysis could still be mapped to GBA, (rs2230288, rs12752133, and rs76763715) and all have been associated to an increased risk of PD[52].

Differential response to symptomatic motor treatment across progression clusters.

When the progression of motor symptoms was compared between the clusters after the initiation of Levodopa and/or Dopamine agonists, a substantial difference in the response to the symptomatic treatment was observed, which could not be explained either by medication dosage or type of therapy. Together with the observed genetic differences between clusters, our results strongly suggest that the identified progression clusters represent an inherent property of the disease. Notably, differential response to symptomatic treatment for PPMI de-novo PD cohort participants with fastest motor progression was also reported in[53], and by Lawton et al. using data from the Tracking Parkinson and Oxford Parkinson's Disease Centre Discovery cohort[54].

Limitations.

When interpreting the genetic data, it should be noted that our SNP inclusion was hypothesis driven based on prior evidence for an association with PD. Nevertheless, the work presented highlights the ability of the models to discriminate between molecular pathways involved in the different clusters, and the importance of genetic data in PD. The availability of larger datasets with attached genome wide genetic data would support a more hypothesis generating approach and potentially uncover novel mechanisms. Further, our approach relies on a clinical diagnosis of PD. While the PD diagnosis of patients was repeatedly confirmed over the several year long follow-up of PPMI, a potential misdiagnosis of patients could bias the results and the retention time of patients in the prodromal phase of PD remains unknown. Finally, PPMI as a primary data source for our analysis is an observational study in which patients are treated according to best clinical routine practice. The treatment itself is not monitored precisely, thus, the entirety of medication taken by patients, their treatment compliance, as well as a potential presence of residual medication effects remain unknown. The minimum 6 h medication washout defined by PPMI might be too short when extended release formulations were administered to patients. However, as the LEDD calculation takes into account the type of formulation of the dopaminergic therapy, as well as the impact of any adjuvant therapy, it is unlikely that this biassed our clustering as no significant difference in either the type of medication nor the LEDD was observed across clusters.

Conclusion

Using our clustering approach, we show that PD patients can be divided into 'slow', 'moderate', and 'fast'-progressors based on the relative change of symptoms over the time course of the study. These groups not only show differences in the progression rates of clinical symptoms but also differ in the rate of dopaminergic cell loss, and importantly respond differently to symptomatic treatment. An analysis of whole genome sequencing data also suggests that genetic and mechanistic differences underpin these groupings. Currently, several agents are being tested in the clinic for their ability to slow disease progression but running such trials in a group of patients containing individuals with very different progression rates is fraught with difficulty. In the PPMI cohort that we used in this work, we identified 124 of 407 patients as slow progressors, and these patients showed no worsening of any symptom for at least 24 months. Given that current disease modifying trials in PD do not exceed two years, one can expect about a third of the patients to show no symptom worsening for the duration of the trial, provided that PPMI can be regarded as a representative PD study. As disease modifying treatments do not aim to improve symptoms but to slow down their worsening then the presence of a significant number of slow progressors who will not deteriorate during the trial will make it very difficult to observe disease slowing in a mixed population even with a highly effective treatment.

Future work is needed to further validate our established PD progression clusters ideally with the help of a larger study where similar data modalities as in PPMI are measured in de-novo PD patients.

Materials and methods

Dataset and patient selection criteria.

We selected 407 de-novo PD patients from the PPMI dataset. Our inclusion criteria were: age older than 30 years, Hoehn and Yahr stage of 1 or 2, recent PD diagnosis, and untreated by anti-PD medication (patient in the off-state according to the PPMI data). Furthermore, we used only patients with at least 48 months of follow-up.PPMI acquired informed consent to data collection and sharing from all participating individuals and got ethical approval. Ethical guidelines on human data collection were adhered to.

Preprocessing by calculating progression scores.

To enable a cluster of patients along their disease progression, we transformed the selected variables into 'progression scores' that capture each variable's change relative to baseline. We calculated these progression scores by subtracting the baseline value from the value measured at each respective time point and dividing the result by the variables standard deviation at baseline. When training the machine learning models, the raw baseline (or month three) measurements were taken and standardised or one-hot-encoded (ie., in contrast to the clustering they were progression agnostic).

Multivariate clustering of clinical trajectories. Optimal hyperparameters for the VaDER model were found following the procedure described in[11]: We evaluated several possible models using a varying set of hyperparameters (including the number of sought clusters) and, finally, selected the hyperparameters which led to the best model performance. The performance of the model was quantified by comparing the prediction strength of the model against a random subtyping of the same data. We selected the smallest number of clusters that showed a significant difference to a random clustering with respect to the achieved prediction performance (Fig. S1). The clustering was repeated 20 times and the final subtypes were assigned based on a consensus clustering across the 20 repeats. Supplementary Section S1 provides further details, including diagnostic plots.

Characterisation of PD progression clusters. *Analysis of dopaminergic deficiency.* DaTSCAN data were analysed for differences between PD clusters over time. Data from baseline up to 48 months was considered. Participants without DaTSCAN screening data (N = 17) were excluded from the analysis, leaving data for 390 participants. The longitudinal progression profile for individual patients in each cluster is shown in Fig. S6 . Details about the statistical analysis are presented in Supplementary Section S3.

Response to symptomatic therapy. Patients were defined as being on symptomatic treatment, if they were taking L-DOPA, or dopamine agonists, with or without other types of motor symptom therapy such as MAO-B inhibitors at a respective visit[19]. Since a relatively highest fraction of patients started treatment at 9 months of follow-up, we focused our analysis on this time point. Altogether 44 in the 'slow' cluster started a symptomatic treatment at 9 months, 67 in the 'moderate', and 16 patients in the 'fast' cluster. The longitudinal progression profile using loess smoothing for individual patients in each cluster is shown in Fig. S8. Details about the statistical analysis including diagnostic plots are presented in Supplementary Section S6.

Analysis of whole genome sequencing data. PPMI provides whole genome sequencing (WGS) data of de novo diagnosed PD patients. To reduce the extreme high dimensionality of the WGS data while taking into account the very limited sample size, we focused only on single nucleotide polymorphisms (SNPs) with putative association to PD. More specifically, we obtained an initial list of 646 PD associated SNPs obtained from GWAS Catalogue[55], PheWas[56], and DisGeNET[13]. This list was subsequently expanded via linkage disequilibrium analysis (LD, $r^2 > 0.8$) using Haploreg[57], which also provides a gene mapping based on proximity. In addition, we employed a cis-eQTL mapping via GTex[58] to associate SNPs to genes expressed in brain tissues. Altogether 14520 SNPs were mapped to 1055 genes. In a second step, the genes were further mapped onto 12 PD specific mechanisms defined in the NeuroMMSig database[16], as well as 36 KEGG[14] and 10 Reactome[15] pathways that were significantly enriched for PD associated genes. How we calculated the pathway scores based on the selected SNPs is presented in the Supplementary Section S4.

Data availability

The authors have no permission to directly share any of the patient-level data as stated by the data usage agreement with the original data owners (the PPMI study). The PPMI data used in this work can be accessed at www.ppmi-info.org after successful access application.

Received: 2 August 2022; Accepted: 14 February 2023
Published online: 18 February 2023

References

1. Postuma, R. B. *et al.* MDS clinical diagnostic criteria for parkinson's disease. *Mov. Disord.* **30**(12), 1591–1601 (2015).
2. Weintraub, D. & Mamikonyan, E. The neuropsychiatry of parkinson disease: A perfect storm. *Am. J. Geriatr. Psychiatry* **27**(9), 998–1018 (2019).
3. Thenganatt, M. A. & Jankovic, J. Parkinson disease subtypes. *JAMA Neurol.* **71**(4), 499–504 (2014).
4. Sieber, B. A. *et al.* Prioritized research recommendations from the National Institute of Neurological Disorders and Stroke Parkinson's Disease 2014 conference. *Annals of neurology* **76**(4), 469–472 (2014).
5. Van Rooden, S. M. *et al.* The identification of parkinson's disease subtypes using cluster analysis: A systematic review. *Mov. Disord.* **25**(8), 969–978 (2010).
6. Fereshtehnejad, S. M. & Postuma, R. B. Subtypes of parkinson's disease: What do they tell us about disease progression?. *Curr. Neurol. Neurosci. Rep.* **17**(4), 34 (2017).
7. Marek, K. *et al.* The parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* **95**(4), 629–635 (2011).
8. Erro, R. *et al.* Clinical clusters and dopaminergic dysfunction in de-novo parkinson disease. *Parkinsonism Relat. Disord.* **28**, 137–140 (2016).
9. Faghri, F., Hashemi, S. H., Leonard, H., Scholz, S. W., Campbell, R. H., Nalls, M. A., & Singleton, A. B. Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning. *bioRxiv*, 338913 (2018).
10. Zhang, X. *et al.* Data-driven subtyping of parkinson's disease using longitudinal clinical records: A cohort study. *Sci. Rep.* **9**(1), 1–12 (2019).
11. de Jong, J. *et al.* Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience* **8**(11), giz134 (2019).
12. Goetz, C. G. *et al.* Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord. Off. J. Mov. Disord. Soc.* **23**(15), 2129–2170 (2008).
13. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**(D1), D845–D855 (2020).
14. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**(suppl_1), D480–D484 (2007).
15. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**(D1), D498–D503 (2020).
16. Domingo-Fernández, D. *et al.* Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): A web server for mechanism enrichment. *Bioinformatics* **33**(22), 3679–3681 (2017).

17. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **22**(2), 231–245. https://doi.org/10.1080/10618600.2012.681250 (2013).
18. Ibanez, L. *et al.* Parkinson disease polygenic risk score is associated with parkinson disease status and age at onset but not with alpha-synuclein cerebrospinal fluid levels. *BMC Neurol.* **17**(1), 1–9 (2017).
19. Simuni, T. *et al.* Longitudinal change of clinical and biological measures in early parkinson's disease: Parkinson's progression markers initiative cohort. *Mov. Dis.* **33**(5), 771–782 (2018).
20. Marras, C. & Lang, A. Parkinson's disease subtypes: Lost in translation?. *J. Neurol. Neurosurg. Psychiatry* **84**(4), 409–415 (2013).
21. Pagonabarraga, J. *et al.* Minor hallucinations occur in drug-naive parkinson's disease patients, even from the premotor phase. *Mov. Disord.* **31**(1), 45–52 (2016).
22. Weil, R. S. & Reeves, S. Hallucinations in parkinson's disease: New insights into mechanisms and treatments. *Adv. Clin. Neurosci. Rehabil. ACNR* **19**(4), 189 (2020).
23. Goetz, C. G. & Stebbins, G. T. Risk factors for nursing home placement in advanced parkinson's disease. *Neurology* **43**(11), 2222–2222 (1993).
24. Pacchetti, C. *et al.* Relationship between hallucinations, delusions, and rapid eye movement sleep behavior disorder in parkinson's disease. *Mov. Dis. Off. J. Mov. Disord. Soc.* **20**(11), 1439–1448 (2005).
25. Fereshtehnejad, S. M. *et al.* New clinical subtypes of parkinson disease and their longitudinal progression: A prospective cohort comparison with other phenotypes. *JAMA Neurol.* **72**(8), 863–873 (2015).
26. Eisensehr, I. *et al.* Reduced striatal dopamine transporters in idiopathic rapid eye movement sleep behaviour disorder: Comparison with parkinson's disease and controls. *Brain* **123**(6), 1155–1160 (2000).
27. Wang, L. *et al.* Association of specific biotypes in patients with parkinson disease and disease progression. *Neurology* **95**(11), e1445–e1460 (2020).
28. Fereshtehnejad, S. M., Zeighami, Y., Dagher, A. & Postuma, R. B. Clinical criteria for subtyping parkinson's disease: Biomarkers and longitudinal progression. *Brain* **140**(7), 1959–1976 (2017).
29. Maetzler, W., Liepelt, I. & Berg, D. Progression of parkinson's disease in the clinical phase: Potential markers. *Lancet Neurol.* **8**(12), 1158–1171 (2009).
30. Latourelle, J. C. *et al.* Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed parkinson's disease: A longitudinal cohort study and validation. *Lancet Neurol.* **16**(11), 908–916 (2017).
31. Josephs, K. A., Matsumoto, J. Y. & Ahlskog, J. E. Benign tremulous parkinsonism. *Arch. Neurol.* **63**(3), 354–357 (2006).
32. Foltynie, T., Brayne, C. & Barker, R. A. The heterogeneity of idiopathic parkinson's disease. *J. Neurol.* **249**(2), 138–145 (2002).
33. Oeda, T. *et al.* Impact of glucocerebrosidase mutations on motor and nonmotor complications in parkinson's disease. *Neurobiol. Aging* **36**(12), 3306–3313 (2015).
34. Bittner, M. A. & Holz, R. W. Kinetic analysis of secretion from permeabilized adrenal chromaffin cells reveals distinct components. *J. Biol. Chem.* **267**(23), 16219–16225 (1992).
35. Schoch, S. *et al.* SNARE function analyzed in synaptobrevin/VAMP knockout mice. *Science* **294**(5544), 1117–1122 (2001).
36. Beilina, A. *et al.* The parkinson's disease protein LRRK2 interacts with the GARP complex to promote retrograde transport to the trans-golgi network. *Cell Reports* **31**(5), 107614 (2020).
37. Cookson, M. R. The role of leucine-rich repeat kinase 2 (LRRK2) in parkinson's disease. *Nat. Rev. Neurosci.* **11**(12), 791–797 (2010).
38. Agliardi, C. *et al.* Oligomeric α-Syn and SNARE complex proteins in peripheral extracellular vesicles of neural origin are biomarkers for parkinson's disease. *Neurobiol. Dis.* **148**, 105185 (2021).
39. Zhang, X. *et al.* Balance between dopamine and adenosine signals regulates the PKA/Rap1 pathway in striatal medium spiny neurons. *Neurochem. Int.* **122**, 8–18 (2019).
40. Axelsen, T. M. & Woldbye, D. P. Gene therapy for parkinson's disease, an update. *J. Parkinsons Dis.* **8**(2), 195–215 (2018).
41. Ebanks, K., Lewis, P. A. & Bandopadhyay, R. Vesicular dysfunction and the pathogenesis of parkinson's disease: Clues from genetic studies. *Front. Neurosci.* **13**, 1381 (2019).
42. Liu, Y., Deng, J., Liu, Y., Li, W. & Nie, X. FGF, mechanism of action, role in parkinson's disease, and therapeutics. *Front. Pharmacol.* **12**, 1572 (2021).
43. García-Sanz, P., MFGAerts, J. & Moratalla, R. The role of cholesterol in α-synuclein and lewy body pathology in gba1 parkinson's disease. *Mov. Dis.* **36**(5), 1070–1085 (2021).
44. Huang, X. *et al.* Serum cholesterol and the progression of parkinson's disease: Results from DATATOP. *PLoS One* **6**(8), e22854 (2011).
45. Sere, Y. Y., Regnacq, M., Colas, J. & Berges, T. A Saccharomyces cerevisiae strain unable to store neutral lipids is tolerant to oxidative stress induced by α-synuclein. *Free Radical Biol. Med.* **49**(11), 1755–1764 (2010).
46. Kabuto, H., Yamanushi, T. T., Janjua, N., Takayama, F. & Mankura, M. Effects of squalene/squalane on dopamine levels, antioxidant enzyme activity, and fatty acid composition in the striatum of parkinson's disease mouse model. *J. Oleo Sci.* **62**(1), 21–28 (2013).
47. Sánchez-Pernaute, R. *et al.* Selective COX-2 inhibition prevents progressive dopamine neuron degeneration in a rat model of parkinson's disease. *J. Neuroinflammation* **1**(1), 1–11 (2004).
48. Van't Erve, T. J. *et al.* Reinterpreting the best biomarker of oxidative stress: The 8-iso-prostaglandin F2α/prostaglandin F2α ratio shows complex origins of lipid peroxidation biomarkers in animal models. *Free Radical Biol. Med.* **95**, 65–73 (2016).
49. Onodera, Y., Teramura, T., Takehara, T., Shigi, K. & Fukuda, K. Reactive oxygen species induce Cox-2 expression via TAK1 activation in synovial fibroblast cells. *FEBS Open Bio.* **5**, 492–501 (2015).
50. Sturgeon, M., Perry, W. & Cornall, R. SLC41A1 and TRPM7 in magnesium homeostasis and genetic risk for parkinson's disease. *J. Neurol. Neuromed.* **1**(9), 23 (2016).
51. Wang, J., Liu, Y. & Chen, T. Identification of key genes and pathways in parkinson's disease through integrated analysis. *Mol. Med. Rep.* **16**(4), 3769–3776 (2017).
52. Huang, Y., Deng, L., Zhong, Y. & Yi, M. The association between E326K of GBA and the risk of parkinson's disease. *Parkinsons Dis.* **2018**, 1048084 (2018).
53. Tsiouris, K. M., Konitsiotis, S., Koutsouris, D. D. & Fotiadis, D. I. Prognostic factors of Rapid symptoms progression in patients with newly diagnosed parkinson's disease. *Artif. Intell. Med.* **103**, 101807 (2020).
54. Lawton, M. *et al.* Developing and validating parkinson's disease subtypes and their motor and cognitive progression. *J. Neurol. Neurosurg. Psychiatry* **89**(12), 1279–1287 (2018).
55. Buniello, A. *et al.* The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**(D1), D1005–D1012 (2019).
56. Denny, J. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111. https://doi.org/10.1038/nbt.2749 (2013).
57. Ward, L. D. & Kellis, M. HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**(D1), D877-81. https://doi.org/10.1093/nar/gkv1340 (2016).
58. GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**(6), 580–585. https://doi.org/10.1038/ng.2653 (2013).

Acknowledgements

This project was partially funded by the EU-wide ERAPerMed project DIGIPD (01KU2110) and the European Union's Horizon 2020 research and innovation program under grant agreement No. 826421, "TheVirtualBrain-Cloud". Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database www.ppmi-info.org/data. PPMI—a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners. A list of names of all the PPMI funding partners can be found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/. This work was partially performed, while A.Ah. and H.F. were affiliated with UCB Monheim and A.Av. was affiliated with UCB Braine.

Author contributions

Designed the project: P.D., M.A., H.F.; supervised the project: M.A., H.F., A.Av., P.D.; analysed the data and implemented algorithms: C.B., T.R., N.J.M., A.Ah.; drafted the manuscript: C.B., H.F., M.A., P.D., A.Av., N.J.M.; all authors have read and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

PD and MA are employees of UCB BioPharma. HF, AAh, and AAv were full time employees of UCB BioPharma at the start of this study. NJM is a Veramed statistical consultant for UCB Biopharma.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30038-8.

Correspondence and requests for materials should be addressed to C.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.7 Generation of realistic synthetic data using multimodal neural ordinary differential equations

Reprinted with permission from "Wendland, P.[†], **Birkenbihl, C.[†]**, Gomez-Freixa, M., Sood, M., Kschischo, M., and Fröhlich, H. (2022). Generation of realistic synthetic data using multimodal neural ordinary differential equations. *NPJ Digital Medicine*, 5(1), 122.".

