Deconstructing and Approaching Heterogeneities in the Biomedical Field via Computational Modeling

Kumulative Dissertation
zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Sepehr Golriz Khatami
aus Esfahan, Iran

Bonn, 2022

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

1. Gutachter: Univ.-Prof. Dr. rer. nat. Martin Hofmann-Apitius
2. Gutachter: Univ.-Prof. Dr. rer. nat. Thomas Schultz
Tag der Promotion: 22.11.2022
Erscheinungsjahr: 2022

Abstract

Natural variation between human characteristics as well as differences across collected datasets in disparate medical or research centers on various levels (e.g., semantical and technical) lead to high heterogeneity in terms of patients and data in the biomedical field. These heterogeneities not only impede understanding of disease pathology and clinical diagnosis but also their implications in the treatment of disease are substantial. Moreover, these heterogeneities limit the impact of computational solutions on clinical practice in spite of their high potential in bringing significant advances in the biomedical domain.

In this thesis, we address the aforementioned issues in the context of complex diseases, namely Alzheimer's disease (AD) and multiple types of cancers. First, in an in-depth study, we shed light on hurdles derived from heterogeneities and outline how they can restrict the impact of computational models in clinical practice with special focus in AD. Then, to demonstrate the findings of the preceding work, we present a comparative study on characterizing the order of pathological markers by applying a computational model, more specifically a data-driven one, to multiple independent datasets collected in different research centers. In this work, we investigate how heterogeneity across datasets can result in disparities among the ordering of changes in AD biomarkers and influence the models' impact on clinical practices. Further, to provide a more meaningful biological context into AD pathology, we use a pure knowledge-driven approach to showcase different mechanisms of disease development and progression that genetic variants may cause. Finally, we conclude this thesis by proposing a novel methodology to address heterogeneity among cancer patients in the context of disease treatment. In this publication, with the help of highly predictive machine learning models and an innovative scoring algorithm, we evaluate whether a given sample that was formerly classified as diseased could be predicted as normal after treatment with a given drug taking into account the corresponding molecular signatures of that particular sample.

In summary, this thesis presents the challenges and their implications brought on by heterogeneities in the biomedical domain in order to understand disease pathology and possible treatments, and attempt to uncover avenues to tackle the hindrances. Such advances have numerous applications in the biomedical field, ranging from patient stratification to drug discovery and achieving the ideal of precision medicine.

Acknowledgments

To my wonderful parents, all I am today and hope to become rests on their shoulders. Throughout the time of this Ph.D. and all other difficult moments of my entire life, I have been always reminded that there is a safe haven called Home where I am offered unconditional love and support. Mom and Dad, thank you very much for everything.

To my supervisor, Prof. Dr. Martin Hofmann-Apitius, for giving me the opportunity to contribute to the scientific community and for all the support and guidance I had since the day I started as a young student in the department. Moreover, to Prof. Dr. Thomas Schultz, thank you for acceding to be the second reviewer of this thesis. Finally, to Prof. Dr. Reinhard Klein and Prof. Dr. Diana Imhof, thank you for being in my defense committee.

My sincere gratitude goes out to Dr. Daniel Domingo-Fernández, Colin Brikenbihl, Sarah Mubeen, and Yasamin Salimi who were my teammates through all of this journey and good friends. It's been a pleasure working with you. My thanks and appreciation also go to Alina Enns and Meike Knieps, who helped me with all the administrative work and organizational processes.

Last but not least, a heartfelt thanks to my friends, who I grew up with, but then later I left them; no matter how distant we were, we not only never lost touch but also our friendship grows stronger day by day. Living proof of a possible, significant relationship between miles and friendship lies with you. Thanks for your embrace whenever I come and for all the banter.

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.

Sepehr Golriz Khatami

Publications

Thesis publications

- **Sepehr Golriz Khatami**, Christine Robinson, Colin Birkenbihl, Daniel Domingo-Fernández, Charles Tapley Hoyt and Martin Hofmann-Apitius. "Challenges of Integrative Disease Modeling in Alzheimer's Disease". *Front Mol Biosci*, 6:158, (2020).

  https://doi.org/10.3389/fmolb.2019.00158

- **Sepehr Golriz Khatami**, Yasamin Salimi, Martin Hofmann-Apitius, Neil Oxtoby, and Colin Birkenbihl. "Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease". *Alz Res Therapy*, 14(1), 55, (2022).

  https://doi.org/10.1101/2021.11.14.21266316

- **Sepehr Golriz Khatami**, Daniel Domingo-Fernández, Sarah Mubeen, Charles Tapley Hoyt, Christine Robinson, Reagon Karki, Anandhi Iyappan, Alpha Tom Kodamullil, Martin Hofmann-Apitius. "A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease". *J Alzheimers Dis*, 80(2), pp.831-840, (2021).

  https://doi.org/10.3233/JAD-201397

- **Sepehr Golriz Khatami**, Sarah Mubeen, Vinay Srinivas Bharadhwaj, Alpha Tom Kodamullil, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. "Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures". *npj Syst Biol Appl* 7(1), 40, (2021).

  https://doi.org/10.1038/s41540-021-00199-1

Other publications

- **Sepehr Golriz Khatami**, Sarah Mubeen, and Martin Hofmann-Apitius. "Data science in neurodegenerative disease: its capabilities, limitations, and perspectives". *Current opinion in neurology* 33(2), 249-254, (2020).

  https://doi.org/10.1097/WCO.0000000000000795

- **Sepehr Golriz Khatami**, Maria Francesca Russo, Daniel Domingo-Fernández, Andrea Zaliani, Sarah Mubeen, Yojana Gadiya, Astghik Sargsyan, Reagon Karki, Stephan Gebel, Ram Kumar Ruppa Surulinathan, Vanessa Lage-Rupprecht, Saulius Archipovas, Geltrude Mingrone, Marc Jacobs, Carsten Claussen, Martin Hofmann-Apitius, Alpha Tom Kodamullil and the COPERIMOplus consortium. "Curating, collecting, and cataloguing global COVID-19 datasets for the aim of predicting personalized risk". *bioRxiv*, (2022).

  https://doi.org/10.1101/2021.11.14.21265797

- Philipp Wegner, Geena Mariya Jose, Vanessa Lage-Rupprecht, **Sepehr Golriz Khatami** , Bide Zhang, Stephan Springstubbe, Marc Jacobs, Thomas Linden, Cindy Ku, Bruce Schultz, Martin Hofmann-Apitius, Alpha Tom Kodamullil, and the COPERIMOplus consortium. "Common data model for COVID-19 datasets". *Submitted*. (2022).

- Anandhi Iyappan, **Sepehr Golriz Khatami**, and Martin Hofmann-Apitius. "Complexity across scales: a walkthrough to linking neuro-imaging readouts to molecular processes". *J Syst Integr Neurosci* 3, (2017).

  https://doi.org/10.15761/JSIN.1000151

Contents

1 Introduction

Heterogeneity is formally defined as ‘diversity in character or content’ [1]. Extrapolating this definition to the biomedical domain, heterogeneity appears as two main categories: i) diversity of patients, known as “patient heterogeneity”, and ii) disparities across collected datasets of different medical or research centers on various levels including semantical, statistical, technical, and clinical [2]. While the former category (i.e., patient heterogeneity) originates from the complexity of the human body and natural variation between certain characteristics of patients such as age, sex, beliefs, attitudes, disease pathology, and genetic profile [3, 4], the latter category is derived from varied factors including differences in the distribution of measured variables and recording methods (e.g., using inconsistent units and labels), study-specific patient recruitment processes (e.g., different inclusion and exclusion criteria defined based on study goal), data measurement tools (e.g., use of different machines from different manufacturers), and medical practices across countries [5]. On the one hand, the two types of heterogeneity are appreciated, each of which for a different reason. First, heterogeneity among humans makes individuals unique in which heeding to this uniqueness helps to identify subpopulations, base medical decisions on individual sample characteristics and lay the fundaments for precision medicine. Further, individual datasets with different characteristics can potentially provide exclusive, additional, and complementary information, and their combination may help to gain more comprehensive insights into the investigated study. On the other hand, disregarding or mishandling any of these types of heterogeneities poses unique challenges in the biomedical field. First, not heeding to differences among individuals may lead to poor biological resolution, inaccurate clinical conclusions in diagnosis, and ineffective diseases treatments that all ultimately add to disease burdens [6]. This challenge is becoming especially relevant with respect to large heterogeneities which are known to exist in complex disorders (e.g., neurodegenerative diseases and cancers). Second, the characteristics of individual datasets can propagate into their signals and bias the models derived from that particular dataset. This, in turn, impedes validation, reproducibility, and generalizability of the models which are crucial to make a model trustworthy, ensure robust scientific insights, and applicable in clinical practice [7]. This challenge is becoming particularly pertinent with reference to increasing popularity of computational-driven approaches, more specifically data-driven-based solutions during recent years.

The publications in this thesis are centered around the aforementioned Heterogeneity - driven issues by focusing on three major topics, i) explanation of how disparities across datasets limit the impact of computational models on clinical practice, ii) advocating for heterogeneity handling across datasets in the context of validation, reproducibility, and generalizability crises in parallel with utilizing this heterogeneity to provide more comprehensive insights into disease pathology, and iii) developing a methodology to contend with heterogeneity among patients in the context of disease treatment in order to customize treatment of individuals based on their specific disease characteristics and bring the concept of precision medicine into reality.

Before presenting the publications contained in this thesis, the background is given on several topics, including the complex diseases investigated in the course of heterogeneity analyses, their respective biomarkers, heterogeneities and related causes in them, the need for taking into account the heterogeneities, and finally, the state-of-the-art of computational models utilized for the comprehension of disease pathologies and treatment.

1.1 Neurodegenerative and cancer disorders

Neurodegenerative disease (NDD) and cancers are among the top five causes of death in the world [8]. NDDs are a class of disorders characterized by the progressive degeneration of neurons and associated cell types in the nervous system [9]. On the other hand, cancers are a group of conditions that are characterized by the uncontrolled proliferation of cells and boosted resistance to cell death [10]. These two groups of diseases impose immense social and economic burdens by not only impacting the patients but also affecting their families, caregivers, and healthcare systems. For example, dementia, an age-associated condition, alone affects over 55 million individuals worldwide, necessitating the annual investment of 18 billion hours of care by more than 17 million healthcare personnel at a cost of more than one trillion dollars to address dementia-related issues [11]. Extrapolating these statistics into the coming decades, an immeasurable socio-economic impact of dementia worldwide is expected, given the societal ageing trend. The following subsections introduce the most common types of dementia (i.e., Alzheimer's disease) and four of the most deadliest cancers (i.e., kidney, liver, breast, and prostate) which this thesis focuses on.

1.1.1 Alzheimer's Disease

Alzheimer's disease (AD), the most common form of neurodegenerative disease, is a multifaceted complex disease, characterized by progressive decline of thinking, remembering, reasoning, and behavioral abilities to such an extent that it can disrupt a person's daily activities. While many hypotheses have been proposed about the disease etiology, amyloid-beta (Aβ) and neurofibrillary tangles (NFTs) are the two most widely known pathological hallmarks of AD and are considered as major causes for neurodegeneration in AD. It has been demonstrated that these neuropathological changes, which start to occur up to 10 to 20 years before the onset of symptoms, impede the proper function of neurons by restricting their communication, which ultimately leads to neuronal death [12, 13]. It has been shown that neuronal deterioration first starts in the hippocampus, the primary brain region for learning and memory, and later expands to the cerebral cortex, which plays a major role in language processes and social behaviors [14]. Over time, neuronal degeneration spreads to other parts of the brain, and loss of daily living skills is gradually experienced by the patient [12, 15].

1.1.2 Kidney Cancer

Kidney cancer also called renal cancer, is the most common urological disease with an estimated incidence of more than 400,000 cases annually [16]. Kidney cancer has different types including, renal cell carcinoma (RCC), renal transitional cell carcinoma, Wilms tumor, and renal sarcoma. RCC is the most common type of kidney cancer, accounting for 90% of all cases, whereas renal sarcoma is rare and accounts only for approximately 1% of all kidney cancers [17]. RCC arises from the renal tubular epithelial cells and a range of risk factors from genetics to hypertension and lifestyle (e.g., smoking and obesity) contribute to the disease initiation and progression [18]. It has been revealed that mutation in multiple genes, including VHL, MET, FLCN, BAP1, FLCN, TSC1, TSC2, TFE3, TFEB, MITF, and PTEN increases the risk of RCC [19]. These genes are involved in pathways that respond to metabolic stress or nutrient stimulation (e.g., changes in oxygen, iron, nutrients, or energy) and thus, kidney cancer can be fundamentally considered as a metabolic disorder [20]. Similarly, it has been established that tobacco use and excess body weight can also predispose persons to the development of RCC [21, 22]. Patients with renal cancer have no symptoms in the early stages, however as the tumor grows larger, different symptoms such as blood in urine, lump in the abdomen, and anemia may appear.

1.1.3 Liver Cancer

Liver cancer is the sixth most commonly diagnosed cancer with an estimated incidence of more than one million cases annually [23]. Typically, liver cancer is classified into primary and secondary types, each of which incorporates several subtypes. While in the primary types, such as hepatocellular carcinoma (HCC) and hepatoblastoma, cancer initiates in the liver, in the secondary class, such as hemangioma and hepatic adenoma, the tumors are not liver cancers, but rather, have spread to the liver from other parts of the body, e.g., the pancreas, colon, or stomach [24]. HCC is the main form of liver cancer that accounts for 90% of all liver cancers [25, 26]. During the last few years, extensive research for unraveling the risk factors and molecular profiles of HCC has been conducted. It has been discovered that HCC develops from chronic liver disease caused by various risk factors such as chronic hepatitis B and C virus [27, 28]. Moreover, it has been described that mutation and unexpected activity of genes such as TP53, CTNNB1, ARID1A, and FGF also contribute to HCC development [24]. HCC patients usually experience no symptoms and thus, diagnosis of HCC is often made with advanced disease when patients already have some degree of liver impairment. This, in turn, results in no effective treatments that would improve the survival of HCC patients.

1.1.4 Breast Cancer

Breast cancer is the most common cancer type and the second leading cause of cancer death in females, worldwide [29]. In general, breast cancer is classified into invasive and non-invasive types, each of which includes different subtypes. In non-invasive types, such as lobular carcinoma in situ and ductal carcinoma in situ, the abnormal cells have not spread beyond the lobule or ducts where it is located. In contrast, in invasive types such as infiltrating lobular carcinoma and infiltrating ductal carcinoma, the abnormal cells spread from lobules or ducts to close proximity with breast tissue [29]. Invasive breast cancer, more specifically, infiltrating ductal carcinoma is the most common type — accounting for approximately 80% of all cases. A wide range of risk factors from sex, age, lifestyle, and family history to estrogen and gene mutations increase the possibility of initiation and progression of breast cancer. For example, it has been observed that mutation and abnormal activity of BRCA1, BRCA2, HER2, EGFR, and c-Myc genes contribute to breast cancer development. Similarly, it has been shown that undue alcohol drinking and excessive dietary fat intake can increase the risk of breast cancer [30]. While different symptoms such as change in size or shape of the breast, nipple discharge, and change in breast skin texture are defined as early indications of breast cancer, the majority of patients with breast cancer do not have any symptoms when they are first diagnosed with the disease.

1.1.5 Prostate Cancer

Prostate cancer is the most common cancer type and the fourth leading cause of cancer death in males, worldwide [31, 32]. Prostate cancer has different subtypes including, adenocarcinomas, interstitial cell carcinoma, and neuroendocrine carcinomas in which adenocarcinomas is by far the most common type of prostate cancer, diagnosed in up to 95 percent of cases [33]. Although the exact cause of prostate cancer has not yet been fully discovered, a broad range of risk factors from endogenous (e.g., genetics and ethnicity) to exogenous ones (e.g., diet and occupation) have been established contributing to the initiation and progression of the disease [34, 35]. For example, evidence has been shown that over 100 single nucleotide polymorphisms and genes such as HPC1, PMS2, HPCX, CAPB, and BRIP1 have been associated with an increased risk of prostate cancer [36, 37]. Similarly, it is believed that a higher saturated fat intake and a higher vitamin A level may contribute to prostate cancer [38, 39]. Unlike the other cancer types, prostate cancer grows very slowly and does not expand to other body organs rapidly. This not only explains why prostate cancer may not cause any symptoms for a long time but also justifies why it can often be managed well even after it has spread to other parts of the body. Nevertheless, the most common prostate cancer symptoms are frequent urination, weak or interrupted urine flow or the need to strain to empty the bladder, an urge to urinate at night, and blood in the urine are the most common prostate cancer symptoms.

1.2 Biomarkers in Alzheimer's disease and cancers

Biomarker, an amalgamation of "biological marker", is an indicator of biological or pathological processes, or a response to a therapeutic intervention that could be objectively measured and evaluated [40]. Biomarkers can be classified based on different parameters including their characteristics or their application. For example, based on characteristics, biomarkers can be categorized as non-molecular biomarkers (e.g., imaging biomarkers) or molecular biomarkers (e.g., lipids metabolites). Similarly, based on applications, biomarkers can be grouped including diagnostic, and prognostic, as well as biomarkers for investigation of the response to a therapeutic intervention. While various biomarkers have been established for the five aforementioned disorders, the commonly used biomarkers are explored for each of them in the following.

- **Alzheimer's disease**: a wide range of biomarkers throughout different biological scales has been established in AD, including, i) fluid-based biomarkers which are either extracted from the blood,such as glial fibrillary acidic protein [41] or from cerebrospinal fluids (CSF), such as Aβ and hyperphosphorylated tau protein [42], ii) imaging-based biomarkers, including magnetic resonance imaging (MRI)-based biomarkers which measure the volume of different brain regions as well as the structural integrity of the brain [43], and positron emission tomography-based biomarkers which can be used to investigate and monitor various brain systems such as neurotransmitter systems [44], iii) cognitive-based biomarkers which are used to measures mental performances such as quantifying attention or episodic memory [45], and iv) genetic-based biomarkers such as amyloid precursor protein (APP), presenilin 1 (PSEN1), and Apolipoprotein E (APOE) [46, 47].
- **Cancers**: Various biomarker types especially genetic ones have been es-

tablished in breast [48], prostate [49], kidney [50], and liver [51] cancers. Although these gene-based biomarkers improve our understanding of the underlying molecular disease mechanisms, they are highly unstable and their clinical usage is limited due to multiple reasons, the first being the heterogeneity among patients as well as tumors. Second, genes, do not act in isolation, but through complex biological pathways. Therefore, it has been suggested to map the data at the genetic level to functional modules i.e., pathways, which are not only interpretable but also more stable and optimize patient-specific therapeutic strategies. Different pathway-based biomarkers have been realized in these cancers. As an illustration, lectin-induced complement pathway, peptide ligand-binding receptors, immune-related pathways including the inflammatory response, and metabolic-related pathways such as oxidative phosphorylation signaling pathways are known pathway-based biomarkers to recognize patients with breast cancers [52, 53]. Furthermore, the growth hormone receptor signaling pathway, and the JAK-STAT cascade involved in the growth hormone signaling pathway are widely-known biomarkers in liver cancer [54]. Changes in the cell cycle and the p53 signaling pathway are two major signatures that are known in patients with prostate cancer [55]. Finally, glycolysis, propanoate metabolism, pyruvate metabolism, urea cycle, and arginine/proline metabolism, as well as the non-metabolic p53 and FAS pathways have been established for early diagnosis and treatment in kidney cancer [51].

1.3 Heterogeneity in Alzheimer's disease and cancers

AD and cancers are notoriously heterogeneous — each of which is present in a variety of subgroups and has its unique set of histopathological and biological characteristics. The subtypes, corresponding characteristics, and the related causes are explored in the following subsections within the frame of each disease.

1.3.1 Heterogeneity in Alzheimer's disease

Evidences indicate that there are variances between individual patients with AD in terms of genetics, neuropathology and pattern of brain atrophy, pathways of disease development and progression, clinical manifestation, and the rate of disease progression. For example, studies on genetics of AD patient revealed that there are two main subtypes namely early-onset (early-onset familial AD) and late-onset (late-onset AD). While the former type presents in young patients (before age 65), the latter type manifests at divergent ages (usually after age 65) [56]. Based on neuroimaging and neuropathological studies, there are three different patterns of brain atrophy, namely typical (balanced NFT counts in the hippocampus and cortex with balanced atrophy in both regions), limbic-predominant (more NFTs in the hippocampus with more atrophy in the hippocampus), and hippocampal-sparing (more NFTs in the cortex and more atrophy in cortical) [57, 58]. Recently, in the extension of previous studies, the fourth subtype called "minimal atrophy" has been identified which is characterized by a lower level of atrophy and higher frequency of NFTs distribution compared to other subtypes [59, 60]. Furthermore, given the well-established association of AD with various pathophysiologic mechanisms including tau-mediated neurodegeneration, Aβ, neuroinflammation, synaptic signaling, and immune activity [61, 62], molecular-based subtyping studies have identified different subtypes — each of which corresponds to dysregulation of one mechanism (e.g., tau-predominant or Aβ-predominant subtype) or combinations of multiple mechanisms in parallel such as decreased synaptic signaling and increased immune response [63, 64]. Moreover, based on the clinical manifestation, there exist two AD subtypes, namely amnestic and non-amnestic ones. While in the amnestic subtype episodic memory loss is more prominent, in the non-amnestic subtype, difficulties with language or visuospatial/perception deficits are more predominant [65]. Additionally, it has been observed that the disease progresses at different rates among patients with AD. while the slow progression form of AD proceeds slowly, with an average survival time of 8 years and a mean cognitive reduction of 3 Mini-Mental State Examination (MMSE) points per year, the rapid form progresses fast with survival shorter than 4 years and MMSE score decreases of more than 5 points per year [66].

The heterogeneity between patients with AD is attributed to different determinants including, 1) risk factors such as age, sex, genetics and family history, 2) protective factors such as education as a proxy of cognitive reserve, and 3) concomitant non-AD pathologies such as different forms of cerebrovascular disease [67, 68]. For instance, it has been explained that heritability for AD is up to 80% and genetic influences on timing of the disease [69]. Gatz *et al.* [69] have demonstrated that early-onset familial type usually characterized by mendelian inheritance and rare mutations in three autosomal dominant causal genes (i.e., amyloid precursor protein, presenilin1, and presenilin 2), while there is no consistent mode of transmission in late-onset type and APOE is considered to be the main responsible gene in this type. Similarly, it has been revealed that the hippocampal-sparing subtype is more frequent in non-amnestic, males, APOE $\epsilon4$ noncarriers, and younger age patients, while the limbic-predominant subtype is more frequent in amnestic, female, APOE $\epsilon4$ carriers, and older age patients. On the other hand, it has been shown that the patients with hippocampal-sparing subtype have the highest level of education and disease-related clinical symptoms are not manifested as early as patients with minimal atrophy subtype which have the lowest level of education. It has been explained that patients with a higher level of education have a more pathology-robust brain network which avoids aggregation of NFTs compared to patients with a lower level of education [70]. This explains why patients with less education develop symptoms earlier in the minimal atrophy subtype when the NFTs start to form and before atrophy can be identified on MRI. In addition, it has been demonstrated that cerebral amyloid angiopathy, a form of cerebrovascular disease, makes a stronger contribution to hippocampal-sparing, whereas hypertensive arteriopathy, another form of cerebrovascular disease, may make a stronger contribution to limbic-predominant AD [71]. However, how non-AD coexisting pathologies contribute to AD heterogeneity remains unanswered.

Heterogeneity in AD is not merely limited to variations between patients but extends to cohort datasets that have been collected by different research centers as well. This type of heterogeneity can be attributed to different factors including study-specific inclusion and exclusion criteria that are specified in light of the study's objectives and could lead to a disparate distribution of biomarkers or distinct statistical distributions of equivalent biomarkers. For example, the department of defense Alzheimer's Disease Neuroimaging Initiative sites [72] recruited Vietnam veterans aged 60 to 80 years with a documented history of traumatic brain injury (TBI) with or without posttraumatic stress disorder (PTSD) to investigate the associations between a history of TBI and/or current PTSD and brain AD pathology and thus comorbidity variables are the most accentual measurements. However, presymptomatic evaluation of novel or experimental treatments for AD [73] recruited individuals aged 60 years or older with a parental or multiple-sibling history of AD to pursue innovative studies of pre-symptomatic AD and prevention trials in which genetics and family variables are the most highlighted ones. These disparities in key characteristics of cohort datasets impose flaws with regard to interoperability between existing datasets both from a semantical and statistical perspective which further hamper the robustness and reproducibility of results achieved in the course of computational modeling. This together with heterogeneity among patients explains why despite many years of research and investment into AD, only a few computational models can be found that have an impact on contemporary clinical practice and there is still no cure for AD [74] .

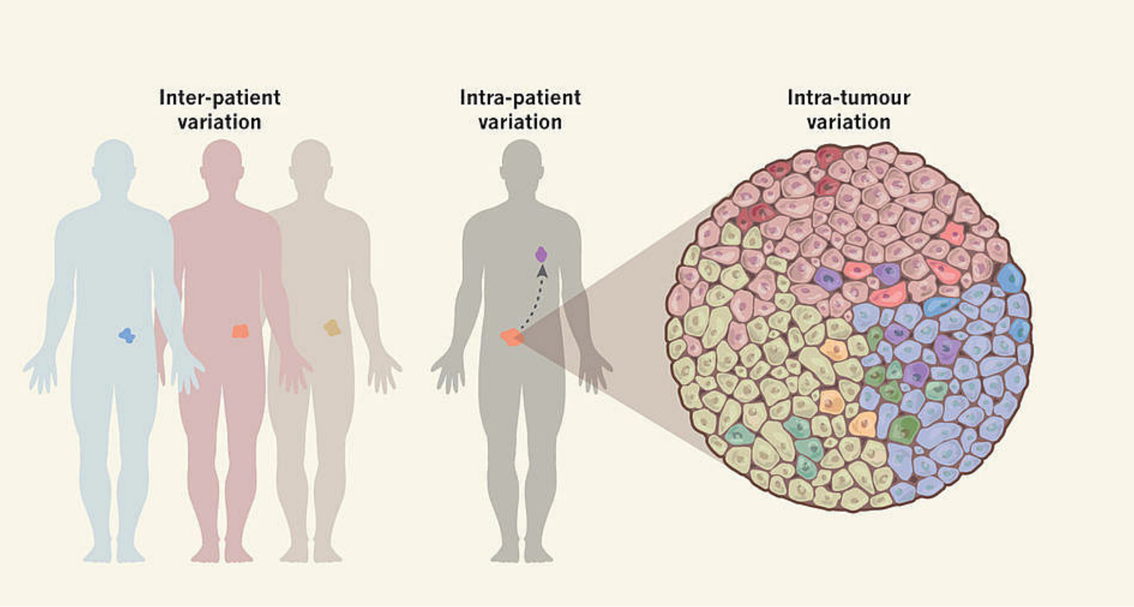


Figure 1: Heterogeneity in cancer. The characteristics of cancers vary between patients (i.e., interpatient heterogeneity), between primary and metastatic tumours in a single patient (i.e., intertumor intrapatient heterogeneity), and between the individual cells of a tumour (i.e., intratumor heterogeneity). This figure was adapted from [76].

1.3.2 Heterogeneity in Cancers

Cancer is a dynamic disease whose development and progression does not follow a fixed course and continues to evolve even after malignant transformation. This ongoing evolution produces a molecularly heterogeneous bulk tumor including cancer cells with distinct molecular, morphological, phenotypic, and particular degrees of sensitivity to antitumor treatment profiles [75]. These heterogeneities are not only observed between different patients (intertumor/interpatient), but also between distinct tumors within an individual patient (intersite/intertumor intrapatient heterogeneity), and within a tumor in one patient (intratumor) (Figure 1).

Each of the heterogeneity types (i.e., intertumor, intersite, intratumor) can be attributed to various factors [77, 78]. For example, it has been shown that interpatient heterogeneity generally results from specific factors of patients such as genomic variations of germline, differences in somatic mutation profiles, and environmental factors [75, 79]. This genetic variation can be passed from generation to generation and accumulation of multiple variation over many years not only adds to heterogeneity but also increase the chance for developing certain cancers (e.g., breast and prostate) in individual lifetime [80]. It has been estimate that up to 10% of cancers are hereditary[81]. Furthermore, among various factors that give rise to intratumor heterogeneity, genomic instability is the most prominent factor [79]. Genomic instability refers to the high frequency of mutation in the genome during the life cycle of cells which is attributed to various factors such as hypoxia and therapeutic intervention [82–84]. It has been described that hypoxia promotes genomic instability by disrupting genomic integrity through impeding replication errors correction (i.e., it downregulates DNA mismatch repair system) [85]. Similarly, it has been revealed that therapeutic intervention (e.g., chemotherapy) contributes to genomic instability by increasing the tumor mutational burden and selective pressure towards resistant cancerous cells populations [86, 87].

Heterogeneity has been observed in many cancers including prostate, breast, kidney, and liver cancers, each of which is classified into various subtypes based on genomic profiling. For example, based on specific gene fusions and mutations, prostate cancer is categorized into seven subgroups including ERG, ETV, ETV4, FLI1 SPOP, FOXA1, and IDH1 [88]. Similarly, breast cancer based on genomic and transcriptomic profiling, expression pattern of hormone receptors (estrogen and/or progesterone receptors; ER/PR), and epidermal growth factor receptor 2 (HER2/Neu) is subtyped into six different types namely: normal breast-like, luminal A (ER+/PR+ and Ki67-low), luminal B (ER+/PR+ and HER2+ or HER2–, and Ki67-high), HER2-enriched (HER2+), basal-like and claudin-low [89]. Furthermore, there are well-established genetic mutations that cause the four subtypes of kidney cancer, including clear cell kidney cancer (mutations of VHL gene), papillary kidney cancer (mutation of MET and fumarate hydratase), chromophobe kidney cancer (mutations in folliculin), and translocation kidney cancer (translocation in TFEB/TFE3 genes) [90]. Additionally, investigations on the mutational landscape of liver cancer recognized five subtypes including mutation in TERT, CTNNB1, TP53, ARID1A/2, and AXIN1 genes.

Despite considerable progress in tumor heterogeneity research, its origin and consequences remain poorly understood and heterogeneity is still a great barrier to the successful treatment of cancer. Moreover, according to current medication approval requirements in Europe and the United States, the experimental treatment group must show a clear statistically significant benefit as compared to the control group for a medicine to be approved instead of its approval based on a subpopulation of responders in a clinical trial [91]. These imply that it could be the right time to take a step back, heed the heterogeneity and propose avenues to approach it before starting yet another (destined to fail?) study.

1.4 Need for addressing the heterogeneities

Heeding the two types of heterogeneities in the biomedical domain is of utmost importance, each for a different reason(s) which are explored in the following subsections within the frame of heterogeneity type.

- **Heterogeneity among patients:** There exist two important reasons to address patient heterogeneity. First, it enhances the accuracy of disease diagnosis. For example, it has been evidenced that memory problems are the most common first cognitive symptom experienced at any age in patients with AD, while non-memory symptoms (e.g., visuospatial problems) are more prevalent in younger patients (roughly one of three patients with the disease onset before the age of 65 years represents atypical indication of AD) [92, 93]. Nevertheless, diagnosis is frequently missed in young AD patients who experience visuospatial problems or behavioral phenotypes, as many clinicians do not think of AD when they examine a young patient with non-memory cognitive symptoms. However, taking this heterogeneity into account can improve the accuracy of disease diagnosis and help to identify individuals at risk of developing symptoms which ultimately leads to better individualized disease management. Second, it helps to prescribe a more effective treatment as it has been shown that a response triggered by a drug in a given patient may differ if administered in another patient. For instance, it has been shown that the luminal subtypes of breast cancer benefit from treatment with hormones. However, the HER2 subtype not only does not experience any benefit from hormone therapy but also cancer cells become resistant to treatment [94]. This drug resistance remains the primary stumbling block to cancer therapies [95]. Therefore, gaining insight into molecular makeup of individuals and prescribing treatment based on their characteristics, rather than base treatment on the prevailing clinical diagnosis alone improves treatment efficacy and eliminates the misuse of ineffective and potentially harmful treatment [96].

- **Disparities across the datasets:** Given the high dependency of computational solutions, specifically data-driven ones, on data, approaching disparities across datasets strengthens the impact of these approaches on clinical practices and further supports clinical decision-making. It has been demonstrated that being generalizable and reproducible is one of the criteria that a data-driven model should have to be able to impact clinical practices [97]. However, disparities across datasets on different levels (e.g., semantical and statistical) impede the interoperability of datasets which further hamper the validation, generalizability, and reproducibility of the derived results. Considering heterogeneity across datasets ensures the transferability of models and results across disease (sub)populations and enhances the potential impact of models and cutting-edge technologies on clinical practices.

1.5 Translational research: applying computational models to the clinic

The emergence of big biodata and its generated knowledge, machine learning, and artificial intelligence brings steep hopes that these cutting-edge technologies lead to considerable progress in the biomedical field [98]. Computational methods enable us to provide a holistic view of the biological system and better insights into diseases by employing high throughput *omics* (e.g., genome, proteome, transcriptome) as well as personalized clinical data (e.g., imaging, digital device data), and capturing complex relationships among them. Computational methods can be placed in two primary categories namely, knowledge-driven and data-driven methods — each of which is explored in the following subsections within the frame of each disease.

1.5.1 Knowledge-based modeling

The advent of big data in the biomedical field generates an enormous amount of information and knowledge which offers us the opportunity to investigate biological systems at a high degree of granularity. However, leveraging existing knowledge and information requires the formalization and assembly of these in a computable form which ultimately leads to the construction of the biological system models. These models not only facilitate the explanation of relevant biological mechanisms and how components of biological systems, interact but also predict the system's behavior perturbed by internal factors (e.g., mutations) or external ones (e.g., therapeutic interventions or environmental changes) [99].

While different biological knowledge-based modeling approaches such as systems biology graphical notation and proteomics standards initiative-molecular interaction are available, we only survey conceptual modeling ones as these approaches organize biological complexity that we aimed for in this thesis by capturing the important characteristics of a biological system and structuring our conceptualizations into the relevant entities and their relationships. Resource Description Framework (RDF), Systems Biology Markup Language (SBML), Biological Pathways Exchange (BioPax), and Biological Expression Language (BEL) are the most common conceptual knowledge formats in the biomedical domain, each of which is briefly discussed in the following.

- **Resource Description Framework (RDF):** is a standard format derived from a semantic web domain to represent resources and the relations that connect them and is used for storing, managing, and modeling knowledge. RDF consists of triples, each includes a subject, a predicate, and an object. The subject is the acting resource, the predicate is the linking relationship, and the object is the resource that is affected by the subject. RDF allows the subject and object to be represented as a Uniform Resource Identifier (URI) which is the string of characters used to identify subject and object. This flexibility enables data merging, although the structures which form their basis may differ as opposed to other formats including Extensible Markup Language (XML). Additionally, the use of triples as semantic units prompts linking data across different resources.

- **Systems Biology Markup Language (SBML):** is an XML-based format that was originally designed to represent biochemical reaction networks; however, it can be used to describe other biological processes such as metabolic pathways, gene regulatory networks, cell signaling pathways, and disease models [100]. Furthermore, SBML enables users to incorporate quantitative information appearing as equations such as chemical interactions. Real entities are designated species and processes are denoted reactions. They can be ciphered as models that, when deciphered, firmly simulate chemical reaction equations. This capability of SBML not only makes it suitable for simulations of stochastic kinetic models, considering the dynamics of multiscale interactions of biological systems but also endorses the trade of biochemical networks' quantitative models between different simulation tools [101].

- **Biological Pathways Exchange (BioPAX):** is a standard language that uses web ontology language formats to describe biological pathways at the cellular and molecular levels. BioPax can capture and index a broad range of metabolic, signaling, molecular, and gene regulatory networks. BioPAX specifies five top-level classes (i.e., entities, genes, physical entities, interactions, and pathways) to endorse the pathways representation. BioPAX has been exploited by different databases to illustrate pathways interactions in various organisms in a computable form which facilitates the exchange of information between pathway users and databases and promotes integration, visualization, analysis, and interpretation of pathway data. However, in contrast to SBML, BioPAX is unable to simulate the dynamic and quantitative components of biological processes [102].

- **Biological Expression Language (BEL):** is a high-level knowledge-based systems biology modeling language that enables users to describe causal and correlative relations between biological entities. BEL can function as a semantic platform that facilitates capturing, integrating, and analyzing a wide range of mechanistic details of biological phenomena, from the molecular to organism scale, and thus it is a perfect candidate for modeling complex disease biology. Similar to RDF, BEL describes the relationship between biological components in the form of triple (subject-predicate-object) in a context-specific manner across multi-scales. Objects in a BEL triple can be the subject of one or multiple other triples which help to develop a knowledge assembly in the form of a conceptual graph [103]. This knowledge assembly further can be subjected to graph algorithms for analyses and reasoning such as network perturbation amplitudes [104] and reverse causal reasoning [105]. A simple example of a BEL statement capturing BDNF infusion increases phosphorylation of the mitogen-activated protein (MAP) kinases [106] is depicted in Figure 2.