Correction to the publication: During the editorial process, errors were introduced into the publication. The "Figure 3" presented in the article was intended to display Figure 4 shown below.

npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Generation of realistic synthetic data using Multimodal Neural Ordinary Differential Equations

Philipp Wendland[1,2,4], Colin Birkenbihl[1,3,4], Marc Gomez-Freixa[3], Meemansa Sood[1,3], Maik Kschischo[2] and Holger Fröhlich[1,3]✉

Individual organizations, such as hospitals, pharmaceutical companies, and health insurance providers, are currently limited in their ability to collect data that are fully representative of a disease population. This can, in turn, negatively impact the generalization ability of statistical models and scientific insights. However, sharing data across different organizations is highly restricted by legal regulations. While federated data access concepts exist, they are technically and organizationally difficult to realize. An alternative approach would be to exchange synthetic patient data instead. In this work, we introduce the Multimodal Neural Ordinary Differential Equations (MultiNODEs), a hybrid, multimodal AI approach, which allows for generating highly realistic synthetic patient trajectories on a continuous time scale, hence enabling smooth interpolation and extrapolation of clinical studies. Our proposed method can integrate both static and longitudinal data, and implicitly handles missing values. We demonstrate the capabilities of MultiNODEs by applying them to real patient-level data from two independent clinical studies and simulated epidemiological data of an infectious disease.

npj Digital Medicine (2022)5:122 ; https://doi.org/10.1038/s41746-022-00666-x

INTRODUCTION

Patient-level data build the foundation for a plethora of healthcare research endeavors such as drug discovery, clinical trials, biomarker discovery, and precision medicine[1]. Collecting such data is extremely time-consuming and cost-intensive, and additionally access-restricted by ethical and legal regulations in most countries. Individual organizations, such as hospitals, pharmaceutical companies, and health insurance providers are currently limited in their ability to collect data that are fully representative of a disease population. This issue is especially pronounced in clinical studies, where patients are usually recruited based on predefined inclusion and exclusion criteria that introduce cohort-specific statistical biases[2]. These biases, in turn, can negatively impact the generalization ability of machine learning models, since the usual i.i.d. assumption is violated[3]. A naive idea to counteract this issue might be to build up large data repositories pooling diverse clinical studies from several organizations. However, here, a major obstacle is that sharing patient-level data across different organizations is exceedingly difficult due to legal restrictions, as formulated, for example, in the General Data Protection Rule of the European Union.

The idea we propagate in this paper is to learn a continuous-time generative machine learning model from clinical study data. Given the distribution of the real training data was appropriately learned by such a model, the generated synthetic datasets maintain the real data signals, such as variable interdependencies and time-dependent trajectories. Furthermore, these synthetic datasets can overcome crucial limitations of their real counterparts like missing values or irregular assessment intervals, hence opening the opportunity to make at least subsets of variables from different studies statistically comparable. A further strong motivation for generating synthetic datasets is the aim to use the generated data as an anonymized version of its real-world counterpart and thereby mitigate the increased restrictions for sharing human data[4–6]. However, synthetic patient-level datasets open opportunities that reach far beyond data sharing. For example, trained generative models could be used for synthesizing control arms for clinical trials based on data from previously conducted trials, or from real-world clinical routine data[7]. This helps addressing major ethical concerns in disease areas, such as cancer, where it is impossible to leave patients untreated. Both, the American Food and Drug Administration and the European Medicines Agency have recognized this issue and taken initiatives to allow for synthetic control arms[7].

Over the last years, generative models (mostly generative adversarial networks [GANs]) have found notable success, mostly in the medical imaging domain[8–13]. However, GANs are often found to show a collapse in the statistical mode of a distribution, which raises concerns regarding coverage of the real patient distribution by synthetic data. Moreover, these methods are not necessarily suited to cope with the complex nature of clinical data collected in observational, longitudinal cohort studies, which is the main focus of our work: In addition to the previously mentioned issue of irregular measurement frequencies and missing values not at random (e.g., due to participant drop-out), clinical studies often comprise several modalities combining time-dependent variables (e.g., measures of disease severity) and static information (e.g., biological sex). One approach specifically designed for the joint modeling and generation of multimodal, time-dependent, and static patient-level data containing missing values is the recently introduced Variational Autoencoder Modular Bayesian Networks (VAMBN)[4]. However, VAMBN only operates on a discrete time scale while relevant clinical indicators such as, for example, disease progression expressed through a cognitive decline or rising inflammatory markers, are intrinsically time continuous. Recently, Neural Ordinary Differential Equations (NODEs) have been introduced as a hybrid approach fusing neural networks and Ordinary Differential Equations (ODEs)[14]. While NODEs are time

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany. [2]Department of Mathematics and Technology, University of Applied Sciences Koblenz, Remagen 53424, Germany. [3]Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany. [4]These authors contributed equally: Philipp Wendland, Colin Birkenbihl. ✉email: holger.froehlich@scai.fraunhofer.de

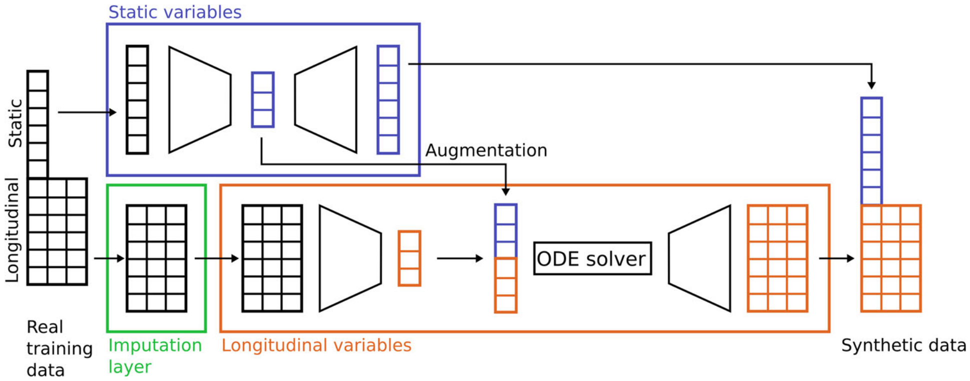


Fig. 1 Conceptual framework of MultiNODEs. Blue box: HI-VAE for the encoding and generation of static variables. Orange box: NODEs that learn and generate longitudinal trajectories. Green box: the imputation layer that can handle missing data implicitly during model training.

continuous and thus enable smooth interpolation between observed data points and extrapolation beyond the observations in the data, they are not able to integrate static variables.

In this work, we present the Multimodal Neural Ordinary Differential Equations (MultiNODEs) as an extension of the NODEs. MultiNODEs allow learning a generative model from multimodal longitudinal and static data that may contain missing values not at random. To demonstrate MultiNODEs' generative capabilities, we applied the model to clinical, patient-level data from an observational Parkinson's disease (PD) cohort study (the Parkinson's Progression Markers Initiative [PPMI][15]) and, additionally, a longitudinal Alzheimer's disease (AD) data collection (National Alzheimer's Coordination Center [NACC][16]). We compared the generated trajectories and correlation structure with the real counterpart. In this context, we additionally evaluated MultiNODEs' performance against the previously published VAMBN approach. Furthermore, we assessed MultiNODEs' interpolation and extrapolation performance. Finally, we investigated the influence of sample size, noisiness of the data, and longitudinal assessment density on the training of MultiNODEs in a systematic benchmark on data simulated from a mathematical model well-known in the epidemiology field.

RESULTS

Conceptual introduction of the MultiNODEs

MultiNODEs represent an extension of the original NODEs framework[14] that overcomes the limitations of its predecessor such that an application to incomplete datasets consisting of both static and time-dependent variables becomes feasible. Conceptually, MultiNODEs build on three key components (Fig. 1): (1) latent NODEs, (2) a variational autoencoder (more specifically a Heterogenous Incomplete Variational Autoencoder [HI-VAE], designed to handle multimodal data with missing values[17]), and (3) an implicit imputation layer[18]. The latent NODEs enable the learning and subsequent generation of continuous longitudinal variable trajectories. The longitudinal properties of the initial condition (i.e., the starting point for the ODE system solver of the latent NODEs) are defined by the output of a recurrent variational encoder that embeds the longitudinal input data into a latent space (Fig. 1, orange box). To allow for an additional influence of static variables on the estimation of the longitudinal variable trajectories, the second component, a HI-VAE, is introduced (Fig. 1, blue box). This component transforms the static information into a distinct latent space and the resulting embedding is used to augment the latent starting condition of the NODEs by concatenating the static variable embedding and the latent representation of the longitudinal variables (Fig. 1, "augmentation"). The HI-VAE component itself holds generative properties and conducts the synthesis of the static variables when MultiNODEs are applied in a generative setting. Conclusively, MultiNODEs integrate static variables (e.g., biological sex or genotype information) both to inform the learning of longitudinal trajectories, and in the generative process. Finally, to mitigate the original NODEs' incapability of dealing with missing values, we introduced the imputation layer which implicitly replaces missing values during model training with learned estimates (Fig. 1, green box). For further details on the model architecture, training, and hyperparameter optimization, we refer to the Method section and Supplementary material, respectively.

Synthetic data generation using MultiNODEs

Generating synthetic data using MultiNODEs starts by randomly sampling a latent representation for both the static and longitudinal variables, respectively. The longitudinal variables in data space are then generated by first constructing the initial conditions of the latent ODE system (i.e., concatenating the static latent representation to the longitudinal one), followed by solving the ODE system given these initial conditions, and finally by decoding the result into data space. The static variables are generated by directly transforming their sampled latent representation into data space using the HI-VAE decoder.

MultiNODEs support two different approaches for the initial sampling of the latent representations, namely sampling from the prior distribution employed during model training and sampling from the learned posterior distribution of the input data.

During the posterior sampling procedure, the reparameterization trick[19] is applied to draw a latent representation from the posterior distribution learned from the training data. The amount of noise added in this process can be tuned, whereas greater noise will lead to a wider spread of the generated marginal distributions of the synthetic data. Alternatively, the latent representations can be sampled from the prior distributions imposed on the latent space during variational model training. We ensure statistical dependence between static and longitudinal variables by drawing their values from a Bayesian network that connects both latent representations such that the longitudinal variables are conditionally dependent on the static variables. More detailed descriptions of both generation procedures are provided in the Method section.

Application cases: Parkinson's disease and Alzheimer's disease

We applied MultiNODEs to longitudinal, multimodal data from two independent clinical datasets with the goal of generating realistic synthetic datasets that maintain the real data properties. Details about the data preprocessing steps are described in the Supplementary material.

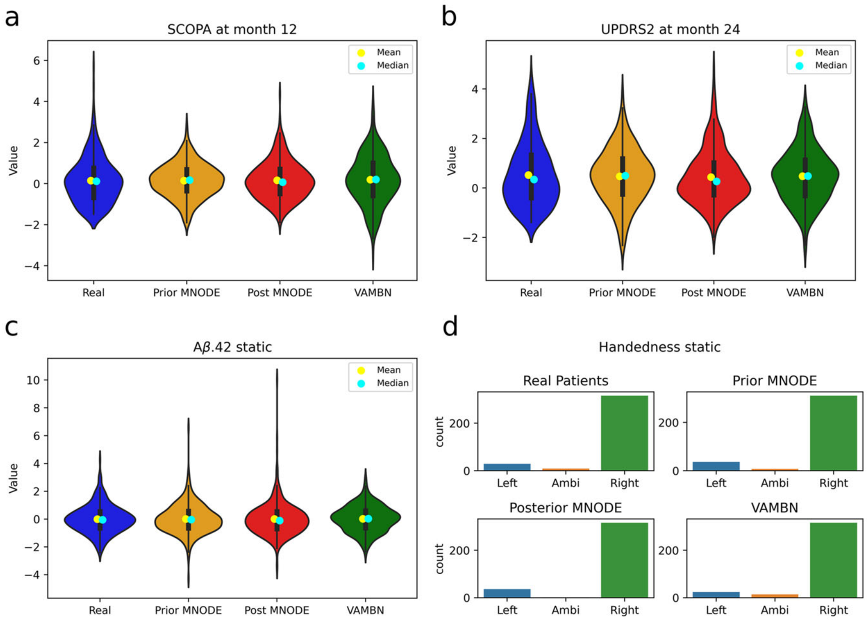


Fig. 2 Marginal distributions of real and synthesized data for multiple variables. Mean, standard deviation, and KL-divergence for the displayed variables can be found in Supplementary Table 1. Equivalent results for the NACC data are presented in Supplementary Fig. 5. **a** Time-dependent variable "SCOPA" at month 12. **b** Time-dependent variable "UPDRS2" at month 24. **c** Static variable "Aβ.42". **d** Categorical static variable "Handedness".

The first dataset was the PPMI, an observational clinical study containing 354 de-novo PD patients who participated in a range of clinical, neurological, and demographic assessments which form the variables of the dataset. In total, a set of 25 longitudinal and 43 static variables was investigated.

Furthermore, as a second example, we applied MultiNODEs to longitudinal, multimodal data from the NACC. NACC is a database storing patient-level AD data collected across multiple memory clinics. After preprocessing, the dataset used in this study contained 2284 patients, and a set of three longitudinal and four static variables was investigated.

In the following sections, we will focus on the results achieved on the PPMI data and refer to the equivalent experiments based on the NACC data that are presented in the Supplementary material.

MultiNODEs generate realistic synthetic patient-level datasets

We applied prior as well as posterior sampling for comparison purposes. With each method, we generated the same number of synthetic patients as encountered in the real dataset to allow for a fair comparison. To assess whether the generated data followed the real data characteristics, we conducted thorough comparisons of the marginal distributions using qualitative and visual assessments and further, quantitatively compared the Jensen–Shannon divergence (JS-divergence) between the generated data and real distributions. The JS-divergence is bound between 0 and 1 with 0 indicating equal distributions. In addition, we investigated the underlying correlation structure of the measured variables. Finally, we trained a machine learning classifier (Random Forest) that evaluated whether real and synthetic patients showed similar clinical characteristics when compared to real healthy control individuals from their respective studies. Across all these aspects, we evaluated MultiNODEs' performance in comparison to the previously published VAMBN approach[4].

The synthetic data generated using MultiNODE generally exhibited marginal distributions that bore high similarity to their corresponding real counterparts (Fig. 2, Supplementary Table 1, and Supplementary Fig. 1; equivalent figures for the NACC data are presented in Supplementary Fig. 4). The average JS-divergences between the real and synthetic distributions calculated across all variables and timepoints amounted to 0.018 ± 0.015 and 0.011 ± 0.009 for the PPMI data generated from the prior and posterior, respectively. For NACC the average JS-divergence was 0.071 ± 0.055 and 0.029 ± 0.031 for prior and posterior sampling, respectively. With respect to PPMI, data generation from the posterior distribution resulted in synthetic data that resembled the real data significantly closer than those generated from the prior distribution (Mann–Whitney U test, $p < 0.02$).

Compared to VAMBN, the prior sampling method seemed to be inferior with respect to the average JS-divergence when using NACC (U test, $p = 0.038$). However, no statistically significant difference in the performance of VAMBN compared to MultiNODE's posterior sampling could be observed (U test, $p = 0.80$). For PPMI, no significant differences were found between VAMBN and any of MultiNODEs' generation approaches (U test, $p = 0.31$ for the prior approach; U test, $p = 0.24$ for the posterior).

In order to evaluate whether MultiNODEs learned not only to reproduce univariate distributions but actually captured their interdependencies accurately, we compared the correlation structure of the generated data to that of the real variables. Visualizations of the Spearman rank correlation coefficients showed that both the prior and posterior sampling generated synthetic data which successfully reproduced the real variables' interdependencies (Fig. 3). Comparing the results against VAMBN-generated data revealed that both generation procedures of MultiNODEs were significantly better at reproducing the real data characteristics: the Frobenius norm of real data correlation matrix resulted in 45.3, and with a Frobenius norm of 25.66 the VAMBN-generated data placed substantially further from the real data than the MultiNODEs approaches with 62.63 and 56.47 for the prior and posterior sampling, respectively. This shows that MultiNODEs slightly overestimated the present correlations, while VAMBN underestimated them. Concordantly, the relative error (i.e., the deviation of the respective synthetic dataset's correlation matrix from the real one normalized by the norm of the real correlation matrix), was 0.81, 0.62, and 0.46, respectively, for

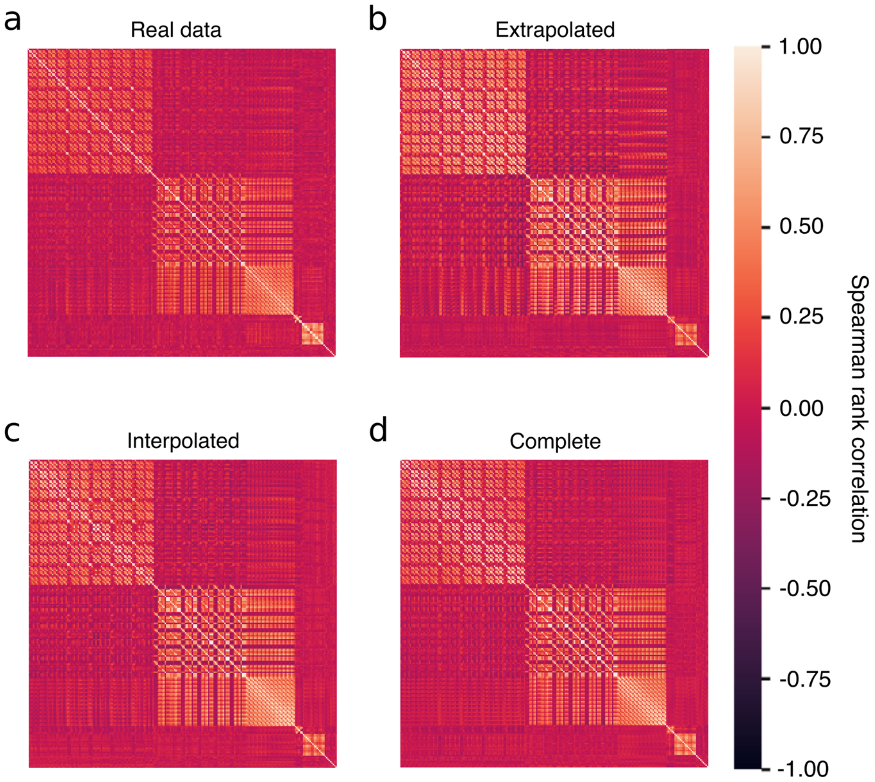


Fig. 3 Correlation structure of real and synthetic data expressed as Spearman rank correlation coefficients. Equivalent results for the NACC data are shown in Supplementary Fig. 6. **a** Real data. **b** Posterior sampling from MultiNODEs. **c** Prior sampling from MultiNODEs. **d** VAMBN-generated data.

Table 1. Performance (AUC) of machine learning classifiers differentiating between real healthy control subjects and real as well as synthetic patients, respectively.

	PPMI	Trained on synthetic PPMI tested on real	NACC	Trained on synthetic NACC tested on real
Real patients	0.97 ± 0.02		0.90 ± 0.01	
Synthetic (prior sampling)	0.97 ± 0.02	0.97 ± 0.002	0.96 ± 0.01	0.85 ± 0.002
Synthetic (posterior sampling)	0.97 ± 0.01	0.98 ± 0.002	0.93 ± 0.01	0.87 ± 0.002
Synthetic (VAMBN)	0.96 ± 0.01	0.98 ± 0.004	0.88 ± 0.01	0.89 ± 0.001

Values represent the average and standard deviation across a 10-time repeated 5-fold cross-validation.

VAMBN and MultiNODEs' prior and posterior sampling, leaving MultiNODEs with a substantially lower error than the VAMBN approach.