1.5.2 Data-based modeling

In coincident with the advent of big biodata, data-based modeling using cutting-edge technologies such as artificial intelligence and machine learning has grown rapidly as well. The data-based modeling can be largely classified into two categories: i) models that utilize more traditional approaches such as linear regression and ii) models that employ more advanced artificial intelligence and machine learning (ML) approaches [107]. Traditional models have the advantage of being simple to comprehend and tend to include a small number of clinically relevant variables [107]. However, the latter property may lead to overlooking the

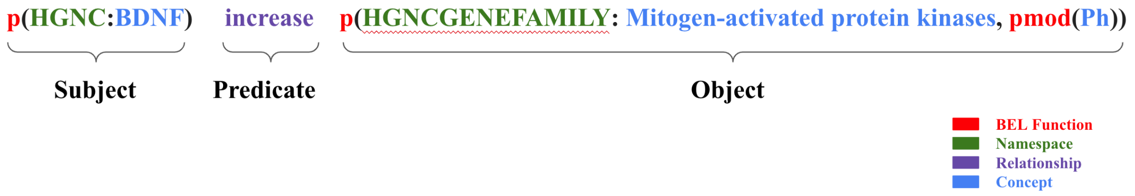


Figure 2: Biological Expression Language. The triple represents that BDNF protein increases the phosphorylation of the MAPK. 'BDNF' is a subject, 'increases' is a predicate, and 'phosphorylation' is an object. A molecular entity (e.g.,genes, proteins,an abstract class such as biological processes, disorders, biochemical events) can be used for the subject and object. The type of relationship between the subject and the object is represented by the predicate. External namespaces, such as the HUGO Gene Nomenclature Committee (HGNC) utilized in this example, are allowed to be used in BEL for the formal description of concepts.

interdependency among the involved features within or across the various biological scale participating in diseases, specifically complex diseases such as AD and cancers. Furthermore, traditional-based models rely on *a priori* assumptions which do not often match clinical practice [107]. On the other hand, advanced ML-based models have a high degree of flexibility, are devoid of *a priori* assumptions, and allow inference and conclusions to be made directly from samples. Moreover, more advanced ML-based models are able to integrate data from various modalities (e.g., *omics*, imaging) which helps us to improve our understanding of disease pathology through capturing complex relationships between the features which are contributing to the diseases.

While potentially such models provide us valuable insights into the research areas where they are exploited, for example, biomarker discovery [108], disease diagnosis [109], and drug repurposing [110] , a remarkable number of contemporary models either discount heterogeneity or employ ad hoc methods to account for heterogeneity. In the following, different types of models (i.e., those that account for heterogeneity as well as those that discount it) are explored in the context of disease progression modeling, drug repurposing, and predicting drug response, the applications that this thesis focuses on.

- **Disease progression models:** are data-driven models that can be categorized into two main classes: i) the models that help to identify subjects who are at higher risk of converting to patients (e.g., normal or mild cognitive impairment subjects convert to AD), and ii) the models that serve to understand the temporal evolution of multimodal disease-related measurements (i.e., biomarkers). Both classes can take advantage of cross-sectional or longitudinal observations (i.e., repeatedly measured biomarkers over a period of time), traditional or advanced ML methods, and a single or combination of biomarkers. In the last decade, numerous disease progression models in each class have been developed by utilizing different types of data and data-based approaches. For example, [111, 112] exploited cross-sectional data and traditional approaches while in contrast [113, 114] applied a more complicated ML approach on longitudinal data to determine subjects with a higher risk of conversion to AD patients. Similarly, [115, 116], employed cross-sectional data and traditional approaches, whereas [117, 118] utilized longitudinal data and more advanced ML approaches to model the temporal evolution of biomarker trajectories in AD. Furthermore, [119, 120] used different biomarkers (i.e., neuroimaging, lab results, and neuropsychological) while, [121, 122] employed only neuropsychological biomarkers to detect subjects who are at high risk of developing AD. In contrast to previous models in which heterogeneity among patients was disregarded, [123, 124] developed a model with the capability of handling patient heterogeneity to represent distinct temporal progression patterns using cross-sectional patient studies. Despite the establishment of different AD disease progression models with the capability of handling heterogeneity among patients, none of them have reached the level of clinical accuracy that allows researchers to develop targeted clinical trials and facilitate personalized health care.

- **Drug repurposing and response prediction:** a wide range of data-driven drug repurposing models have been developed. These models can be classified into different categories based on the input data including i) transcriptomics - based ii) electronic health record-based (EHR-based), iii) structure-based models, and iv) phenotype-based [125, 126]. As an example, transcriptomics - based models take advantage of large-scale transcriptomics data in which together with machine learning approaches suggest new disease targets and next-generation treatments. [127] combined cancer-specific gene expression data and related them to drug response using the deep neural network method and found a novel use for the chemotherapeutic drug vinorelbine in titin-mutated tumors. Furthermore, EHR-based approaches leverage real-world patient data to identify new applications for approved drugs, outside the scope of the original medical indication. For instance, [128] improved cancer survival rates by linking two large EHR databases and developing a machine learning framework that showed metformin, an anti-diabetic medication, reduced mortality following a cancer diagnosis. Moreover, structure-based methods use target protein-derived information to find new targets for drugs. These approaches anticipate binding poses and drug activity by virtual screening current medications against a wide number of clinically relevant targets. [129] developed a convolutional neural network that takes the 3D representation of a protein-ligand interaction as input, learns key features of the interactions and discriminates correct and incorrect binding poses as well as known binders and non-binders. Additionally, phenotype-based models use the data yielded from studying the phenotypic effects of drugs on cells, complex tissues with dose-response, cell death, and apoptosis assays. [130] investigated the effect of kinase inhibition data on neurite growth, a vital process for neural regeneration in central nervous system injury by utilizing support vector machines. Using support vector, they could find both kinase targets whose inhibition promoted neurite growth, as well as kinase “anti-targets” whose inhibition blocked growth.

Besides, in recent years, various models have been established to predict drug response, each of which utilized different types of data-driven approaches and input data. For example, [131] used cohort genomic, chemical structure, and target information together with a network-based method, named HNMDRP, to accurately predict cell line-drug associations through incorporating relationships among cell lines, drug, and target. In another study, [132] developed a deep learning model that predicts anticancer drug responsiveness based on genomic profiles of human cancer cell lines and drug structural profiles. [133] used pathway signatures derived from cell lines as input to kernelized Bayesian matrix factorization. While different drug response prediction models have been developed, these methods thus far fail to account for heterogeneity among patients and until now the precision medicine concept is far from becoming reality.

1.6 Outline of the thesis

This thesis focuses on multiple important questions raised by the two types of heterogeneities (i.e., heterogeneity among patients and disparities across cohort datasets) including, i) how can heterogeneities limit the impact of computational models (i.e., data-driven models) on clinical practices, ii) in light of the need for reproducibility and generalizability of data-driven models how can cohort-specific models be comparable despite existing disparities across cohort datasets, iii) how can exclusive and additional information provided by individual cohort datasets be harnessed to deliver a more comprehensive picture of disease development and progression, and iv) how can heterogeneity among patients be addressed in the context of disease treatment and to bring the concept of precision medicine into reality. To approach these questions, the thesis is structured as follows:

Chapter 2 investigates the heterogeneity-driven flaws which we encounter in the course of developing computational models to further our understanding of disease pathophysiology. We inspect a wide range of computational models developed in the context of AD, compare their strengths and weaknesses, and discuss how the heterogeneities-based bottlenecks limit the clinical impact of these models. This work can serve as a checklist of bottlenecks for researchers which have to be taken into account while initiating a study or establishing computational models.

Chapter 3 presents the utilization of a data-driven model to explore how the pattern of biomarker changes vary not only among different patients but also between plentiful cohort datasets with reference to AD. This empowers us to approach reproducibility and generalizability as we also discuss in the last chapter. Additionally, we develop a novel algorithm that exploits the exclusive and additional information available in cohort datasets unique to each study and provides more complete insight into disease development and progression than insights brought by single datasets. These two aspects of heterogeneities are important to address both the transferability of models across AD (sub)populations and to improve our understanding of disease progression.

Chapter 4 introduces disease-specific knowledge assembly building and its applications in decoding biologically interesting problems in that disease. We have curated the AD knowledge assembly [134] to enrich it with multiscale information (genetic to neuroimaging) from scientific literature and different biological databases such as DisGeNET [135]. Using this multiscale enriched AD knowledge assembly, we prioritize multiple critical mechanisms in AD where the genetic layer may have an impact on the neuroimaging layer. We show such a knowledge assembly does help us to characterize how different biological scales interact through diverse biological processes as well as empowers us to establish druggable mechanisms in AD.

Chapter 5 demonstrates how, with the help of highly predictive machine learning models and an innovative scoring algorithm that calibrates a samples' pathway activity scores we can simulates a drug response in individual cancer patients. In other words, we examine whether a given sample that was formerly classified as a patient could be predicted as normal after treatment with a given drug. This technique works as a proxy for the identification of potential drug candidates for a particular sample and addresses the heterogeneity among patients which help to prevent drug resistance as well as brings the precision medicine concept into reality.

The final chapter summarizes the core message of 'heterogeneities in the biomedical field', presents the limitations, and discusses possible future directions of this work, serving as a conclusion of the thesis.

2 Challenges of Integrative Disease Modeling in Alzheimer's Disease

This section presents the following publication (see Appendix 7.1):

Sepehr Golriz Khatami, Christine Robinson, Colin Birkenbihl, Daniel Domingo-Fernández, Charles Tapley Hoyt and Martin Hofmann-Apitius. "Challenges of Integrative Disease Modeling in Alzheimer's Disease". *Front Mol Biosci*, 6:158, (2020).

Sepehr Golriz Khatami's contributions in this chapter is writing the manuscript.

Summary

Given the exponential advances of technologies during the last decades, numerous clinical studies have been conducted by different medical or research centers. Large amount of data(sets) have been collected and a wide range of computational models have been developed utilizing these collected data(sets). While these state-of-the-art models have prompted great breakthroughs in exploring many uncharted territories in biomedicine (e.g., understanding disease pathophysiology) and have generated significant insights, there exist only a few examples that impact current clinical practice. Given the high dependency of computational models on data, this limited number of impactful models can be largely attributed to one of the two main types of heterogeneity in the biomedical field, namely data(sets) heterogeneity, which is observed across data(sets) on various levels including semantical, statistical, and clinical. This heterogeneity imposes multiple challenges in the deployment of these state-of-the-art models and can be summarized into three main aspects including, i) insufficient performance of developed models, ii) difficulties in interpretation, and iii) difficulties in validation and reproducibility.

In this study, we mainly discussed these challenges and explained how data deficiency and disparities across data(sets) result in these challenges through the investigation of recently developed computational models in the context of AD. First, we placed the models into two main categories, namely hypothetical models (i.e., those which relied on reasoning over findings of previously published studies) and data-driven models (i.e., those which were informed directly by patient-level data), the latter of which is divided into two main classes. The first contains traditional statistical methods of generally lower complexity, such as linear models, and the second covers advanced AI/ML models. Then, we compared the dedicated models of each group together, enumerated their strengths and weaknesses, and discussed how imperfect data and heterogeneties across data(sets) lead to the challenges that limit the impact of these models on clinical practice. As an illustration, we have discussed how given that most observational cohorts are not representative of the general AD population, to corroborate findings it is important to validate the resulting models with an independent cohort study. However, due to disparities across cohort studies (e.g., different inclusion and exclusion criteria defined based on study goal), interoperability between datasets is limited and thus validating the resulting models with an independent cohort study is a non-trivial task. We thus recommended the annotation of datasets using controlled vocabularies to address the challenges associated with dataset interoperability, among other possible solutions. Furthermore, we proposed avenues to address

several of the other challenges discussed in this work.

In summary, this study enumerated various challenges that limit the influence of computational models on clinical practice by reviewing developed models that utilize biodata and state-of-the-art technologies. The challenges were explained in the context of a specific disease (i.e., AD), however, they can potentially be extrapolated to other diseases such as cancers. Essentially, this work can be considered as an inventory of hindrances for researchers which have to be taken into consideration whilst initiating a clinical study or developing a computational model.

3 Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease

This section presents the following publication (see Appendix 7.2):

Sepehr Golriz Khatami, Yasamin Salimi, Martin Hofmann-Apitius, Neil Oxtoby, and Colin Birkenbihl. "Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease". *Alz Res Therapy*, 14(1), 55, (2022).

Sepehr Golriz Khatami's contributions in this chapter are: conceptualization, implementing the methodology, preprocessing the data, running experiments, and writing the manuscript.

Summary

The publication presented in the previous chapter mainly enumerated the challenges (e.g., difficulties in validation and reproducibility) that are imposed by data(set) heterogeneity in the context of AD. In line with the previous study, in this chapter, we showed how disparities across data(sets) may limit validation, reproducibility and generalizability of computational models in the context of the same disease in more detail by investigating two scenarios. First, we examined whether the results obtained from a dataset are consistent across the other cohort datasets despite their disparities. In other words, we tested whether the results obtained by a model from a cohort dataset are robust and reproducible if fitted to an independent cohort dataset. Second, we analyzed whether a fitted model on a cohort dataset learns potential cohort-specific characteristics that could hamper the reproducibility and generalizability of results.

Alongside the challenges that disparities across data(sets) present in different contexts, such as reproducibility, in theory, they allow for developing unique models from individual data(sets) which, when aggregated, may help to gain more comprehensive insights into the investigated study as individual datasets potentially provide exclusive, additional, and complementary information. In another endeavour in this chapter, we explored whether aggregating results across datasets can harness this complementary information and provide a more complete picture of the disease.

So far, only limited studies in the AD domain, such as [136–138], have applied their models to data from other cohorts besides the discovery cohort, although they were mostly to diagnose and predict patient outcome. Furthermore, to the best of our knowledge, only a single study [139] focused on integrating information from individual studies while addressing the same biological question to arrive at a more reliable and comprehensive picture of the disease specifically, to identify and rank potential driver genes of AD.

In light of this shortcoming, we first conducted a systematic, in-depth investigation concerning the validation, robustness, and reproducibility of results obtained from different independent AD landmark cohort studies in the context of sequences of pathological marker (i.e., biomarker) changes. To do so, a probabilistic generative model called an event-based model [116] was deployed. We fit the model to ten independent AD cohort datasets and compared the results. While we observed general consistency over the changes of biomarkers across all cohorts (i.e., changes start with abnormality in Aβ, followed by tauopathy, memory impairment, and ultimately brain deterioration), slight variation in the position of these core features was identified. We explained that this variation could be caused by i) distinct statistical biases of the cohorts, for example introduced through specific recruitment criteria, ii) distinct prevalence of AD disease progression subtypes that follow different disease mechanisms, or iii) mixed neuropathologies. Furthermore, we developed a novel rank aggregation method to combine the obtained sequences of biomarkers in the previous step which enabled us to utilize exclusive and complementary information unique to each study. By doing so, we could generate a sequence of biomarker changes that is highly multimodal and more comprehensive than sequences built from individual datasets. Similar to the sequence of biomarker changes obtained from individual cohorts, the changes of biomarkers in the aggregated sequence started with abnormality in cerebrospinal fluid biomarkers (i.e., Aβ and tauopathy), followed by memory impairment, and ultimately brain atrophy.

Essentially, in this study, we demonstrated that in light of the challenges in model validity and reproducibility, it is critical to explore beyond single data sources, validate obtained results across different cohort studies, and continuously develop and assess data-driven methodologies. To that end, we identified general consistency across data-driven sequences of biomarker changes derived from multiple independent cohorts using the event-based model and only minor differences in the position of the main biomarkers that were available in all cohorts were observed. In addition, the novel aggregation method developed harnesses the heterogeneity in cohort study designs and measurements and generates a meta-sequence that provides a more complete, and robust, picture of the sequence of biomarker changes to improve our understanding of disease progression.

4 A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease

This section presents the following publication (see Appendix 7.3):

Sepehr Golriz Khatami, Daniel Domingo-Fernández, Sarah Mubeen, Charles Tapley Hoyt, Christine Robinson, Reagon Karki, Anandhi Iyappan, Alpha Tom Kodamullil, Martin Hofmann-Apitius. "A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease". *J Alzheimers Dis*, 80(2), pp.831-840, (2021).

Sepehr Golriz Khatami's contributions in this chapter are: conceptualization, implementing the methodology, interpreting the results, and writing the manuscript.

Chapter 4

Summary

Ever-growing data in the biomedical field, including imaging and genetics, set the stage for new opportunities to understand disease pathophysiology, specifically in complex disorders such as AD. Given the effect that the genetic layer (e.g., genes and genetic variants) has on brain structure and function, linking these two disparate layers is one of the main avenues that pave the way to accomplish this feat (i.e., unravel disease pathophysiology). While linking molecular mechanisms to clinical readouts is non-trivial, various initiative such as Enhancing NeuroImaging Genetics through Meta Analysis (ENIGMA) [140] as well as numerous studies have been carried out to examine the association of the genetic layer with brain structure and function by utilizing a wide range of methods from genome-wide association studies (GWAS) [141], to imaging genetics [142], differential equations [143], and the development of an an innovative data-driven framework [144]. However, these studies either only calculate statistical associations between genotype-phenotype and lack mechanistic insight into interaction between these two layers [141, 142] or fail to deal with the multitude of variables that are needed to represent the pathophysiological phenomenon involved in a multifactorial disorder, such as AD [143]. The inadequacies of these methodologies prompted us to develop a new method for interpreting how a certain genetic variant may affect neuroimaging feature changes through sequences of molecular causalities in AD.

This work explores the potential of knowledge assembly as a consolidated and computable collection of domain-specific knowledge, in investigating biological phenomena in AD. It leverages the semi-automatically curated domain knowledge around the disorder to investigate how genetic polymorphisms can cause functional changes in intermediate molecular features, which can then affect neuroimaging markers over a series of biological processes at many scales. First, we extract knowledge pertaining to single nucleotide polymorphisms (SNPs) and imaging readouts from the literature using natural language processing. Then, the genes corresponding to or associated with the SNPs are identified. Accordingly, a gene was selected for further study, and a corpus covering its role in AD was enriched with knowledge about multiscale biological processes. As the next step, manually extracted relations from this corpus were encoded in BEL to enable computer-aided reasoning. We demonstrated our method in a case study that suggests KANSL1 as a potential gene for the clinically observed link between genetic variations and hippocampus shrinkage. We discovered that the workflow prioritizes multiple mechanisms documented in the literature by which the gene may influence hippocampus atrophy, including cell proliferation, synaptic

plasticity, and metabolic processes.

Essentially, this study enabled us to explore the association of genotype to phenotype from a disease biology perspective instead of solely investigating their statistical associations (i.e., GWAS). Furthermore, this study manifested that integration of different biological entities throughout all modalities (i.e., from the molecular level to tissue and organ level) to knowledge assemblies helps to enhance the understanding of disease etiology and may shed light on heterogneity of pathways of disease development and progression — each of which can potentially be posited as attractive therapeutic targets for pharmaceutical intervention (see chapter 5). Moreover, the semi-automatic curation workflow developed during the course of this work has the potential to act as an instruction to enrich the knowledge assemblies for capturing and representing knowledge. Additionally, this study has indicated that formalization and capturing of knowledge in a computable form facilitates the development of tools for understanding, mapping, and representing the existing knowledge about a particular domain, and subsequently enables novel interpretation of biomedical data. Eventually, while the method has been used to demonstrate the mechanism behind the associations between genotype and phenotype in the context of neurodegenerative diseases, other diseases can also benefit from the approach.

5 Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures

This section presents the following publication (see Appendix 7.4):

Sepehr Golriz Khatami, Sarah Mubeen, Vinay Srinivas Bharadhwaj, Alpha Tom Kodamullil, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. "Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures". *npj Syst Biol Appl* 7(1), 40, (2021).

Sepehr Golriz Khatami's contributions in this chapter are: conceptualization, implementing the methodology, running the experiments, interpreting the results, and writing the manuscript.

Summary

The two preceding chapters (i.e., chapter 2 and 3), deconstructed data(set) heterogeneity, one of the two main classes of heterogeneity which exists in the biomedical domain. This chapter, however, approaches the other major heterogeneity class, namely heterogeneity between patients. This type of heterogeneity is not only observed between individual patients in their different characteristics, such as genetics and pathways of disease development and progression, but also is widely seen in their response to treatment. In other words, individuals differ in their response to therapies and a response triggered by a drug in a given patient differs if administered in another. This in turn leads to low efficacy or the failure of therapies and further resistance to medications [145]. This becomes even more evident in cancers where patients show different levels of sensitivity to treatment due to their heterogeneity and distinct molecular signatures [75]. Indeed, failure of therapies and resistance to treatments are the two major causes of death in cancer [146]. Therefore, cancer therapy needs to become more personalized, particular drugs against specific targets for the disorder must be established and more accurate tailoring of remedies to the target populations are needed [147]. To facilitate achieving this goal, accurate prediction of therapy responses in patients based on their molecular and clinical profiles is beneficial. Clinical trials are a means by which investigational therapies can be evaluated for efficacy and safety. However, they are time-consuming and expensive. One path to overcome these costs manifests in the utilization of state-of-the-art technologies (i.e., ML) and patient-specific molecular and clinical profiles (e.g., *omics*). Various studies have been performed in this non-trivial direction. For example, [148] utilized gene-expression profiles along with combination of the genetic algorithm and the k-Nearest Neighbor to predict potential therapeutic drugs in breast cancer. Similarly, [149] employed an established ML approach to build models of drug response based on transcriptomic data from breast cancer samples. Furthermore, [150] deployed epigenomics and three different ML algorithms to predict drug response in different cancers.

While a wide range of *omics* data has been utilized for drug response prediction in cancer research, it has been shown that gene-expression profiling is the most informative data for this purpose [151], Nevertheless, the validity of results based on individual gene markers has been questioned in several recent studies [152, 153]. These studies have enumerated different reasons including, i) small sample size of typical clinical data, ii) inherent noise in high-throughput measurements, and iii) that genes do not act in isolation. To resolve these complications, it has been suggested to interpret *omics* data at the level of functional modules, i.e., pathways, which are not only more easy to interpret but also more stable and optimize patient-specific therapeutic approaches. In addition, by mapping measured *omics* data to the pathway-level, their dimensional complexity is reduced, thus facilitating their utilization by ML models and enhancing interpretive power. Various studies such as [133, 154] used pathway signatures to predict drug response by applying varied ML models, however, the models thus far rarely address heterogeneity among patients.

Inspired by the above-mentioned observations, this work leverages highly predictive machine learning models and pathway signatures to simulate drug response and predict whether a candidate drug could be effective for a given patient. First, we calculated patient-specific pathway activity scores using single-sample gene set enrichment analysis [155]. We then used these scores to train a machine learning model that can accurately classify samples (i.e., disease vs normal). Next, we developed an intuitive scoring algorithm that calibrates the calculated pathway activity scores following the application of a drug. The modified pathway activity scores are then used as an input in the trained model to evaluate whether a sample that was previously labeled as "diseased" now could be predicted as "normal" following drug treatment. This is used as a proxy for the identification of drug candidates for patients. Ultimately, the method was evaluated against several comparable methods to analyze the model performance.

The methodology has been demonstrated on four different cancer datasets (i.e., breast, liver, prostate, kidney) and two independent drug-target datasets (i.e., DrugBank and DrugCentral). The method could successfully prioritize a drug intervention for patients based on their specific pathway scores. For example, given the fact that the Ras/Raf/MAPK pathway is one of the most important pathways that play a role in the development of liver cancer, the method prioritized a tyrosine kinase inhibitor from a class of JAK inhibitors (i.e., sorafenib), an already FDA-approved drug to treat patients with dysregulation of this pathway. Moreover, in addition to the prioritization of FDA-approved medications and drugs in clinical trials, the method was able to prioritize other drugs, suggesting that the drugs may represent promising candidates for repurposing. Additionally, in the evaluation of the approach against several comparable methods, our developed method outperformed similar studies, more specifically yielding a higher proportion of true positives.

Essentially, this study underscores the importance of considering heterogeneity among patients by simulating a drug's effect on an individual patient and offering a method that aims at proposing the most likely beneficial treatment for a given patient. This method could eventually be a valuable tool to support clinical decision-making in personalized medicine.

6 Conclusion and outlook

Heterogeneity is a fundamental property in the biomedical field which is not only observed among patients but also detected across collected data(sets) of different medical or research centers. On one hand, differences among individuals limits biological resolution, clinical conclusions, and efficacy of treatment. On the other hand disparities across collected data(sets) leads to low performance, interpretation, validity, and reproducibility of findings obtained from utilization of cutting-edge technologies (i.e., AI/ML models). Therefore, heeding and handling these heterogeneities is of utmost importance and this becomes especially relevant with respect to large heterogeneities which are known to exist in complex disorders such as AD and cancers.

During the last decades, a wide range of studies from handling the disparities across datasets [156, 157] to addressing heterogeneity among patients in the context of understaing disease pathophysiology [158, 159] and disease treatment [160, 161] have been conducted. However, only a few are free of limitations and have impact on clinical practices. Inspired by these observations, this dissertation demonstrates a set of studies and computational models that were exploited and developed to heed disparities across the data(sets) and address existing heterogeneity among patients.

This dissertation, first, has investigated a wide range of computational models which have been developed for different analytic purposes, enumerated their strengths and limitation, and discussed major heterogeneities-based constraints which limited the impact of computational models on clinical practices. This work has not only explained why current models are incapable of directly benefiting clinical practice but has also proposed avenues to address several of the constraints in order to bring the models closer to their translation into clinical practice. We believe that this work can serve as a checklist of bottlenecks for researchers which have to be taken into account while initiating a study or developing computational models. Another endeavor of this thesis focused on the generalizability and reproducibility crises that are derived from disparities across the clinical studies and their collected data. This work has shown that it is crucial to look beyond single data resources, validate achieved results across multiple clinical studies, and constantly develop and evaluate data-driven methods to ensure robust scientific insights. In addition, this work has described that despite the limitations that disparities across clinical studies bring, how can the additional available information unique to each clinical study be used to further provide a more comprehensive intuition over the investigated study. Besides the two previous endeavors, this thesis addressed another crucial subject in the biomedical field by describing the need for having organized domain knowledge in the field of AD to increase the understanding and explanation capability of biological systems. The enrichment of an AD knowledge assembly [134] with information from different modalities revealed that computable organized knowledge can undeniably demonstrate different biological phenomena in complex disorders such as AD and this explanatory competence can further be utilized to propose mechanism-based drug target candidates. Beyond the capability of the work in explaining biological processes spanning across several diverse biological scales, the work is part of NeuroMMSig [162] which represents the knowledge around well-established disease-specific mechanisms involved in AD and has a wide range of applications from drug discovery [163] to precision medicine [164]. Finally, given that a drug's response in one patient may differ from what it elicits in another as well as drug resistance driven by this heterogeneity are the biggest impediments in cancer treatment, our predictive machine learning model could effectively simulate the response triggered by a drug in a given sample and evaluate whether the sample that was previously labeled as diseased could be projected to be normal following drug treatment. This work enables us to prioritize drugs and set up medical decisions for individual patients based on their respective biological signatures which not only paves the way for precision medicine but also leads to the early detection of ineffective or dangerous off-target effects in medications which in turn reduce the cost of research and development spent on establishing new therapeutic medication. Moreover, this work allows us to not merely uncover sets of dysregulated pathways, but deconvolute a drugs' mechanism of action as well which subsequently allows for better determination of the correct therapeutic dosage by keeping track of the drug's impact on the patient's targeted pathway.

In order to address the described challenges, multiple new algorithmic approaches were developed in this dissertation. First, we have designed and applied a novel rank aggregation algorithm that combined exclusive, but complementary pathological marker changes derived from individual cohort datasets to provide a more complete insight into disease development and progression than an insight that is brought by single datasets. This robust and flexible algorithm enables users to include other changes of pathological markers coming from any new public or proprietary clinical study without requiring any substantial efforts for adaptation. This is crucial for guaranteeing model transferability and results across disease (sub)populations, and for providing a more comprehensive overview of disease progression. Second, we have developed an algorithm that simulates the effect of a given drug at the pathway level by modifying pathway activity scores of disease samples through an intuitive scoring algorithm. While the developed algorithm in this work has been applied in the cancer domain, in principle it can be employed to any disease domain with little (e.g., fine-tuning the weights for other datasets) or no further adaptations to yield promising results. Above all, all generated scripts, pipelines, and resources in the framework of this thesis are made openly accessible through online web tools and GitHub repositories in compliance with open and reproducible science, allowing more researchers to replicate the works and perform their analyses.

While this dissertation presented thriving implementation and promising results, it is not without limitations. First, in order to build a robust meta-model for changes of pathological marker in the course of disease progression and development, the designed rank aggregation algorithm needs to have the information which is presented in at least some of the individual order of biomarker changes to allow for meaningful distance calculations. Furthermore, the high amounts of missing data occurring when multiple data modalities were considered in each clinical study led to a substantial decrease in the number of available participants per study. This could have led to more noise in the order of biomarker changes distributions. Second, knowledge assemblies are prone to bias, and they may only represent data that has been deliberately selected. This is because researchers may tend to draw more heavily from authors or subfields with which they are more familiar or established well enough [165]. This could provoke a decisive problem in hypothesis-driven investigations of the data owing to knowledge unbalance among different subdomains while data interpretation. Third, the developed methodology for simulating drug response relies on pathway signatures derived from transcriptomics data. This indicates that this methodology is inherently limited to conditions where pathway activity score is highly predictive. In other words, pathway activity scores must effectively discriminate between disease and normal samples in the disease we analyze. This is because the methodology needs a highly predictive model to assure that the change in the predicted class label is solely caused by the drug simulation step and not by the model inaccuracy. Thus, it would be less practical in conditions where pathway activity score has restricted prediction power to separate between normal and disease samples, such as Parkinson's disease [166].

Finally, this thesis could serve as a good starting point for future efforts. As an illustration, given the increasing number of clinical studies in AD — each of which contains exclusive measurements, a larger number of such studies could be incorporated into the rank aggregation algorithms which in turn can potentially provide a more comprehensive insight into disease development and progression. Moreover, within a project context, the expanded AD knowledge assembly with knowledge relating to multiple biological scales and speculated mechanisms might be proposed as therapeutic target candidates and tested experimentally in-situ and in-vitro. Furthermore, the utilized predictive machine learning model for drug response simulation doesn't take time into consideration. As a future effort, drug administration could be simulated in an ML model that takes into consideration temporal dimensions (e.g., survival analysis [167]). As an additional future effort, the presented drug response simulation approach can be customized towards drug discovery by combining brute-force and reverse engineering approaches to identify the most effective pathway score that should be targeted by a drug for any given indication. In addition, while in the initial effort of drug response simulation the analysis was restricted to a single pathway database since it was enough to exploit a predictive ML model for the particular classification task, in future efforts pathway information from other databases or drug-target information from different databases such as ExCAPE-DB43 [168] could be incorporated into the ML model. Such incorporation not only can increase the total number and coverage of pathways to potentially reveal additional pathway targets but also broaden the chemical space which may lead to the identification of new candidates. Looking forward, translation of these efforts into clinical practice would pave the way to bring the precision medicine concept into reality.

Bibliography

1. OxfordUniversityPress. *Oxfordlanguages* `http://oxforddictionaries.com/`. **May 2013**.
2. Wallis, W. A., Roberts, H. V. The nature of statistics (Courier Corporation, **2014**).
3. Grutters, J. P. *et al.* Acknowledging patient heterogeneity in economic evaluation. *Pharmacoeconomics* **2013**, 31 (2), 111–123 **2013**.
4. Ramaekers, B. L., Joore, M. A., Grutters, J. P. How should we deal with patient heterogeneity in economic evaluation: a systematic review of national pharmacoeconomic guidelines. *Value in health* **2013**, 16 (5), 855–862 **2013**.
5. Cahan, E. M., Khatri, P. Data Heterogeneity: The Enzyme to Catalyze Translational Bioinformatics? *Journal of Medical Internet Research* **2020**, 22 (8), e18044 **2020**.
6. Schlee, W. *et al.* Visualization of global disease burden for the optimization of patient management and treatment. *Frontiers in medicine* **2017**, 4, 86 **2017**.
7. 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* **2020**, 11 (3), 367–376 **2020**.
8. Ritchie, H., Roser, M. Causes of Death. *Our World in Data* **2018**. https://ourworldindata.org/causes-of-death **2018**.
9. Dugger, B. N., Dickson, D. W. Pathology of neurodegenerative diseases. *Cold Spring Harbor perspectives in biology* **2017**, 9 (7), a028035 **2017**.
10. Sun, Y., Peng, Z. Programmed cell death and cancer. *Postgraduate medical journal* **2009**, 85 (1001), 134–140 **2009**.
11. Wimo, A. *et al.* The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & Dementia* **2017**, 13 (1), 1–7 **2017**.

12. Serrano-Pozo, A., Frosch, M. P., Masliah, E., Hyman, B. T. Neuropathological alterations in Alzheimer disease. *Cold Spring Harbor perspectives in medicine* **2011**, 1 (1), a006189 **2011**.
13. Holtzman, D. M., Morris, J. C., Goate, A. M. Alzheimer's disease: the challenge of the second century. *Science translational medicine* **2011**, 3 (77), 77sr1–77sr1 **2011**.
14. Gold, C. A., Budson, A. E. Memory loss in Alzheimer's disease: implications for development of therapeutics. *Expert review of neurotherapeutics* **2008**, 8 (12), 1879–1891 **2008**.
15. McLaughlin, T. *et al.* Dependence as a unifying construct in defining Alzheimer's disease severity. *Alzheimer's & Dementia* **2010**, 6 (6), 482–493 **2010**.
16. Huang, J. *et al.* A global trend analysis of kidney cancer incidence and mortality and their associations with smoking, alcohol consumption, and metabolic syndrome. *European Urology Focus* **2022**, 8 (1), 200–209 **2022**.
17. Chow, W.-H., Dong, L. M., Devesa, S. S. Epidemiology and risk factors for kidney cancer. *Nature Reviews Urology* **2010**, 7 (5), 245–257 **2010**.
18. Rini, B. I., Campbell, S. C., Escudier, B. Renal cell carcinoma. *The Lancet* **2009**, 373 (9669), 1119–1132 **2009**.
19. Haas, N. B., Nathanson, K. L. Hereditary kidney cancer syndromes. *Advances in chronic kidney disease* **2014**, 21 (1), 81–90 **2014**.
20. Linehan, W. M., Ricketts, C. J. The metabolic basis of kidney cancer. In *Seminars in cancer biology* **23** (**2013**), 46–55.
21. Calle, E. E., Kaaks, R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nature Reviews Cancer* **2004**, 4 (8), 579–591 **2004**.
22. On the Evaluation of Carcinogenic Risks to Humans, I. W. G., Organization, W. H., for Research on Cancer, I. A. Tobacco smoke and involuntary smoking (Iarc, **2004**).
23. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **2018**, 68 (6), 394–424 **2018**.
24. Recio-Boiles, A., Waheed, A., Babiker, H. M. Cancer, liver. **2017** **2017**.
25. Watson, J., Hydon, K., Lodge, P. Primary and secondary liver tumours. *InnovAiT* **2016**, 9 (8), 477–482 **2016**.

26. Llovet, J. M. *et al.* Hepatocellular carcinoma (Primer). *Nature Reviews: Disease Primers* **2021**, 7 (1) **2021**.

27. Yang, J. D. *et al.* A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nature reviews Gastroenterology & hepatology* **2019**, 16 (10), 589–604 **2019**.

28. Ringehan, M., McKeating, J. A., Protzer, U. Viral hepatitis and liver cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2017**, 372 (1732), 20160274 **2017**.

29. Alkabban, F. M., Ferguson, T. Cancer, breast. **2018 2018**.

30. Jung, S. *et al.* Alcohol consumption and breast cancer risk by estrogen receptor status: in a pooled analysis of 20 studies. *International journal of epidemiology* **2016**, 45 (3), 916–928 **2016**.

31. Jemal, A., Center, M. M., DeSantis, C., Ward, E. M. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology and Prevention Biomarkers* **2010**, 19 (8), 1893–1907 **2010**.

32. Leslie, S. W., Soon-Sutton, T. L., Sajjad, H., Siref, L. E. Prostate Cancer.

33. Alizadeh, M., Alizadeh, S. Survey of clinical and pathological characteristics and outcomes of patients with prostate cancer. *Global Journal of Health Science* **2014**, 6 (7), 49 **2014**.

34. Bostwick, D. G. *et al.* Human prostate cancer risk factors. *Cancer: Interdisciplinary International Journal of the American Cancer Society* **2004**, 101 (S10), 2371–2490 **2004**.

35. Pienta, K. J., Esper, P. S. Risk factors for prostate cancer. *Annals of internal medicine* **1993**, 118 (10), 793–803 **1993**.

36. Rebbeck, T. R. Prostate cancer genetics: variation by race, ethnicity, and geography. In *Seminars in radiation oncology* **27** (**2017**), 3–10.

37. Tan, S.-H., Petrovics, G., Srivastava, S. Prostate cancer genomics: Recent advances and the prevailing underrepresentation from racial and ethnic minorities. *International journal of molecular sciences* **2018**, 19 (4), 1255 **2018**.

38. Epstein, M. M. *et al.* Dietary fatty acid intake and prostate cancer survival in Örebro County, Sweden. *American journal of epidemiology* **2012**, 176 (3), 240–252 **2012**.

39. Soni, M. G. *et al.* Safety of vitamins and minerals: controversies and perspective. *Toxicological sciences* **2010**, 118 (2), 348–355 **2010**.

40. Strimbu, K., Tavel, J. A. What are biomarkers? *Current Opinion in HIV and AIDS* **2010**, 5 (6), 463 **2010**.

41. Teunissen, C. E. *et al.* Blood-based biomarkers for Alzheimer's disease: towards clinical implementation. *The Lancet Neurology* **2022**, 21 (1), 66–77 **2022**.

42. Blennow, K., Hampel, H. CSF markers for incipient Alzheimer's disease. *The Lancet Neurology* **2003**, 2 (10), 605–613 **2003**.

43. Salvatore, C. *et al.* Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. *Frontiers in neuroscience* **2015**, 9, 307 **2015**.

44. Nordberg, A., Rinne, J. O., Kadir, A., Långström, B. The use of PET in Alzheimer disease. *Nature Reviews Neurology* **2010**, 6 (2), 78–87 **2010**.

45. Weintraub, S., Wicklund, A. H., Salmon, D. P. The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor perspectives in medicine* **2012**, 2 (4), a006171 **2012**.