Assessment of the utility of generated synthetic patients for machine learning

To evaluate whether the generated synthetic patients could be reliably used in a machine learning context, we built a Random Forest classifier that aimed to distinguish between healthy individuals and diseased patients. The classifier was trained within a five-fold cross-validation scheme once using real and once using synthetic diseased patients. In addition, we trained a classifier on each respective synthetic dataset (comprising synthetically generated diseased and healthy subjects) and evaluated their performance on the real data (Table 1). As predictors, we used clinical symptoms and genetic markers that are characteristic of the disease in question. For PD (PPMI), these were the UPDRS scores that describe a series of motor and non-motor symptoms commonly encountered in PD patients, for AD (NACC), we predominantly used cognitive assessments and a genetic risk factor. Technical details about the classifiers can be found in the Supplementary material. Distinguishing real patients from healthy control subjects was possible with a 10 times repeated five-fold cross-validated performance of 0.97 ± 0.02 area under the receiver operator curve (AUC) and 0.90 ± 0.01 AUC for PPMI and NACC, respectively. On PPMI, all evaluated generative methods achieved almost equal performance, indicating that clinical characteristics of synthetic patients followed the same patterns as in real patients. In addition, the most relevant features were the same across the real and all synthetic data-trained classifiers (Supplementary Fig. 13).

For NACC, some deviations were found between a classifier's cross-validated performance on real data and the synthetic-data-based performances. Here, MultiNODEs' posterior and VAMBN showed similar deviations in opposite directions, with the posterior slightly overperforming and VAMBN slightly underperforming. The performance on the data generated via Multi-NODE's prior sampling method deviated the most (Table 1). When trained on synthetic data and evaluated on real data, all trained

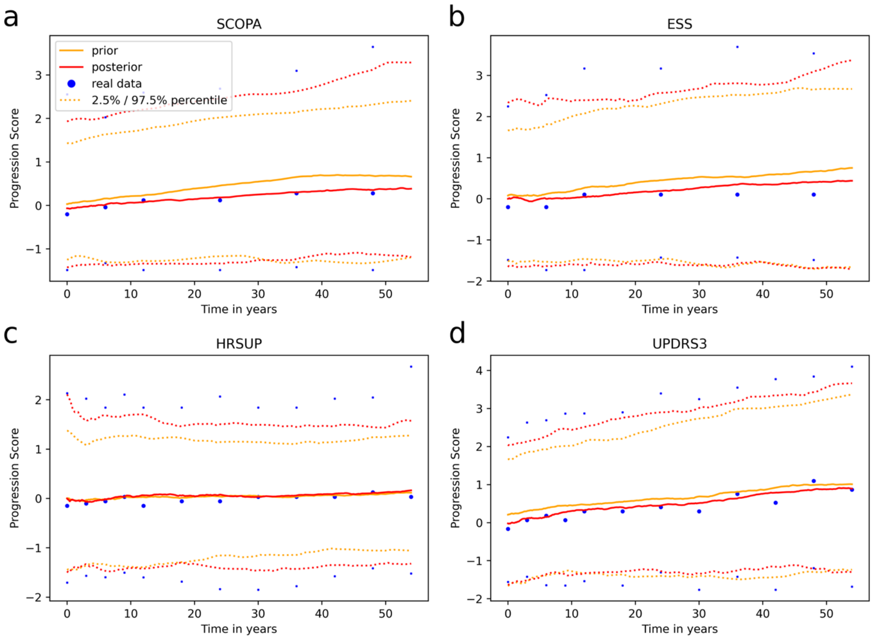


Fig. 4 Comparison of median trajectories including the 2.5%/97.5% quantiles of longitudinal variables from synthetic and real PPMI data. Additional examples are provided in Supplementary Fig. 2. A corresponding example for the NACC dataset is shown in Supplementary Fig. 7. **a**–**d** depict different longitudinal variables from the PPMI dataset.

classifiers underperformed compared to classifiers trained on real data. The feature importances of predictors were highly similar between the real data-trained and the respective synthetic data-trained classifiers.

Generating data in continuous time through smooth interpolation and extrapolation

One particular strength of MultiNODEs, that sets it apart from alternative approaches such as VAMBN, is its ability to model variable trajectories in continuous time. The latent ODE system allows for the estimation of variable trajectories at any arbitrary timepoint and thereby opens possibilities for (1) the generation of smooth trajectories, (2) overcoming panel-data limitations through interpolation, and finally, (3) extrapolation beyond the time span covered in training data themselves. Again, we evaluated these capabilities based on the PPMI and NACC datasets (for brevity, NACC results are presented in the Supplementary material). For the following, we only focused on the MultiNODE posterior sampling approach to generate synthetic subjects.

Comparing the median trajectories of variables from the real data to those generated using MultiNODEs revealed that MultiNODEs accurately learned and reproduced the longitudinal dynamics exhibited in the real data (Fig. 4). Generation from both the prior and posterior distribution led to synthesized median trajectories that closely resembled the real median trajectories. Equivalently, also the 97.5% and 2.5% quantiles of the synthetic data approximated the corresponding real quantiles closely, indicating a realistic distribution of the synthetic data across the observed timepoints. This observation held true for most of the time-dependent variables (plots for all variables are linked in the Supplementary material).

We further assessed the interpolation and extrapolation capabilities of MultiNODEs. For interpolation, one timepoint was excluded from model training and subsequently, data were generated for all timepoints including the one left-out. Contrasting the interpolated/imputed values against the corresponding real values showed that MultiNODEs accurately reproduced the longitudinal dynamics of a variable, even for unobserved timepoints (Fig. 5a, c). In this context, we further compared the interpolated values against synthetic data that was generated based on the complete real data trajectory. We observed that the mean JS-divergence calculated across all variables between the interpolated data and the real data was slightly higher (0.025 ± 0.011) than that of the real data and the synthetic data generated after training MultiNODEs on the complete trajectory (0.016 ± 0.011). Similarly, the relative error between the interpolated correlation matrix and the real data was again only marginally higher than between the complete data and the real data (0.48 and 0.46, respectively; Supplementary Fig. 4).

In order to test MultiNODEs' extrapolation capabilities, only the first 24 months of assessment follow-up and the static variables were used during model training. The trained model was then applied to generate data for the remaining, left-out timepoints of the longitudinal variables. In this course, 77 values were extrapolated while not every variable had the same number of follow-up assessments after month 24. Comparing the extrapolated synthetic data to the left-out real data demonstrated reliable extrapolation beyond the training data (Fig. 5b, d). As in the interpolation setting, we also compared the average JS-divergence between the extrapolated data and the real data with that between the real data and synthetic data that were generated after training MultiNODEs on the complete trajectory. As expected, we could see a larger difference between the JS-divergences compared to the interpolation setting with 0.037 ± 0.024 for the extrapolated data and 0.016 ± 0.009 for the synthetic data based on the complete trajectory. The correlation structure in the extrapolation culminated in a relative error of 0.64 compared to 0.46 when using the complete trajectory for training MultiNODEs (Supplementary Fig. 4).

In addition, the marginal distributions at both the interpolated and extrapolated timepoints also followed those of the real data (Fig. 5c, d).

Systematic model benchmarking on simulated data

To explore the learning properties of MultiNODEs more systematically, we investigated how alternating training conditions with

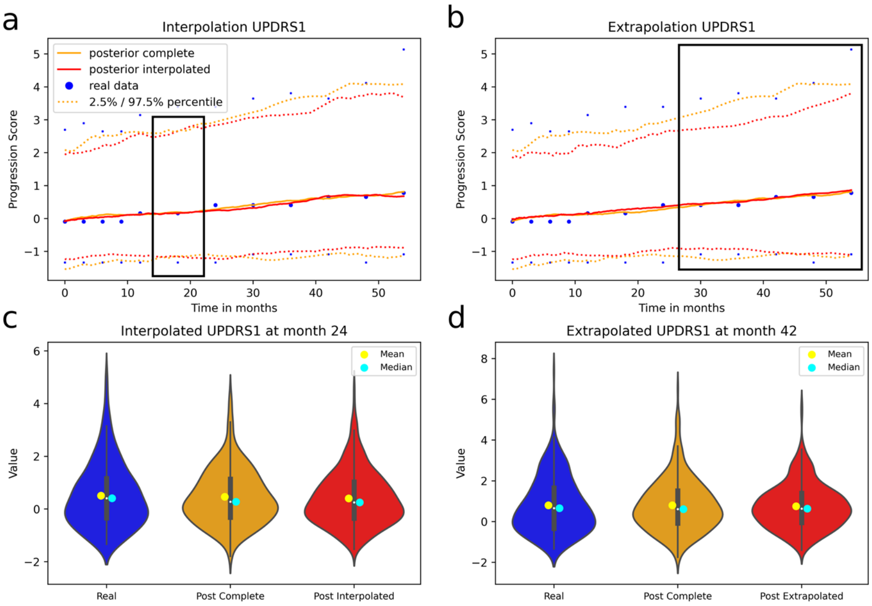


Fig. 5 Time-continuous interpolation and extrapolation of exemplary PPMI variables. The black box indicates the interpolated and extrapolated sections. Plots for additional variables are presented in Supplementary Fig. 3. A corresponding example for the NACC dataset is shown in Supplementary Fig. 8. **a** Interpolation of the UPDRS1 variable at month 24. **b** Extrapolation of the last five assessments of the UPDRS1 variable. **c** Distribution of the interpolated values for UPDRS1 at visit 24. **d** Distribution of the extrapolated values for UPDRS1 at month 42.

respect to measurement frequency, sample size, and noisiness of the data influence MultiNODEs' generative performance.

The benchmarking data was simulated via the well-established Susceptible-Infected-Removed (SIR) model that is often used to describe the spread of infectious diseases and follows a highly nonlinear structure: Let $S(t)$ be the number of susceptible individuals at a timepoint t, $I(t)$ be the number of infectious individuals at a timepoint t and $R(t)$ be the number of removed or recovered individuals at a timepoint t. With β as transmission rate, γ as mean recovery/death rate, and $N = S(t) + I(t) + R(t)$ as fixed population size the SIR model can be defined by the ODE system presented in Eq. (1):

$$\begin{aligned} \frac{dS}{dt} &= \frac{-\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1}$$

Details about the SIR parameter settings are described in the Supplementary material.

As baseline settings for each investigation, we simulated 1000 data points with 10 equidistant assessment timepoints each, distributed over a span of 40 time intervals, and added 5% Gaussian noise to each measurement. That means we added a normally distributed variable with the standard deviation set to 5% of the theoretical range of each of the variables $S(t)$, $I(t)$, and $R(t)$. During the benchmarking, we individually alternated the sample size, timepoints, and noise level. For the timepoint investigation, we compared MultiNODEs trained on 5, 10, and 100 equidistant assessments; for the sample size we considered 100, 1000, and 5000 samples; and for the noise level, we tested 50%, 75%, and 100% of the maximum encountered value added as noise.

Alternating the amount of equidistant, longitudinal timepoints exposed a strong dependency of MultiNODEs on the longitudinal coverage of the time-dependent process (Fig. 6a). While the general trends in the data were appropriately learned for all explored assessment frequencies, the position of the observations in time influenced how close the learned function approximated the true data-underlying process. Especially the peak of the "Infected"-function represented a challenge for MultiNODEs if no data point was located close to it (Fig. 6a, "Infected"). Similarly, the start of the decline in the "Susceptible"-function and the incline in the "Removed"-function were shifted, depending on the positioning of measurements. In conclusion, and as expected, a higher observation frequency of the data-underlying the time-dependent process significantly increased the fit of MultiNODEs to the process, although, general trends could already be approximated for lower assessment frequencies.

Investigating the effect of the sample size on training MultiNODEs, we observed that an increase of the sample size led to an expected improvement of the model fit to the SIR dynamics (Fig. 6b). While the general trends could again be learned from limited data ($n = 100$), sample sizes of 1000 or 5000 substantially reduced the model's deviation from the true SIR model. With 1000 samples, the learned dynamic is less stable than when trained on 5000 samples, where a smooth dynamic was learned that closely resembled the true underlying process. In conclusion, MultiNODEs can already learn longitudinal dynamics based on only a few data points, however, they tend to underfit under these circumstances and benefit from larger sample sizes.

Adding an increasing noise level to the SIR training data revealed that MultiNODEs remain very robust (Fig. 6c). Only when introducing 100% of the maximal encountered value as additional noise, a clear deviation from the underlying true model could be observed.

DISCUSSION

In this work, we presented MultiNODEs, a hybrid AI approach to generate synthetic patient-level datasets. MultiNODEs are specifically designed to consider the characteristics of clinical studies, extend its predecessor, the Neural ODEs, and enable the application of the latent ODE system to multimodal datasets comprising both time-dependent and static variables with values missing not at random. MultiNODEs learn a latent, continuous time trajectory from observed data. This concept fits well with processes like disease progression, where relevant observations

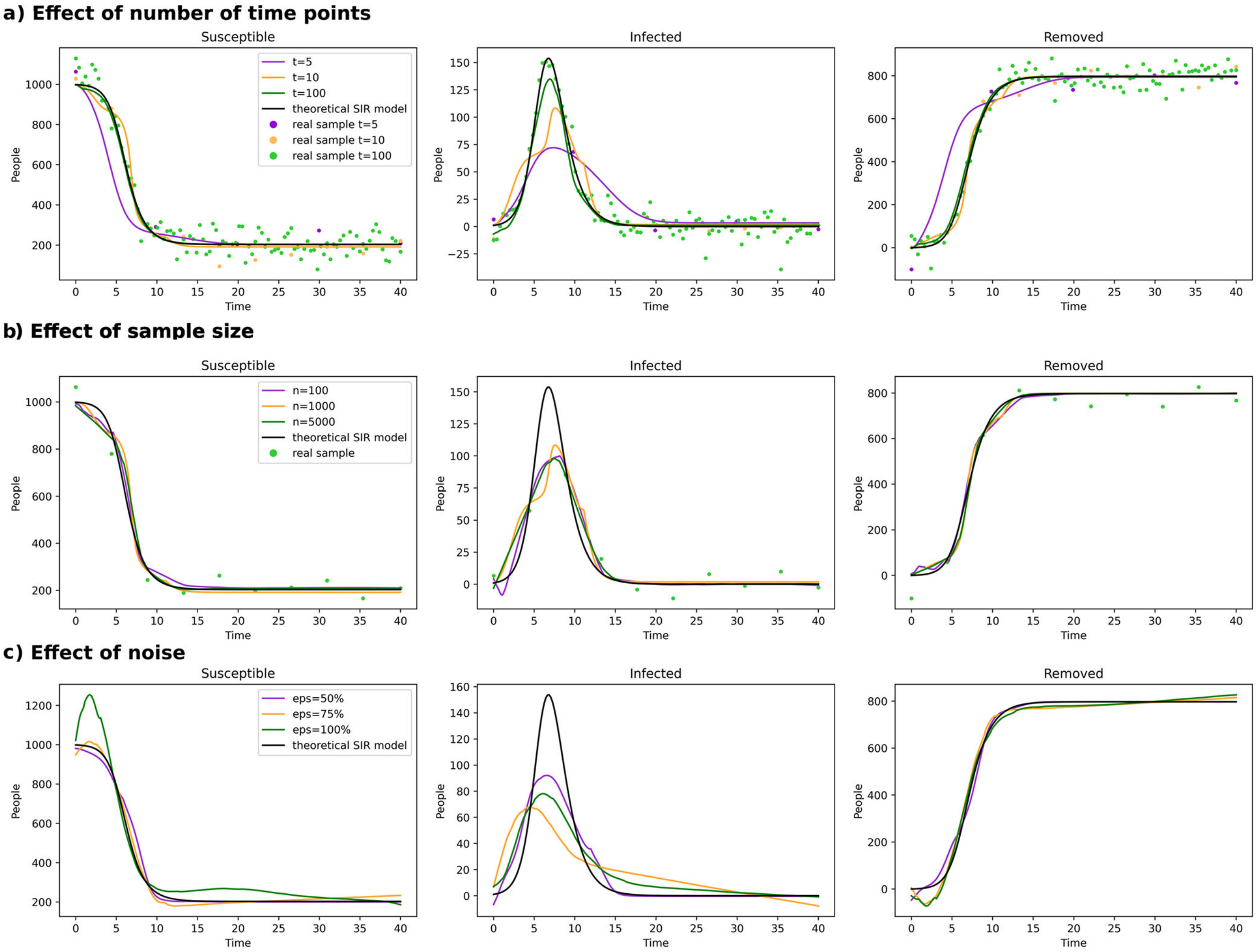


Fig. 6 Model benchmarking on simulated data from the SIR model. Each panel (**a**–**c**) represents the evaluation of another parameter (assessment frequency, sample size, and noise level, respectively).

(e.g., biomarkers and disease symptoms) only indirectly mimic the true, underlying disease mechanism. Consequently, MultiNODEs are well suited for an application to heterogeneous datasets holding complex signals as encountered, for example, in biomedical research.

Our evaluations showed that MultiNODEs successfully generated complex, synthetic medical datasets that accurately reproduced the characteristics of their real-world counterparts. In a direct comparison MultiNODEs' outperformed the state-of-the-art VAMBN approach, most notably with respect to the integrity of the correlation structure. This finding implies that the single data instances generated using MultiNODEs exhibit more realistic properties and that the real data characteristics are not only reproduced at the population level. Out of MultiNODEs two generative methods, the posterior sampling expectedly led to more realistic synthetic patients; however, generating from the prior distribution comes with the benefit that the model itself can be shared and used for data generation without needing any real data points in the process.

Machine learning classifiers that discriminated between real healthy controls and diseased subjects showed almost equal performance when trained on data from synthetic and real diseased subjects, respectively. Here, we only observed small deviations from the performance on real data for the NACC dataset, where classifiers trained and tested on synthetic patients and real healthy controls within a cross-validation setting showed a slightly increased performance to those trained on real data. Interestingly, at the same time, we found a lower prediction performance compared to real data when we trained on synthetic subjects and evaluated on the real data. A possible explanation is that synthetic data can contain noise that is introduced during the generation of synthetic data points (e.g., through overestimated correlations between variables). Therefore, synthetically generated diseased patients are better discriminated against real healthy controls than real diseased patients. At the same time, this situation leads to the fact that a classifier trained on synthetic data (synthetic patients as well as healthy controls) shows a slightly lower prediction performance on real data compared to a classifier trained entirely on real data. Altogether our results demonstrate that synthetically generated subjects share patterns of real patients, but they are not completely identical.

Besides the reproduction of marginal distributions and synthesis of realistic data instances, MultiNODEs most prominent strength lies in the generation of smooth longitudinal data. The latent ODE system allows MultiNODEs to learn dynamics that are continuous in time and cover the unobserved time intervals of real-world data. Here, both the prior and posterior sampling approach resulted in synthetic trajectories that obey real variables' dynamics.

Furthermore, the time-continuous generative capabilities of MultiNODEs create opportunities to fill gaps in the real data through interpolation and go beyond the observation time by extrapolating the longitudinal dynamics. Hence, MultiNODEs could be used to support the design of longitudinal clinical studies, in which the maximum observation period, as well as visit frequency, is always a crucial decision to make. Here, the question of how patients might develop between two visits or after the last one determines the optimal follow-up time, to demonstrate, for example, the most significant treatment effect. Furthermore, synthetic disease trajectories generated based on data from one clinical study can be compared to those generated based on other studies, even if the visit intervals employed in the real studies were not identical.

Our benchmark experiments on the simulated SIR model data demonstrated that MultiNODEs are applicable under a variety of different data settings. While the general trends of a data-underlying process could already be learned from a relatively limited dataset, similar to any machine learning task, the accuracy and trustworthiness of the model critically depends on the available data. Especially for complex, nonlinear processes, a sufficiently high observation frequency should be considered. Here, the position of the observation timepoints relative to the true underlying process is crucial for MultiNODEs to accurately learn nonlinear dynamics. The sample size of the training data mainly impacts how well MultiNODEs fitted the data dynamics and we observed that lower sample sizes can lead to underfitting and rather rigid ODE systems. On the other hand, only severe noise levels led to a model deviation from the true data-underlying process, and thus, with respect to noise, MultiNODEs proved to be highly robust. In conclusion, MultiNODEs' requirements toward the training data ultimately depend on the complexity of the data-underlying process, whereas the learning of more complex processes requires more frequent observations and larger sample size, while more linear systems can already be learned from rather limited datasets.