46. Bateman, R. J. *et al.* Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. *Alzheimer's research & therapy* **2011**, 3 (1), 1–13 **2011**.

47. Van Cauwenberghe, C., Van Broeckhoven, C., Sleegers, K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in Medicine* **2016**, 18 (5), 421–430 **2016**.

48. Duffy, M. J., Walsh, S., McDermott, E. W., Crown, J. Biomarkers in breast cancer: where are we and where are we going? *Advances in clinical chemistry* **2015**, 71, 1–23 **2015**.

49. Pinsky, P. F., Prorok, P. C., Kramer, B. S. Prostate Cancer Screening-A Perspective on the Current State of the Evidence. *The New England journal of medicine* **2017**, 376 (13), 1285–1289 **2017**.

50. Bratu, O. *et al.* Renal tumor biomarkers. *Experimental and therapeutic medicine* **2021**, 22 (5), 1–7 **2021**.

51. Malaguarnera, G. *et al.* Serum markers of hepatocellular carcinoma. *Digestive diseases and sciences* **2010**, 55 (10), 2744–2755 **2010**.

52. Zhang, F., Deng, Y., Wang, M., Cui, L., Drabier, R. Pathway-based biomarkers for breast cancer in proteomics. *Cancer Informatics* **2014**, 13, CIN–S14069 **2014**.

53. Liu, X., Su, L., Li, J., Ou, G. Identification of Pathway-Based Biomarkers with Crosstalk Analysis for Overall Survival Risk Prediction in Breast Cancer. *Frontiers in genetics* **2021**, 1997 **2021**.

54. Liu, G. *et al.* Prognostic gene biomarker identification in liver cancer by data mining. *American Journal of Translational Research* **2021**, 13 (5), 4603 **2021**.

55. Song, Z. *et al.* The identification of potential biomarkers and biological pathways in prostate cancer. *Journal of Cancer* **2019**, 10 (6), 1398 **2019**.

56. Tanzi, R. E. The genetics of Alzheimer disease. *Cold Spring Harbor perspectives in medicine* **2012**, 2 (10), a006296 **2012**.

57. Murray, M. E. *et al.* Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *The Lancet Neurology* **2011**, 10 (9), 785–796 **2011**.

58. Whitwell, J. L. *et al.* Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. *The Lancet Neurology* **2012**, 11 (10), 868–877 **2012**.

59. Ferreira, D. *et al.* Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Scientific reports* **2017**, 7 (1), 1–13 **2017**.

60. Ferreira, D., Pereira, J. B., Volpe, G., Westman, E. Subtypes of Alzheimer's disease display distinct network abnormalities extending beyond their pattern of brain atrophy. *Frontiers in neurology* **2019**, 10, 524 **2019**.

61. Du, X., Wang, X., Geng, M. Alzheimer's disease hypothesis and related therapies. *Translational neurodegeneration* **2018**, 7 (1), 1–7 **2018**.

62. Mohandas, E., Rajmohan, V., Raghunath, B. Neurobiology of Alzheimer's disease. *Indian journal of psychiatry* **2009**, 51 (1), 55 **2009**.

63. Neff, R. A. *et al.* Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Science advances* **2021**, 7 (2), eabb5398 **2021**.

64. Zheng, C., Xu, R. Molecular subtyping of Alzheimer's disease with consensus non-negative matrix factorization. *Plos one* **2021**, 16 (5), e0250278 **2021**.

65. Dubois, B. *et al.* Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *The Lancet Neurology* **2014**, 13 (6), 614–629 **2014**.

66. Thalhauser, C. J., Komarova, N. L. Alzheimer's disease: rapid and slow progression. *Journal of the Royal Society Interface* **2012**, 9 (66), 119–126 **2012**.

67. Ferreira, D., Wahlund, L.-O., Westman, E. The heterogeneity within Alzheimer's disease. *Aging (Albany NY)* **2018**, 10 (11), 3058 **2018**.

68. Ferreira, D., Nordberg, A., Westman, E. Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology* **2020**, 94 (10), 436–448 **2020**.

69. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Archives of general psychiatry* **2006**, 63 (2), 168–174 **2006**.

70. Stern, Y. Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology* **2012**, 11 (11), 1006–1012 **2012**.

71. Ferreira, D. *et al.* The contribution of small vessel disease to subtypes of Alzheimer's disease: a study on cerebrospinal fluid and imaging biomarkers. *Neurobiology of aging* **2018**, 70, 18–29 **2018**.

72. Weiner, M. W. *et al.* Effects of traumatic brain injury and posttraumatic stress disorder on development of Alzheimer's disease in Vietnam veterans using the Alzheimer's Disease Neuroimaging Initiative: preliminary report. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **2017**, 3 (2), 177–188 **2017**.

73. Tremblay-Mercier, J. *et al.* Open science datasets from PREVENT-AD, a longitudinal cohort of pre-symptomatic Alzheimer's disease. *NeuroImage: Clinical* **2021**, 31, 102733 **2021**.

74. Kodamullil, A. T. *et al.* Trial watch: Tracing investment in drug development for Alzheimer disease. *Nature reviews. Drug discovery* **2017**, 16 (12), 819 **2017**.

75. Dagogo-Jack, I., Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology* **2018**, 15 (2), 81–94 **2018**.

76. Fox, E. J., Loeb, L. A. One cell at a time. *Nature* **2014**, 512 (7513), 143–144 **2014**.

77. Jamal-Hanjani, M., Quezada, S. A., Larkin, J., Swanton, C. Translational implications of tumor heterogeneity. *Clinical cancer research* **2015**, 21 (6), 1258–1266 **2015**.

78. Grzywa, T. M., Paskal, W., Włodarski, P. K. Intratumor and intertumor heterogeneity in melanoma. *Translational oncology* **2017**, 10 (6), 956–975 **2017**.

79. El-Sayes, N., Vito, A., Mossman, K. Tumor heterogeneity: A great barrier in the age of cancer immunotherapy. *Cancers* **2021**, 13 (4), 806 **2021**.

80. Marusyk, A., Polyak, K. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **2010**, 1805 (1), 105–117 **2010**.

81. Anand, P. *et al.* Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical research* **2008**, 25 (9), 2097–2116 **2008**.

82. Kim, Y.-s. *et al.* Clinical implications of APOBEC3A and 3B expression in patients with breast cancer. *PloS one* **2020**, 15 (3), e0230261 **2020**.

83. Li, M., Zhang, Z., Li, L., Wang, X. An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. *Communications biology* **2020**, 3 (1), 1–19 **2020**.

84. Bristow, R. G., Hill, R. P. Hypoxia and metabolism: Hypoxia, DNA repair and genetic instability. *Nature Reviews Cancer* **2008**, 8 (3) **2008**.

85. Jiménez, F. Hypoxia causes down-regulation of Mismatch Repair System and genomic instability in Stem Cells. **2008 2008**.

86. Findlay, J. M. *et al.* Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nature communications* **2016**, 7 (1), 1–13 **2016**.

87. Loeb, L. A. Mutator phenotype may be required for multistage carcinogenesis. *Cancer research* **1991**, 51 (12), 3075–3079 **1991**.

88. Abeshouse, A. *et al.* The molecular taxonomy of primary prostate cancer. *Cell* **2015**, 163 (4), 1011–1025 **2015**.

89. Lüönd, F., Tiede, S., Christofori, G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *British Journal of Cancer* **2021**, 125 (2), 164–175 **2021**.

90. Beksac, A. T. *et al.* Heterogeneity in renal cell carcinoma. In *Urologic Oncology: Seminars and Original Investigations* **35** (**2017**), 507–515.

91. Embracing patient heterogeneity. *Nature Medicine* **2014**, 20 (7), 689–689 **2014**.

92. Barnes, J. *et al.* Alzheimer's disease first symptoms are age dependent: evidence from the NACC dataset. *Alzheimer's & dementia* **2015**, 11 (11), 1349–1357 **2015**.

93. Koedam, E. L. *et al.* Early-versus late-onset Alzheimer's disease: more than age alone. *Journal of Alzheimer's Disease* **2010**, 19 (4), 1401–1408 **2010**.

94. Onitilo, A. A., Engel, J. M., Greenlee, R. T., Mukesh, B. N. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. *Clinical medicine & research* **2009**, 7 (1-2), 4–13 **2009**.

95. Wang, X., Zhang, H., Chen, X. Drug resistance and combating drug resistance in cancer. *Cancer drug resistance (Alhambra, Calif.)* **2019**, 2, 141 **2019**.

96. Cao, J., Hou, J., Ping, J., Cai, D. Advances in developing novel therapeutic strategies for Alzheimer's disease. *Molecular neurodegeneration* **2018**, 13 (1), 1–20 **2018**.

97. Niven, D. J. *et al.* Reproducibility of clinical research in critical care: a scoping review. *BMC medicine* **2018**, 16 (1), 1–12 **2018**.

98. Fröhlich, H. *et al.* From hype to reality: data science enabling personalized medicine. *BMC medicine* **2018**, 16 (1), 1–15 **2018**.

99. Heiner, M., Gilbert, D. Biomodel engineering for multiscale systems biology. *Progress in biophysics and molecular biology* **2013**, 111 (2-3), 119–128 **2013**.

100. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **2003**, 19 (4), 524–531 **2003**.

101. Resat, H., Petzold, L., Pettigrew, M. F. Kinetic modeling of biological systems. *Computational systems biology* **2009**, 311–335 **2009**.

102. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nature biotechnology* **2010**, 28 (9), 935–942 **2010**.

103. Lieu, C., Elliston, K. Applying a causal framework to system modeling. *Systems Biology* **2007**, 139–152 **2007**.

104. Martin, F. *et al.* Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. *BMC systems biology* **2012**, 6 (1), 1–18 **2012**.

105. Catlett, N. L. *et al.* Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC bioinformatics* **2013**, 14 (1), 1–14 **2013**.

106. Ying, S.-W. *et al.* Brain-derived neurotrophic factor induces long-term potentiation in intact adult hippocampus: requirement for ERK activation coupled to CREB and upregulation of Arc synthesis. *Journal of Neuroscience* **2002**, 22 (5), 1532–1540 **2002**.

107. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., Fanos, V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina* **2020**, 56 (9), 455 **2020**.

108. Challis, E. *et al.* Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage* **2015**, 112, 232–243 **2015**.

109. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *nature* **2017**, 542 (7639), 115–118 **2017**.

110. Emon, M. A., Domingo-Fernández, D., Hoyt, C. T., Hofmann-Apitius, M. PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures. *BMC bioinformatics* **2020**, 21 (1), 1–21 **2020**.

111. Frölich, L. *et al.* Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. *Alzheimer's research & therapy* **2017**, 9 (1), 1–15 **2017**.

112. Dickerson, B. C., Wolk, D. Biomarker-based prediction of progression in MCI: comparison of AD-signature and hippocampal volume with spinal fluid amyloid-β and tau. *Frontiers in aging neuroscience* **2013**, 5, 55 **2013**.

113. Huang, L. *et al.* Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of aging* **2016**, 46, 180–191 **2016**.

114. Franzmeier, N. *et al.* Predicting sporadic Alzheimer's disease progression via inherited Alzheimer's disease-informed machine-learning. *Alzheimer's & Dementia* **2020**, 16 (3), 501–511 **2020**.

115. 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* **2012**, 60 (3), 1880–1889 **2012**.

116. Young, A. L. *et al.* A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* **2014**, 137 (9), 2564–2577 **2014**.

117. Fisher, C. K., Smith, A. M., Walsh, J. R. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Scientific reports* **2019**, 9 (1), 1–14 **2019**.

118. Lee, G., Nho, K., Kang, B., Sohn, K.-A., Kim, D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports* **2019**, 9 (1), 1–12 **2019**.

119. Spasov, S. *et al.* A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage* **2019**, 189, 276–287 **2019**.

120. Westman, E., Muehlboeck, J.-S., Simmons, A. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. *Neuroimage* **2012**, 62 (1), 229–238 **2012**.

121. Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., Garrard, P. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease* **2014**, 42 (s3), S3–S17 **2014**.

122. Liu, L., Zhao, S., Chen, H., Wang, A. A new machine learning method for identifying Alzheimer's disease. *Simulation Modelling Practice and Theory* **2020**, 99, 102023 **2020**.

123. Young, A. L. *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature communications* **2018**, 9 (1), 1–16 **2018**.

124. Young, A. L. *et al.* Multiple orderings of events in disease progression. In *International Conference on Information Processing in Medical Imaging* (**2015**), 711–722.

125. Issa, N. T., Stathias, V., Schürer, S., Dakshanamurthy, S. Machine and deep learning approaches for cancer drug repurposing. In *Seminars in cancer biology* **68** (**2021**), 132–142.

126. Paranjpe, M. D., Taubes, A., Sirota, M. Insights into computational drug repurposing for neurodegenerative disease. *Trends in pharmacological sciences* **2019**, 40 (8), 565–576 **2019**.

127. Chiu, Y.-C. *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC medical genomics* **2019**, 12 (1), 143–155 **2019**.

128. Xu, H. *et al.* Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association* **2015**, 22 (1), 179–191 **2015**.

129. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling* **2017**, 57 (4), 942–957 **2017**.

130. Al-Ali, H. *et al.* Rational polypharmacology: systematically identifying and engaging multiple drug targets to promote axon growth. *ACS chemical biology* **2015**, 10 (8), 1939–1951 **2015**.

131. Zhang, F., Wang, M., Xi, J., Yang, J., Li, A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Scientific reports* **2018**, 8 (1), 1–9 **2018**.

132. Chang, Y. *et al.* Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports* **2018**, 8 (1), 1–11 **2018**.

133. Ammad-Ud-Din, M. *et al.* Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* **2016**, 32 (17), i455–i463 **2016**.

134. Kodamullil, A. T., Younesi, E., Naz, M., Bagewadi, S., Hofmann-Apitius, M. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimer's & Dementia* **2015**, 11 (11), 1329–1339 **2015**.

135. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* **2016**, gkw943 **2016**.

136. Qiu, S. *et al.* Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain* **2020**, 143 (6), 1920–1933 **2020**.

137. Bron, E. E. *et al.* Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage: Clinical* **2021**, 31, 102712 **2021**.

138. Licher, S. *et al.* External validation of four dementia prediction models for use in the general community-dwelling population: a comparative analysis from the Rotterdam Study. *European journal of epidemiology* **2018**, 33 (7), 645–655 **2018**.

139. Mukherjee, S. *et al.* Identifying and ranking potential driver genes of Alzheimer's disease using multiview evidence aggregation. *Bioinformatics* **2019**, 35 (14), i568–i576 **2019**.

140. Thompson, P. M. *et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior* **2014**, 8 (2), 153–182 **2014**.

141. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **2018**, 562 (7726), 210–216 **2018**.

142. Wachinger, C. *et al.* A longitudinal imaging genetics study of neuroanatomical asymmetry in Alzheimer's disease. *Biological psychiatry* **2018**, 84 (7), 522–530 **2018**.

143. Stefanovski, L. *et al.* Linking molecular pathways and large-scale computational modeling to assess candidate disease mechanisms and pharmacodynamics in Alzheimer's disease. *Frontiers in computational neuroscience* **2019**, 54 **2019**.

144. Beam, E., Potts, C., Poldrack, R. A., Etkin, A. A data-driven framework for mapping domains of human neurobiology. *Nature neuroscience* **2021**, 24 (12), 1733–1744 **2021**.

145. McGranahan, N., Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell* **2015**, 27 (1), 15–26 **2015**.

146. Holohan, C., Van Schaeybroeck, S., Longley, D. B., Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* **2013**, 13 (10), 714–726 **2013**.

147. Ashley, E. A. The precision medicine initiative: a new national effort. *Jama* **2015**, 313 (21), 2119–2120 **2015**.

148. Li, Y. *et al.* Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. *BMC genomics* **2021**, 22 (1), 1–18 **2021**.

149. Gruener, R. F. *et al.* Facilitating drug discovery in breast cancer by virtually screening patients using in vitro drug response modeling. *Cancers* **2021**, 13 (4), 885 **2021**.

150. Yuan, R., Chen, S., Wang, Y. Computational prediction of drug responses in cancer cell lines from cancer omics and detection of drug effectiveness related methylation sites. *Frontiers in genetics* **2020**, 917 **2020**.

151. Graim, K., Friedl, V., Houlahan, K. E., Stuart, J. M. PLATYPUS: A Multiple—View Learning Predictive Framework for Cancer Drug Sensitivity Prediction. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium* (**2018**), 136–147.

152. Braga-Neto, U. M., Dougherty, E. R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **2004**, 20 (3), 374–380 **2004**.

153. Ntzani, E. E., Ioannidis, J. P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *The Lancet* **2003**, 362 (9394), 1439–1444 **2003**.

154. Smith, A. M. *et al.* Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC bioinformatics* **2020**, 21 (1), 1–18 **2020**.

155. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **2005**, 102 (43), 15545–15550 **2005**.

156. Basu, A. *et al.* Call for data standardization: lessons learned and recommendations in an imaging study. *JCO clinical cancer informatics* **2019**, 3, 1–11 **2019**.

157. Spjuth, O. *et al.* Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *European Journal of Human Genetics* **2016**, 24 (4), 521–528 **2016**.

158. Sirkis, D. W., Bonham, L. W., Johnson, T. P., La Joie, R., Yokoyama, J. S. Dissecting the clinical heterogeneity of early-onset Alzheimer's disease. *Molecular Psychiatry* **2022**, 1–15 **2022**.

159. Habes, M. *et al.* Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods. *Biological psychiatry* **2020**, 88 (1), 70–82 **2020**.

160. Rybinski, B., Yun, K. Addressing intra-tumoral heterogeneity and therapy resistance. *Oncotarget* **2016**, 7 (44), 72322 **2016**.

161. Fustero-Torre, C. *et al.* Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome medicine* **2021**, 13 (1), 1–15 **2021**.

162. Domingo-Fernández, D. *et al.* Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics* **2017**, 33 (22), 3679–3681 **2017**.

163. Hoyt, C. T., Domingo-Fernández, D., Balzer, N., Güldenpfennig, A., Hofmann-Apitius, M. A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities. *Database* **2018**, 2018 **2018**.

164. 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* **2018**, 8 (1), 1–13 **2018**.

165. Andersen, F., Anjum, R. L., Rocca, E. Philosophy of Biology: Philosophical bias is the one bias that science cannot avoid. *Elife* **2019**, 8, e44929 **2019**.

166. Chen-Plotkin, A. S. Blood transcriptomics for Parkinson disease? *Nature Reviews Neurology* **2018**, 14 (1), 5–6 **2018**.

167. Tibshirani, R. The lasso method for variable selection in the Cox model. *Statistics in medicine* **1997**, 16 (4), 385–395 **1997**.

168. Sun, J. *et al.* ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of cheminformatics* **2017**, 9 (1), 1–9 **2017**.

7 Appendixes

Chapter 7

7.1 Challenges of Integrative Disease Modeling in Alzheimer′s Disease

Reprinted with permission from "Golriz Khatami, S. *et al.*. "Challenges of Integrative Disease Modeling in Alzheimer′s Disease". *Front Mol Biosci*, 6:158, (2020).

frontiers
in Molecular Biosciences

REVIEW
published: 14 January 2020
doi: 10.3389/fmolb.2019.00158

Challenges of Integrative Disease Modeling in Alzheimer's Disease

Sepehr Golriz Khatami[1,2*], *Christine Robinson*[1,2], *Colin Birkenbihl*[1,2], *Daniel Domingo-Fernández*[1,2], *Charles Tapley Hoyt*[1,2] *and Martin Hofmann-Apitius*[1,2]

[1] *Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany,* [2] *Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany*

OPEN ACCESS

Edited by:
Daniel Garrido,
Pontifical Catholic University of Chile, Chile

Reviewed by:
Lin Tan,
Qingdao Municipal Hospital, China
Owen T. Carmichael,
Pennington Biomedical Research Center, United States

****Correspondence:***
Sepehr Golriz Khatami
sepehr.golriz.khatami@scai-extern.fraunhofer.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *02 September 2019*
Accepted: *18 December 2019*
Published: *14 January 2020*

Citation:
Golriz Khatami S, Robinson C, Birkenbihl C, Domingo-Fernández D, Hoyt CT and Hofmann-Apitius M (2020) Challenges of Integrative Disease Modeling in Alzheimer's Disease. Front. Mol. Biosci. 6:158. doi: 10.3389/fmolb.2019.00158

Dementia-related diseases like Alzheimer's Disease (AD) have a tremendous social and economic cost. A deeper understanding of its underlying pathophysiologies may provide an opportunity for earlier detection and therapeutic intervention. Previous approaches for characterizing AD were targeted at single aspects of the disease. Yet, due to the complex nature of AD, the success of these approaches was limited. However, in recent years, advancements in integrative disease modeling, built on a wide range of AD biomarkers, have taken a global view on the disease, facilitating more comprehensive analysis and interpretation. Integrative AD models can be sorted in two primary types, namely hypothetical models and data-driven models. The latter group split into two subgroups: (i) Models that use traditional statistical methods such as linear models, (ii) Models that take advantage of more advanced artificial intelligence approaches such as machine learning. While many integrative AD models have been published over the last decade, their impact on clinical practice is limited. There exist major challenges in the course of integrative AD modeling, namely data missingness and censoring, imprecise human-involved priori knowledge, model reproducibility, dataset interoperability, dataset integration, and model interpretability. In this review, we highlight recent advancements and future possibilities of integrative modeling in the field of AD research, showcase and discuss the limitations and challenges involved, and finally, propose avenues to address several of these challenges.

Keywords: Alzheimer's disease, challenges, integrative disease modeling, hypothetical, data-driven

INTRODUCTION

Alzheimer's Disease (AD) manifests in a collection of symptoms including the deterioration of cognition, memory, and behavior which often leads to interference with activities of daily living. In 2017, AD ranked among the top five causes of death worldwide, with 2.44 million (4.5%) deaths from AD[1,2]. Worldwide, there are currently around 50 million people living with AD, and every 3 s a person develops this condition. It is estimated that only a quarter of those living with AD are diagnosed, and more than 17 million healthcare workers annually invest 18 billion hours of care, at a cost of more than one trillion US dollars to tackle AD-associated problems[3,4]. Extrapolating these statistics to the coming decades suggests the immense socioeconomic impact of AD on all involved

[1] https://ourworldindata.org/causes-of-death
[2] https://www.thestreet.com/world/leading-causes-of-death-world-14869811
[3] https://www.alz.co.uk/research/statistics
[4] https://ourworldindata.org/causes-of-death

parties: patients, caregivers, healthcare systems, and indirectly, the economy. Thus, strategies to reduce the global emotional and financial burden of AD are of great importance. To develop such strategies, a deeper understanding of the pathophysiology underlying AD is necessary and may lead to opportunities for earlier detection and therapeutic interventions.

In general, AD progression is categorized into three clinical disease stages: (i) During the pre-symptomatic phase, individuals may have already developed pathological changes that underlie AD, but remain cognitively normal, (ii) in the prodromal phase, often referred to as mild cognitive impairment (MCI), the first cognitive symptoms, commonly episodic memory deficits, appear. These symptoms can be acute, but they do not yet meet the criteria for dementia, (iii) in the dementia stage, impairments are severe enough to interfere with daily life (Jack et al., 2010).

Understanding of the etiology of AD is complicated due to the existence of dysregulations at different biological scales, ranging from genetic mutations to structural and functional alterations of the brain (Aisen et al., 2017). For this reason, significant efforts have been made in recent years to discover candidate markers for disease-related pathological changes throughout all modalities, including neuro-imaging, cerebrospinal fluid (CSF) samples and a broad variety of -omics data. Studies have successfully identified multiple biomarkers for neurodegeneration and AD (Blennow and Zetterberg, 2018). However, effectively translating extensive biomarker screenings into clinical application remains a challenging task, because individual biomarkers can only provide a highly incomplete view on such a multifactorial disease (Younesi and Hofmann-Apitius, 2013). For instance, while multiple associations between genetic variants and AD have been established (Jansen et al., 2019; Kunkle et al., 2019), none of these associations fully describe disease pathogenesis. As a result, one of the major challenges in AD research is translating diverse biomarker signals available into multimodal, multiscale models of disease pathogenesis.

In recent years, a new translational research paradigm called "integrative disease modeling" has emerged, to address this challenge (Younesi and Hofmann-Apitius, 2013). It aims at modeling heterogeneous measurements across different biological scales, in order to provide a holistic picture of biomarker intercorrelations in the disease of study. To this end, advanced high-throughput technologies and neuroimaging procedures are being used to collect data from multiple modalities. These diverse data need to be integrated, that is, combined in a way that preserves the structure and meaning in the data, using computational algorithms. Only then can they provide a solid basis for further analysis such as reasoning, simulation, and visualization. In order to contribute to understanding of the complex pathophysiology of the disease, the results should be actionable and thus must be interpretable. Integrative disease modeling, by collecting, integrating, analyzing, and ultimately interpreting the measurements, facilitates the understanding of the pathophysiology of complex diseases like AD (Hampel et al., 2017).

Existing integrative models in the context of AD can be placed in two primary categories, namely hypothetical models and data-driven models (**Table 1**). Hypothetical models are non-numerical and rely on reasoning over findings of previously published studies (Jack et al., 2010), rather than large amounts of data. By including this prior knowledge, these models try to detail the temporal changes of AD biomarkers relative to each other as well as to clinical disease stages and trial endpoints.

TABLE 1 | Organization of and references for data-driven integrative AD models.

Data-driven integrative AD models			References
Traditional			Caroli and Frisoni, 2010; Jack et al., 2011, 2012
Machine learning	Generative		Fonteijn et al., 2012; Chen et al., 2016; Khanna et al., 2018; Oxtoby et al., 2018; Basu et al., 2019; De Jong et al., 2019; Gootjes-Dreesbach et al., 2019; Martinez-Murcia et al., 2019
	Discriminative	Supervised	Hinrichs et al., 2010; Magnin et al., 2010; Rao et al., 2011; Zhang et al., 2011; Da et al., 2013; Li et al., 2013
		Unsupervised	Nettiksimmons et al., 2014; Gamberger et al., 2017; Toschi et al., 2019

We subdivide data-driven integrative AD models which into two subgroups. While the first group uses simple statistical approaches (e.g., simple linear models), the second group uses more advanced techniques (e.g., machine learning). The advanced machine learning models include generative and discriminative models, the latter of which can be classified as either supervised or unsupervised models.

By contrast, data-driven integrative models take advantage of developments in computational approaches and big data. For the sake of this review, we will distinguish between two subcategories of data-driven models. The first covers traditional statistical methods of generally lower complexity, such as linear models. Often, these models are used to estimate biomarker trajectories by regressing measured data against a prespecified dependent variable, such as a clinical readout or the disease stage (Bateman et al., 2012). The second subtype exploits more advanced artificial intelligence approaches such as machine learning. Within this subtype, models can be characterized as discriminative or generative. Discriminative models are designed to discriminate between groups (e.g., cases and controls) and can be further described as supervised or unsupervised, depending on whether they rely on labeled (Hinrichs et al., 2011; Da et al., 2013) or unlabeled (Toschi et al., 2019) data. Generative models contribute to disease understanding by automatically learning the inherent distribution of a dataset and its feature interdependencies (Oxtoby et al., 2018). An exemplary application is the extraction of disease progression signatures as demonstrated by the ensemble of Bayesian networks developed by Khanna et al. (2018).

Integrative AD modeling faces many challenges. Hypothetical models, by their nature, are time-intensive to construct and require specialist knowledge. Their primary role in AD research is to provide ideas for future experiments. Likewise in data-driven modeling, several challenges at each step of the process (i.e., collection, integration, analysis, and interpretation) must be addressed. Data missingness and data censoring are significant bottlenecks in data collection as well as analysis and interpretation. Meanwhile, the heterogeneity and complexity of biological data are major impediments to data integration, which forms the basis for all data-driven approaches. Furthermore, data mapping, data labels, and biased data are additional barriers to robust data analysis and interpretation. Finally, insufficient numbers of subjects restrict the statistical power of data-driven integrative AD models. These fundamental challenges explain why, at this point in time, although many integrative AD models have been published over the last decade, their impact on clinical practice is limited.

In this review, we highlight recent advancements and future possibilities of integrative modeling, discuss the limitations and challenges involved, and finally, propose avenues to address several of these challenges, in the context of AD research.

INTEGRATIVE AD MODELS

As already mentioned, integrative AD models can be characterized as either hypothetical or data-driven, each of which has strengths and weaknesses. In the following, we compare different models of each type and discuss their benefits and limitations. Finally, we elaborate on how associated limitations and challenges could be handled.

Hypothetical Models

In hypothetical modeling, a model is generated about an object of study, direct knowledge of which is difficult to obtain. These models provide hypotheses about the object (Gladun, 1997). In integrative AD modeling, researchers develop so-called cascade models, in which the measurements of a set of biomarkers are normalized and their trajectories are plotted on a common time scale, aligned to disease stages (Jack et al., 2010, 2013). These models are typically developed by reviewing the available knowledge and reasoning over observations from previously published studies. They are not directly informed by measured data.

One of the first hypothetical integrative AD models was developed by Jack et al. (2013) [revised from a previous model (Jack et al., 2010)]. This model hypothesized the temporal changes of the five most studied biomarkers of AD pathology in relation to estimated years from expected symptom onset and in relation to other biomarkers. These biomarkers are CSF amyloid-beta protein (CSF $A\beta_{1-42}$) and tau protein (CSF tau) levels, amyloid-beta PET imaging (PET $A\beta$), Fluorodeoxyglucose-PET imaging, and structural MRI readouts. In this cascade model, the authors presumed that biomarker trajectories should exhibit a sigmoid-shaped curve. This imposition is a direct result of the limited sensitivity of measurements at time extremes, which the authors addressed by taking the floor of the measurements at early timepoints, and the ceiling of the measurements at late timepoints. The authors hypothesized that the two amyloid-beta ($A\beta$) biomarkers (i.e., CSF $A\beta_{1-42}$ and PET $A\beta$ imaging) gradually approach an abnormal state while the subject remains in a cognitively normal state. After a lag period, the length of which varies from patient to patient, and in later disease stages, CSF tau, Fluorodeoxyglucose-PET, and structural MRI biomarkers follow the same pattern and begin the transition to an abnormal state. Similarly, Frisoni et al. (2010) established a theoretical progression of cognitive and biological markers (primarily imaging features) based not only on the clinical disease stages, but also patient age at AD diagnosis and time since diagnosis. Although both models captured earliest detectable changes in amyloid markers, Frisoni et al. (2010) additionally theorized that these changes plateau by the MCI stage, when the individuals are no longer cognitively normal. Furthermore, they suggested that F-fluorodeoxyglucose PET is abnormal by the MCI stage and continues to change well into the dementia stage. Structural changes appear later, following a temporal pattern mirroring tau pathology deposition, which slightly differs from the Jack et al. models (Jack et al., 2010, 2013).

While hypothetical models cannot be directly applied, they can be used to suggest directions for future experiments that themselves would address diagnosis, prediction, or decision making tasks (Gladun, 1997). However, there are a number of challenges relating to the construction of hypothetical models. In the following, we discuss these challenges and propose ways to address some of them.

Challenges of Hypothetical Models

The exclusive reliance of hypothetical models on literature presents several challenges. First, relevant literature must be identified. Second, the scientific knowledge contained in the literature must be extracted in a meaningful form. Finally, the knowledge has to be modeled.

In order to build a hypothetical model, a researcher must identify a set of relevant publications, called a literature corpus, which accurately reflects AD knowledge. This corpus should be representative of the relevant aspects of AD, contain the most up-to-date publications, and not be biased toward subfields or trends. However, the number of new AD publications has increased each year since 2005, and there were nearly 15,000 such publications in 2017 alone (Dong et al., 2019). With such publication rates, it is challenging for researchers to manually create high quality corpora (Rodriguez-Esteban, 2015), Moreover, manual generation of these corpora is susceptible to bias, because researchers may tend to draw more heavily from authors or subfields with which they are more familiar (Atkins et al., 1992). The size of a corpus will also be limited by the time and resources available to the researchers. However, text mining has been used effectively to automatically classify relevant literature, based on titles and abstracts (e.g., see Simon et al., 2018), and to prioritize texts (Singh et al., 2015). Publications identified by this classification can be directly taken as the corpus or used as a more manageable set of publications from which the domain experts can appropriately select. Hypothetical models are susceptible to biases present in the literature (Boutron and Ravaud, 2018), but a well-designed, computationally selected corpus can mitigate the effects of those biases.

Once the corpus has been identified, the challenge of knowledge extraction remains. The goal here is to recover the knowledge contained in the publications in a meaningful way. Conducting this task manually is a time-consuming process that requires a high degree of domain knowledge. Here, text mining poses the opportunity to extract knowledge in a computable form (Gyori et al., 2017; Lamurias and Couto, 2019). Moreover, it significantly reduces the amount of time required to read publications, which enables significantly larger corpora to be used in the building of hypothetical models.

Finally, in order to build hypothetical models, the information gleaned from the literature corpus must be organized in a coherent way. The entities and the relationships between them should all be represented. Mind maps provide a non-automated way of generating a knowledge model, driven by domain-expert knowledge (Kudelic et al., 2011). However, if automated information extraction strategies were used on the literature corpus, then knowledge graphs are well-suited for storing the extracted knowledge (Gyori et al., 2017). A major advantage of this strategy is that the knowledge graph is computable, meaning downstream machine learning tasks can be carried out for knowledge discovery. Furthermore, knowledge graphs support hypothesis generation by enabling researchers to assess whether their hypotheses are compatible with existing knowledge (Humayun et al., 2019).

Automated methods of corpus identification, knowledge extraction, and knowledge modeling provide a means of mitigating the challenges of hypothetical modeling. They reduce the time burden, mitigate the risk of bias in manual methods, and generate computable knowledge representations. This can yield more reliable hypothetical AD models.

Hypothetical models are non-numerical and rely exclusively on qualitative information, gleaned from a review of previous findings. This limits their usability solely to eliciting hypotheses for future experiments. They are neither predictive nor can they be used for analysis of any kind of data. They are meant to represent a kind of "typical" AD progression, without reflecting individual deviations from that. Given the broad biological heterogeneity observed among AD subjects, and the increasing relevance of personalized medicine (Reitz, 2016), there is a need for models that are capable of achieving this.

Data-driven models built on data collected in longitudinal cohort studies can serve to support or challenge hypotheses generated by hypothetical models (Petrella et al., 2019). Data-driven models are appropriate for a wide range of tasks that lie beyond the scope of what hypothetical models are designed for. For example, using data models can capture individual subject particularities that hypothetical models cannot (see e.g., Young et al., 2015). In the following, we discuss data-driven models and their challenges in depth.

Data-Driven Models

In contrast to hypothetical models, data-driven integrative models are directly derived from datasets comprising readouts of multiple biomarkers. Such models can be applied to a broad variety of tasks ranging from predictive modeling e.g., predicting patient diagnosis (Ding et al., 2018) or age at disease onset (Chuang et al., 2016; Peng et al., 2016) to discovering patterns in the data that shed light on biomarker interdependencies and disease underlying mechanisms. Since these models use extensive data, they are not limited by preconceived notions in the way that hypothetical integrative models are.

Data-driven AD models can be classified into two primary subtypes based on the statistical approaches and algorithms applied (**Table 1**). The first subtype use traditional statistical methods such as linear modeling, and the second employs artificial intelligence and more specifically machine learning approaches.

Traditional Statistical Models

In AD modeling, traditional statistical approaches, such as linear mixed-effects models, are often used to estimate biomarker trajectories (Caroli and Frisoni, 2010; Jack et al., 2011, 2012). In these models, measured data, are regressed against a prespecified variable, such as disease stage, to detail the temporal changes of AD biomarkers during the course of disease. Essentially, these models provide empirical testing of hypothetical multiple biomarker trajectory plots.

Jack et al. (2012) used linear mixed-effects models to investigate the shape of five important AD biomarker trajectories (i.e., Aβ_{42}, tau, amyloid, fluorodeoxyglucose PET, and structural MRI) as a function of a cognitive test score, the Mini-Mental State Exam (MMSE). This model parameterization enabled them to assess within-subject rates of biomarker changes with respect to changes of the MMSE score. They found that lower baseline MMSE scores are correlated with worse baseline biomarker values and that higher rates of biomarker change were associated with worsening MMSE score. This model constructed the biomarker trajectories without making any assumptions about the shapes of the trajectories. This contrasts with the authors' earlier hypothetical biomarker cascade model, which imposed a sigmoid trajectory curve.

While the shapes of the trajectories in this data-driven model agree with the assumptions made in the hypothetical exemplar, the model has several limitations, pertaining to model design choices and deficiencies in the data. The authors chose to use the MMSE score as the independent variable. This choice was made because the MMSE score provides a linear measure of disease progression that was available across all datasets. However, this introduces challenges in the estimation of trajectories in early disease stages, because MMSE scores in cognitively normal patients are relatively stable over time (Tombaugh, 2005), yielding only a narrow range of values. Moreover, especially when studying early disease stages, the model additionally suffers from possible absence of information on future disease developments of a subject. This absence of data on future disease outcome is related to data censoring, which will be addressed in more detail later.

In their data-driven model (Jack et al., 2011), Jack et al. aimed to unravel the temporal order of biomarker trajectories becoming abnormal, rather than only describing the shape of their trajectories. They used the prevalence of biomarker abnormalities at different disease stages to empirically assess the temporal ordering of their trajectories. They employed generalized estimating equations, a generalized linear model for longitudinal data that can deal with correlated observations, to evaluate and compare the proportion of abnormal observations per biomarker. The proper choice of a cut-off defining when biomarker measures are considered to be abnormal is a point of debate and making this choice requires critical judgement. To differentiate between normal and abnormal biomarkers, Jack et al. (2011) determined a cut-off by looking at an independent post-mortem cohort. However, since, by construction, results were highly sensitive to the selected cut-off for each biomarker, the temporal resolution of the model is limited.