One limitation of MultiNODEs in their current form only allows static categorical variables. This is because the variational encoder for longitudinal data maps trajectories to a latent Gaussian distribution. Sampling from this distribution (even, if conditioned on the distribution of the static data) and decoding will result in real valued features rather than categorical ones. In future work, we will thus explore whether a recurrent version of the HI-VAE encoder can be used instead of a recurrent variational long-short term memory (LSTM) encoder.

In addition, MultiNODEs are sensitive to several hyperparameters that should be optimized for optimal performance. The training process and all relevant hyperparameters are explained in the Method section.

Synthetic data generated using models trained on sensitive personal information can bear a risk of information disclosure (e.g., attribute disclosure or dataset membership disclosure), if an attacker has information about properties of real patients that are similar to a synthetic subject. Therefore, before synthetic data are distributed, it must be assured that the probability of private information disclosure remains within task-appropriate boundaries[20]. Disclosure risk often stands in a direct trade-off with data utility and a sensible compromise should be taken balancing the two according to the application in question. Several approaches are described in the literature that can reduce the risk of information disclosure[21], one of which is based on the concept of differential privacy[4]. MultiNODEs themselves provide a way to tune the deviation from the real data when sampling from the posterior distribution by changing the amount of noise injected in the latent space. We would like to mention that a rigorous quantification of the re-identification risk is a non-trivial and challenging task for its own requiring several assumptions and is thus beyond the scope of this paper.

METHODS

Application case datasets

Both datasets, namely PPMI and NACC, are well-known staples in their respective fields and can be accessed after successful data access applications. For PPMI see https://www.ppmi-info.org/. For NACC we refer to https://naccdata.org/. More details on the investigated variables are presented in the Supplementary material.

Both studies retrieved informed consent from their participants for data collection and sharing and followed the declaration of Helsinki to ensure ethical data collection. Both studies got ethical approval from their respective review boards. We followed their employed regulations and thus did not seek further ethical approval, as we did not work with human participants ourselves.

Neural ODEs (NODEs)

NODEs are a hybrid of neural networks and ODEs[14]. They can be seen as an extension of a ResNet[22], which does not rely on a discrete sequence of hidden layers, but on a continuous hidden dynamical system defined by an ODE.

For $0 < t < M$ and $z_0 \in R^D$ the dynamics of the hidden layer of a NODE are given as Eq. (2).

$$\begin{aligned} \tfrac{dz(t)}{dt} &= f(z(t), t, \theta) \\ z(0) &= z_0 \end{aligned} \tag{2}$$

where $z(0)$ may be interpreted as the first hidden layer and $z(T)$ as the solution to the initial value problem at timepoint T. Importantly, f is a feed-forward neural network parameterized by θ.

NODEs as generative latent time series models. As demonstrated by the authors in their publication, NODEs can be trained as a continuous time Variational Autoencoder. The basic idea is to learn the initial conditions z_0 of the dynamical system in Eq. (2) from observed time series data using a variational LSTM recurrent encoder[23]. Hence, Eq. (2) now describes the dynamics of a latent system, resulting in a classical state-observation model. Accordingly, a feed-forward neural network decoder is required to project the solution of Eq. (2) back to observed data at defined timepoints (Supplementary Fig. 10).

Overall NODEs are trained at once by maximizing the evidence lower variational bound (ELBO): let the training data be $D = \{(x_{t_i}^n, t_i) | n = 1, \ldots, N, i = 1, \ldots, M\}$, where N is the number of patients and $t_{i_1}, \ldots, t_{i_M}$ the observed timepoints / patient visits. That means $x_{t_i}^n \in R^p$ is the p-dimensional vector of measurements taken for the nth patient at visit t_i. The ELBO for NODEs is then given as Eq. (3).

$$\begin{aligned} ELBO^{NODE} &= \tfrac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{M} -D_{KL}\left(q\left(z_{t_0}^n | \left\{x_{t_i}^n, t_i\right\}_i\right) \| p\left(z_{t_0}^n\right)\right) \\ &\quad + E_{q\left(z_{t_0}^n | \left\{x_{t_i}^n, t_i\right\}_i\right)}\left(log\left(p\left(x_{t_i}^n | z_{t_i}^n\right)\right)\right) \end{aligned} \tag{3}$$

where $p\left(z_{t_0}^n\right) = N(0, I)$, as usual. For details, we refer to Chen et al.[14].

Multimodal Neural ODEs (MultiNODEs)

Handling missing values. To handle missing values (potentially not at random) in longitudinal clinical data we build on our previously published work, in which we introduced an imputation layer to implicitly estimate missing values during neural network training[18]: let $A := \left\{x_{t_i,j}^n | x_{t_i,j}^n \text{ is not missing}\right\}$, 1_A be the indicator function on set A with cardinality $|A|$. The imputation layer can be defined as a data transformation $\tilde{x}_{t_i,j}^n = x_{t_i,j}^n \times 1_A\left(x_{t_i,j}^n\right) + b_{t_i,j} \times \left(1 - 1_A\left(x_{t_i,j}^n\right)\right)$, where parameters $b_{t_i,j}$ are trainable weights. That means missing values in a patient's data vector $x_{t_i,j}^n$ are replaced by $b_{t_i,j}$. The accordingly completed data is subsequently mapped through a recurrent neural network encoder to a static, lower dimensional vector, which is interpreted as the initial condition of the latent ODE system (Supplementary Fig. 11).

To learn parameters $b_{t_i,j}$ the NODEs' loss function needs to be adapted. More specifically, we use the modified ELBO criterion presented in Eq. (4).

$$\begin{aligned} ELBO_{IMP}^{NODE} &= \tfrac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{M} -D_{KL}\left(q\left(z_{t_0}^n | \left\{x_{t_i}^n, t_i\right\}_i\right) \| p\left(z_{t_0}^n\right)\right) \\ &\quad + \tfrac{DM}{A} \sum_{n=1}^{N} \sum_{i=1}^{M} \sum_{j=1}^{D} 1_A\left(x_{t_i,j}^n\right)\left(x_{t_i,j}^n - \hat{x}_{t_i,j}^n\right)^2 \end{aligned} \tag{4}$$

where $\hat{x}^n_{t_i,j}$ denotes the reconstructed data. Note that we only aim for reconstructing the observed data, but not the imputed one. Due to the layer-wise architecture of a neural network $\hat{x}^n_{t_i,j}$ implicitly depends on $b_{t_i,j}$.

In practice, we initialize $b_{t_i,j}$ for neural network training as $\frac{1}{N}\sum_{n=1}^{N} x^n_{t_i,j}$.

Dealing with multimodal data. In addition to implicit missing value imputation, the second main idea of MultiNODEs is to complement NODEs with a HI-VAE encoder[17] for static variables (Supplementary Fig. 11). A HI-VAE is an extension of a Variational Autoencoder that can implicitly impute missing values via an input drop-out model and handle heterogeneous multimodal data, including categorical data and count data, via an accordingly factorized generative model. In addition, a HI-VAE uses a Gaussian Mixture Model (GMM) as a prior distribution rather than a single Gaussian. We refer to Nazabal et al.[17] for details.

The HI-VAE results in a lower dimensional latent representation z_{stat} of static variables, which can be used to augment the initial conditions z_{t0} learned from time series data. Consequently, we arrive at the following formulation of the latent ODE system given in Eq. (5).

$$\begin{aligned} \tfrac{d}{dt} z^{aug}(t) &= \tfrac{d}{dt}\begin{bmatrix} z(t) \\ \tilde{z}(t) \end{bmatrix} = f\left(\begin{bmatrix} z(t) \\ \tilde{z}(t) \end{bmatrix}, t, \theta_f^{aug}\right) \\ z_{t_0}^{aug} &= \begin{bmatrix} z_{t_0} \\ z_{stat} \end{bmatrix} \end{aligned} \tag{5}$$

This approach resembles the Augmented Neural ODEs by Dupont et al.[24]. In contrast to our work, in their work no additional features were added during the augmentation step, i.e., $z_{stat} = 0$. According to Dupont et al. the purpose of Augmented Neural ODEs is to smoothen *f*, whereas we focus here on multimodal data integration.

For training MultiNODEs, we have to jointly consider $ELBO^{NODE}_{IMP}$ as well as $ELBO^{HI-VAE}$. After bringing both quantities on a comparable numerical scale, we use a weighted sum as our final training objective (see Eqs. (6) and (7)):

$$\begin{aligned} \widetilde{ELBO}^{NODE}_{IMP} &= \tfrac{ELBO^{HI-VAE}}{ELBO^{HI-VAE}+ELBO^{NODE}_{IMP}} ELBO^{NODE}_{IMP} \\ \widetilde{ELBO}^{HI-VAE} &= \tfrac{ELBO^{NODE}_{IMP}}{ELBO^{HI-VAE}+ELBO^{NODE}_{IMP}} ELBO^{HI-VAE} \end{aligned} \tag{6}$$

$$ELBO^{MultiNODE} = \widetilde{ELBO}^{NODE}_{IMP} + \beta \widetilde{ELBO}^{HI-VAE} \tag{7}$$

Where β is a tunable hyperparameter. Details about hyperparameter optimization are described in the Supplementary material.

Generating synthetic subjects

We tested two methods to generate synthetic subjects with MultiNODEs:

a. The first option is drawing a sample of latent static and longitudinal representations from the respective prior distributions $z_{t_0} \sim N(0, I)$and $z_{stat} \sim GMM(\pi)$. To assure that interdependencies between static and longitudinal variables are conserved, we model their joint distribution $P(z_{t0}, z_{stat})$ using a Bayesian network. This network contains three nodes (random variables) representing (1) the GMM mixture coefficients π for the static data used by the HI-VAE, (2) the latent static representations $Z_{stat} = GMM(\pi)$, and (3) the latent longitudinal representations $Z_{t0} = N(0, I)$, respectively. The network is constrained such that directed edges can only go from s_i to Z_{stat} and from there to Z_{t0}. After randomly sampling a mixture component s_i from a multinomial distribution $multinom(\pi)$, we can conditionally sample $z_{stat} \sim Z_{stat}|s_i$ and finally $z_{t0} \sim Z_{t0}|Z_{stat}$. Subsequently, we concatenate $z_0 = [z_{t_0}, z_{stat}]$ into a vector forming the initial conditions for the latent ODE system, solve the ODE system, and decode the solution. We call this approach "prior sampling".

b. A second option is to draw $z_{t_0} q(z^n_{t_0}|\{x^n_{t_i}, t_i\}_i) = N(\lambda(x^n_{t_i}, t_i), \sigma(x^n_{t_i}, t_i))$ for the longitudinal data and $z_{stat} q(z^n_{stat}|x^n_{stat}, \pi) = N(\lambda(x^n_{stat}, s^n),$ $\sigma(x^n_{stat}, s^n)), s^n Categorical(\pi(x^n_{stat}))$ for the static data. That means we generate a blurred / noisy version of the original *n*th patient. We call this approach "posterior sampling" and recommend this sampling procedure for data generation. In our experiments, we doubled the posterior variance during sampling because we found the synthetic data otherwise to lie too close to the real data. Tuning the added noise can provide one option to balance identification risk versus data utility.

c. Synthetic data can not only be generated for observed visits, but also for definable timepoints in between (interpolation) and after the end of the study (extrapolation). This is possible because the latent ODE system is continuous in time.

Data preprocessing

Few steps are required to preprocess the clinical data before MultiNODEs can be applied. First, the data must be organized into a three-dimensional tensor of the shape samples × timepoints × variables for the longitudinal variables, and samples × variables for the static ones. Furthermore, the longitudinal variables are then transformed into a progression score by subtracting the baseline value and normalizing them by the standard deviation of this variable at baseline.

Calculating the relative error for correlation matrices

The relative error between correlation matrices is calculated as the norm of the matrix describing the difference between the real correlation matrix and synthetic data correlation matrix divided by the norm of the real correlation matrix.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The PPMI dataset is available under: https://www.ppmi-info.org/. The NACC data are available under: https://naccdata.org/. The data are shared by the data owners after successful application. The data generated for this study cannot be shared by the authors due to the signed data usage agreements with the data owners of the corresponding real data (i.e., PPMI and NACC).

CODE AVAILABILITY

The code for MultiNODEs is available at https://github.com/philippwendland/MultiNODEs.

Received: 3 November 2021; Accepted: 25 July 2022;
Published online: 20 August 2022

REFERENCES

1. Fröhlich, H. et al. From hype to reality: data science enabling personalized medicine. *BMC Med.* **16**, 150 (2018).
2. Birkenbihl, C., Salimi, Y. & Fröhlich, H. Japanese Alzheimer's Disease Neuroimaging Initiative; Alzheimer's Disease Neuroimaging Initiative Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimers Dement.* **18**, 251–261 (2022).
3. Birkenbihl, C. et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia – lessons for translation into clinical practice. *EPMA J.* **11**, 367–376 (2020).
4. Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M. & Fröhlich, H. Variational Autoencoder Modular Bayesian Networks for simulation of heterogeneous clinical study data. *Front. Big Data* **3**, 16 (2020).
5. Sood, M. et al. Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders. *Sci. Rep.* **10**, 10971 (2020).
6. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).
7. Thorlund, K., Dron, L., Park, J. J. & Mills, E. J. Synthetic and external controls in clinical trials – a primer for researchers. *Clin. Epidemiol.* **12**, 457–467 (2020).
8. Lei, Y. et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med. Phys.* **46**, 3565–3581 (2019).
9. Yang, G. et al. DAGAN: Deep De-Aliasing Generative Adversarial Networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imaging* **37**, 1310–1321 (2018).
10. Lin, Z., Jain, A., Wang, C., Fanti, G. & Sekar, V. Using GANs for sharing networked time series data: challenges, initial promise, and open questions. in *Proceedings of*

the ACM Internet Measurement Conference 464–483 (ACM, 2020). https://doi.org/10.1145/3419394.3423643.

11. Bae, H., Jung, D., Choi, H.-S. & Yoon, S. AnomiGAN: Generative Adversarial Networks for anonymizing private medical data. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **25**, 563–574 (2020).
12. Jordon, J. & Yoon, J. PATE-GAN: generating synthetic data with differential privacy guarantees. in *International Conference on Learning Representations* 21 (2019).
13. Beaulieu-Jones, B. K. et al. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* **12**, e005122 (2019).
14. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. in *Advances in Neural Information Processing Systems* (eds Bengio, S. et al.) vol. 31 (Curran Associates, Inc., 2018).
15. Marek, K. et al. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).
16. Besser, L. et al. Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. *Alzheimer Dis. Assoc. Disord.* **32**, 351–358 (2018).
17. Nazabal, A., Olmos, P. M., Ghahramani, Z. & Valera, I. Handling incomplete heterogeneous data using VAEs. Preprint at *ArXiv180703653 Cs Stat* (2020).
18. de Jong, J. et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience* **8**, giz134 (2019).
19. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at http://arxiv.org/abs/1312.6114 (2014).
20. Goncalves, A. et al. Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* **20**, 108 (2020).
21. Park, N. et al. Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.* **11**, 1071–1083 (2018).
22. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). https://doi.org/10.1109/CVPR.2016.90.
23. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735–1780 (1997).
24. Dupont, E., Doucet, A. & Teh, Y. W. Augmented neural ODEs. in *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826421, "TheVirtualBrainCloud" and from the Deutsche Forschungsgemeinschaft (DFG) funded project "NFDI4Health" (project number 442326535). Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI—a public–private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners [list of the full names of all of the PPMI funding partners can be found at www.ppmi-info.org/fundingpartners]. The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), and P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

AUTHOR CONTRIBUTIONS

H.F., C.B., P.W. and M.K. conceived the project. P.W. implemented the method. P.W., M.G.-F. and M.S. performed the experiments. C.B. and H.F. wrote the manuscript. P.W., M.K., and M.S. revised the manuscript. C.B. and H.F. supervised the work.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00666-x.

Correspondence and requests for materials should be addressed to Holger Fröhlich.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

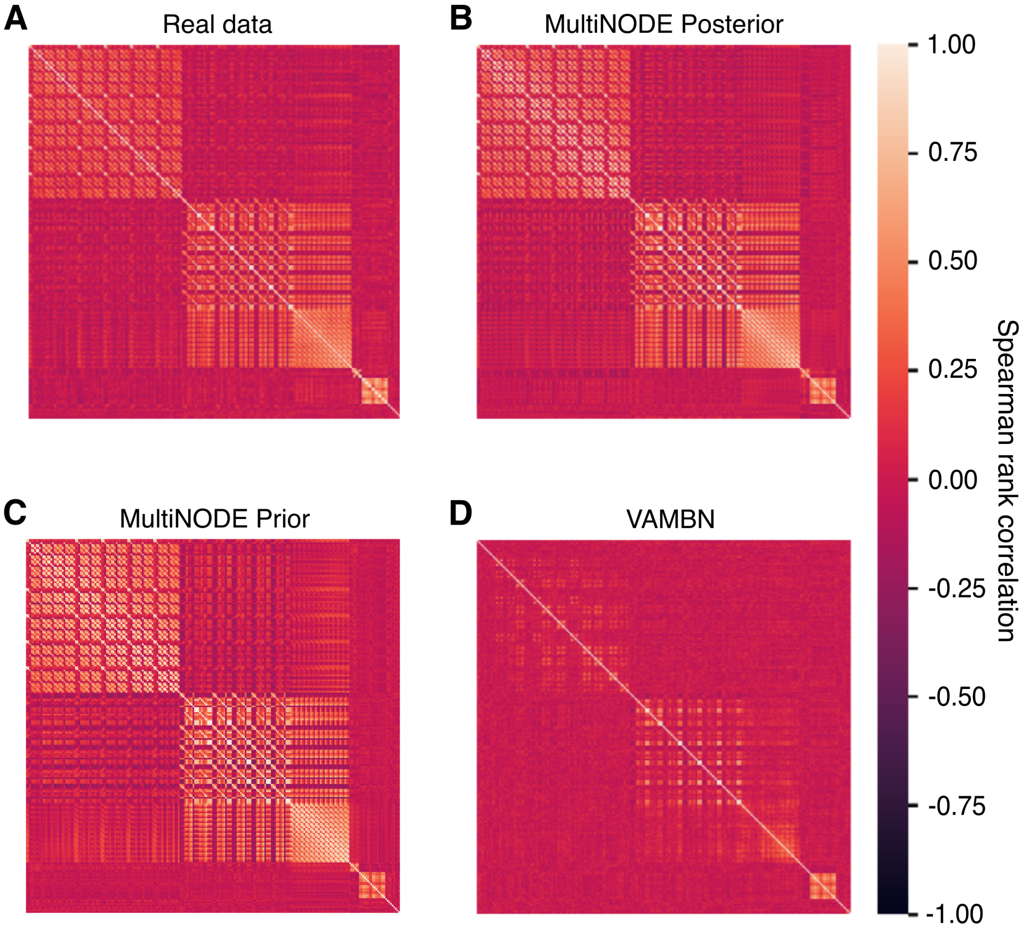


Figure 4: Equivalent results for the NACC data are shown in Supplementary Fig. 6. a Real data. b Posterior sampling from MultiNODEs. c Prior sampling from MultiNODEs. d VAMBN-generated data.