While the proportion of patients with abnormal biomarker values might seem an unnatural choice for comparing biomarkers, alternative strategies also have drawbacks. Caroli and Frisoni (2010) computed Z-scores based on values of each biomarker and fitted them against Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-cog) scores, comparing linear and sigmoidal fits. Their investigation showed that a sigmoid curve fit the observed data significantly better than a linear one for most of the biomarkers, and thereby might be able to characterize the time course of those biomarkers. These results were consistent with the hypothetical model proposed by Jack et al. (2010) and Jack et al. (2013). However, the biomarker trajectories cannot be directly compared with the data-driven model developed by Jack et al. (2011), since different scales were employed in both studies. While standardization of values by converting them into Z-scores resolves this problem, it introduces a new one: by definition, the arithmetic mean of each biomarker will be 0. This makes it impossible to reasonably compare biomarker distributions based on their means using standard statistical procedures like, for example, t-tests (Jack et al., 2011; Moeller, 2015).

The arbitrariness of defining a cut-off for abnormality of a biomarker will always pose a limitation on statistical approaches relying on biomarkers. While such cut-offs simplify the interpretation of the biomarker, there is no universally correct cut-off for a given biomarker. Rather, appropriate cut-offs heavily depend on the population, and even the individual, on which a biomarker will be used. Covariates such as an individual's age, genetic risk factors, and family history of AD must be considered. For these reasons, there is no single optimal cut-off for any given biomarker (Bartlett et al., 2012; Anne and Fagan, 2014). To address this, a less rigid technique has been developed, that designates an intermediate range using two cut-offs, one permissive and the other conservative (Klunk et al., 2012; Jack et al., 2016a,b; Bzdok, 2017). The permissive point can be used for earliest detectable evidence of AD pathologic changes and the conservative one for high diagnostic certainty. Moreover, different statistical approaches, like Youden's index and the receiver operating characteristic (ROC) curve, can be applied to help determine an appropriate cut-off.

Linear traditional models are ill-equipped to handle the increasingly high-dimensional data being collected in AD studies. Thanks to recent technological advancements, the granularity of AD datasets with respect to information resolution, feature size, and complexity of meta-information have increased. For example, improved neuro-imaging techniques generate datasets with higher resolution than previously available. This information distributed over voxels, a 3D imaging unit, is hard to capture using linear models (Bzdok, 2017). Therefore, more advanced data-driven models have been developed based on machine learning. These models are generally more flexible and compatible with the complex datasets encountered in biology research (Bzdok, 2017).

Machine Learning Models

Machine learning models can be characterized as generative or discriminative. As previously mentioned, discriminative models are designed to differentiate between groups, while generative models provide better disease understanding by learning inherent properties from datasets, such as feature interdependencies.

Generative models

Generative modeling relies on the use of statistics and probability to extract patterns from data and learn the underlying distribution. In the following, three types of generative integrative AD models are reviewed: event-based models, Bayesian network learning, and autoencoders.

Event-based models. Event-based models estimate the most probable sequence of events based on the assessment of a probability density function for a particular event order. Fonteijn et al. (2012), Chen et al. (2016), and Oxtoby et al. (2018), used this method to learn the sequence of AD events based on imaging and non-imaging measurements from a clinical study. The authors first fitted simple mixture models (e.g., gaussian mixture models) to individual biomarkers in order to calculate the likelihood of the normality or abnormality status per biomarker. Given these likelihoods, by multiplication of the probabilities, the likelihoods for each possible order of events was calculated. The order with the highest probability was then selected using a greedy Markov Chain Monte Carlo algorithm to describe the temporal correlation of the biomarker trajectories over the course of AD progression.

The models developed by Fonteijn et al. (2012) and Chen et al. (2016) simplified the sequence of biomarker abnormalities over the course of the disease progression by relying on the assumption that all subjects follow a single event sequence. However, AD is highly heterogeneous and includes distinct subgroups (Ferreira et al., 2018). To account for this, Young et al. (2015) established their event-based models with two extensions: a Mallows model and a Dirichlet process mixture of generalized Mallows models. The first extension allows subjects to deviate from the main event sequence, and the latter clusters subjects according to different event sequences.

In principle, the event sequence proposed in the hypothetical model is similar to that observed using traditional and event-based models. Changes in CSF measures are the earliest events, followed by regional brain atrophies and finally succeeded by diminished cognitive scores. However, the event sequence in the hypothetical and traditional models is constructed based on predefined clinical assessments and often imprecise or subjective cut-offs. By contrast, in generative models, the sequence of events, as well as the clustering of biomarkers into normal and abnormal classes, is directly extracted from the data (e.g., the onset of a new symptom, like memory performance decline). Thus, event-based models explain the changes without a priori biases. Moreover, generative models are able to characterize uncertainty in the event ordering arising from heterogeneity in the population and thus, can address individual deviations from the generic model.

Bayesian network learning. Extensive research efforts have been made to uncover the relationships between individual biomarkers and AD. Yet the number of studies that investigated the interplay between multiple biomarkers themselves is comparably limited. Khanna et al. (2018) and Ding et al. (2018) built Bayesian network models covering different biological scales and time points to uncover the interplay amongst sets of biomarkers. Ding et al. (2018) considered the ApoE allele, PET and MRI imaging data, scores from psychological and functional tests, and the medical history of patients with respect to neurological diseases. Using a variety of feature selection metrics, they determined the most relevant features with respect to the clinical dementia rating and modeled these heterogeneous measurements using a Bayesian network to determine their probabilistic interdependencies. However, these models only capture conditional probabilities between predictor variables and clinical outcomes. They are unable to provide a causal mechanistic understanding of an observed phenomenon. Such hypothesized pathophysiological mechanisms are important for making reliable predictions and having confidence in the practical application of data-driven models. To this end, Khanna et al. (2018) employed a combination of data-driven probabilistic and knowledge-driven mechanistic approaches. They modeled clinical variables, genetic variants, pathways, and neuro-imaging readouts using Bayesian network learning to estimate dependencies between disease relevant features. Together with a cause-and-effect knowledge model derived from scientific literature, they partially reconstructed biological mechanisms that could play a role in the conversion of normal/MCI into AD pathology.

Autoencoders. The last type of generative model discussed in this review is autoencoders. In essence, an autoencoder is a neural network that aims to encode the input data into a lower dimensional representation and from that decode it again, reconstructing the original input. It has successfully been applied for different tasks on AD cohorts (Basu et al., 2019; Martinez-Murcia et al., 2019). The two main applications of this approach in the field consist of classifying patients based on AD diagnosis (Basu et al., 2019) and clustering of patient trajectories into subgroups (De Jong et al., 2019). These strategies are especially interesting for patient classification and stratification tasks in datasets where information is sparse. However, another novel and promising task for autoencoders is the generation of synthetic data from real patient level data (Gootjes-Dreesbach et al., 2019). This, in turn, could be used to circumvent legal and ethical constraints that restrict data sharing.

Discriminative models

Discriminative models are a class of models generally used for classification. Discriminative models that rely on labeled data are called supervised models, while unsupervised models use unlabeled data.

Supervised discriminative models. Diverse supervised discriminative methods such as support vector machines (SVM; Magnin et al., 2010), and multiple-kernel SVM (MKL; Hinrichs et al., 2010; Zhang et al., 2011) have been used to classify AD patients, MCI subjects, and controls. However, studies that used multiple-kernel SVM reported superior classification performance, because the use of multiple kernels facilitates the integration of multimodal biomarker data (Zhang et al., 2011). Additionally, MKL are well-suited for dealing with very high dimensional data (Young et al., 2013). MKL also enable individual weighting of biomarker modalities. This offers more flexibility for kernel combination and thus, a better integration of the data. For example Hinrichs et al. (2010), applied MKL in combination with MRI and PET imaging to differentiate between AD subjects and controls. Their method showed high classification performance, achieving 92.4% accuracy. Similarly, Zhang et al. (2018) combined MRI, PET, and CSF biomarkers to discriminate between healthy controls and AD/MCI. After integrating all biomarker data using a MKL, they deployed a linear SVM for the actual classification task, which resulted in 93.2% accuracy for classifying AD and healthy controls and 76.4% for discriminating between MCI and healthy controls. Both studies applied a similar method for classification, yet the latter one achieved a slightly higher accuracy. Comparing the approaches applied in Zhang et al. (2018) and Hinrichs et al. (2010) it becomes clear that the major reason for the difference in performance is the feature selection process. Depending on the available sample size, other methods might prove more promising (Liu et al., 2012). Moreover, Zhang et al. (2018) benefits from employing three biomarker modalities, namely, CSF measurements and two imaging modalities, compared to Hinrichs et al. (2010) who only use the two imaging modalities.

While the above kernel-based pattern recognition approaches yield categorical class decisions, Young et al. (2013) used gaussian process classification, which is a probabilistic classification algorithm. This study integrated imaging, CSF, neuropsychological, and genetic biomarkers to classify MCI subjects who remained stable and MCI patients who converted to AD within 3 years. In contrast to MKL, the probabilistic classification afforded by the gaussian process approach provides the opportunity to position the subjects according to disease stage, to stratify patients, and to model the sequence order of biomarker abnormality.

Another type of discriminative model is disease risk models. This type of supervised model can be used to predict the time to AD diagnosis for normal/MCI patients. Multiple approaches have been used to develop risk models for AD (Da et al., 2013; Li et al., 2013). Li et al. (2013) used a combination of cox regression analyses and time-dependent ROC approaches to evaluate prognostic utility and performance stability of candidate biomarkers. The authors deduced that both baseline volumetric MRI and cognitive measures can predict progression from MCI to AD. However, in participants' follow-up visits, only cognitive measurements remained predictive. Da et al. (2013) employed the cox proportional hazards models to compare the magnitudes of the relative association between predictors (patterns of brain atrophy, cognitive assessments, genetics, and CSF biomarkers) and time to conversion from MCI to AD. They concluded that brain atrophy and cognitive assessments in combination offer the highest predictive power of conversion from MCI to AD.

Although the results in both studies were similar, the time-dependent ROC curve used by Li et al. (2013) enabled them to predict disease risk as a function of time. Thus, this method provides clear benefit for a progressive disease such as AD, in which both the disease status and biomarker measurements change over time (Kamarudin et al., 2017).

The data labeling which enables supervised discriminative models to determine decision boundaries for distinguishing classes of interest can also introduce errors. Inaccurate labels will negatively affect the performance of the classifier. Such mislabeling is not uncommon in AD, due to the absence of a clear diagnostic biomarker (Fischer et al., 2017). Instead, diagnosis is currently made based on symptoms (Schott and Petersen, 2015) Furthermore, integrative data analysis is further complicated by the fact that the diagnostic criteria for MCI have changed over the years, and MCI is not consistently defined across clinical studies. While one study relies on assessing only a single cognitive domain for MCI diagnosis, such as speech or memory, others base their diagnoses on performance on cognitive tests for multiple domains. Apart from that, there are multiple pathologies for MCI; AD is just one of them. Thus, unified clear disease definitions are crucial, since the MCI classification accuracy can influence outcomes of research and clinical practice (Jak et al., 2010).

Unsupervised Discriminative Models. Unsupervised discriminative models use a variety of clustering techniques on unlabeled data, avoiding the challenges of data label accuracy. These techniques use properties of each data point to iteratively form groups, called clusters. This ultimately leads to a discrimination of the data into several clusters of highly similar data points. Given the observed biological heterogeneity among normal control subjects, Nettiksimmons et al. (2014) hypothesized that different subgroups may also be found among the MCI subjects. Using agglomerative hierarchical clustering, they sorted subjects based on MRI volumes, CSF measurements, and cognitive tests. Next, the resulting clusters were explored with regard to longitudinal atrophy, conversion time, and cognitive trajectories. Four clusters with unique biomarker patterns resulted: (i) a cluster biologically similar to normal controls. MCI patients from that cluster rarely converted to AD, (ii) one cluster with early AD pathology characteristics, (iii) another cluster of subjects with hardly any tau abnormality, but a high proportion of AD converters, and (iv) and finally one cluster with pre-AD symptoms wherein almost all subjects converted to AD. Based on these findings, they hypothesized that clusters ii and iv reflected the amyloid cascade pattern (Ricciarelli and Fedele, 2017) since both clusters presented lower CSF Aβ levels and elevated tau proteins. However, the tau level in cluster iv was higher, and more severe atrophy as well as cognitive impairment were detected. The authors concluded that more tau accumulation may lead to more cognitive decline. One of the intrinsic limitations of their clustering approach is that the number of clusters must be predefined. The maximum gap statistic is one approach to determine this number (Tibshirani et al., 2001). However, specifying the number of clusters beforehand will always bias the clustering to some extent, and choosing a reasonable number is no trivial task given the broad variety of subtypes found among AD subjects.

Toschi et al. (2019) used Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Thanh et al., 2013), which does not require pre-specifying the number of clusters. They integrated five validated CSF biomarkers in order to cluster a cohort where symptomatic patients presented diagnoses ranging from self-perceived cognitive decline (Zhang et al., 2011) to MCI to AD. In contrast to the previous study, Toschi et al. (2019) adjusted all biomarker values for age, sex and their interactions to exclude them as confounders (Pourhoseingholi et al., 2012). Moreover, Toschi et al. (2019) used t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of biomarkers space, since defining the distance between the data points in a high dimensional space of biomarkers is notoriously difficult (Domingos, 2012). Finally, they applied DBSCAN on this lower dimensional representation. DBSCAN defines a high data density region based on two parameters: (i) the radius of the neighborhood, and (ii) the minimum number of points within the radius. These values are determined by a nearest neighbor method, in which the distance of each point to their nearest n points is calculated. Afterwards, results are sorted, plotted and the value with most pronounced change is selected as the optimal value. Using DBSCAN, Toschi et al. (2019) characterized five biological clusters which were not significantly bound to the original distinct clinically phenotyped diagnostic groups. They explained that the clusters included all phenotypic groups and were not homogeneous enough to be considered as a specific AD pathophysiology. Moreover, contrary to general belief that $A\beta_{1-42}$ is linearly associated with the progression of AD and cognitive decline (Sperling et al., 2011a; Samtani et al., 2013), their findings suggest that $A\beta_{1-42}$ is less likely to contribute to phenotypic discrimination.

The dimensionality reduction technique, t-SNE, used by Toschi et al. (2019) enabled them to better separate the data and hence, to enhance cluster identification, in comparison to directly running a clustering algorithm on a high dimensional data as Nettiksimmons et al. (2014). However, their main limitation is that clustering results are highly sensitive to two parameters necessary for DBSCAN. Moreover, they did not include other biomarkers, such as imaging and genetics biomarkers, which could enhance their clustering, as previously reported by Young et al. (2013, 2018).

Unsupervised clustering algorithms are ideal for identifying subgroups and non-linear associations between individuals based on a multidimensional profile, regardless of the individual labels, in contrast to supervised algorithms. This allows the grouping of individuals based on shared pathophysiological drivers and triggers and, possibly, similar longitudinal disease trajectories. This is an advantage in the AD field due to the prevalence of unreliable labels stemming from misdiagnosis and to the biological heterogeneity of AD subjects. On the other hand, most unsupervised clustering algorithms perform better with a larger sample size than is often obtainable in AD studies (Oxtoby and Alexander, 2017). Therefore, the smaller size inherent to AD cohorts may lead to clustering instability.

To this point, we have reviewed a broad variety of data-driven integrative AD models and elaborated on their associated limitations and challenges. In the following, we enumerate more general challenges researchers encounter in the course of data-driven integrative AD modeling and suggest how these could be addressed.

Challenges of Data-Driven Modeling

Although there exists a wide range of data-driven integrative modeling approaches, not all of them are well-suited for every analytic task and each has its own strengths and weaknesses. Still, there are some challenges which affect all data-driven approaches to some degree: data collection, reproducibility of findings, and interpretability of models and results.

Data Collection

Collecting patient level data, the basis for all data-driven modeling, is a time-consuming and costly process. Additionally, it is a source of major challenges and limitations of these models. In particular, data "censoring" and "missingness," can impede modeling, bias models, or even make certain modeling techniques unfeasible.

Data censoring describes the condition in which a particular event (here AD diagnosis) is not observed for certain study participants during the study runtime. This censoring can occur in two ways: if AD diagnosis occurred before the start of the study; or if the patient drops out of the study, or the study ends without occurrence of the AD diagnosis event. A significant number of patients enrolled in clinical studies have already received a diagnosis before the beginning of the study, indicating that they are in a progressed stage of the disease (Ellis et al., 2009). It is therefore not possible to obtain indications of early disease onset in such patients. The second form of censoring arises from two sources. First, all observational cohort studies experience participant dropout for a variety of reasons, including the participation burden on caregivers or medical problems (Coley et al., 2008). Second, subjects that remain healthy throughout study runtime could still develop the disease after the study ended, meaning they were in a prodromal disease stage. It is thus impossible to know if or when the patient would eventually receive an AD diagnosis. This form of censoring is common in longitudinal AD studies, because AD is a slow-progressing disease, while the studies are typically quite short (Lawrence et al., 2017), due to limited funding (Prabhakaran and Bakshi, 2018).

Disease onset is a critical point for clinical intervention (Sperling et al., 2011b), so it is subject to extensive research efforts. It is here, however, where data censoring impedes data analysis the most. Data censoring can result in over- or under-sampling of early and advanced disease stages. This, in turn, leads to models biased toward specific disease stages (Ning et al., 2010). Various methods, such as complete data analysis (Xiang et al., 2013), imputation (Fisher et al., 2019), or analysis based on dichotomized data (Donohue et al., 2011), have been established to address censored data. Yet all of these methods may introduce error and impose complexities and biases on other integrative modeling steps, such as model interpretation, and thus need to be used with care (Prinja et al., 2010).

The complete absence of a value for variables in the observation of interest likewise poses a significant challenge to data-driven modeling. This missing data in AD cohort studies occurs for several reasons, including unwillingness of patients to undergo invasive tests like lumbar punctures, and the high cost of measuring a particular variable, such as imaging scans (Engelborghs et al., 2017). The implications of such a scenario include a loss of statistical power of the study and may bias the conclusions that can be drawn (Hughes et al., 2019). Over the past decades, novel statistical methods (Molenberghs et al., 2014) and software (Quartagno and Carpenter, 2016; Moreno-Betancur et al., 2017) have been developed for analyzing data with missing values. However, analysis restricted to individuals with complete data is generally preferred, if feasible.

Despite the challenges in collecting complete and uncensored data, the value of data in strengthening disease understanding is clear. Several large-scale AD patient datasets have been collected for use in a variety of studies (Lawrence et al., 2017) including, for example, Alzheimer's Disease Neuroimaging Initiative (ADNI; Mueller et al., 2005), Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL; Ellis et al., 2009), the Dominantly Inherited Alzheimer Network (DIAN; Moulder et al., 2013), and European Prevention of Alzheimer's Dementia (EPAD; Vermunt et al., 2018). However, these classical observational studies are subject to bias, resulting from the inclusion and exclusion criteria used to select participants (Miksad and Abernethy, 2018).

The use of electronic medical records (EMRs) has been proposed as a potential solution to reduce the bias of classical clinical trials. They provide an alternative view on patient measurements (Fröhlich et al., 2018), so, a collection of EMRs can provide a more representative view on patient measurements. However, EMRs are largely phenotypic: molecular phenomena such as genomic variants are not reflected in the data. Moreover, extracting information from EMRs requires natural language preprocessing, which itself currently remains a difficult and error-prone process.

Reproducibility

The ability to reproduce the findings of a study using different subjects is an important part of scientific research. This is particularly the case in integrative AD modeling, since the tendency of AD datasets is not to fully reflect the diversity of AD patients. Inclusion-exclusion criteria in clinical studies can lead to significant under-representation of some populations. For example, the landscape of data-driven AD models is currently dominated by only a few cohorts which are made up largely of White Caucasians, and, to a lesser extent, are constrained by geographic location (Lawrence et al., 2017). Since most observational cohorts are not representative of the general AD population (Ferreira et al., 2017), it is important to validate the resulting models with an independent cohort study. While this external validation is a necessary step to corroborate findings, it is complicated by data interoperability and sample size.

Interoperability

The ability to map the data coming from one study to data from another study is known as data interoperability[5]. Each of the major AD clinical studies was established with a specific sample and feature characterization. Since they might not be directly interoperable, extensive curation is needed before the external validation of a model can be carried out. Otherwise, the training cohort and the validation cohort would be based on different populations, and would contain different measurements. Thus, before validation, researchers must map and assess the "comparability" of both features and subjects.

Feature mapping requires specifying relationships between data elements from different data models and standardizing the terms used to represent the features in the two datasets. This is due to the fact that controlled vocabularies are not used to annotate the datasets. Thus, even if the same biomarker has been collected in two studies, it is usually referred to by different terms, impeding a direct comparison of the datasets. For example, the hippocampus is one of the earliest sites of AD pathology, and hippocampal volume is measured in ADNI and EPAD. However, ADNI identifies this biomarker as "Hippocampus," while EPAD refers to it as "lhvr" (right hemisphere) and "lhvl" (left hemisphere).

Moreover, the subject populations in each study must be comparable. For instance, if the biological sex distributions in two AD studies differ significantly, then the cognitive impairment scores of the cohorts cannot be directly compared, because female AD patients have been shown to have greater cognitive impairment than men in comparable stages of the disease (Laws et al., 2016).

There are several strategies to overcome the lack of interoperability between datasets at both feature and subject level. At the feature level, interoperability can be attained by annotating datasets according to a standard controlled vocabulary. Several such vocabularies (e.g., NIFT Iyappan et al., 2017 and PTS Iyappan et al., 2016) have been established, but significant improvements in interoperability will only come with widespread adoption (Neu et al., 2012). The most prominent example might be the AD specific standard developed by the Clinical Data Interchange Standards Consortium (CDISC; Neville et al., 2017). At the subject level, mapping between training and validation cohorts can be accomplished by identifying, in the validation cohort, a subset of subjects that is statistically comparable to the training cohort. Finally, in order to assess the comparability of subjects from different studies, techniques such as statistical matching can be used (Austin, 2011).

Sample size

The relatively small sample sizes of AD clinical studies also contributes to the challenge of reproducibility in AD integrative modeling. Many AD studies contain fewer than a thousand patients, and the longitudinal follow-up is limited. In addition, typically not all of the subjects were screened for the complete biomarker set, leading to sparse subsets of patients for whom the study contains complete data. As a result, models generated from these studies have a high margin of error and low statistical power, meaning they struggle to detect small effects.

The integration of different datasets into a larger dataset can overcome some of the challenges related to small sample sizes (Gomez-Cabrero et al., 2014). Integrated datasets provide more comprehensive data, and the resulting models have greater statistical power. However, current approaches for data integration were developed for the analysis of single-data-type datasets, and only subsequently adapted to handle datasets with multiple data types. For this reason, data integration methodologies can be ill-suited to manage the computational challenges arising from the variety of different data sizes, formats, and dimensionalities present in AD datasets, as well as their noisiness, complexity, and the level of agreement between datasets (Gomez-Cabrero et al., 2014; Gligorijević et al., 2015). Furthermore, even data acquired by analogous technologies are not necessarily integrable. For example, neuroimaging data acquired from similar scanners and similar modalities may still be stored in different formats and have different metadata content (Goble and Stevens, 2008).

Several strategies could be applied to address the interoperability challenges arising from data integration. The first strategy is to normalize and standardize data across all platforms (O'Bryant et al., 2015). However, scientific independency and freedom for innovation, as well as uniqueness of databases, must be respected. The second strategy is to collect a standardized set of biomarkers across different studies. Finally, the ideal solution would be performing a systematic longitudinal clinical and -omics follow-up of each individual in a large and rigorously characterized cohort since this would provide a statistically sufficient number of measurements in the context of subjects and variables. The Deep and Frequent Phenotyping study from Lawson et al. (2017) showed that such a cohort, in theory, is feasible. Yet, including a sufficient number of participants in such an ambitious study is costly.

Interpretability

In order for an AD model to have clinical impact, its findings must be interpretable. There are several barriers to AD model interpretability. Machine learning models often act as "black boxes"; it may be impossible to uncover the reasons for the predictions made by the model (Rudin, 2019). Indeed, as the number of features and the complexity of the computational processes used in models increases, this interpretability problem will worsen. Moreover, data-driven models are not causal and typically capture non-linear correlations between predictor and explanatory variables. While prior understanding of cause–effect relationships and detailed mechanisms might prove helpful to well-performing models, it is not necessarily required. Lack of mechanistic explanations for model prediction complicates the interpretation of data-driven findings and reduces acceptance by physicians (Fröhlich et al., 2018). Thus, the translation of data-driven models into a biomedical knowledge context is a major challenge in integrative AD modeling.

[5] https://library.ahima.org/doc?oid=65895#.Xdl-iZPYrOQ

Combining available mechanistic knowledge with machine learning-based sub-models, so-called hybrid modeling could bridge the gap between experimental biological and computational research by improving interpretability (Fröhlich et al., 2018). For example, Bayesian networks which built on causal knowledge graphs constitute such a hybrid model (Arora et al., 2019). They shed light on interdependencies across features, which can be on different scales (e.g., clinical, genetic, and molecular), and allow for predicting the outcome of purely hypothetical clinical interventions. Similarly, other recent deep learning methodologies use knowledge-derived biological networks to define the layers of neural networks in order to improve interpretability (Fortelny and Bock, 2019).

CONCLUSION

In the era of extensive biomarker profiling, big data, and artificial intelligence, integrative AD modeling comes with high promises. By integrating multi-scale, multimodal, and longitudinal patient data, such modeling approaches aim to provide a holistic picture of disease pathophysiology and progression. However, as we have discussed in this review, while integrative models have generated significant insights, and thus proved to be valuable in research, existing models do not yet fully describe critical aspects of AD.

The construction of hypothetical models simultaneously benefits and suffers from the vast amount of published knowledge. Prioritization of articles and computational text mining of literature corpora are reasonable approaches to identify a greater quantity of relevant knowledge while designing hypothetical models. In the field of data-driven integrative AD modeling, we highlighted several major ongoing challenges throughout the whole modeling process of data collection, integration of disparate data sources, data analysis, and model interpretation. Data missingness and data censoring are major bottlenecks in data collection as well as analysis and interpretation. Heterogeneity and complexity in biological data are major impediments to data integration, which is central to data-driven integrative modeling and validation. Data mapping, imprecise diagnostic stages, and biased data are barriers that hamper data analysis and interpretation. Furthermore, there is an insufficient number of subjects in studies, which restricts the statistical power of data-driven integrative AD models. Because of these challenges, to the best of our knowledge, at this point in time, there are no integrative AD models which have been used in clinical practice.

While in theory, certain existing integrative models are capable of predicting AD diagnosis and progression, they are not used in clinical practice. We see a number of steps that could bring us closer to the goal of precision medicine and that could enable patient diagnosis through integrative disease models in a clinical context. First, we, the AD research community, need to establish valid, informative biomarkers and clear criteria for AD diagnosis. This would result in reliable predictors that could be included in modeling approaches, as well as fewer diagnostic errors, which in turn reduce the effect of mislabeled data. Second, a global data schema that could support the normalization and standardization of data across measurements would ultimately facilitate improved data integration. If future cohort studies would adhere to such a schema, data integration would be straightforward and the cumulative time saved for researchers working with it would be enormous. Finally, innovative modeling approaches, such as causal inference techniques and hybrid modeling, which go beyond current state-of-the-art data-driven models by linking prior knowledge with data-driven models, need to be developed and made more robust. Overall, novel computational modeling approaches that surmount the current integrative AD modeling challenges may hold the potential to play an increasing role in the planning of medical interventions and practice.

AUTHOR CONTRIBUTIONS

SG drafted the manuscript. CR, CB, DD-F contributed to the final version of the manuscript. CH and MH-A reviewed the final version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support from the B-IT foundation.

REFERENCES

Aisen, P. S., Cummings, J., Jack, C. R. Jr., Morris, J. C., Sperling, R., Frölich, L., et al. (2017). On the path to 2025: understanding Alzheimer's disease continuum. *Alzheimers Res. Ther.* 9:60. doi: 10.1186/s13195-017-0283-5

Anne, M., and Fagan. (2014). CSF biomarkers of Alzheimer's disease: impact on disease concept, diagnosis, and clinical trial design. *Adv. Geriatr.* 2014:302712. doi: 10.1155/2014/302712

Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., and Druzdzel, M. J. (2019). Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Val. Health* 22, 439–445. doi: 10.1016/j.jval.2019.01.006

Atkins, S., Clear, J., and Ostler, N. (1992). Corpus design criteria. *Lit. Ling. Comput.* 7, 1–16. doi: 10.1093/llc/7.1.1

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46, 399–424. doi: 10.1080/00273171.2011.568786

Bartlett, J. W., Frost, C., Mattsson, N., Skillbäck, T., Blennow, K., Zetterberg, H., et al. (2012). Determining cut-points for Alzheimer's disease biomarkers: statistical issues, methods and challenges. *Biomark. Med.* 6, 391–400. doi: 10.2217/bmm.12.49

Basu, S., Wagstyl, K., Zandifar, A., Collins, L., Romero, A., and Precup, D. (2019). "Early prediction of alzheimer's disease progression using variational autoencoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, eds D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan (Cham: Springer), 205–213. doi: 10.1007/978-3-030-32251-9_23

Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and biomarker changes in dominantly inherited

Alzheimer's disease. *N. Engl. J. Med.* 367, 795–804. doi: 10.1056/NEJMoa1202753

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

Boutron, I., and Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2613–2619. doi: 10.1073/pnas.1710755115

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* 11:543. doi: 10.3389/fnins.2017.00543

Caroli, A., and Frisoni, G. B. (2010). The dynamics of Alzheimer's disease biomarkers in the Alzheimer's disease neuroimaging initiative cohort. *Neurobiol. Aging* 31, 1263–1274. doi: 10.1016/j.neurobiolaging.2010.04.024

Chen, G., Shu, H., Chen, G., Ward, B. D., Antuono, P. G., Zhang, Z., et al. (2016). Staging Alzheimer's disease risk by sequencing brain function and structure, cerebrospinal fluid, and cognition biomarkers. *J. Alzheimers Dis.* 54, 983–993. doi: 10.3233/JAD-160537

Chuang, Y. F., An, Y., Bilgel, M., Wong, D. F., Troncoso, J. C., O'Brien, R. J., et al. (2016). Midlife adiposity predicts earlier onset of Alzheimer's dementia, neuropathology and presymptomatic cerebral amyloid accumulation. *Mol. Psychiatry* 21, 910–915. doi: 10.1038/mp.2015.129

Coley, N., Gardette, V., Toulza, O., Gillette-Guyonnet, S., Cantet, C., Nourhashemi, F., et al. (2008). Predictive factors of attrition in a cohort of Alzheimer disease patients. *Neuroepidemiology* 31, 69–79. doi: 10.1159/000144087

Da, X., Toledo, J. B., Zee, J., Wolk, D. A., Xie, S. X., and Ou, Y. (2013). Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *Neuroimage Clin.* 4, 164–173. doi: 10.1016/j.nicl.2013.11.010

De Jong, J., Emon, M. A., Wu, P., Karki, R., Sood, M., Godard, P., et al. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience* 8:giz134. doi: 10.1093/gigascience/giz134

Ding, X., Bucholc, M., Wang, H., Glass, D. H., Wang, H., Clarke, D. H., et al. (2018). A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data. *Sci. Rep.* 8:9774. doi: 10.1038/s41598-018-27997-8

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi: 10.1145/2347736.2347755

Dong, R., Wang, H., Ye, J., Wang, M., and Bi, Y. (2019). Publication trends for Alzheimer's disease worldwide and in China: a 30-year bibliometric analysis. *Front. Hum. Neurosci.* 13:259. doi: 10.3389/fnhum.2019.00259

Donohue, M. C., Gamst, A. C., Thomas, R. G., Xu, R., Beckett, L., Petersen, R. C., et al. (2011). The relative efficiency of time-to-threshold and rate of change in longitudinal data. *Contemp. Clin. Trials* 32, 685–693. doi: 10.1016/j.cct.2011.04.007

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., et al. (2009). 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.* 21, 672–687. doi: 10.1017/S1041610209009405

Engelborghs, S., Niemantsverdriet, E., Struyfs, H., Blennow, K., Brouns, R., Comabella, M., et al. (2017). Consensus guidelines for lumbar puncture in patients with neurological diseases. *Alzheimers Dement.* 8:111–126. doi: 10.1016/j.dadm.2017.04.007

Ferreira, D., Hansson, O., Barroso, J., Molina, Y., Machado, A., Hernández-Cabrera, J. A., et al. (2017). The interactive effect of demographic and clinical factors on hippocampal volume: a multicohort study on 1958 cognitively normal individuals. *Hippocampus* 27, 653–667. doi: 10.1002/hipo.22721

Ferreira, D., Wahlund, L., and Westman, E. (2018). The heterogeneity within Alzheimer's disease. *Aging* 10, 3058–3060. doi: 10.18632/aging.101638

Fischer, C. E., Qian, W., Schweizer, T. A., Ismail, Z., Smith, E. E., Millikin, C. P., et al. (2017). Determining the impact of psychosis on rates of false-positive and false-negative diagnosis in Alzheimer's disease. *Alzheimers Dement.* 3, 385–392. doi: 10.1016/j.trci.2017.06.001

Fisher, C. K., Smith, A. M., and Walsh, J. R. (2019). Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci. Rep.* 9:13622. doi: 10.1038/s41598-019-49656-2

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* 60, 1880–1889. doi: 10.1016/j.neuroimage.2012.01.062

Fortelny, N., and Bock, C. (2019). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. doi: 10.1101/794503

Frisoni, G. B., Fox, N. C., Jack, C. R. Jr, Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneurol.2009.215

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., et al. (2018). From hype to reality: data science enabling personalized medicine. *BMC Med.* 16:150. doi: 10.1186/s12916-018-1122-7

Gamberger, D., Lavrač N., Srivatsa, S., Tanzi, R. E., and Doraiswamy, P. M. (2017). Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease. *Sci. Rep.* 7:6763. doi: 10.1038/s41598-017-06624-y

Gladun, V. P. (1997). Hypothetical modeling: methodology and application. *Cybern. Syst. Anal.* 33, 7–15. doi: 10.1007/BF02665935

Gligorijević V., and PrŽulj, N. (2015). Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface* 12:20150571. doi: 10.1098/rsif.2015.0571

Goble, C., and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *J. Biomed. Inform.* 41, 687–693. doi: 10.1016/j.jbi.2008.01.008

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8(Suppl. 2):I1. doi: 10.1186/1752-0509-8-S2-I1

Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., and Fröhlich, H. (2019). Variational Autoencoder Modular Bayesian Networks (VAMBN) for simulation of heterogeneous clinical study data. *bioRxiv* 760744. doi: 10.1101/760744

Gyori, B. M., Bachman, J. A., Subramanian, K., Muhlich, J. L., Galescu, L., and Sorger, P. K. (2017). From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.* 13:954. doi: 10.15252/msb.20177651

Hampel, H., O'Bryant, S. E., Durrleman, S., Younesi, E., Rojkova, K., Escott-Price, V., et al. (2017). A precision medicine initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling. *Climacteric* 20, 107–118. doi: 10.1080/13697137.2017.1287866

Hinrichs, C., Singh, V., Xu, G., and Johnson, S. (2010). MKL for robust multi-modality AD classification. *Med. Image Comput. Comput. Assist. Interv*. 12(Pt 2), 786–794. doi: 10.1007/978-3-642-04271-3_95

Hinrichs, C., Singh, V., Xu, G., and Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* 55, 574–589. doi: 10.1016/j.neuroimage.2010.10.081

Hughes, R. A., Heron, J., Sterne, J. A. C., and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int. J. Epidemiol.* 48, 1294–1304. doi: 10.1093/ije/dyz032

Humayun, F., Domingo-Fernandez, D., George, A. A. P., Hopp, M. T., Syllwasschy, B. F., Detzel, M., et al. (2019). A computational approach for mapping heme biology in the context of hemolytic disorders. *bioRxiv* 804906. doi: 10.1101/804906

Iyappan, A., Gündel, M., Shahid, M., Wang, J., Li, H., Mevissen, H. T., et al. (2016). Towards a pathway inventory of the human brain for modeling disease mechanisms underlying neurodegeneration. *J. Alzheimers Dis.* 52, 1343–1360. doi: 10.3233/JAD-151178

Iyappan, A., Younesi, E., Redolfi, A., Vrooman, H., Khanna, S., Frisoni, G. B., et al. (2017). Neuroimaging feature terminology: a controlled terminology for the annotation of brain imaging features. *J. Alzheimers Dis.* 59, 1153–1169. doi: 10.3233/JAD-161148

Jack, C. R. Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., et al. (2016a). A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology* 2, 539–547. doi: 10.1212/WNL.0000000000002923

Jack, C. R. Jr., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* 12, 207–216. doi: 10.1016/S1474-4422(12)70291-0

Jack, C. R. Jr., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* 9, 119–128. doi: 10.1016/S1474-4422(09)70299-6

Jack, C. R. Jr., Vemuri, P., Wiste, H. J., Weigand, S. D., Aisen, P. S., Trojanowski, J. Q., et al. (2011). Evidence for ordering of Alzheimer disease biomarkers. *Arch. Neurol.* 68, 1526–1535. doi: 10.1001/archneurol.2011.183

Jack, C. R. Jr., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Lowe, V., et al. (2012). Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Arch. Neurol.* 69, 856–867. doi: 10.1001/archneurol.2011.3405

Jack, C. R. Jr., Wiste, H. J., Weigand, S. D., Therneau, T. M., Lowe, V. J., and Knopman, D. S. (2016b). Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement.* 13, 205–216. doi: 10.1016/j.jalz.2016.08.005

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., et al. (2010). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 17, 368–375. doi: 10.1097/JGP.0b013e31819431d5

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* 51, 404–413. doi: 10.1038/s41588-018-0311-9

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., and Fröhlich, H. (2018). Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. *Sci. Rep.* 8:11173. doi: 10.1038/s41598-018-29433-3

Klunk, W., Cohen, A., Bi, W., Weissfeld L,Aizenstein, H., McDade, E., et al. (2012). Why we need two cutoffs for amyloid imaging: early versus Alzheimer's-like amyloid-positivity. *Alzheimers Dement.* 8, P453–P454. doi: 10.1016/j.jalz.2012.05.1208

Kudelic R., Konecki, M., and Malekovic, M. (2011). "Mind map generator software model with text mining algorithm," in *Proceedings of the ITI 2011, 33rd International Conference on Information Technology Interfaces* (Cavtat), 487–494.

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* 51, 414–430. doi: 10.1038/s41588-019-0358-2

Lamurias, A., and Couto, F. M. (2019). "Text mining for bioinformatics using biomedical literature," in *Encyclopedia of Bioinformatics and Computational Biology*, eds S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Elsevier), 602–611. doi: 10.1016/B978-0-12-809633-8.20409-3

Lawrence, E., Vegvari, C., Ower, A., Hadjichrysanthou, C., De Wolf, F., and Anderson, R. M. (2017). A systematic review of longitudinal studies which measure alzheimer's disease biomarkers. *J. Alzheimers Dis.* 59, 1359–1379. doi: 10.3233/JAD-170261

Laws, K. R., Irvine, K., and Gale, T. M. (2016). Sex differences in cognitive impairment in Alzheimer's disease. *World J. Psychiatry* 6, 54–65. doi: 10.5498/wjp.v6.i1.54

Lawson, J., Murray, M., Zamboni, G., Koychev, I. G., Ritchie, C. W., Ridha, B. H., et al. (2017). Deep and frequent phenotyping: a feasibility study for experimental medicine in dementia. *J Alzheimers Dement.* 13, p1268–1269. doi: 10.1016/j.jalz.2017.06.1897

Li, S., Okonkwo, O., Albert, M., and Wang, M. C. (2013). Variation in variables that predict progression from MCI to AD dementia over duration of follow-up. *Am. J. Alzheimers Dis.* 2, 12–28. doi: 10.7726/ajad.2013.1002

Liu, M., Zhang, D., Yap, P. T., and Shen, D. (2012). Tree-guided sparse coding for brain disease classification. *Med. Image Comput. Comput. Assist. Interv.* 15(Pt 3), 239–247. doi: 10.1007/978-3-642-33454-2_30

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., and Sarazin, M. (2010). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* 51, 73–83. doi: 10.1007/s00234-008-0463-x

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J. M., Ramirez, J., and Castillo-Barnes, D. (2019). Studying the manifold structure of Alzheimer's Disease: a deep learning approach using convolutional autoencoders. *IEEE J. Biomed. Health Inform.* 1-1. doi: 10.1109/JBHI.2019.2914970

Miksad, R. A., and Abernethy, A. P. (2018). Harnessing the Power of Real-World Evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin. Pharmacol. Ther.* 103, 202–205. doi: 10.1002/cpt.946

Moeller, J. (2015). A word on standardization in longitudinal studies: don't. *Front. Psychol.* 6:1389. doi: 10.3389/fpsyg.2015.01389

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. New York, NY: Chapman and Hall/CRC. doi: 10.1201/b17622

Moreno-Betancur, M., Leacy, F. P., Tompsett, D., and White, I. (2017). *mice: The NARFCS Procedure for Sensitivity Analyses.*

Moulder, K. L., Snider, B. J., Mills, S. L., Buckles, V. D., Santacruz, A. M., Bateman, R. J., et al. (2013). Dominantly inherited Alzheimer network: facilitating research and clinical trials. *Alzheimers Res. Ther.* 5:48. doi: 10.1186/alzrt213

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N Am.* 15, 869–877. doi: 10.1016/j.nic.2005.09.008

Nettiksimmons, J., DeCarli, C., Landau, S., and Beckett, L. (2014). Biological heterogeneity in ADNI amnestic mild cognitive impairment. *Alzheimers Dement.* 10, 511–521.e1. doi: 10.1016/j.jalz.2013.09.003

Neu, S. C., Crawford, K. L., and Toga, A. W. (2012). Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories. *Front. Neuroinform.* 6:8. doi: 10.3389/fninf.2012.00008

Neville, J., Kopko, S., Romero, K., Corrigan, B., Stafford, B., LeRoy, E., et al. (2017). Accelerating drug development for Alzheimer's disease through the use of data standards. *Alzheimer's Dement.* 3, 273–283. doi: 10.1016/j.trci.2017.03.006

Ning, J., Qin, J., and Shen, Y. (2010). Nonparametric tests for right-censored data with biased sampling. *J. R. Stat. Soc. Series B Stat. Methodol.* 72, 609–630. doi: 10.1111/j.1467-9868.2010.00742.x

O'Bryant, S. E., Gupta, V., Henriksen, K., Edwards, M., Jeromin, A., Lista, S., et al. (2015). Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. *Alzheimers Dement.* 11, 549–560. doi: 10.1016/j.jalz.2014.08.099

Oxtoby, N. P., and Alexander, D. C. (2017). Imaging plus X: multimodal models of neurodegenerative disease. *Curr. Opin. Neurol.* 30, 371–379. doi: 10.1097/WCO.0000000000000460

Oxtoby, N. P., Young, A. L., Cash DM Benzinger, T. L. S., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141, 1529–1544. doi: 10.1093/brain/awy050

Peng, D., Shi, Z., Xu, J., Shen, L., Xiao, S., Zhang, N., et al. (2016). Demographic and clinical characteristics related to cognitive decline in Alzheimer's disease in China: a multicenter survey from 2011 to 2014. *Medicine* 95:26. doi: 10.1097/MD.0000000000003727

Petrella, J. R., Hao, W., Rao, A., and Doraiswamy, P. M. (2019). Computational causal modeling of the dynamic biomarker cascade in Alzheimer's disease. 2019:6216530 *Comput. Math. Methods Med.* doi: 10.1155/2019/6216530

Pourhoseingholi, M. A., Baghestani, A. R., and Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterol. Hepatol. Bed Bench* 5, 79–83.

Prabhakaran, G., and Bakshi, R. (2018). Analysis of structure and cost in an American longitudinal study of Alzheimer's disease. *J. Alzheimers Dis. Parkinsonism* 8:411. doi: 10.4172/2161-0460.1000411

Prinja, S., Gupta, N., and Verma, R. (2010). Censoring in clinical trials: review of survival analysis techniques. *Indian J. Community Med.* 35, 217–221. doi: 10.4103/0970-0218.66859

Quartagno, M., and Carpenter, J. (2016). *jomo: A Package for Multilevel Joint Modelling Multiple Imputation.* R package version 2.

Rao, A., Lee, Y., Gass, A., and Monsch, A. (2011). Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 4, 499–502. doi: 10.1109/IEMBS.2011.6091115

Reitz, C. (2016). Toward precision medicine in Alzheimer's disease. *Ann. Transl. Med.* 4:107. doi: 10.21037/atm.2016.03.05

Ricciarelli, R., and Fedele, E. (2017). The amyloid cascade hypothesis in Alzheimer's disease: it's time to change our mind. *Curr. Neuropharmacol.* 2017, 926–935. doi: 10.2174/1570159X15666170116143743

Rodriguez-Esteban, R. (2015). Biocuration with insufficient resources and fixed timelines. *Database* 2015:bav116. doi: 10.1093/database/bav116

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Samtani, M. N., Raghavan, N., Shi, Y., Novak, G., Farnum, M., and Lobanov, V. (2013). Disease progression model in subjects with mild cognitive impairment from the Alzheimer's disease neuroimaging initiative: CSF biomarkers predict population subtypes. *Br. J. Clin. Pharmacol.* 75, 146–161. doi: 10.1111/j.1365-2125.2012.04308.x

Schott, J. M., and Petersen, R. C. (2015). New criteria for Alzheimer's disease: which, when and why? *Brain* 138(Pt 5), 1134–1137. doi: 10.1093/brain/awv055

Simon, C., Davidsen, K., Hansen, C., Seymour, E., Barnkob, M. B., and Olsen, L. R. (2018). BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* 19:57. doi: 10.1186/s12859-019-2607-x

Singh, M., Murthy, A., and Singh, S. (2015). Prioritization of free-text clinical documents: a novel use of a bayesian classifier. *JMIR Med. Inform.* 3:e17. doi: 10.2196/medinform.3793

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011a). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Sperling, R. A., Jack, C. R. Jr., and Aisen, P. S. (2011b). Testing the right target and right drug at the right stage. *Sci. Transl. Med.* 3:111cm33. doi: 10.1126/scitranslmed.3002609

Thanh, N.,Tran, T., Drab, K., and Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *J. Chemom. Intell. Lab. Syst.* 120, 92–96. doi: 10.1016/j.chemolab.2012.11.006

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc.* 63, 411–423. doi: 10.1111/1467-9868.00293

Tombaugh, T. N. (2005). Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Arch. Clin. Neuropsychol.* 20, 485–503. doi: 10.1016/j.acn.2004.11.004

Toschi, N., Lista, S., Baldacci, F., Cavedo, E., Zetterberg, H., Blennow, K., et al. (2019). Biomarker-guided clustering of Alzheimer's disease clinical syndromes. *Neurobiol. Aging* 83, 42–53. doi: 10.1016/j.neurobiolaging.2019.08.032

Vermunt, L., Veal, C. D., Ter Meulen, L., Chrysostomou, C., van der Flier, W., Frisoni, G. B., et al. (2018). European prevention of Alzheimer's dementia registry: recruitment and pre screening approach for a longitudinal cohort and prevention trials. *Alzheimers. Dement.* 14, 837–842. doi: 10.1016/j.jalz.2018.02.010

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., and Ye, J. (2013). "Multi-source learning with block-wise missing data for Alzheimer's disease prediction,". in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 185–193. doi: 10.1145/2487575.2487594

Younesi, E., and Hofmann-Apitius, M. (2013). From integrative disease modeling to predictive, preventive, personalized and participatory (P4) medicine. *EPMA J.* 4:23. doi: 10.1186/1878-5085-4-23

Young, A. L., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N. C., et al. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat. Commun.* 9:4273. doi: 10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Huang, J., Marinescu, R. V., Daga, P., Cash, D. M., et al. (2015). Multiple orderings of events in disease progression. *Inf. Process. Med. Imaging* 24, 711–722. doi: 10.1007/978-3-319-19992-4_56

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., and Ourselin, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *Neuroimage Clin.* 2, 735–745. doi: 10.1016/j.nicl.2013.05.004

Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 55, 856–867. doi: 10.1016/j.neuroimage.2011.01.008

Zhang, J., Zhou, W., Cassidy, R. M., Su, H., Su, Y., Zhang, X. (2018). Risk factors for amyloid positivity in older people reporting significant memory concern. *Comprehensive Psychiatry* 80, 126–131. doi: 10.1016/j.comppsych.2017.09.015

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7.2 Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease

Reprinted with permission from "Golriz Khatami, S. *et al.*. "Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease", *Alz Res Therapy* , 14(1), 55, (2022).

Golriz Khatami *et al.*
Alzheimer's Research & Therapy (2022) 14:55
https://doi.org/10.1186/s13195-022-01001-y

Alzheimer's
Research & Therapy

RESEARCH

Open Access

Comparison and aggregation of event sequences across ten cohorts to describe the consensus biomarker evolution in Alzheimer's disease

Sepehr Golriz Khatami[1,2*], Yasamin Salimi[1,2], Martin Hofmann-Apitius[1,2], Neil P. Oxtoby[3†], Colin Birkenbihl[1,2†], for the Alzheimer's Disease Neuroimaging Initiative, the Japanese Alzheimer's Disease Neuroimaging Initiative and the Alzheimer's Disease Repository Without Borders Investigators

Abstract

Background: Previous models of Alzheimer's disease (AD) progression were primarily hypothetical or based on data originating from single cohort studies. However, cohort datasets are subject to specific inclusion and exclusion criteria that influence the signals observed in their collected data. Furthermore, each study measures only a subset of AD-relevant variables. To gain a comprehensive understanding of AD progression, the heterogeneity and robustness of

*Correspondence: sepehr.golriz.khatami@scai.fraunhofer.de
†Neil P. Oxtoby and Colin Birkenbihl contributed equally to this work.
[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53757 Sankt Augustin, Germany
Full list of author information is available at the end of the article

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 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.

estimated progression patterns must be understood, and complementary information contained in cohort datasets be leveraged.

Methods: We compared ten event-based models that we fit to ten independent AD cohort datasets. Additionally, we designed and applied a novel rank aggregation algorithm that combines partially overlapping, individual event sequences into a meta-sequence containing the complementary information from each cohort.

Results: We observed overall consistency across the ten event-based model sequences (average pairwise Kendall's tau correlation coefficient of 0.69 ± 0.28), despite variance in the positioning of mainly imaging variables. The changes described in the aggregated meta-sequence are broadly consistent with the current understanding of AD progression, starting with cerebrospinal fluid amyloid beta, followed by tauopathy, memory impairment, FDG-PET, and ultimately brain deterioration and impairment of visual memory.

Conclusion: Overall, the event-based models demonstrated similar and robust disease cascades across independent AD cohorts. Aggregation of data-driven results can combine complementary strengths and information of patient-level datasets. Accordingly, the derived meta-sequence draws a more complete picture of AD pathology compared to models relying on single cohorts.

Keywords: Alzheimer's disease, Event-based models, Biomarker ordering, Disease progression, External validation, Meta-sequence

Background

Alzheimer's disease, in combination with its clinical manifestation/syndrome (AD) [1], is a progressive, multifaceted disease whose cognitive symptoms surface years after disease onset [2]. In order to identify crucial opportunities for medical interventions that could potentially prevent or delay symptoms, it is vital to understand the temporal relationship of pathological changes underlying the progressive nature of AD. To this end, cognitive assessments and a wide range of biomarkers, including cerebrospinal fluid (CSF) markers and neuroimaging-derived measures, have been established to monitor the disease's progression. Measuring these markers enables the observation of biochemical, structural, functional, and cognitive changes that occur as the disease progresses [3] and the resulting data can build the basis for data-driven approaches that aim to determine the relative temporal dependencies between biomarkers and cognitive symptoms [4]. Previously, a variety of data-driven models have been developed with the aim of accomplishing this task [5–10].

One model archetype that has found wide success in the context of neurodegenerative diseases [11–14] and AD specifically [15] is the event-based model (EBM) [13]. It is a data-driven probabilistic generative model that characterizes the progression of a disease in the form of a single sequence of events which describes the relative order of measured markers turning from a normal state to a diseased state (i.e., abnormal state). Such event sequences carry the benefit that they are highly interpretable and, although describing disease progression, can already be learned from cross-sectional cohort study data. Previously, EBMs have been used to derive event sequences [13], stage subjects in their disease progression [15], predict conversion from one clinical stage to the other (i.e., cognitively unimpaired (CU) to mild cognitive impairment (MCI), or MCI to AD) [16], and uncover disease phenotypes with distinct temporal progression patterns.

To build an EBM, patient-level data are needed on which the model can be fit. In recent decades, an increasing number of observational cohort studies have released their collected data for research purposes, including the Alzheimer's Disease Neuroimaging Initiative (ADNI) [17], the European Prevention of Alzheimer's Dementia (EPAD) [18], and AddNeuroMed [4]. So far, however, only a few studies in the AD domain have applied EBMs to data from other cohorts besides ADNI [19, 20]. Previous work evaluating data-driven progression modeling based on cohort datasets has shown that the participant recruitment procedures can introduce cohort-specific systematic statistical biases into the collected data [21], which, in turn, can bias the estimation of disease progression [22]. Therefore, it is necessary to replicate and validate data-driven results in independent cohorts to ensure robust conclusions. Consequently, it remains unclear whether event sequences determined from one cohort dataset would generalize beyond the discovery cohort itself and, further, if sequences generated across several cohorts were concordant among each other. Simultaneously, gaining a comprehensive event sequence combining all relevant AD biomarkers, cognitive assessments, and functional scores is infeasible, since cohort studies can only measure a limited set of variables that are often only partially overlapping between them [23]. In theory, however, this allows for an estimation of individual event sequences from distinct cohorts which cover complementary sets of markers. Aggregating results across cohorts would harness this complementary information by assembling a meta-sequence that provides a more complete picture of the development and progression of AD.

In this work, we present a systematic, in-depth comparison of AD event sequences derived from ten independent landmark cohort studies to investigate the generalizability and robustness of EBM-derived AD progression patterns. Furthermore, we designed a novel rank aggregation algorithm which we used to aggregate the event sequences into a single meta-sequence, thereby fusing the complementary information in all variables assessed across the studies. Our work harnesses the heterogeneity in cohort study designs and measurements to produce a meta-sequence providing a more complete, and robust, picture of the temporal order of pathological marker changes in AD progression.

Methods

Investigated cohort datasets

We selected ten independent AD cohort studies for our analysis by systematically exploring suitable datasets using the ADataViewer [23]. The prerequisite for including a cohort into our analysis was that (1) diagnostic staging into CU, MCI, and AD was performed [24]; (2) cross-sectional data was available for at least 10 patients per diagnostic group; and (3) multiple data modalities were collected. The cohorts that were ultimately selected are presented in Table 1. All cohorts followed the NINCDS-ADRDA diagnostic criteria [24].

Variable selection

We aimed at including a wide spectrum of variables to uncover the temporal relationship across multimodal markers of AD pathology that capture, for example, different biochemical, cognitive, or structural changes. In order to include a specific variable, it must have been measured in at least the CU and AD groups of the respective study to allow for later modeling. Furthermore, only a minimal amount of missing values was tolerable, as participants with missing values in any of the ultimately selected variables had to be excluded from the analysis. This led to a trade-off between the inclusion of an increasing number of variables and the amount of participants available for analysis. We present an example of variable inclusion and the effect on sample size in the supplementary material (Table S1). In total, 36 unique variables were selected from different data modalities covering neuropsychological and cognitive tests, CSF markers, and MRI-derived brain region volumes. The complete list of selected biomarkers and their corresponding modality are presented in Table 2. The number of variables per cohort is given in Table 1.

Participants

An available diagnosis of a participant as either CU, MCI, or AD was a prerequisite for inclusion. Furthermore, any participant with a diagnosis of cognitive impairment that was not linked to AD by the respective study's clinicians was excluded. Furthermore, only participants with complete data across all selected biomarkers could be used in our modeling approach. The number of participants per cohort and diagnostic group is described in Table 1.

Progression modeling via event-based models

The EBM derives a probabilistic sequence from patient-level data that describes the temporal order in which measured values of variables turn from a normal to an abnormal state. Each of these transitions is called an

Table 1 Selected cohorts, their number of participants per disease stage, and their number of considered variables

Cohort	Consortium	CU	MCI	AD	Total	Number of CSF, PET, and imaging biomarkers	Number of cognitive tests
ADNI [17]	The Alzheimer's Disease Neuroimaging Initiative	38	63	35	136	9	9
JADNI [25]	Japanese Alzheimer's Disease Neuroimaging Initiative	17	87	10	114	9	9
AIBL [26]	The Australian Imaging, Biomarker Lifestyle Flagship Study of Ageing	92	23	13	128	0	10
NACC [27]	The National Alzheimer's Coordinating Center	24	42	24	90	9	7
ANM [28]	AddNeuroMed	120	161	103	384	6	1
EMIF-1000 [29]	European Medical Information Framework	47	229	53	329	4	5
EDSD [30]	European DTI Study on Dementia	26	34	32	92	5	7
ARWIBO [31]	Alzheimer's Disease Repository Without Borders	214	115	38	367	7	3
OASIS-1 [32], OASIS-2 [33]	Open Access Series of Imaging Studies	135	70	30	235	6	1
WMHAD [34]	White Matter Hyperintensities in Alzheimer's Disease	19	27	42	88	6	7

Table 2 The selected biomarkers and their corresponding abbreviations

Modality	Biomarker	Abbreviation	Number of cohorts containing variable
Clinical assessments	Neuropsychiatric Inventory	NPI	2
	Logical Memory - Delayed Recall Total Number of Story Units Recalled	LDEL	5
	Alzheimer's Disease Assessment Scale (13-items)	ADAS13	2
	Alzheimer's Disease Assessment Scale (11-items)	ADAS11	2
	Logical Memory - Immediate Recall Total Number of Story Units Recalled	LIMM	6
	Trail Making Test-B	TRABS	2
	Digit-Symbol Coding Test	DIGITS	2
	California Verbal Learning Test Delayed Raw Score	LIDE	1
	Category Fluency (animals - fruits/vegetables)	CATFLU	3
	Figure Copy	FIGC	3
	California Verbal Learning Test Recall Raw Score	LIRE	2
	Figure recall	FIGR	2
	C/D Stroop Test Raw	STROOP	1
	Short Term Memory	STM	1
	Language	LANGU	1
	Perceptual Orientation	ORIENT	2
	Mental Manipulation	MENMA	1
	Attention	ATTEN	1
	Clock Drawing Test Total Score	CLKS	2
	Executive Memory	EXECUTIVE	1
	Word List Learning Trial	LICOR	1
	Boston Naming Test Score	BNTS	2
	Digit Symbol Substitution Test	WAIS	2
CSF markers	Amyloid-β	ABETA	4
	Total tau	TAU	4
	Phosphorylated tau (p-Tau)	PTAU	4
Imaging markers	Entorhinal volume	ENTOR	8
	Hippocampal volume	HIPPO	8
	Fusiform volume	FUSIF	8
	Ventricles volume	VENT	8
	Middle temporal volume	MIDTEPM	8
	Accumulated CSF in the brain	CSFVOL	5
	Fluorodeoxyglucose positron emission tomography (FDG PET)	FDG	2

event. In this context, normality or abnormality are defined non-parametrically using kernel density estimation mixture modeling on the empirical values of the modeled cohort's CU and AD populations, respectively [35]. This probabilistic allocation of measurements into two groups allows study participants (in particular, patients) to have a mix of occurred and non-occurred events across all measurements which lays the foundation to estimate the most likely event sequence. Here, the EBM assumes that the biomarkers monotonically change towards abnormality as the disease progresses and that this process is irreversible. Furthermore, there are no a priori assumptions regarding predefined disease stages, cut points determining the abnormality of biomarkers, or the temporal relationship between them. The most likely sequence of events S is then estimated by maximizing the likelihood $(X|S)$ (Eq. 1), where variable measurements are denoted by $x \in X$ for $i \in M$ markers and $j \in N$ indicates the individual samples.

$$\Pr(X|S) = \prod_{j=1}^{N} \left[\sum_{m=0}^{M} \left\{ \prod_{i=1}^{m} \Pr\left(x_{ij}|E_i\right) \prod_{i=m+1}^{M} \Pr\left(x_{ij}|\neg E_i\right) \right\} \right] \tag{1}$$

Here, $Pr(x_{ij}|\, E_i)$ and $Pr(x_{ij}|\, \neg\, E_i)$ describe the probability of observing the value of x given that the event E_i (i.e., variable x turning abnormal) has, or has not, occurred, respectively. For more details, we refer to the Supplementary Material and the original publication of the KDE EBM by Firth et al. [35]. The derived mixture models per cohort and measurement are presented in Fig. S3.

To quantify the similarity of distinct event sequences, we calculated the pairwise Kendall's tau rank correlation coefficient (KTC) across sequences and the Bhattacharrya coefficient (*BC*) for specific events as explained in Oxtoby et al. [12]. The KTCs were calculated pairwise across all cohorts while considering only the relative ranks of variables which were common among the respective two cohorts' sequences. An average KTC that is close to 1 and shows low standard deviation across the cohorts would indicate high concordance. An average BC close to 1 implies high similarity in the positional variance of ranks while the BC amounts to 0 for completely different patterns.

Generating a meta-sequence based on event sequences derived from multiple cohort studies

To generate a meta-sequence, we propose a method that combines individual event sequences (called "base sequences") stemming from independent datasets. We assemble a meta-sequence in a two-step procedure: first, building on the ideas presented in [36] and [37], we generate all possible sequences comprising k variables that are randomly drawn from the union of variables encountered in the base sequences (with $k <$ total number of variables). The generated sequence with the minimum average distance to all base sequences is selected as a starting sequence for the next step. In step 2, this starting sequence is extended by iteratively adding the remaining variables to it (i.e., those not in the k variables of the starting sequence), such that the average distance between the altered sequence and all base sequences remains minimal. Here, the new variable is not necessarily added to the end of the sequence but all possible positions are considered. This process is repeated until all variables have been included into the sequence which finally forms the aggregated meta-sequence. Therefore, the algorithm is deterministic once the base sequences are calculated. Splitting the algorithm in two steps (an exhaustive search for the first k variables followed by the greedy insertions) was necessary, as the search space (i.e., all possible meta-sequences) grows exponentially with the number of variables in the base sequences. Further explanations about the algorithm, the handling of partially overlapping lists, and access to the corresponding python code are provided in the Supplementary Material and Fig. S1.

We designed and applied two algorithms for generating a meta-sequence: one based on the maximum likelihood (ML) sequences presented by EBMs and one relying on bootstrapping. In the former, only the ML base sequences of each cohort were used as an input to our algorithm. Therefore, however, solely the rank of each event is considered while its positional variance within a sequence is not taken into account.

During the bootstrapping approach, all base sequences are resampled b-times with replacement. This means that a new base sequence is generated per cohort based on a sample of that cohort's participants that was randomly drawn with replacement and is of equal size to the original cohort. For each of these b sets of base sequences, one meta-sequence is generated. The resulting consensus over the b meta-sequences is visualized using a positional variance diagram which displays the variation in event ranks exhibited across the generated meta-sequences.

For this work, we generated a meta-sequence considering only variables which were present in at least three cohorts (Table 2) and set k equal to eight. In our bootstrapped version, we drew 500 bootstrap samples. The distance metric chosen was Spearman's footrule distance which takes the absolute difference in positions of variables into account.

Patient staging according to the determined meta-sequence

Once a meta-sequence was determined, one possible way to evaluate its plausibility across cohorts was to evaluate the assignment of subjects of the respective cohorts to the disease stages defined by the meta-sequence. In this process, each participant of a study was assigned to a disease stage which represents the current step in the meta-sequence at which the participant most likely resides. Therefore, stage 0 refers to the absence of any abnormal markers, while the farthest progressed stage m (with m being equal to the length of the sequence) implies that all events occurred for that particular subject. The corresponding equation underlying the patient staging is provided in the Supplemental Material.

Here, we staged only participants from cohorts that contained measurements of all investigated modalities (i.e., ADNI, JADNI, EMIF, and NACC) and were bound to consider only those variables of the meta-sequence that were found in the respectively staged cohort.

Results

Comparing event sequences derived from multiple cohort studies

We observed broad consistency with respect to the position of events across all cohorts' sequences which resulted in an average KTC of 0.69 ± 0.28 (pairwise KTCs are presented in Table S4; sequence similarity is also indicated visually through an approximately diagonal line of the event ranks from top-left to bottom-right in Fig. 1).

In most cohorts' sequences, CSF markers ranked highly, before cognitive impairments, which were again followed by MRI-derived brain volumes in the lower ranks.

The relative order among clinical assessments measuring different cognitive domains (e.g., memory, language, visuospatial, executive) was consistent across most cohorts (see Table S2 for a mapping of tests to cognitive domains). The cognitive impairment in all investigated cohorts started with memory dysfunction detected by logical memory tests (e.g., LDEL and LIMM), proceeded with language impairments exposed by tests such as the BNT and CATFLU. Thereafter, in most cohorts, visual dysfunction identified through the CLKS or FIGC followed, and finally, executive dysfunction recognized by, for example, the DIGIT and WAIS, occurred.

Among the cohorts where CSF biomarkers had been measured (ADNI, JADNI, EMIF, NACC), the relative positions of these biomarkers, in particular of tau (TAU) and phosphorylated tau (PTAU), varied. ABETA consistently placed first in all of these cohorts' sequences, and TAU and PTAU were mainly found in early positions as well (ADNI, JADNI, and EMIF), with the exception of NACC where they placed in the middle of the sequence. However, in all cases except JADNI, PTAU and TAU were direct neighbors, indicating the consistent, direct link between them.

The relative order of the MRI-derived brain volume events was consistent across cohorts, albeit with some variance (average KTC of 0.64 ± 0.29 for MRI variables only). While the volume changes in ADNI, JADNI, ARWIBO, and WMHAD started with ventricular expansion and were then followed by atrophy of the temporal lobe (here, hippocampus, entorhinal, middle temporal, and fusiform gyrus), in other cohorts (ANM, OASIS, NACC, EDSD), atrophy of the temporal lobe regions was the first detected variables of the MRI modality. The position that was taken by each respective brain region varied again among the cohorts. However, in many cases, the probabilistic nature of the EBMs indicated that the order of MRI events could be interchangeable among themselves (average BC of 0.17 ± 0.13 for MRI variables only) and events occurred most probably in close temporal proximity or even simultaneously (Fig. S2), as far as the model could discern from the data.

The position of FDG-PET, another well-established imaging biomarker measuring brain hypometabolism, was consistent in both cohorts it was measured in (ADNI, JADNI). It preceded the MRI marker changes and occurred concurrently with clinical symptoms, being placed after logical memory tests such as the LIMM and LDEL. However, its positioning of FDG-PET related to assessments of executive function differed between the two cohorts.

A multimodal meta-sequence of AD progression

To aggregate and investigate the complementary information from the base sequences in each cohort, we combined them into a single meta-sequence. Here, the position of a variable was determined based on its relative positions in all cohort sequences. Both versions of our algorithm (i.e., ML sequence-based and bootstrapping) were applied.

In the meta-sequence generated based on each cohort's ML sequence (Fig. 2), ABETA was ranked first, followed by PTAU and TAU. The latter were again closely linked and seemingly interchangeable given their ambiguous positioning across the base sequences. In positions four and five, LDEL and LIMM followed respectively, two clinical assessments measuring memory impairment. Next, the volume of CSF in the brain was positioned in the meta-sequence. The later event ranks were covered by MRI markers of brain volume, starting with the temporal lobe (e.g., hippocampus and entorhinal cortex) and ending with the ventricles. The previously described ambiguity in the order of MRI regions is not reflected in the ML-based meta-sequence because the algorithm considers only the ranks, and not the uncertainty estimated by the individual EBMs. However, it seems sensible to consider MRI events as fairly interchangeable in the meta-model. FIGC, an assessment of visual function, positioned before FUSIF and MIDTEMP near the end of the sequence, yet its position with respect to those two variables remained rather indefinite across the base sequences in which it was assessed (ARWIBO, AIBL, EDSD).

The consensus meta-sequence generated using the bootstrapping approach resembled the ML meta-sequence closely (KTC between both meta-sequences: 0.79; Fig. 3). Again, CSF markers placed first in the meta-sequence, were followed by cognitive assessments, and MRI events

(See figure on next page.)

Fig. 1 Individual event sequences estimated from the ten investigated cohorts. To facilitate the comparison of relative event positions, the *y*-axes follow the ADNI sequence. Common events between ADNI and the other cohorts are presented above a dashed green line. The closer the sequences are to the ADNI sequence, the more diagonal the probabilistic position (colored squares) will align from top-left to bottom-right. Lateral shifts due to additional events which were not available in ADNI have to be disregarded (as for example observed in WMHAD and EDSD). Event order 1 corresponds to the first position in the sequence. The shading of squares indicates the positional probability with darker shades corresponding to higher probabilities. The relative sizes of the squares do not encode any information. The event sequences in their original form are presented in Fig. S2

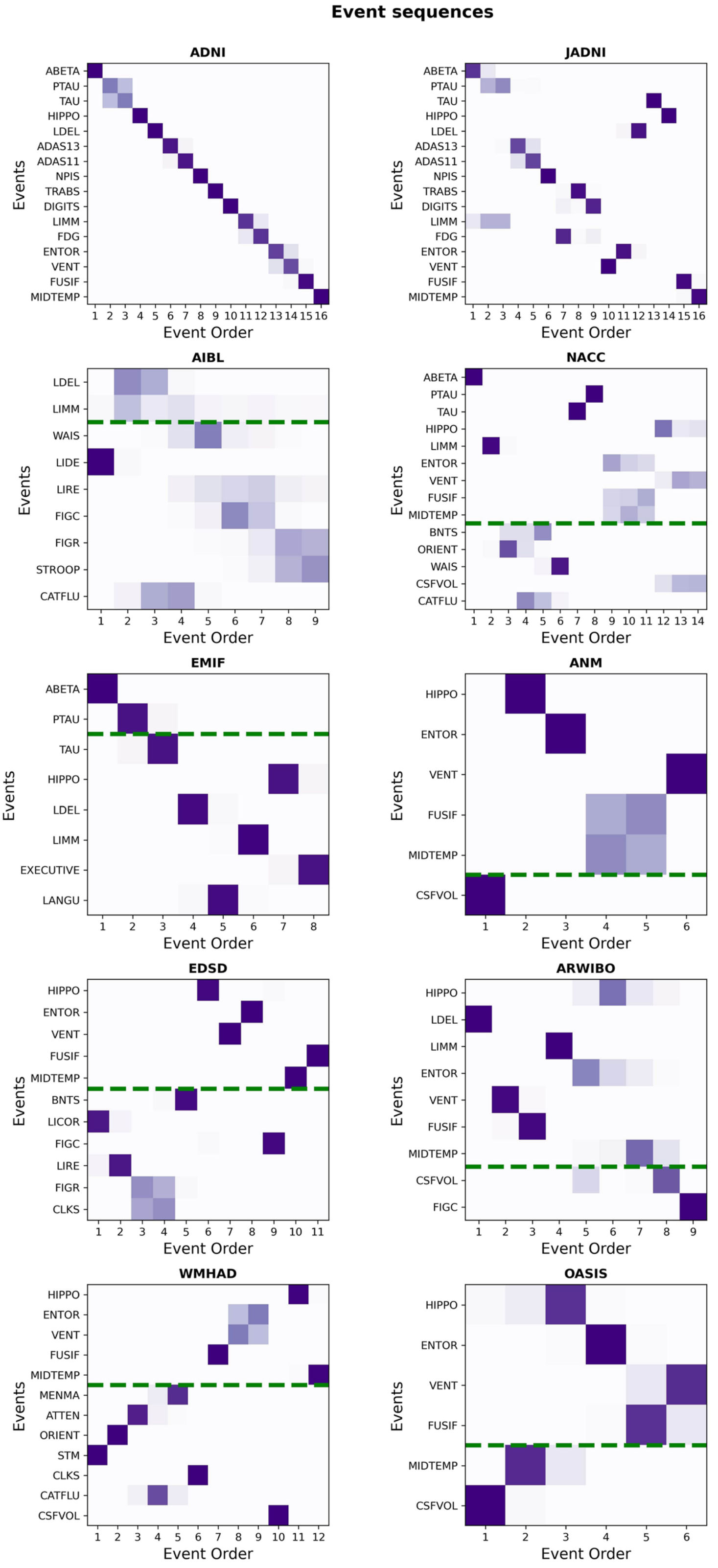


Fig. 1 (See legend on previous page.)

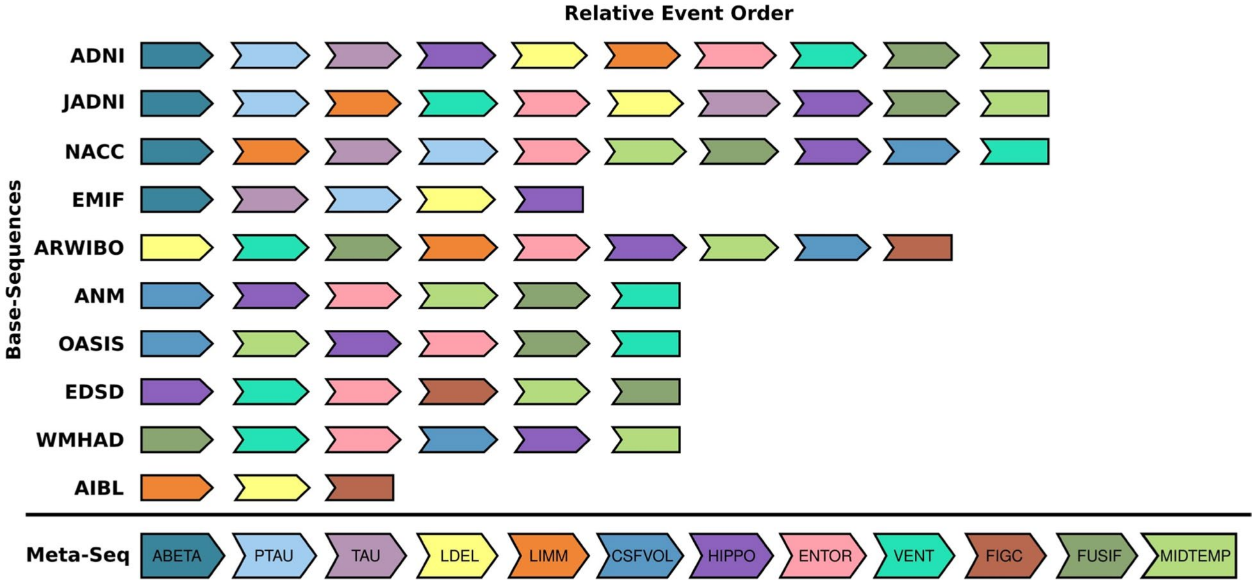


Fig. 2 All ML base sequences from the ten investigated cohorts and the resulting meta-sequence. Due to only partially overlapping lists, the determining factor for an event's position in the meta-sequence was not its absolute position in each base sequence (i.e., rank 1, 2, …, 11), but its relative position to other biomarkers in the same sequence (e.g., ABETA commonly places before MMSE when they were assessed together; thus, it appears before MMSE in the meta-sequence)

started with the temporal lobe and further progressed with the ventricles. The main difference to the ML-based meta-sequence, as well as the major region of model uncertainty, was again found among the MRI variables. This further underlined the impression that the MRI events were fairly interchangeable and probably occurred in close temporal proximity. The highest ambiguity was in the positioning of FIGC which showed a slight tendency towards the last ranks. The average KTC across all bootstrapped meta-sequences was 0.5 ± 0.20, with the highest discordance found among the MRI modality.

Staging the patients of cohorts with available CSF, MRI, and cognitive scores (i.e., ADNI, JADI, NACC, EMIF) revealed a consistent pattern across them (Fig. 4). For all cohorts, the vast majority of CU subjects were assigned to the first stage which corresponds to no event occurrences. As expected, MCI patients were largely staged between CU subjects and AD patients with some overlap in both directions. This suggests that these subjects experienced CSF marker abnormalities and some cognitive symptoms. Finally, the majority of AD patients were assigned to the last stages, indicating their abnormality along CSF markers, cognitive performance, and brain region atrophy.

Discussion

In this work, we used EBMs to investigate AD progression across ten independent cohort studies by evaluating the concurrence of their individually derived event sequences. Furthermore, we proposed an algorithm to combine event sequences estimated from partially overlapping, and thus complementary, sets of variables into a single meta-sequence describing AD progression more comprehensively. Finally, we applied said algorithm on the ten event sequences to estimate a meta-sequence comprising 13 AD variables spanning CSF biomarkers, MRI measures, and clinical assessments of cognitive and functional performance.

Consistent trends across cohorts' event sequences

The derived event sequences proved to be broadly consistent across cohorts, with the most notable variability in the ordering of MRI brain volume events. This could be caused by (1) distinct statistical biases of the cohorts for example introduced through specific recruitment criteria [21], (2) distinct prevalence of AD disease progression subtypes that follow different disease mechanisms [38–40], or (3) mixed neuropathologies.

Inclusion and exclusion criteria of a study shape the demographic compositions of its cohort and thus can directly affect the data-driven disease progression patterns (Table S3). For instance, ADNI held a higher proportion of APOE4 carriers compared to JADNI. Given that it has been repeatedly reported that early TAU depositioning is more prominent in APOE4 carriers [41–43], this difference might explain the earlier positioning of TAU in ADNI's sequence opposed to its relatively lower rank in JADNI's.

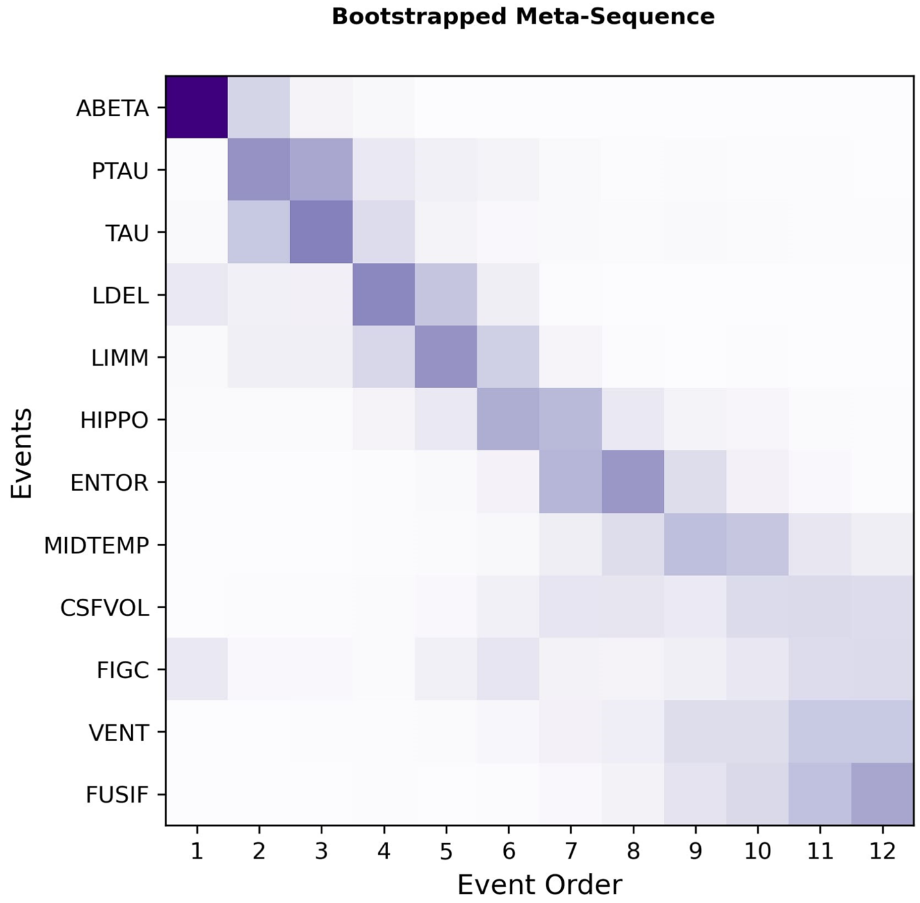


Fig. 3 Bootstrapped meta-sequence generated from 500 samples of the base sequences of the 10 cohorts. Event order 1 corresponds to the first position in the sequence. The shading of squares indicates the positional probability with darker shades corresponding to higher probabilities

Previously, for example, two empirically determined AD progression subtypes called "hippocampal-sparing" and "limbic-predominant" were described and associated with distinct patterns of brain atrophy [38, 44]. While structural changes in the brain start with atrophy in the medial temporal lobe (e.g., entorhinal and hippocampus) for the limbic-predominant subtype, the brain deterioration in the hippocampal-sparing subtype begins with atrophy of the frontal cortex and with the enlargement of ventricles [44]. Given their respective event sequences, this could indicate that OASIS, ADNI, and NACC might have included more patients expressing the limbic-predominant subtype, while the hippocampal-sparing subtype was more dominant among patients from ARWIBO and JADNI.

We observed that CSF biomarkers placed first in all cohorts which measured them. This finding is in concordance with previous biomarker studies that observed the occurrence of both ABETA accumulation and brain atrophy before global cognitive decline [45–48].

Autopsies of AD patients have shown that AD pathology hardly appears in isolation and that patients often suffer from a mixture of brain pathologies [49]. While most studies aim to exclude patients affected by other cognitive diseases, an AD clinical diagnosis is still mainly symptom driven and misclassification errors are possible.

Meta-sequence combines heterogeneous event sequences from multiple cohorts

A particular strength of our meta-sequence algorithm is that it works agnostic towards the differences in variable value representations exhibited across cohorts. A direct comparison of the provided data values often remains challenging without introducing statistical biases since studies differ, for example, in their data collection procedures, employed imaging machinery, and used assays. Using our approach, such semantically equivalent but statistically heterogeneous information can be combined as all computations are performed solely on the base sequences and thus

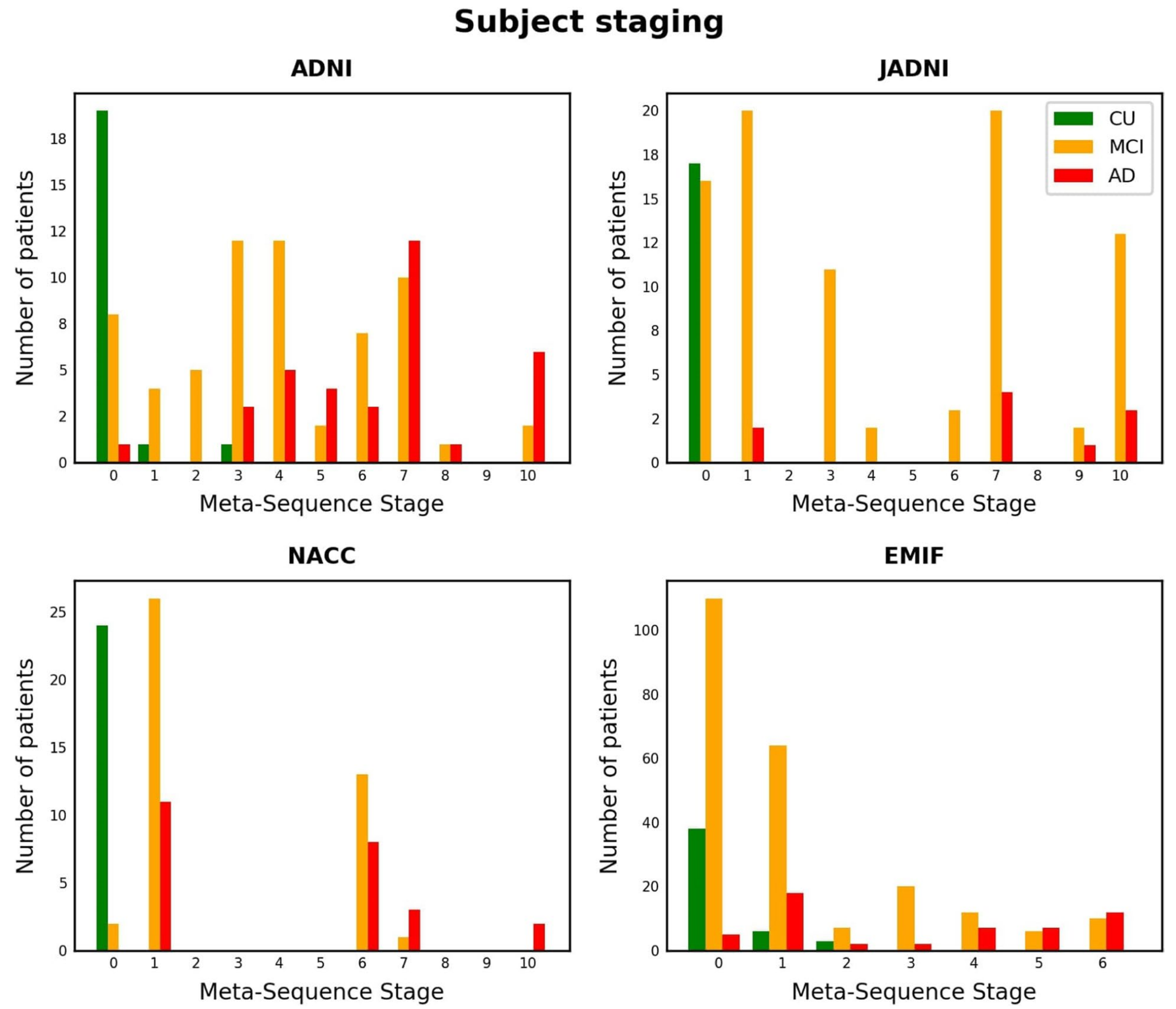


Fig. 4 Number of subjects from each diagnostic group per meta-sequence stage. Each step along the *x*-axis corresponds to the occurrence of a new biomarker abnormality event. Stage 0 corresponds to no event occurrence while the last stage implies abnormality of all variables. Events are ordered according to the bootstrapped meta-sequence, always considering only variables in common between the measurements available in the respective cohort and the meta-sequence

potential across-cohort-biases due to value representations are avoided.

The biggest advantage of the bootstrapping approach compared to ML sequence-based one is that it allows for uncertainty quantification. However, bootstrapped EBM sequences tend to display a substantially higher positional variance (i.e., "fuzziness") than ML derived ones (for an example, see Firth et al. Figures 1 and 2 [35]). Comparing our ML-based meta-sequence to the bootstrapping-based meta-sequence revealed high similarity between them. Observed differences seemed to be within variational limits expressed in the bootstrapped meta-sequence and mainly affected MRI variables.

Generated meta-sequence resembles AD pathology

One possibility to validate the derived meta-sequence was to evaluate its concordance with previous findings describing the temporal relationship between smaller subgroups of variables.

The ordering of CSF biomarkers discovered in previous EBM studies supported our observations in the meta-sequence (ABETA followed by PTAU and TAU) [15]. Our findings were also in line with a recent study [50] which demonstrated that TAU and PTAU become abnormal after ABETA and that their abnormality occurred in close temporal relationship with cognitive decline. The latter was also in concordance with our findings; however, the cognitive assessments we investigated (i.e., LDEL and LIMM) were not directly included in the referenced study. Furthermore, there is a well-established association between cognitive decline and ABETA abnormality and abundant evidence that changes in cognition typically occur after abnormalities related to CSF biomarkers [45, 50, 51].

Our observation that memory function showed abnormality before brain volumes agrees with previous studies which suggested that individual-level brain atrophy rates (not assessed in our study) precede cognitive events; however, MRI-derived brain volumes become abnormal afterwards [15].

In our meta-sequences, changes in MRI biomarkers were ranked after cognitive decline. In agreement with this, for example, Hadjichrysanthou et al. reported that changes in MRI markers appear in close succession with memory decline [52]. Also, the positioning of MRI variables with respect to CSF markers was concordant with previous observations where significant correlations between CSF biomarkers and temporal lobe atrophy were found [53–55]. These studies argue that increases of TAU and PTAU are attributable to the deposition of neurofibrillary tangles in the temporal lobe, including the hippocampus and entorhinal cortex, which we found to be the first brain region volumes turning abnormal. Furthermore, elevated CSF biomarkers predicted future brain atrophy in these regions (i.e., CSF biomarkers became abnormal before brain volumes).

In concordance with the relative positioning of MRI biomarkers in the meta-sequence, various studies have shown that volumetric changes start with the temporal lobe areas, including the hippocampus which preceded the abnormality of the entorhinal cortex, fusiform, and middle temporal, and further proceed to other brain regions such as the ventricles [56–59].

Finally, in agreement with a previous study [60–63] in which visual memory dysfunction was identified as one of the last stages in AD progression, the FIGC test was ranked among the end of the sequences. The fact that it was positioned after the enlargement of ventricles is in agreement with experimental evidence that changes in the ventricles may precede a deficit in visual memory function [64, 65]. Another EBM study [35] also suggested that visual processing becomes impaired after episodic memory in typical AD.

The conducted patient staging provided further evidence that the generated meta-sequence described a sensible cascade of AD progression: participants from the three diagnostic groups were distributed according to their disease severity with CU subjects being staged first, MCI patients spreading around the intermediate stages, and AD cases occupying the later stages of the sequence. Observing MCI subjects at stage 0 could be explained by CSF biomarker values and cognitive scores that were close to the probabilistic event threshold but did not yet exceed it and, consequently, the model considered them to be normal. The few AD cases that were staged early in the sequence were amyloid-negative subjects which potentially indicated their misclassification.

Limitations

To build a robust meta-sequence, each variable had to be present in at least some of the base sequences to allow for meaningful distance calculations. Furthermore, the high amounts of missing data occurring when multiple data modalities are combined led to a substantial decrease of the number of available participants per study. This could have led to more noise in the EBM's reference distributions. Additionally, modeling signals from heterogeneous data sources, such as AD cohort data, as some form of average bears the potential risk that the resulting average will resemble a rather artificial construct that cannot be observed in its specific form in the real world. However, the similarity among the base sequences as well as between base sequences and the final meta-sequence was quite high and our identified meta-sequences were highly concordant with results from both data-driven and experimental studies. Furthermore, the patient staging along the meta-sequence displayed a sensible distribution of CU, MCI, and AD subjects along the disease stages. Consequently, it is improbable that the presented meta-sequence represents such an artificial average. Finally, we want to highlight again that AD was considered primarily from a clinical perspective in all of our investigated cohort studies. As such, there is a chance that misdiagnosed patients were present in the cohorts and therefore included in this analysis as well.

Conclusion

In the light of the reproducibility crisis, it becomes especially important that we look beyond single data resources, validate achieved results across multiple cohort studies, and constantly develop and evaluate data-driven methods. To this end, we revealed general consistency across data-driven event sequences derived from ten independent cohorts using EBMs. Here, only relatively minor differences in the ranking of the core features that were available in all ten cohorts were observed. In addition, our novel algorithm estimated a meta-sequence that exploits the additional information available in other variables unique to each study and thus could assemble an event sequence that is highly multimodal and more comprehensive than sequences built from single datasets. This is important for ensuring the transferability of models and results across AD (sub)populations and for improving our understanding of disease progression.

Abbreviations

ADNI: The Alzheimer's Disease Neuroimaging Initiative; JADNI: Japanese Alzheimer's Disease Neuroimaging Initiative; AIBL: The Australian Imaging, Biomarker Lifestyle Flagship Study of Ageing; NACC: The National Alzheimer's Coordinating Center; ANM: AddNeuroMed; EMIF-1000: European Medical Information Framework; EDSD: European DTI Study on Dementia; ARWIBO: Alzheimer's Disease Repository Without Borders; OASIS: Open Access Series of Imaging Studies; WMHAD: White Matter Hyperintensities in Alzheimer's Disease; CDRSB: Clinical Dementia Rating Sum of Boxes; NPI: Neuropsychiatric Inventory; LDEL: Logical Memory - Delayed Recall Total Number of Story Units Recalled; ADAS13: Alzheimer's Disease Assessment Scale (13-items); ADAS11: Alzheimer's Disease Assessment Scale (11-items); MMSE: Mini-Mental State Examination; LIMM: Logical Memory - Immediate Recall Total Number of Story Units Recalled; TRABS: Trail Making Test-B; DIGITS: Digit-Symbol Coding Test; LIDE: California Verbal Learning Test Delayed Raw Score; CATFLU: Category Fluency (animals - fruits/vegetables); FIGC: Figure Copy; LIRE: California Verbal Learning Test Recall Raw Score; FIGR: Figure recall; STROOP: C/D Stroop Test Raw; STM: Short-Term Memory; LANGU: Language; ORIENT: Perceptual Orientation; MENMA: Mental Manipulation; ATTEN: Attention; CLKS: Clock Drawing Test Total Score; EXECUTIVE: Executive Memory; LICOR: Word List Learning Trial; BNTS: Boston Naming Test Score; WAIS: Digit Symbol Substitution Test; ABETA: Amyloid-β; TAU: Total Tau; PTAU: Phosphorylated Tau (p-Tau); ENTOR: Entorhinal volume; HIPPO: Hippocampal volume; FUSIF: Fusiform volume; VENT: Ventricles volume; MIDTEPM: Middle temporal volume; CSFVOL: Accumulated CSF in the brain; FDG: Fluorodeoxyglucose positron emission tomography (FDG PET); MRI: Magnetic resonance imaging; MCI: Mild cognitive impairment; AD: Alzheimer's disease; CU: Cognitive unimpaired; KTC: Kendall's tau rank correlations; EBM: Event-based model; CSF: Cerebrospinal fluid.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13195-022-01001-y.

Additional file 1.

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 AD research.

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 NeuroImaging at the University of Southern California.

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.

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.

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).

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 826421, "TheVirtualBrain-Cloud," and under grant agreement No. 666992 "EuroPOND". NPO is a UKRI Future Leaders Fellow (MR/S03546X/1). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

De-identified data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (https://adni.loni.usc.edu), the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL) database (https://aibl.csiro.au/), the European Collaboration for the Discovery of Novel Biomarkers for Alzheimer's Disease (AddNeuroMed) (https://www.synapse.org/#!Synapse:syn4988768), Alzheimer's Disease Repository Without Borders (ARWIBO) (https://www.neugrid2.eu/), Open Access Series of Imaging Studies (OASIS) (https://www.neugrid2.eu/), White Matter Hyperintensities in Alzheimer's Disease (WMH-AD) (https://www.neugrid2.eu/), European Diffusion Tensor Imaging Study in Dementia (EDSD) (https://www.neugrid2.eu/), National Alzheimer's Coordinating Center (NACC) (https://naccdata.org/), Japanese Alzheimer's Disease Neuroimaging Initiative (JADNI) (https://humandbs.biosciencedbc.jp/en/hum0043-v1), European Medical Information Framework for Alzheimer's Disease Multimodal Biomarker Discovery (EMIF-AD MBD) (https://emif-catalogue.eu; http://www.emif.eu/about/emif-ad). The authors had no special access privileges others would not have to the data obtained from these resources.

Declarations

Ethics approval and consent to participate

Participants of every cohort dataset that was used in this work gave informed written consent for data collection and sharing. For more details, we refer to the provided references of each cohort, respectively.

Consent for publication

The authors submitted the manuscript to all data owners who require manuscript approval prior to publication and acquired consent.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53757 Sankt Augustin, Germany. [2]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115 Bonn, Germany. [3]Centre for Medical Image Computing and Department of Computer Science, University College London, Gower St, London WC1E 6BT, UK.

Received: 12 November 2021 Accepted: 6 April 2022
Published online: 20 April 2022

References

1. Jack CR Jr, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018;14(4):535–62. https://doi.org/10.1016/j.jalz.2018.02.018.
2. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. Mol Neurodegen. 2019;14(1):32. https://doi.org/10.1186/s13024-019-0333-5.
3. Blennow K, Zetterberg H. Biomarkers for Alzheimer's disease: current status and prospects for the future. Intern Med. 2018;284(6):643–63. https://doi.org/10.1111/joim.12816.
4. 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;1180:36–46. https://doi.org/10.1111/j.1749-6632.2009.05064.x.
5. Lorenzi M, Filippone M, Frisoni GB, Alexander DC, Ourselin S. Alzheimer's Disease Neuroimaging Initiative. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. NeuroImage. 2019;190:56–68. https://doi.org/10.1016/j.neuroimage.2017.08.059.
6. Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, et al. A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort. NeuroImage. 2012;63(3):1478–86. https://doi.org/10.1016/j.neuroimage.2012.07.059.
7. Yang E, Farnum M, Lobanov V, Schultz T, Verbeeck R, Raghavan N, et al. Alzheimer's Disease Neuroimaging Initiative. Quantifying the pathophysiological timeline of Alzheimer's disease. J Alzheimers Dis. 2011;26(4):745–53. https://doi.org/10.3233/JAD-2011-110551.
8. Delor I, Charoin JE, Gieschke R, Retout S, Jacqmin P. Modeling Alzheimer's disease progression using disease onset time and disease trajectory concepts applied to CDR-SOB scores from ADNI. CPT Pharmacometrics Syst Pharmacol. 2013;2(10):e78. https://doi.org/10.1038/psp.2013.54.
9. Villemagne VL, Burnham S, Bourgeat P, Brown B, Ellis KA, Salvado O, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. Lancet Neurol. 2013;12(4):357–67. https://doi.org/10.1016/S1474-4422(13)70044-9.
10. Donohue MC, Jacqmin-Gadda H, Le Goff M, Thomas RG, Raman R, Gamst A, et al. Estimating long-term multivariate progression from short-term data. Alzheimers Dement. 2014;10(5 Suppl):S400–10. https://doi.org/10.1016/j.jalz.2013.10.003.
11. Dekker I, Schoonheim MM, Venkatraghavan V, Eijlers A, Brouwer I, Bron EE, et al. The sequence of structural, functional and cognitive changes in multiple sclerosis. NeuroImage Clin. 2021;29:102550. https://doi.org/10.1016/j.nicl.2020.102550.
12. Oxtoby NP, Leyland LA, Aksman LM, Thomas G, Bunting EL, Wijeratne P, et al. Sequence of clinical and neurodegeneration events in Parkinson's disease progression. Brain. 2021;144(3):975–88. https://doi.org/10.1093/brain/awaa461.
13. Fonteijn HM, Modat M, Clarkson MJ, Barnes J, Lehmann M, Hobbs NZ, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage. 2012;60(3):1880–9. https://doi.org/10.1016/j.neuroimage.2012.01.062.
14. Wijeratne PA, Young AL, Oxtoby NP, Marinescu RV, Firth NC, Johnson E, et al. An image-based model of brain volume biomarker changes in Huntington's disease. Ann Clin Transl Neurol. 2018;5(5):570–82. https://doi.org/10.1002/acn3.558.
15. Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, et al. A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain. 2014;137(Pt 9):2564–77. https://doi.org/10.1093/brain/awu176.
16. Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. Nat Commun. 2018;9(1):4273. https://doi.org/10.1038/s41467-018-05892-0.
17. 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.
18. Solomon A, Kivipelto M, Molinuevo JL, Tom B, Ritchie CW. European prevention of Alzheimer's dementia longitudinal cohort study (EPAD LCS): study protocol. Prev Alzheimers Dis. 2018;8(12):e021017.
19. Oxtoby NP, Young AL, Cash DM, Benzinger T, Fagan AM, Morris JC, et al. Data-driven models of dominantly-inherited Alzheimer's disease progression. Brain. 2018;141(5):1529–44. https://doi.org/10.1093/brain/awy050.
20. Archetti D, Ingala S, Venkatraghavan V, Wottschel V, Young AL, Bellio M, et al. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in Alzheimer's disease. NeuroImage. 2019;24:101954. https://doi.org/10.1016/j.nicl.2019.101954.
21. 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.
22. Birkenbihl C, Emon MA, Vrooman H, Westwood S, Lovestone S, 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. 2020;11(3):367–76. https://doi.org/10.1007/s13167-020-00216-z.
23. Salimi Y, Domingo-Fernandez D, Bobis-Alvarez C, Hofmann-Apitius M, Vasculature I, Birkenbihl C, et al. ADataViewer: exploring semantically harmonized Alzheimer's disease cohort datasets. medRxiv. 2021.
24. 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):939–44. https://doi.org/10.1212/wnl.34.7.939.
25. Iwatsubo T. Japanese Alzheimer's Disease Neuroimaging Initiative: present status and future. Alzheimer Dement. 2010;6(3):297–9. https://doi.org/10.1016/j.jalz.2010.03.011.
26. 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. https://doi.org/10.1017/S1041610209009405.
27. 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. https://doi.org/10.1097/WAD.0000000000000279.
28. 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. https://doi.org/10.3233/JAD-200948.
29. 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):1–9. https://doi.org/10.1186/s13195-018-0396-5.
30. 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;144:305–8. https://doi.org/10.1016/j.neuroimage.2016.03.067.
31. 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. https://doi.org/10.1016/j.jalz.2009.04.1235.
32. 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. https://doi.org/10.1162/jocn.2007.19.9.1498.
33. 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. https://doi.org/10.1162/jocn.2009.21407.
34. 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. https://doi.org/10.3233/JAD-180982.
35. Firth NC, Primativo S, Brotherhood E, Young AL, Yong K, Crutch SJ, 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. Alzheimers Demen. 2020;16(7):965–73. https://doi.org/10.1002/alz.12083.
36. DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R. Combining results of microarray experiments: a rank aggregation approach. Stat Appl Genet Mol Biol. 2006;5:Article15. https://doi.org/10.2202/1544-6115.1204.
37. Lin S, Ding J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. Biometrics. 2009;65(1):9–18. https://doi.org/10.1111/j.1541-0420.2008.01044.x.
38. Ferreira D, Nordberg A, Westman E. Biological subtypes of Alzheimer disease: a systematic review and meta-analysis. Neurology. 2020;94(10):436–48. https://doi.org/10.1212/WNL.0000000000009058.
39. Whitwell JL, Jack CR Jr, Przybelski SA, Parisi JE, Senjem ML, Boeve BF, et al. Temporoparietal atrophy: a marker of AD pathology independent of clinical diagnosis. Neurobiol Aging. 2011;32(9):1531–41. https://doi.org/10.1016/j.neurobiolaging.2009.10.012.
40. Piaceri I, Nacmias B, Sorbi S. Genetics of familial and sporadic Alzheimer's disease. Front Biosci. 2013;5(1):167–77. https://doi.org/10.2741/e605.
41. Lemprière S. APOE4 provokes tau aggregation via inhibition of noradrenaline transport. Nat Rev Neurol. 2021;17(6):328. https://doi.org/10.1038/s41582-021-00511-x.
42. Baek MS, Cho H, Lee HS, Lee JH, Ryu YH, Lyoo CH. Effect of APOE ε4 genotype on amyloid-β and tau accumulation in Alzheimer's disease. Alzheimer's Res Ther. 2020;12(1):1–12. https://doi.org/10.1186/s13195-020-00710-6.
43. Benson GS, Bauer C, Hausner L, Couturier S, Lewczuk P, Peters O, et al. Don't forget about tau: the effects of ApoE4 genotype on Alzheimer's disease cerebrospinal fluid biomarkers in subjects with mild cognitive impairment—data from the Dementia Competence Network. J Neural Transm. 2022:1–10. https://doi.org/10.1007/s00702-022-02461-0.
44. Ferreira D, Verhagen C, Hernández-Cabrera JA, Cavallin L, Guo CJ, Ekman U, et al. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. Sci Rep. 2017;7:46263. https://doi.org/10.1038/srep46263.
45. Iturria-Medina Y, Sotero RC, Toussaint PJ, Mateos-Pérez JM, Evans AC, et al. Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. Nat Commun. 2016;7:11934. https://doi.org/10.1038/ncomms11934.
46. Chen G, Shu H, Chen G, Ward BD, Antuono PG, Zhang Z, et al. Staging Alzheimer's disease risk by sequencing brain function and structure, cerebrospinal fluid, and cognition biomarkers. J Alzheimers Dis. 2016;54(3):983–93. https://doi.org/10.3233/JAD-160537.
47. Mormino EC, Kluth JT, Madison CM, Rabinovici GD, Baker SL, Miller B, et al. Episodic memory loss is related to hippocampal-mediated beta-amyloid deposition in elderly subjects. Brain. 2009;132(Pt 5):1310–23. https://doi.org/10.1093/brain/awn3.
48. Wang F, Gordon BA, Ryman DC, Ma S, Xiong C, Hassenstab J, et al. Cerebral amyloidosis associated with cognitive decline in autosomal dominant Alzheimer disease. Neurology. 2015;85(9):790–8. https://doi.org/10.1212/WNL.0000000000001903.
49. Schneider JA, Arvanitakis Z, Bang W, Bennett DA. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. Neurology. 2007;69(24):2197–204. https://doi.org/10.1212/01.wnl.0000271090.28148.24.
50. Luo J, Agboola F, Grant E, Masters CL, Albert MS, Johnson SC, et al. Sequence of Alzheimer disease biomarker changes in cognitively normal adults: a cross-sectional study. Neurology. 2020;95(23):e3104–16. https://doi.org/10.1212/WNL.0000000000010747.
51. Ellis KA, Lim YY, Harrington K, Ames D, Bush AI, Darby D, et al. Decline in cognitive function over 18 months in healthy older adults with high amyloid-β. J Alzheimers Dis. 2013;34(4):861–71. https://doi.org/10.3233/JAD-122170.
52. Hadjichrysanthou C, Evans S, Bajaj S, Siakallis LC, McRae-McKee K, de Wolf F, et al. The dynamics of biomarkers across the clinical spectrum of Alzheimer's disease. Alzheimer's Res Ther. 2020;12(1):1–16.
53. Armstrong NM, An Y, Shin JJ, Williams OA, Doshi J, Erus G, et al. Associations between cognitive and brain volume changes in cognitively normal older adults. Neuroimage. 2020;223:117289. https://doi.org/10.1016/j.neuroimage.2020.117289.
54. Herukka SK, Pennanen C, Soininen H, Pirttilä T. CSF Abeta42, tau and phosphorylated tau correlate with medial temporal lobe atrophy. J Alzheimers Dis. 2008;14(1):51–7. https://doi.org/10.3233/jad-2008-14105.
55. Granadillo E, Paholpak P, Mendez MF, Teng E. Visual ratings of medial temporal lobe atrophy correlate with CSF tau indices in clinical variants of early-onset Alzheimer disease. Dement Geriatr Cogn Disord. 2017;44(1-2):45–54. https://doi.org/10.1159/000477718.
56. Bouwman FH, Schoonenboom SN, van der Flier WM, van Elk EJ, Kok A, Barkhof F, et al. CSF biomarkers and medial temporal lobe atrophy predict dementia in mild cognitive impairment. Neurobiolaging. 2007;28(7):1070–4. https://doi.org/10.1016/j.neurobiolaging.2006.05.006.
57. Younes L, Albert M, Miller MI, BIOCARD Research Team. Inferring changepoint times of medial temporal lobe morphometric change in preclinical Alzheimer's disease. NeuroImage Clin. 2014;5:178–87. https://doi.org/10.1016/j.nicl.2014.04.009.
58. Coupé P, Manjón JV, Lanuza E, Catheline G. Lifespan changes of the human brain in Alzheimer's disease. Sci Rep. 2019;9(1):3998. https://doi.org/10.1038/s41598-019-39809-8.
59. Scahill RI, Schott JM, Stevens JM, Rossor MN, Fox NC. Mapping the evolution of regional atrophy in Alzheimer's disease: unbiased analysis of fluid-registered serial MRI. Proc Natl Acad Sci U S A. 2002;99(7):4703–7. https://doi.org/10.1073/pnas.052587399.
60. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82(4):239–59. https://doi.org/10.1007/BF00308809.
61. Storey E, Slavin MJ, Kinsella GJ. Patterns of cognitive impairment in Alzheimer's disease: assessment and differential diagnosis. Front Biosci. 2002;7:e155–84. https://doi.org/10.2741/A914.
62. Breteler MM, van Amerongen NM, van Swieten JC, Claus JJ, Grobbee DE, van Gijn J, et al. Cognitive correlates of ventricular enlargement and cerebral white matter lesions on magnetic resonance imaging. The Rotterdam Study. Stroke. 1994;25(6):1109–15. https://doi.org/10.1161/01.str.25.6.1109.
63. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage Clin. 2013;2:735–45.
64. Ferreira D, Pereira JB, Volpe G, Westman E. Subtypes of Alzheimer's disease display distinct network abnormalities extending beyond their pattern of brain atrophy. Front Neurol. 2019;10:524. https://doi.org/10.3389/fneur.2019.00524.
65. Birkenbihl C, Salimi Y, Domingo-Fernándéz D, Lovestone S, AddNeuroMed consortium, Fröhlich H, et al. Evaluating the Alzheimer's disease data landscape. Alzheimer's Dementia: Translat Res Clin Interv. 2020;6(1):e12102.

Publisher's Note

Supplementary File

Data preprocessing

While in some of the cohorts brain volumes were calculated as a sum of the two respective hemispheres, in others they were measured individually per hemisphere and thus we had to sum them up to make the measurements consistent across all cohorts. In addition, the brain region volumes of individual cohorts were normalized across the subjects based on their whole brain volume to correct the variations of head size among individuals by dividing the regional volumes through the intracranial volume.

Example of variable inclusion and relation to diminishing number of patients

Below we provide an example to illustrate the decrease in sample size when integrating multimodal AD cohort data and considering only complete cases (ie. no missing data in any variable). We also excluded MCI patients in this example, as only CU and AD were the crucial diagnostic groups for fitting our EBMs. The example is based on a potential inclusion of amyloid PET. **Table S1** provides an overview about the stepwise decrease in participants available for analysis.

Out of the 10 cohorts we analyzed, only ADNI, EMIF, AIBL, and NACC reported measures of amyloid PET, already reducing the number of potentially analyzable cohorts to 4 out of 10. NACC only reported binary values (0, 1) which could not be modeled using our approach. Focusing only on our selected cognitive variables at baseline, ADNI had complete measurements for 229 CU and 183 AD participants, EMIF for 140 CU, 114 AD, and AIBL for 92 CU, 13 AD. Now, combining this data with MRI variables available in each cohort reduced the number of CU/AD to 204 / 130, 67 / 87, and no remaining participants, respectively. Further adding CSF measurements again decreased the sample size to 38 CU and 35 AD for ADNI, 47 CU and 53 AD for EMIF, and no remaining participants for AIBL.

Cohort	ADNI		EMIF		AIBL	
Diagnosis	CU	AD	CU	AD	CU	AD
Total	813	389	230	184	803	181
Cognitive Scores	229	183	140	114	92	13
+ MRI	204	130	67	87	0	0
+ CSF	38	35	47	53	0	0
+ Amyloid PET	0	0	47	53	0	0

Table S1: Decrease in sample size when aiming for a multimodel analysis of amyloid PET measurements and stepwise adding additional modalities while recording only complete cases (ie. no missing values in any selected variable).

Event-based models

While prior versions of event-based models (EBMs) mainly integrated parametric mixture models (i.e. Gaussian mixture models; GMM) [1-3], we leveraged its latest installment which incorporates nonparametric mixture models by employing a kernel density estimation (KDE) to determine the probability density function [4]. KDE estimates the probability density of independent and identically distributed samples (x_1, x_2, …, x_n) drawn from a distribution with an unknown density by

$$\widehat{f}(x) = \frac{1}{Nh} \sum_{j=1}^{N} K(\frac{x-x_j}{h}) \qquad \textit{\textbf{Equation 2}}$$

where *K* and *h* are kernel function and bandwidth, respectively **[5]**. The kernel we used was Gaussian and the bandwidth for estimating the components of the mixture models was determined by Scott's normal reference rule **[6]**.

Meta-sequence generating algorithm

Our proposed method for generating a meta-sequence from multiple, complementary base sequences represents an algorithm addressing the rank aggregation of partial lists **[7]**. It essentially solves an optimization problem (i.e., finding the meta-sequence with the smallest distance to all base-sequences) by combining an exhaustive search for the initial k-length starting sequence with a greedy search procedure adding missing variables into the respective position in the sequence where the average distance of the altered sequence to all base-sequences remains minimal. The reason for combining these two steps lies in the combinatorial explosion of the search space (i.e., the set of all theoretically possible meta-sequences) when including an increasing number of variables. Therefore, an exhaustive search is often computationally infeasible and heuristic approaches have to be considered instead. In such cases, rank aggregation approaches often rely on Monte Carlo sampling to test a set of random meta-sequences and then opt for the one with the lowest distance **[8]**. However, we found that these approaches often end with suboptimal meta-sequences for large search spaces and that the proposed approach mixing an exhaustive search for a starting sequence with subsequent greedy refinements leads to more robust and plausible results, both biologically and in comparison to the base sequences. In theory, multiple meta-sequences could be identified that share the minimum distance to the base sequences (Figure S1). The python code for running this algorithm can be found under (https://github.com/sepehrgolriz/EBM-MultiCohort).

The bootstrapping-based version of our algorithm follows the same logic, however, the process is repeated b times, where the base sequences are determined based on a bootstrap sample of each respective cohort (ie. a sampling with replacement of the cohort of equal size as the original data).

Meta-sequence algorithm

The following pseudocode outlines the algorithm used to generate a meta-sequence out of multiple partially overlapping event-sequences.

```
Require:
    Set of all base sequences, $S$
    Event space defined over the union of events observed in the base sequences, $E$
    Set of all possible sequences of k-length drawn from $E$, $P$
    Distance metric of choice, $D(.)$
1:  function DISTANCE TO ALL SEQUENCES $(p,S)$; with $p$ denoting an arbitrary sequence
2:    initialize set for storing distances $dist_{p,S}=\{\}$
3:    for each $s\in S$ do
4:      $CE=$ find common events between $p$ & $s$
5:      unless $CE=\{\}$
6:        $dist_{p,s}=D(s_{CE},p_{CE})$; with $_{CE}$ denoting a sub-sequence containing only events $\in CE$
7:        add $dist_{p,s}$ to $dist_{p,S}$
8:  return average$(dist_{p,S})$

9:  function FIND START SEQUENCE $(P,S)$
10:   for each $p\in P$ do
11:     $dist_{p,S}=$ DISTANCE TO ALL SEQUENCES $(p,S)$
12: return $p_{min}=p$ for which $dist_{p,S}$ is minimal

13: function ADD REMAINING EVENTS $(R,p,S)$, also depends on global variable $MS$ outlined below
14:   if $R=\{\}$
15:     $dist_{p,S}=$ DISTANCE TO ALL SEQUENCES $(p,S)$
16:     store $p$ and $dist_{p,S}$ in global set $MS$
17:     end function without return
18:   $e=$ first element of $R$
19:   for all possible positions $i$ in $p$ do
20:     $p_{inserted,i}=$ insert $e$ into $p$ at $i$
21:     $dist_{p_{inserted,i},S}=$ DISTANCE TO ALL SEQUENCES $(p_{inserted,i},S)$
22:   get all possible positions for $e$, $I=\{i|dist_{p_{inserted,i},S}$ is minimal $\}$
23:   get new missing events excluding $e$, $R_{new}=\{x|x\in R,x\neq e\}$
24:   for all $i\in I$
25:     $p_{new}=$ insert $e$ into $p$ at $i$
26:     ADD REMAINING EVENTS $(R_{new},p_{new},S)$

Generate Meta-sequence:
27: $p_{min}=$ FIND START SEQUENCE $(P,S)$
28: get events not yet in $p_{min}$, $R=\{x|x\notin p_{min},x\in E\}$
29: initialize set for storing meta-sequences and their distances, $MS=\{\}$
30: ADD REMAINING EVENTS $(R,p_{min},S)$ to $MS$
31: print $ms\in MS$ with minimal $dist_{ms,S}$
```

Figure S1. Proposed algorithm for determining a meta-sequence from multiple potentially only partially overlapping base sequences. Line 23: All possible orderings in which to add the remaining events are tested.

Handling partially overlapping lists

Distance calculations have to be performed in the same mathematical space which, in this case, is defined by the variables in the sequences to be compared. Calculating the distance between two sequences which share the same variables is therefore straight forward. However, since individual base-sequences are often only partially overlapping, such distance calculations are impeded. There are two solutions for this problem: 1) penalizing the absence of variables in either sequence such that the distance increases with a higher number of uncommon variables, or 2) ignore variables that are only present in one of the sequences when calculating the distance. In the context of clinical cohort data, whether a specific variable was assessed depends on the study's goals and funding and, as such, its absence does seldomly hold biological meaning. Therefore, in this case, penalizing the absence of variables would bias the constructed meta-sequence.

Distance metrics

Depending on the focus of the study, different distance metrics can be used in the proposed algorithm. Intuitive choices are Spearman's footrule distance or Kendall's tau. The former takes the magnitude of the rank differences into account, while the latter is only counting how many rank discrepancies are found between two compared sequences, ignoring their specific position. A decision on which metric should be used depends on the emphasis of the study. In this study, we used Spearman's footrule distance because it takes the absolute difference in positions of variables into account which should be informative in our biomedical context.

Patient staging

$$argmax_j\, Pr(X_j|MS, d) = argmax_j\, P(d) \sum_{m=0}^{M} \{ \prod_{i=1}^{m} Pr(x_{ij}|E_i) \prod_{i=m+1}^{M} Pr(x_{ij}|\neg E_i) \quad \text{Equation 3}$$

$\Pr(x_{ij}|E_i)$ and $\Pr(x_{ij}|\neg E_i)$ denote the probability of observing the value of x given that event E_i did, or did not, occurred, respectively. It is assumed that the probability of being at stage *d* is uniform. The final assignment of a particular subject to a certain stage *d* remains a probabilistic assignment and is not a definite description that this participant is exactly at that stage in the disease cascade.

Supplementary Table

Cohort	Memory	Executive	Language	Visuospatial	Global cognitive
ADNI	LIMM LDEL	DIGIT TRABS	-	-	ADAS11 ADAS13
JADNI	LIMM LDEL	DIGIT TRABS	-	-	ADAS11 ADAS13
NACC	LIMM	WAIS	BNTS CATFLU	-	-
AILB	LIMM LDEL	WAIS STROOP	LIRE LIDE LICOR CATFLU	FIGC FIGR	-
EMIF	LIMM LDEL	EXECUTIVE	LANG	-	-
ANM	-	-	-	-	-
ARWIBO	LIMM LDEL	-	-	FIGC	-
OASIS	-	-	-	-	-
EDSD	-	-	LIRE LICOR BNTS	FIGC FIGR CLKS	-
WMAHD	STM	-	CATFLU	CLKS	-

Table S2. Cohort-specific cognitive tests composing each cognitive domain.

Cohort	**Variables**	Years of education (mean±std)	Age (mean±std)	APOE4% (at least one e4 allele)	Female%
ADNI	CU	16 ± 2	74 ± 5	31	32
	MCI	16 ± 2	73 ± 7	62	32
	AD	15 ± 2	75 ± 7	79	40
	Total	15 ± 2	74 ± 6	58	33
JADNI	CU	14 ± 4	69 ± 6	30	56
	MCI	13 ± 4	73 ± 5	51	54
	AD	12 ± 2	74 ± 3	62	55
	Total	13 ± 3	72 ± 6	47	55
ARWIBO	CU	8 ± 2	52 ± 7	20	57
	MCI	6 ± 4	71 ± 7	39	64
	AD	5 ± 3	71 ± 8	42	81
	Total	6 ± 2	64 ± 7	33	67
OASIS	CU	6 ± 2	52 ± 12	-	71
	MCI	5 ± 2	74 ± 6	-	55

	AD	3 ± 2	78 ± 4	-	66
	Total	5 ± 1	68 ± 9	-	64
EMIF	CU	13 ± 5	69 ± 7	48	51
	MCI	11 ± 5	68 ± 7	51	49
	AD	10 ± 5	66 ± 7	52	50
	Total	11 ± 7	67 ± 7	51	50
ANM	CU	10 ± 3	75 ± 4	24	59
	MCI	7 ± 3	78 ± 7	38	54
	AD	7 ± 3	79 ± 5	57	62
	Total	8 ± 3	77 ± 6	39	58
AIBL	CU	12 ± 3	-	23	59
	MCI	12 ± 7	-	52	48
	AD	12 ± 2	-	52	48
	Total	12 ± 2	-	42	51
EDSD	CU	13 ± 5	69 ± 3	31	51
	MCI	11 ± 3	71 ± 5	47	43
	AD	11 ± 5	73 ± 5	56	51
	Total	12 ± 3	72 ± 6	44	48

WMHAD	CU	8 ± 2	73 ± 4	40	35
	MCI	8 ± 2	77 ± 6	60	50
	AD	8 ± 2	77 ± 6	50	80
	Total	8 ± 2	75 ± 6	50	55
NACC	CU	16 ± 5	75 ± 4	38	71
	MCI	16 ± 4	73 ± 4	47	40
	AD	13 ± 5	73 ± 4	61	42
	Total	16 ± 2	63 ± 4	48	51

Table S3. The table above summarizes the demographic characteristics as well as the total number of participants in each diagnostic group for the investigated cohort datasets. CU: Cognitively unimpaired. MCI: Mild cognitive impairment. AD: Alzheimer's disease. Age: The average age of participants in each dataset. Years of Education: The average years of education of participants in each dataset. Female %: The percentage of female participants within each dataset. APOE4 Positive %: the percentage of participants with at least 1 APOE e4 allele. SD: Standard deviation. The years of education listed for OASIS participants seems irritatingly low, however, we found those values listed in the data.

	AIBL	JADNI	ANM	WMHAD	ARWIBO	EMIF	OASIS	ADNI	EDSD	NACC
AIBL	1	-	-	-	0.81	-	-	-	0.81	0.71
JADNI	-	1	0.78	0.73	0.60	0.801	0.73	0.72	0.78	0.86
ANM	-	0.78	1	0.82	0.85	-	0.70	0.90	0.90	0.62
WMHAD	-	0.73	0.82	1	0.81	-	0.73	0.90	0.91	0.67
ARWIBO	0.81	0.60	0.85	0.81	1	1	0.72	0.78	0.62	0.55
EMIF	-	0.80	-	-	1	1	-	0.85	-	0.90
OASIS	-	0.73	0.70	0.73	0.72	-	1	0.69	0.76	0.87
ADNI	-	0.72	0.90	0.90	0.78	0.84	0.69	1	0.90	0.73
EDSD	0.81	0.78	0.90	0.91	0.62	-	0.76	0.90	1	0.81
NACC	0.71	0.86	0.62	0.68	0.55	0.91	0.87	0.73	0.81	1

Table S4. Pairwise Kendall’s tau rank correlation coefficients

Supplementary Figure

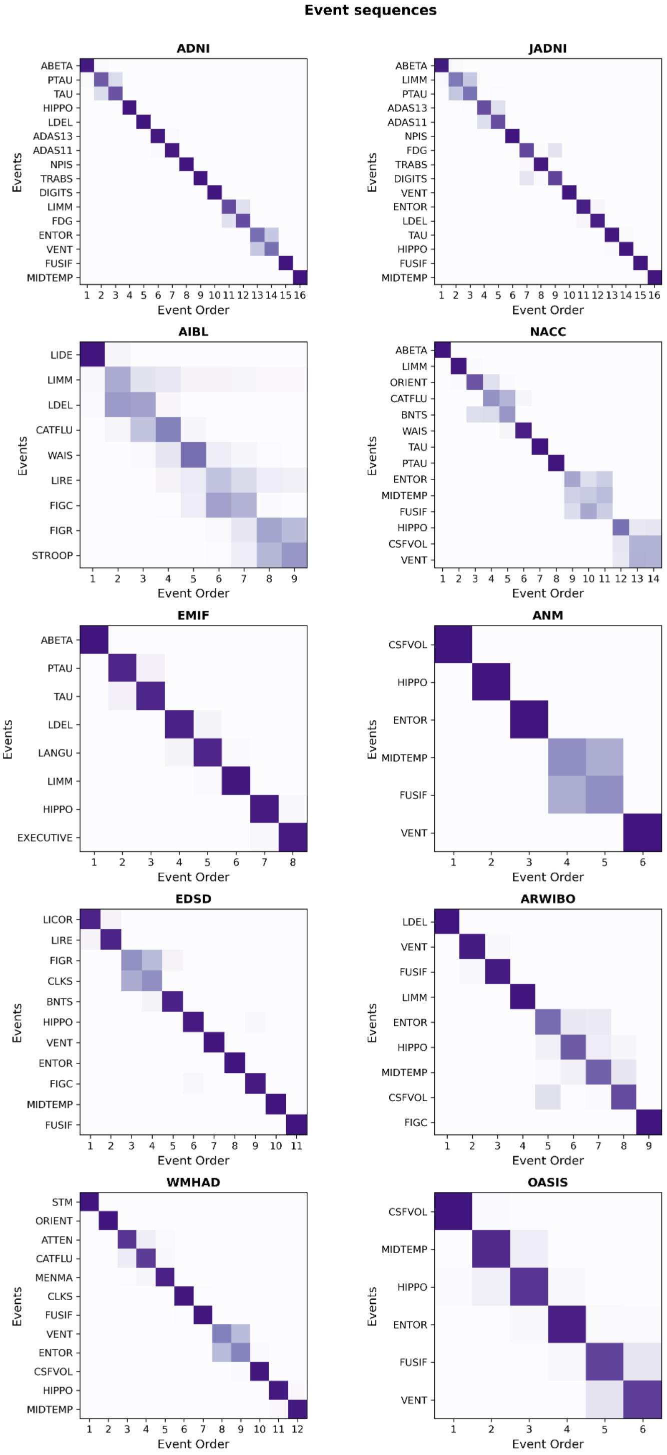
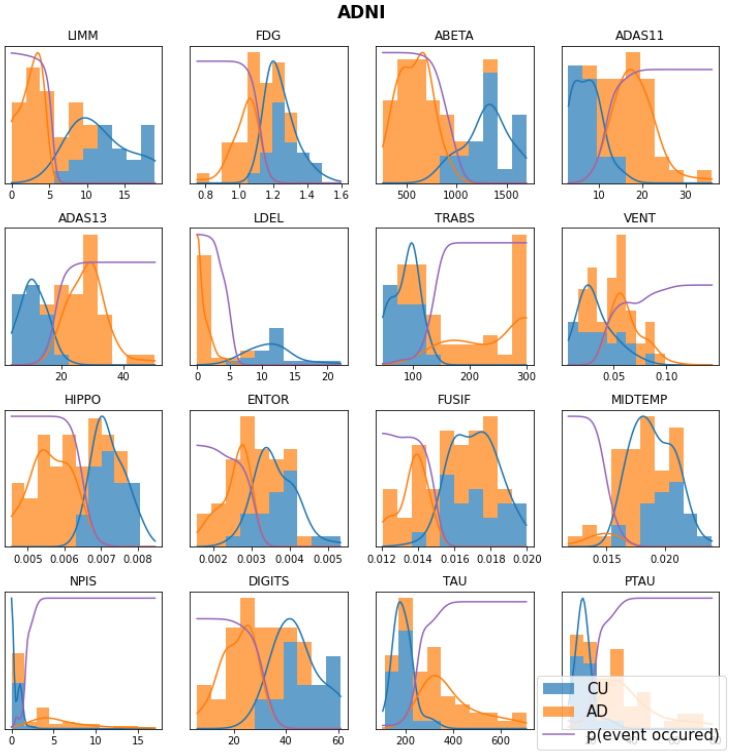
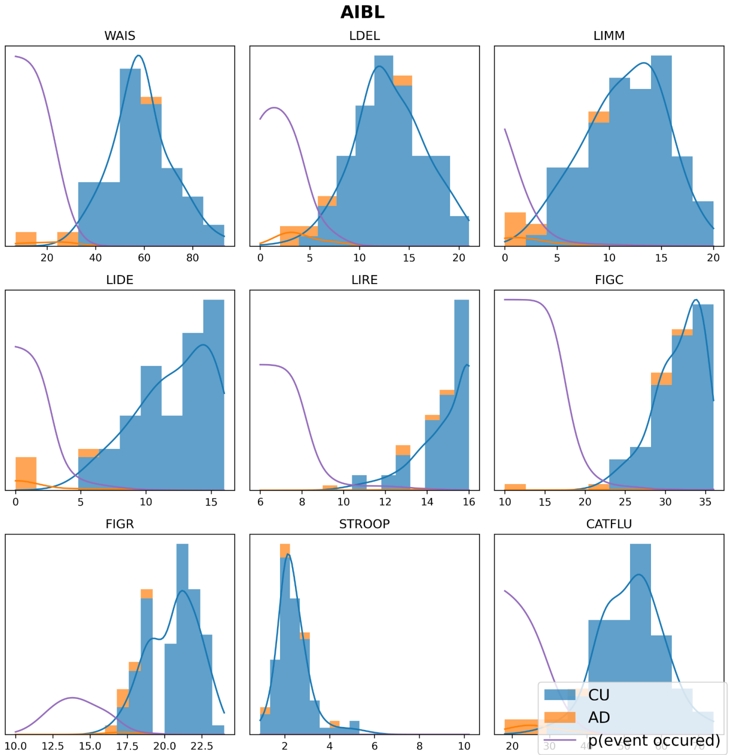


Figure S2. The original individual event sequences (independent y-axes) derived from the ten investigated cohorts. Event order 1 corresponds to the first position in the sequence. The shading of squares indicates the positional probability with darker shades corresponding to higher probabilities. The relative sizes of the squares do not encode any information.

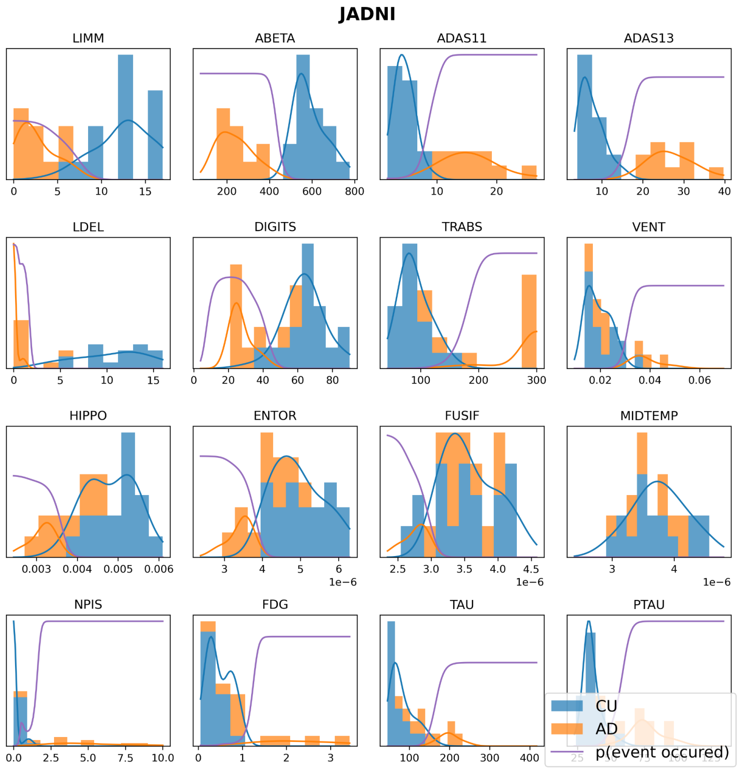
a)


ADNI
LIMM
0 5 10 15
FDG
0.8 1.0 1.2 1.4 1.6
ABETA
500 1000 1500
ADAS11
10 20 30
ADAS13
20 40
LDEL
0 5 10 15 20
TRABS
100 200 300
VENT
0.05 0.10
HIPPO
0.005 0.006 0.007 0.008
ENTOR
0.002 0.003 0.004 0.005
FUSIF
0.012 0.014 0.016 0.018 0.020
MIDTEMP
0.015 0.020
NPIS
0 5 10 15
DIGITS
20 40 60
TAU
200 400 600
PTAU
CU
AD
p(event occured)

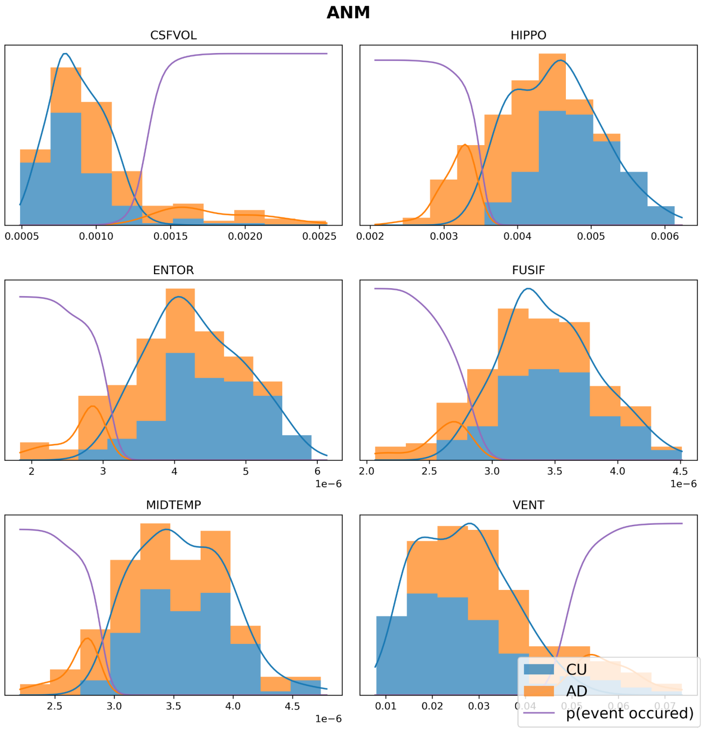
b)


AIBL
WAIS
20 40 60 80
LDEL
0 5 10 15 20
LIMM
0 5 10 15 20
LIDE
0 5 10 15
LIRE
6 8 10 12 14 16
FIGC
10 15 20 25 30 35
FIGR
10.0 12.5 15.0 17.5 20.0 22.5
STROOP
2 4 6 8 10
CATFLU
20
CU
AD
p(event occured)

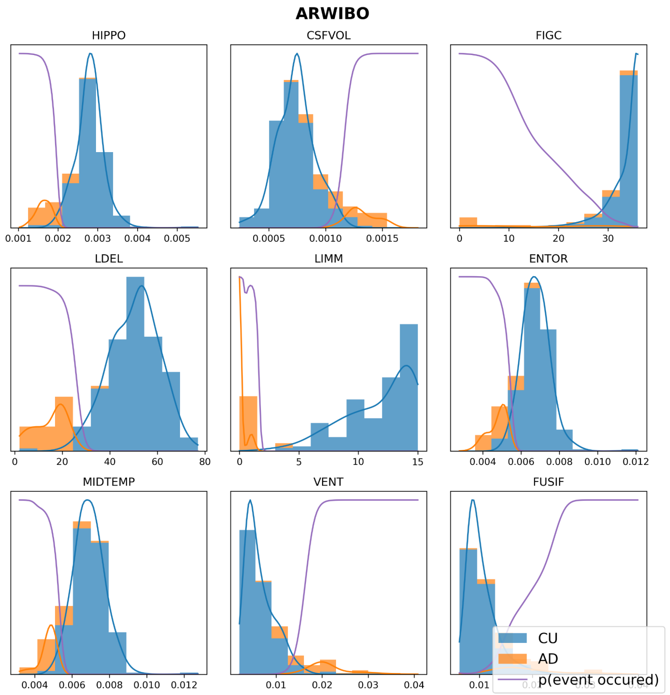
c)


JADNI
LIMM
0
5
10
15
ABETA
200
400
600
800
ADAS11
10
20
ADAS13
10
20
30
40
LDEL
0
5
10
15
DIGITS
0
20
40
60
80
TRABS
100
200
300
VENT
0.02
0.04
0.06
HIPPO
0.003
0.004
0.005
0.006
ENTOR
3
4
5
6
1e−6
FUSIF
2.5
3.0
3.5
4.0
4.5
1e−6
MIDTEMP
3
4
1e−6
NPIS
0.0
2.5
5.0
7.5
10.0
FDG
0
1
2
3
TAU
100
200
300
400
PTAU
CU
AD
p(event occured)

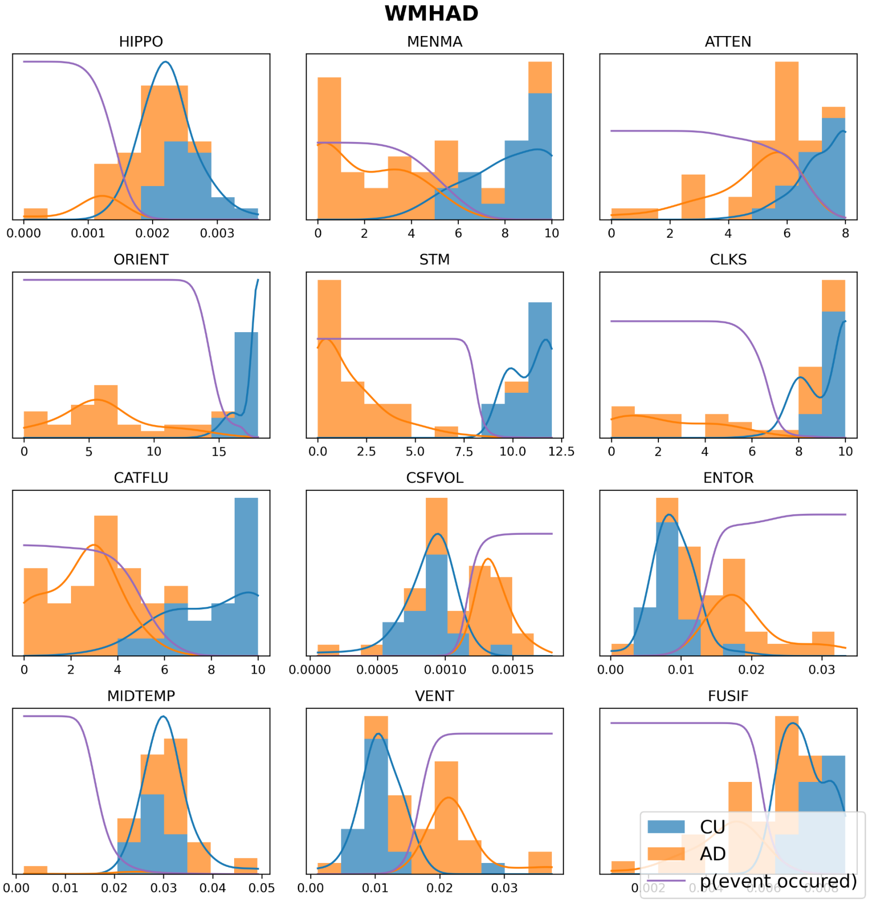
d)


ANM
CSFVOL
0.0005
0.0010
0.0015
0.0020
0.0025
HIPPO
0.002
0.003
0.004
0.005
0.006
ENTOR
2
3
4
5
6
1e−6
FUSIF
2.0
2.5
3.0
3.5
4.0
4.5
1e−6
MIDTEMP
2.5
3.0
3.5
4.0
4.5
1e−6
VENT
0.01
0.02
0.03
CU
AD
p(event occured)

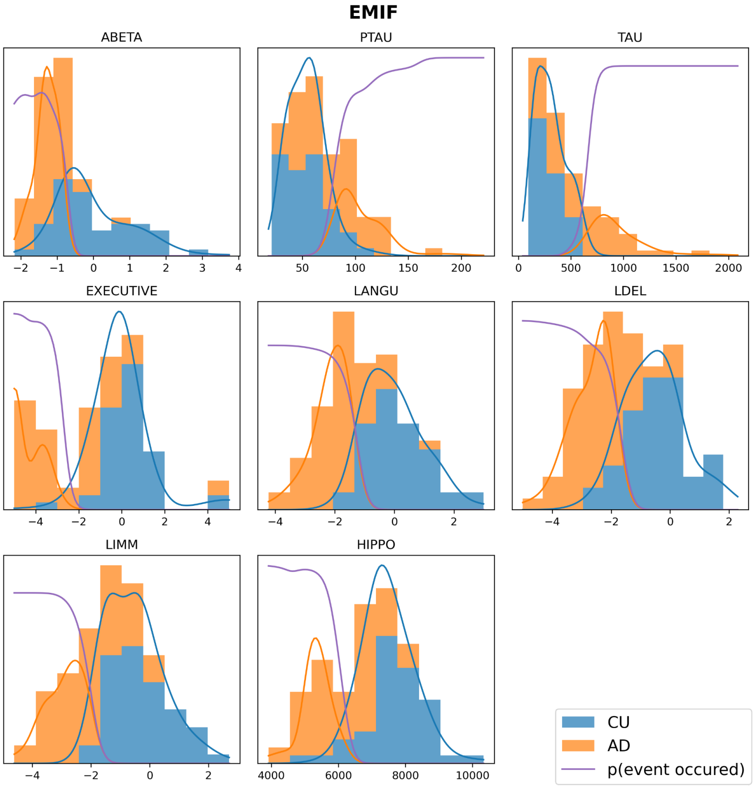
e)


ARWIBO
HIPPO
0.001 0.002 0.003 0.004 0.005
CSFVOL
0.0005 0.0010 0.0015
FIGC
0 10 20 30
LDEL
0 20 40 60 80
LIMM
0 5 10 15
ENTOR
0.004 0.006 0.008 0.010 0.012
MIDTEMP
0.004 0.006 0.008 0.010 0.012
VENT
0.01 0.02 0.03 0.04
FUSIF
0.01
CU
AD
p(event occured)

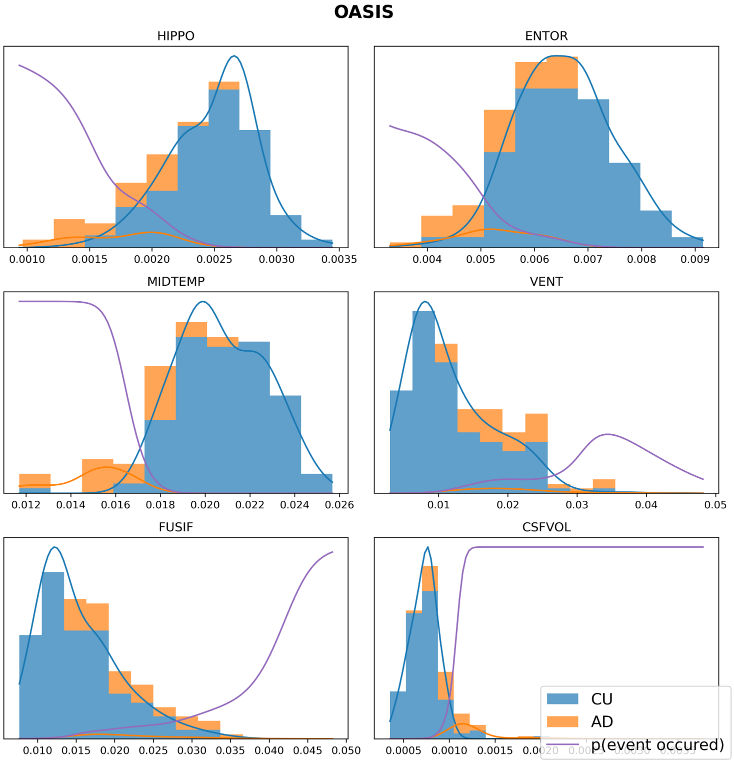
f)


WMHAD
HIPPO
0.000 0.001 0.002 0.003
MENMA
0 2 4 6 8 10
ATTEN
0 2 4 6 8
ORIENT
0 5 10 15
STM
0.0 2.5 5.0 7.5 10.0 12.5
CLKS
0 2 4 6 8 10
CATFLU
0 2 4 6 8 10
CSFVOL
0.0000 0.0005 0.0010 0.0015
ENTOR
0.00 0.01 0.02 0.03
MIDTEMP
0.00 0.01 0.02 0.03 0.04 0.05
VENT
0.00 0.01 0.02 0.03
FUSIF
0.002
CU
AD
p(event occured)

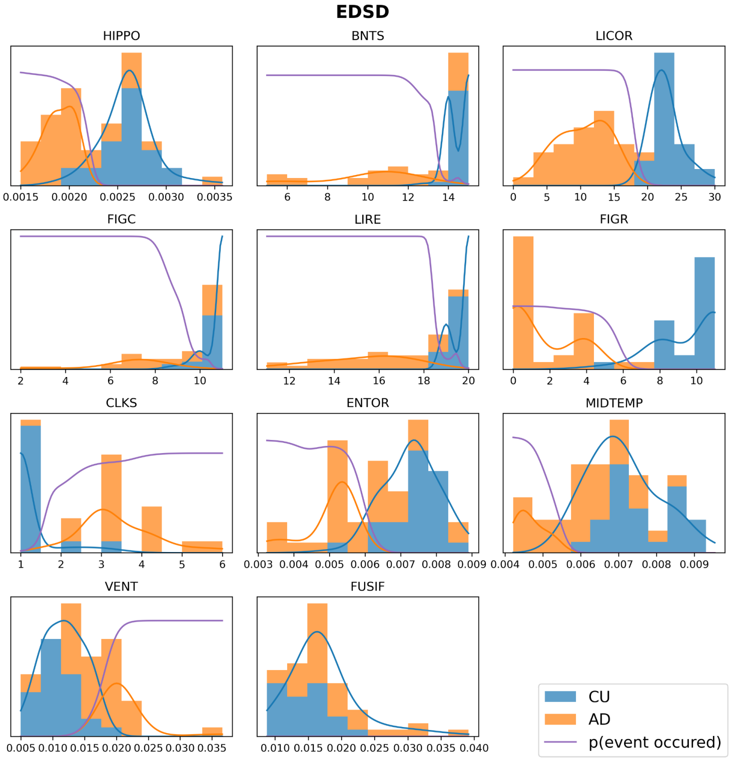
g)


EMIF
ABETA
−2 −1 0 1 2 3 4
PTAU
50 100 150 200
TAU
0 500 1000 1500 2000
EXECUTIVE
−4 −2 0 2 4
LANGU
−4 −2 0 2
LDEL
−4 −2 0 2
LIMM
−4 −2 0 2
HIPPO
4000 6000 8000 10000
CU
AD
p(event occured)

h)


OASIS
HIPPO
0.0010 0.0015 0.0020 0.0025 0.0030 0.0035
ENTOR
0.004 0.005 0.006 0.007 0.008 0.009
MIDTEMP
0.012 0.014 0.016 0.018 0.020 0.022 0.024 0.026
VENT
0.01 0.02 0.03 0.04 0.05
FUSIF
0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045 0.050
CSFVOL
0.0005 0.0010 0.0015 0.0020
CU
AD
p(event occured)

i)



j)

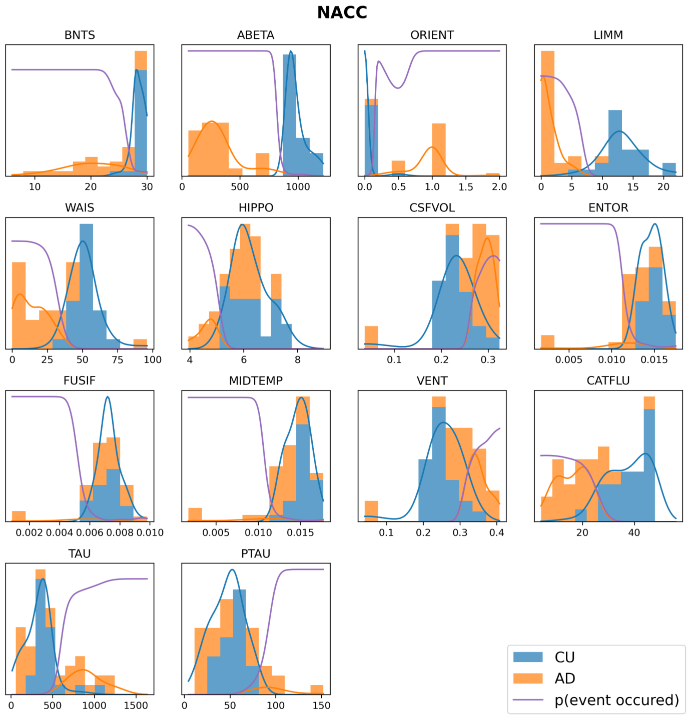


Figure S3. The derived mixture models for each cohort (a-j).

References

1. Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., *et al*. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage* 2012, 60(3), 1880–1889. https://doi.org/10.1016/j.neuroimage.2012.01.062

2. Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., Schott, J. M., *et al*. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 2014; 2564–2577. https://doi.org/10.1093/brain/awu176

3. Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T., Fagan, A. M., Morris, J. C., Bateman, R. J., *et al*. Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 2018; 141(5), 1529–1544. https://doi.org/10.1093/brain/awy050

4. Firth, N. C., Primativo, S., Brotherhood, E., Young, A. L., Yong, K., Crutch, S. J., *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. *Alzheimers dement* 2020, 16(7), 965–973. https://doi.org/10.1002/alz.12083

5. Zhang, W., Zhang, Z., Chao, HC. *et al*. Kernel mixture model for probability density estimation in Bayesian classifiers. *Data Min Knowl Disc* 2018; 32, 675–707. https://doi.org/10.1007/s10618-018-0550-5

6. Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* 1979, 66(3), 605-610.

7. Li, X., Wang, X., & Xiao, G. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in bioinformatics* 2019, 20(1), 178–189. https://doi.org/10.1093/bib/bbx101

8. Lin, S.. Rank aggregation methods. Wiley Interdisciplinary Reviews: Computational Statistics 2010, 2(5), 555-570.

7.3 A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease

Reprinted with permission from "Golriz Khatami, S. *et al.*. "A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease". *J Alzheimers Dis*, 80(2), pp.831-840, (2021).

Journal of Alzheimer's Disease 80 (2021) 831–840
DOI 10.3233/JAD-201397
IOS Press

A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease

Sepehr Golriz Khatami[a,b,*], Daniel Domingo-Fernández[a], Sarah Mubeen[a,b], Charles Tapley Hoyt[a], Christine Robinson[a,b], Reagon Karki[a,b], Anandhi Iyappan[a], Alpha Tom Kodamullil[a] and Martin Hofmann-Apitius[a,b]

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

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

Accepted 7 January 2021
Pre-press 2 February 2021

Abstract.

Background: Neuroimaging markers provide quantitative insight into brain structure and function in neurodegenerative diseases, such as Alzheimer's disease, where we lack mechanistic insights to explain pathophysiology. These mechanisms are often mediated by genes and genetic variations and are often studied through the lens of genome-wide association studies. Linking these two disparate layers (i.e., imaging and genetic variation) through causal relationships between biological entities involved in the disease's etiology would pave the way to large-scale mechanistic reasoning and interpretation.

Objective: We explore how genetic variants may lead to functional alterations of intermediate molecular traits, which can further impact neuroimaging hallmarks over a series of biological processes across multiple scales.

Methods: We present an approach in which knowledge pertaining to single nucleotide polymorphisms and imaging readouts is extracted from the literature, encoded in Biological Expression Language, and used in a novel workflow to assist in the functional interpretation of SNPs in a clinical context.

Results: We demonstrate our approach in a case scenario which proposes KANSL1 as a candidate gene that accounts for the clinically reported correlation between the incidence of the genetic variants and hippocampal atrophy. We find that the workflow prioritizes multiple mechanisms reported in the literature through which KANSL1 may have an impact on hippocampal atrophy such as through the dysregulation of cell proliferation, synaptic plasticity, and metabolic processes.

Conclusion: We have presented an approach that enables pinpointing relevant genetic variants as well as investigating their functional role in biological processes spanning across several, diverse biological scales.

Keywords: Alzheimer's disease, genetic variants, knowledge graph, neuroimaging, systems biology

INTRODUCTION

As aging populations continue to grow, age-associated disorders such as Alzheimer's disease (AD) have become increasingly prevalent [1, 2]. AD is a slow-progressing, complex, idiopathic disorder in which early diagnosis is challenging because patients

*Correspondence to: Sepehr Golriz Khatami, Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53757, Germany. E-mail: sepehr.golriz.khatami@scai.fraunhofer.de.

do not initially present symptoms [3]. Emerging neuroimaging techniques are a versatile, non-invasive approach for the high-resolution, *in vivo* investigation of the underlying pathophysiology of AD that may provide an opportunity for earlier detection and therapeutic intervention.

Neuroimaging techniques quantitatively measure markers of brain structure and function that are considered as endophenotypes, measurable intermediate phenotypes that link molecular changes to organ-specific pathophysiological contexts [4]. One of the numerous neuroanatomical markers considered as an endophenotype is medial temporal atrophy. This well-established AD marker is an intermediate phenotype that implicates the aggregation of hyperphosphorylated tau protein (a well-known molecular change) as a causative biological process of memory decline [5]. The diversity of markers prompted the cataloging and organizing of their information in order to better link clinical readouts to underlying molecular changes. As a first attempt in addressing this need, Iyappan et al. curated the terms used in the literature to describe structural and functional brain information in the Neuroimaging Feature Terminology (NIFT) [6].

Elucidating the effect of genes and genetic variations (e.g., single nucleotide polymorphisms (SNPs)) on brain structure and function often begins with genome-wide association studies (GWASs). However, this type of study only calculates statistical associations between SNPs and traits and ignores mechanistic insights. More robust approaches aimed at addressing the mechanistic shortcomings of GWAS are referred to as imaging genetics [7]. For example, Wachinger et al. [8] studied genetic influences on neuroanatomical shape asymmetries associated with AD progression. Although their findings on the association of genetic variants (i.e., BIN1, CD2AP, ZCWPW1, and ABCA7 genes) to neuroanatomical structures had been reported in previous studies [9–12], here the authors were able to provide an explanation for the observed effect, specifically that alterations in the expression level of the aforementioned genes can affect cellular homeostasis, thus leading to changes in brain symmetry. A common issue facing many imaging genetics approaches is small sample size, which leads to a lack of statistical power, limited replicability, and stratification effects [13, 14]. Alternatively, Stefanovski et al. [15] studied the connection between molecular changes and neuronal population dynamics using differential equations. For example, this study provided a possible mechanistic explanation of how local amyloid beta-mediated synaptic function disinhibition leads to diminishing neural signaling. However, such mathematical models thus far fail to handle the number of variables that are necessary to represent the pathophysiological phenomenon involved in a multifactorial disorder such as AD.

The limitations and lack of mechanistic insights provided by these previously mentioned techniques prompted us to develop a new approach to interpret how a particular genetic variant may have an impact on neuroimaging feature changes through sequences of molecular causalities in the context of AD. Our approach captures knowledge from the literature pertaining to SNPs and imaging readouts in a causal model encoded in Biological Expression Language (BEL) [16] to support the functional interpretation of SNPs in a clinical context. In a case scenario, we propose KANSL1 as a candidate gene mediating the connection between the genetic variants and hippocampal atrophy. We then hypothesized that variants of this gene dysregulate biological processes related to cell proliferation, synaptic plasticity, and energy metabolism that ultimately leads to hippocampal atrophy. These dysregulated biological processes are early events in AD, and they have been posited as attractive therapeutic targets for pharmaceutical intervention [17, 18]. Thus, by garnering these mechanistic insights, it may be possible to reveal novel therapeutic options in the future.

MATERIALS AND METHODS

In order to support the interpretation of the functional impact of SNPs on the alteration of neuroimaging features, associations between SNPs and imaging readouts were extracted using natural language processing. Linkage disequilibrium (LD) analysis was used to identify co-occurring SNPs and their corresponding or associated genes. These genes were then ranked by how often they appear in the literature in the context of AD. This workflow is described in Fig. 1A.

Based on this analysis, one gene (KANSL1) was selected for further investigation and a corpus explaining its role in AD was enriched with knowledge pertaining to multi-scale biological processes. To enable computer-aided reasoning, manually-extracted relations from this corpus were encoded in BEL. The resulting KANSL1 knowledge assembly was validated using PyBEL [19] and integrated into NeuroMMSig [20]. Finally, NeuroMMSig was then

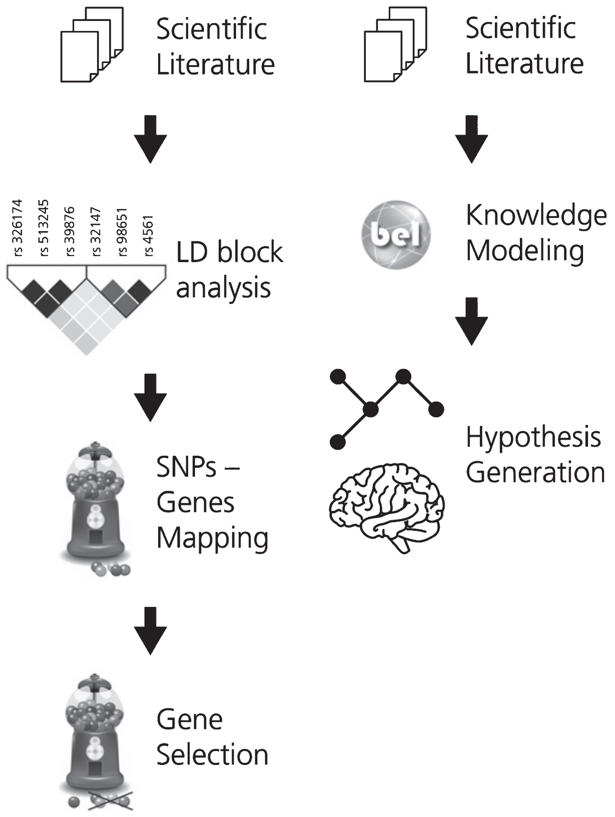


Fig. 1. The two workflows developed for (A) gene prioritization and for (B) generating the mechanistic knowledge assembly around the effect of genetic variants on neuroimaging features in AD. In workflow A, the first step involves the selection of a corpus of relevant scientific literature. Next, the SNPs extracted from this corpus were subjected to LD block analysis and the subsequently obtained SNPs were mapped to their corresponding or associated genes. KANSL1, a novel AD gene, was selected from this pool of mapped genes for further investigation. In workflow B, corpus for the selected gene is extracted and translated into BEL to generate a knowledge assembly model for hypothesis generation.

used to investigate the putative role of KANSL1 in neuroimaging feature alteration, namely hippocampal atrophy. For the sake of reproducibility, we have made the workflow publicly available through GitHub (https://github.com/sepehrgolriz/GeVa_NeIF) under the MIT License. This workflow is described in Fig. 1B. Additionally, to investigate the concordance of knowledge around the KANSL1 gene, pathways from three well-known pathway databases were queried to determine those in which the gene is implicated.

Generation of a SNP-neuroimaging corpus

A corpus enriched with neuroimaging features and SNPs in the context of AD was generated using SCAIView v0.3.3 (https://academia.scaiview.com) on MEDLINE using the following query: "*(([MeSH Disease: "Alzheimer Disease"]) AND [Neuroimaging Feature]) AND [SNP]))"*. The corpus comprised 568 documents with a total of 2215 SNP-neuroimaging feature associations (Supplementary Table 1), including 126 unique neuroimaging features (Supplementary Table 2) and 745 unique SNPs (Supplementary Table 3).

Identification of related SNPs via linkage disequilibrium blocks

Over time, dependencies between genetic variants are developed across populations [21]. This phenomenon, described as LD, implies that correlations between genetic variants and traits are caused by the aggregated effect of multiple variants [22, 23]. However, SNP-trait associations identified in the literature are obtained by analyzing thousands of SNPs individually (the "single-marker" approach). Therefore, we performed LD block analysis using HaploReg v4.1 [24] to identify a total of 6,070 SNPs that occur with the SNPs extracted from the literature and further mapped them to their corresponding or associated genes (Supplementary Table 4).

Gene selection

DisGeNET [25] was used to identify diseases associated with the genes obtained from HaploReg v4.1. After filtering out genes not associated with AD, the remaining genes were categorized as either well-known risk variants (supported by a minimum of 5 literature evidence which are enriched with observational studies, such as case-control studies) or as emerging genetic biomarkers (those supported by few or no published evidence) (Supplementary Table 5). Since the involvement of well-known risk variants has been sufficiently described in the literature, this study investigated novel genes which may contribute to AD development.

The study of genetics in the context of multiple phenotypes, such as physiological traits or diseases, can provide a holistic overview of gene functions in a biological system. For this reason, DisGeNET was used to investigate gene-disease associations of genes that are not clearly linked to AD [26]. Although the genes are associated with a broad range of diseases, from autoimmune disorders to different types of cancer, we focused on enriching the mechanistic context surrounding genes linked to conditions, such as Parkinson's disease (PD), which have substantial

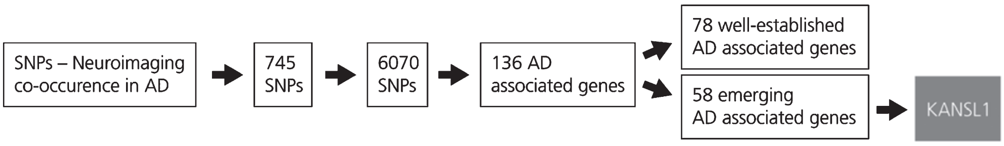

SNPs – Neuroimaging co-occurence in AD
745 SNPs
6070 SNPs
136 AD associated genes
78 well-established AD associated genes
58 emerging AD associated genes
KANSL1

Fig. 2. This figure shows the results obtained from the LD block analysis and gene mapping. The generation of the SNP-Neuroimaging corpus yielded 745 SNPs. Following LD block analysis, 6,070 SNPs that occur with the SNPs extracted from the literature were identified and located on 136 unique AD associated genes. These genes were then classified according to the number of evidences which are available in the scientific literature. The first group, incorporating 78 AD associated genes, comprises well-known genes characterized by a high number of publications in the AD context. The second group, that includes 58 AD associated genes, comprises emerging genes in the context of AD. From the latter group, KANSL1 was selected.

genetic, pathological, and clinical overlap with AD [27]. While it is believed that cancer and autoimmune diseases are less prevalent in AD patients, 25 to 33 percent of AD patients show concomitant PD pathology [28, 29]. Of the 25 PD-associated genes acquired, we selected KANSL1 for further study of its putative pathogenic role in AD as it had the highest number of literature evidence and its functionality can thus be better understood [30, 31] (Supplementary Table 6).

Corpus generation, relation extraction, and mechanism enrichment for KANSL1

Using the same strategy and resources as the previous corpus, a new corpus describing the role of KANSL1 in the context of AD was generated using the following SCAIView query: "*(([Human Genes/ Proteins:"KANSL1"]) AND [MeSH Disease: "Alzheimer Disease"]) AND [Neuroimaging Features]*". The resulting gene-neuroimaging interaction information was then enriched with further causal relations from the literature using manual relation extraction in order to bridge the knowledge gap between genetics and clinical endpoints.

Knowledge modeling

Manually generating mechanistic hypotheses by linking genetic variants to neuroimaging markers is a daunting task. Therefore, in order to empower computer-aided reasoning, the extracted knowledge assembly was encoded in BEL. Both the syntax and semantics of BEL encoded in the knowledge assembly were validated using the PyBEL framework.

Knowledge was extracted from the selected corpus using the official BEL curation guidelines from https://biological-expression-language.github.io as well as additional guidelines from https://github.com/pharmacome/curation.

Evidence from the selected corpus was manually translated into BEL statements together with their contextual information (e.g., brain regions, brain cell types). For instance, the evidence "BDNF infusion led to rapid phosphorylation of the mitogen-activated protein (MAP) in the adult hippocampus" corresponds to the following BEL statement:

SET MeSHAnatomy = "Hippocampus"
p(HGNC:BDNF) - - p(HGNCGENEFAMILY: "Mitogen-activated protein kinases", pmod(Ph))

The resulting knowledge assembly was then integrated into NeuroMMSig, a web server for mechanism enrichment that allows querying over genes, SNPs, and neuroimaging features in the context of a specific disease. Finally, NeuroMMSig was used to identify the mechanistic model representing the putative role of KANSL1 in hippocampal atrophy.

Comparison of the mechanistic model to pathway knowledge

Several manually curated and highly-cited pathway databases are available to deduce biologically relevant pathways. We used three major ones, namely KEGG [32], Reactome, [33] and WikiPathways [34], in order to determine whether knowledge on the KANSL1 gene has yet to be integrated into these resources. Hence, we queried KANSL1 as well as all other proteins from our mechanistic model in pathways from the three databases.

RESULTS

While KANSL1 has been associated with changes in gene expression levels in the hippocampus [35], its mechanism of action remains elusive. In order to better understand KANSL1's involvement in hippocampal dysfunction, we queried NeuroMMSig to investigate the downstream effects of this gene. Then, reasoning over the knowledge assembly led us to the interpretation described below. Finally, we report the results of querying KANSL1 and other genes from our mechanistic model in pathways from three major pathway databases to determine which pathways the gene may be implicated in.

The putative role of KANSL1 in hippocampal atrophy

The transcription and expression of the genes promoting cell proliferation (e.g., BTG2) and synaptic plasticity (e.g., BDNF) as well as metabolic processes (e.g., cell energy production) are both of paramount importance for hippocampal function [36] (Fig. 3). KANSL1 is a protein-coding gene involved in chromatin modification through histone acetylation [37, 38], one of the mechanisms orchestrating gene transcription and expression [39–44]. While histone acetylation transforms the condensed structure of the chromatin into a relaxed architecture enhancing RNA transcription and gene expression, its hypoacetylation causes it to behave adversely [45–47].

KANSL1 and hippocampal neurogenesis

KANSL1 is required for the acetylation of p53 [41], a transcription factor modulating BTG2 expression and a vital protein for hippocampal neurogenesis (i.e., while KANSL1-dependent p53 acetylation induces BTG2 expression, p53 hyperacetylation leads to the overexpression of BTG2) [42, 48, 49]. BTG2 negatively controls the cell cycle since its overexpression results in cell growth rate decline [42, 50]. Through BTG2 binding to Ras (the signaling event mediator), the Ras/MAPK signaling cascade is activated, leading to tau hyperphosphorylation [48]. Tau is a microtubule-associated protein that promotes the assembly and stabilization of cytoskeleton microtubules, both of which are required for cell division (i.e., mitosis). However, tau hyperphosphorylation reduces its capability to bind the microtubules, giving rise to dynamic instability, mitosis impairment, cell cycle deterioration, elimination of proliferating newborn neurons, and ultimately to apoptotic processes [51]. In summary, KANSL1 dysfunction disturbs the expression of cell cycle regulatory genes, leading to the perturbation of cell proliferation processes [46, 50] (Fig. 3A).

KANSL1 and hippocampal metabolic processes

The functional crosstalk between KANSL1 and the metabolic processes occurring in the mitochondria

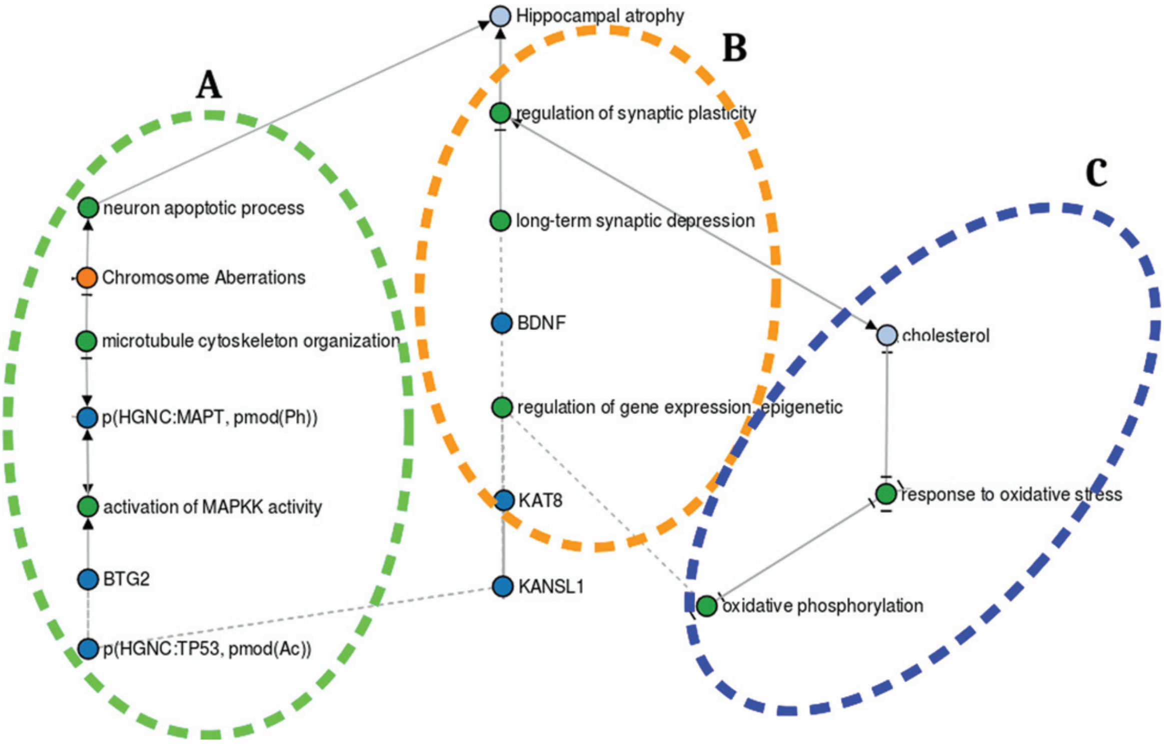

Hippocampal atrophy
A
B
C
neuron apoptotic process
Chromosome Aberrations
microtubule cytoskeleton organization
p(HGNC:MAPT, pmod(Ph))
activation of MAPKK activity
BTG2
p(HGNC:TP53, pmod(Ac))
regulation of synaptic plasticity
long-term synaptic depression
BDNF
regulation of gene expression, epigenetic
KAT8
KANSL1
cholesterol
response to oxidative stress
oxidative phosphorylation

Fig. 3. The putative role of KANSL1 in hippocampal atrophy. A) KANSL1 role in hippocampal neurogenesis. B) KANSL1 function in hippocampal metabolic processes. C) KANSL1 role in hippocampal synaptic plasticity. [https://nbviewer.jupyter.org/github/sepehrgolriz/GeVa_NeIF/blob/master/Semi_automatic_developed_pipeline/Exploring%20KANSL1%20putative%20role%20graph%20in%20hippocampal%20atrophy.ipynb].

(e.g., oxidative phosphorylation) is key for the regulation of hippocampal synaptic plasticity [52–55]. KANSL1 is highly expressed in the mitochondria, where it regulates mitochondrial DNA (mtDNA) transcription and the subsequent translation of genes involved in Oxidative phosphorylation—a set of complex mechanistic processes that form adenosine triphosphate (cell energy currency) by oxidizing nutrients [36, 55, 56]. Oxidative phosphorylation produces potentially harmful reactive oxygen species whose production and detoxification are balanced in normal mitochondria [57]. However, KANSL1 deficiency promotes the downregulation of mtDNA transcription and translation of genes involved in Oxidative phosphorylation, causing reactive oxygen species accumulation. Oxidative stress then occurs, leading to cholesterol metabolism perturbation [58]. Cholesterol homeostasis dysregulation increases cholesterol concentration in cells, leading to synaptic plasticity impairment and ultimately hippocampal shrinkage [59–63] (Fig. 3C).

KANSL1 and hippocampal synaptic plasticity

Long-term potentiation (LTP) is one of the major cellular processes involved in memory formation [64]. BDNF, a member of the neurotrophin family of growth factors, plays a role in LTP [65–67]. One of the mechanisms governing the regulation of BDNF expression is histone acetylation, where KANSL1 contributes significantly as a histone acetyltransferase complex. KANSL1 deficiency might severely affect BDNF expression, which further promotes long-term potentiation impairment and synaptic plasticity. Both are considered to play an important role in memory formation [68] (Fig. 3B).

Pathways implicating genes from mechanistic model

The investigation on the presence, or lack thereof, of KANSL1 in pathways from KEGG, Reactome, and WikiPathways revealed that the KANSL1 gene is largely absent in the major pathway databases. While KANSL1 does participate in the "Chromatin Organization (Homo Sapiens)" and "Pathways Affected in Adenoid Cystic Carcinoma (Homo sapiens)" pathways from WikiPathways, no interaction information for this gene is provided. Moreover, KANSL1 is altogether absent in pathways from KEGG and Reactome. Similarly, we queried pathways from the three databases for all other genes from our mechanistic model (Supplementary Table 7). Unsurprisingly, well-studied genes yielded a higher number of pathways which they participate in (e.g., BDNF was found in 33 pathways across KEGG, Reactome, and WikiPathways), while genes with fewer literature evidence were scarcely present (e.g., KAT8 was found in one pathway across KEGG, Reactome and WikiPathways, however lacked interaction information). Furthermore, these pathway resources do not yet capture SNPs nor image features. Accordingly, the mechanism by which KANSL1 may be implicated in hippocampal atrophy can thus far only be inferred through dedicated modeling approaches, such as the one we have presented in this work.

Assessment of putative KANSL1-mediated mechanism with experimental databases

The putative KANSL1-dependent hippocampal atrophy mechanisms identified through systematically harvested knowledge is based on qualitative information. To further support the mechanisms of action exerted by KANSL1 in the nervous system, we screened evidence from experimental databases containing data sets on knockout mouse models in the Mouse Genome Informatics database [69]. In this database, we queried for KANSL1 and nervous system and found two knockout mice studies which investigated how KANSL1/MAPT dysregulation may cause hippocampal shrinkage [70, 71]. These studies associated tau hyperphosphorylation coupled with impaired microtubule binding of tau with reduction in synaptic transmission and altered synaptic plasticity. Furthermore, the authors argue that these mechanisms may lead to neuronal apoptosis and hippocampal shrinkage (Fig. 3A).

Additionally, with respect to SNPs that occur in the non-coding regions of the gene, we used RegulomeDB [72] to functionally annotate the 60 SNPs associated with KANSL1. RegulomeDB scores SNPs based on transcription factor binding sites, position weight matrix for transcription factor binding, DNase footprinting, open chromatin and chromatin states, expression quantitative trait loci (eQTL), and validated functional SNPs. Moreover, it calculates a score that represents the probability of being a regulatory variant based on functional genomics features along with continuous values such as ChIP-seq signal, DNase-seq signal, information content change, and DeepSEA scores for each SNPs [73]. From the 60 SNPs, our analysis suggested that 11 of them are located in the functional region of KANSL1 (Supplementary Table 8).

DISCUSSION

While the exact mechanism of action of KANSL1 remains obscure, the proposed methodology was able to identify the mechanisms through which it may have an impact on hippocampal atrophy. This demonstrates how the mechanism enrichment approach offers improved interpretation of molecular mechanisms involved in disease pathobiology. Ultimately, the hypotheses derived from such approaches can foster research by identifying unexplored links that have not been validated in the laboratory.

We observed that the information pertaining to different biological scales is not equally distributed in the literature. For example, there is a paucity of results reported at the phenotypic level, compared to those at the molecular or organ level. Shortcomings in knowledge representation at different scales are also reflected in pathway databases which currently do not contain information on SNPs or neuroimaging features. Consequently, linking molecular mechanisms to clinical readouts is one of the great challenges in biomedical informatics.

The results presented in this work are hypotheses that require further investigation. We have shown that despite the scarcity of knowledge from the literature around KANSL1, our approach was able to reveal interesting hypotheses. This sparsity of information surrounding KANSL1 combined with its manifestation as a novel AD associated gene motivates future updates of the knowledge assembly as new information becomes available. Furthermore, in our attempt to validate our hypothesis, we did not find any of the genetic variations of KANSL1 in major AD cohorts, such as Alzheimer's Disease Neuroimaging Initiative and AddNeuroMed [74, 75]. Thus, future work can include measurements of these genetic variations as well as their expression in these and other independent cohorts. Using these quantitative measurements, if available, several tools can be employed to elucidate pathway signatures in disease as well drug-perturbed states, which can then be used to prioritize drug candidates relevant to the particular disease under investigation when these signatures are anti-correlated [76]. Similarly, gene expression measurements paired with a network containing prior knowledge on drug-disease data can also be used for drug candidate identification [77]. Finally, looking ahead, the presented strategy can be applied to other AD genes or across disease domains such as psychiatric diseases.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support from the B-IT foundation.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-1397r2).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD201397.

REFERENCES

[1] Heemels MT (2016) Neurodegenerative diseases. *Nature* **539**, 179.

[2] Ahmad K, Hassan Baig M, Mushtaq G, Amjad Kamal M, Greig NH, Choi I (2017) Commonalities in biological pathways, genetics, and cellular mechanism between Alzheimer's disease and other neurodegenerative diseases: An in silico-updated overview. *Curr Alzheimer Res* **14**, 1190-1197.

[3] Rajput AH, Rozdilsky B, Rajput A (1993) Alzheimer's disease and idiopathic Parkinson's disease coexistence. *J Geriatr Psychiatry Neurol* **6**, 170-176.

[4] Braskie MN, Ringman J M, Thompson PM (2011) Neuroimaging measures as endophenotypes in Alzheimer's disease. *Int J Alzheimers Dis* **2011**, 490140.

[5] Scheltens P, van de Pol L (2012) Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: Diagnostic value and neuropsychological correlates. *J Neurol Neurosurg Psychiatry* **83**, 1038-1040.

[6] Iyappan A, Younesi E, Redolfi A, Vrooman H, Khanna S, Frisoni GB, Hofmann-Apitius M (2017) Neuroimaging feature terminology: A controlled terminology for the annotation of brain imaging features. *J Alzheimers Dis* **59**, 1153-1169.

[7] Meyer-Lindenberg A, Nicodemus KK, Egan MF, Callicott JH, Mattay V, Weinberger DR (2008) False positives in imaging genetics. *Neuroimage* **40**, 655-661.

[8] Wachinger C, Nho K, Saykin AJ, Reuter M, Rieckmann A, Alzheimer's Disease Neuroimaging Initiative (2018) A longitudinal imaging genetics study of neuroanatomical asymmetry in Alzheimer's disease. *Biol Psychiatry* **84**, 522-530.

[9] Biffi A, Anderson CD, Desikan RS, Sabuncu M, Cortellini L, Schmansky N, Salat D, Rosand J; Alzheimer's Disease Neuroimaging Initiative (ADNI) (2010) Genetic variation and neuroimaging measures in Alzheimer disease. *Arch Neurol* **67**, 677-685.

[10] Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW,

Grenier-Boley B, Russo G, Thorton-Wells TA, Jones N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Gerrish A, Schmidt H, Kunkle B, Dunstan ML, Ruiz A, Bihoreau MT, Choi SH, Reitz C, Pasquier F, Cruchaga C, Craig D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Morón FJ, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fiévet N, Huentelman MW, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuiness B, Larson EB, Green R, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossú P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Sanchez-Garcia F, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Mancuso M, Matthews F; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Bullido M, Panza F, Caffarra P, Nacmias B, Gilbert JR, Mayhaus M, Lannefelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH Jr, Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltuenen M, Martin ER, Schmidt R, Rujescu D, Wang LS, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nöthen MM, Graff C, Psaty BM, Jones L, Haines JL, Holmans PA, Lathrop M, Pericak-Vance MA, Launer LJ, Farrer LA, van Duijn CM, Van Broeckhoven C, Moskvina V, Seshadri S, Williams J, Schellenberg GD, Amouyel P (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genet* **45**, 1452.

[11] Ruiz A, Heilmann S, Becker T, Hernández I, Wagner H, Thelen M, Mauleón A, Rosende-Roca M, Bellenguez C, Bis JC, Harold D, Gerrish A, Sims R, Sotolongo-Grau O, Espinosa A, Alegret M, Arrieta JL, Lacour A, Leber M, Becker J, Lafuente A, Ruiz S, Vargas L, Rodríguez O, Ortega G, Dominguez MA; IGAP, Mayeux R, Haines JL, Pericak-Vance MA, Farrer LA, Schellenberg GD, Chouraki V, Launer LJ, van Duijn C, Seshadri S, Antúnez C, Breteler MM, Serrano-Ríos M, Jessen F, Tárraga L, Nöthen MM, Maier W, Boada M, Ramírez A (2014) Follow-up of loci from the International Genomics of Alzheimer's Disease Project identifies TRIP4 as a novel susceptibility gene. *Transl Psychiatry* **4**, e358.

[12] Chan S L, Kim W S, Kwok J B, Hill A F, Cappai R, Rye K A, Garner B (2008) ATP binding cassette transporter A7 regulates processing of amyloid precursor protein *in vitro*. *J Neurochem* **106**, 793-804.

[13] Casey BJ, Soliman F, Bath KG, Glatt CE (2010) Imaging genetics and development: Challenges and promises. *Hum Brain Mapp* **31**, 838-851.

[14] Turner BO, Paul EJ, Miller MB, Barbey AK (2018) Small sample sizes reduce the replicability of task-based fMRI studies. *Commun Biol* **1**, 62.

[15] Stefanovski L, Triebkorn P, Spiegler A, Diaz-Cortes MA, Solodkin A, Jirsa V, McIntosh AR, Ritter P; Alzheimer's Disease Neuroimaging Initiative (2019) Linking molecular pathways and large-scale computational modeling to assess candidate disease mechanisms and pharmacodynamics in Alzheimer's Disease. *Front Comput Neurosci* **13**, 54.

[16] Slater T (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov Today* **19**, 193-198.

[17] Neve RL, McPhie DL (2006) The cell cycle as a therapeutic target for Alzheimer's disease. *Pharmacol Ther* **111**, 99-113.

[18] Cenini G, Voos W (2019) Mitochondria as potential targets in Alzheimer's disease therapy: An Update. *Front Pharmacol* **10**, 192.

[19] Hoyt CT, Konotopez A, Ebeling C (2018) PyBEL: A computational framework for Biological Expression Language. *Bioinformatics* **34**, 703-704.

[20] Domingo-Fernández D, Kodamullil AT, Iyappan A, Naz M, Emon MA, Raschka T, Karki R, Springstubbe S, Ebeling C, Hofmann-Apitius M (2017) Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): A web server for mechanism enrichment. *Bioinformatics* **33**, 3679-3681.

[21] Li H, Roossinck MJ (2004) Genetic bottlenecks reduce population variation in an experimental RNA virus population. *J Virol* **78**, 10582-10587.

[22] Lewis C M, Knight J (2012) Introduction to genetic association studies. *Cold Spring Harb Protoc* **2012**, 297-306.

[23] Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2542.

[24] Ward LD, Kellis M (2012) HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, 930-934.

[25] Piñero J, Bravo Á, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI (2017) DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* **45**, 833-839.

[26] Tyler AL, Crawford DC, Pendergrass SA (2016) The detection and characterization of pleiotropy: Discovery, progress, and promise. *Brief Bioinform* **17**, 13-22.

[27] Nussbaum RL, Ellis CE (2003) Alzheimer's disease and Parkinson's disease. *N Engl J Med* **348**, 1356-1364.

[28] Catalá-López F, Crespo-Facorro B, Vieta E, Valderas JM, Valencia A, Tabarés-Seisdedos R (2014) Alzheimer's disease and cancer: Current epidemiological evidence for a mutual protection. *Neuroepidemiology* **42**, 121-122.

[29] Rosen AR, Steenland NK, Hanfelt J, Factor SA, Lah JJ, Levey AI (2007) Evidence of shared risk for Alzheimer's disease and Parkinson's disease using family history. *Neurogenetics* **8**, 263-70.

[30] Vacic V, Ozelius LJ, Clark LN, Bar-Shira A, Gana-Weisz M, Gurevich T, Gusev A, Kedmi M, Kenny EE, Liu X, Mejia-Santana H, Mirelman A, Raymond D, Saunders-Pullman R, Desnick RJ, Atzmon G, Burns ER, Ostrer H, Hakonarson H, Bergman A, Barzilai N, Darvasi A, Peter I, Guha S, Lencz T, Giladi N, Marder K, Pe'er I, Bressman SB, Orr-Urtreger A (2014) Genome-wide mapping of IBD segments in an

Ashkenazi PD cohort identifies associated haplotypes. *Hum Mol Genet* **23**, 4693-4702.

[31] Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjeide BM, Schjeide LM, Meissner E, Zauft U, Allen NC, Liu T, Schilling M, Anderson KJ, Beecham G, Berg D, Biernacka JM, Brice A, DeStefano AL, Do CB, Eriksson N, Factor SA, Farrer MJ, Foroud T, Gasser T, Hamza T, Hardy JA, Heutink P, Hill-Burns EM, Klein C, Latourelle JC, Maraganore DM, Martin ER, Martinez M, Myers RH, Nalls MA, Pankratz N, Payami H, Satake W, Scott WK, Sharma M, Singleton AB, Stefansson K, Toda T, Tung JY, Vance J, Wood NW, Zabetian CP; 23andMe Genetic Epidemiology of Parkinson's Disease Consortium; International Parkinson's Disease Genomics Consortium; Parkinson's Disease GWAS Consortium; Wellcome Trust Case Control Consortium, Young P, Tanzi RE, Khoury MJ, Zipp F, Lehrach H, Ioannidis JP, Bertram L (2012) Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. *PLoS Gene* **8**, e1002548.

[32] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2016) KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, 353-361.

[33] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P (2018) The reactome pathway knowledgebase. *Nucleic Acids Res* **46**, 649-655.

[34] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL (2018) WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* **46**, 661-667.

[35] Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, Naj AC, Kunkle BW, Wang LS, Bis JC, Bellenguez C, Harold D, Lunetta KL, Destefano AL, Grenier-Boley B, Sims R, Beecham GW, Smith AV, Chouraki V, Hamilton-Nelson KL, Ikram MA, Fievet N, Denning N, Martin ER, Schmidt H, Kamatani Y, Dunstan ML, Valladares O, Laza AR, Zelenika D, Ramirez A, Foroud TM, Choi SH, Boland A, Becker T, Kukull WA, van der Lee SJ, Pasquier F, Cruchaga C, Beekly D, Fitzpatrick AL, Hanon O, Gill M, Barber R, Gudnason V, Campion D, Love S, Bennett DA, Amin N, Berr C, Tsolaki M, Buxbaum JD, Lopez OL, Deramecourt V, Fox NC, Cantwell LB, Tárraga L, Dufouil C, Hardy J, Crane PK, Eiriksdottir G, Hannequin D, Clarke R, Evans D, Mosley TH Jr, Letenneur L, Brayne C, Maier W, De Jager P, Emilsson V, Dartigues JF, Hampel H, Kamboh MI, de Bruijn RF, Tzourio C, Pastor P, Larson EB, Rotter JI, O'Donovan MC, Montine TJ, Nalls MA, Mead S, Reiman EM, Jonsson PV, Holmes C, St George-Hyslop PH, Boada M, Passmore P, Wendland JR, Schmidt R, Morgan K, Winslow AR, Powell JF, Carasquillo M, Younkin SG, Jakobsdóttir J, Kauwe JS, Wilhelmsen KC, Rujescu D, Nöthen MM, Hofman A, Jones L; IGAP Consortium, Haines JL, Psaty BM, Van Broeckhoven C, Holmans P, Launer LJ, Mayeux R, Lathrop M, Goate AM, Escott-Price V, Seshadri S, Pericak-Vance MA, Amouyel P, Williams J, van Duijn CM, Schellenberg GD, Farrer LA (2016) A novel Alzheimer's disease locus located near the gene encoding tau protein. *Mol Psychiatry* **21**, 108-117.

[36] Todorova V, Blokland A (2017) Mitochondria and synaptic plasticity in the mature and aging nervous system. *Curr Neuropharmacol* **15**, 166-173.

[37] Cai Y, Jin J, Swanson SK, Cole MD, Choi SH, Florens L, Washburn MP, Conaway JW, Conaway RC (2010) Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex. *J Biol Chem* **285**, 4268-4272.

[38] Huang J, Wan B, Wu L, Yang Y, Dou Y, Lei M (2012) Structural insight into the regulation of MOF in the male-specific lethal complex and the non-specific lethal complex. *Cell Res* **22**, 1078-1081.

[39] Gregory PD, Wagner K, Hörz W (2001) Histone acetylation and chromatin remodeling. *Exp Cell Res* **265**, 195-202.

[40] Peixoto L, Abel T (2013) The role of histone acetylation in memory formation and cognitive impairments. *Neuropsychopharmacology* **38**, 62-76.

[41] Bahari-Javan S, Sananbenesi F, Fischer A (2014) Histone-acetylation: A link between Alzheimer's disease and post-traumatic stress disorder? *Front Neurosci* **8**, 160.

[42] Farioli-Vecchioli S, Saraulli D, Costanzi M, Leonardi L, Ciná I, Micheli L, Nutini M, Longone P, Oh SP, Cestari V, Tirone F (2009) Impaired terminal differentiation of hippocampal granule neurons and defective contextual memory in PC3/Tis21 knockout mice. *PLoS One* **4**, e8339.

[43] Leal G, Bramham CR, Duarte CB (2017) BDNF and hippocampal synaptic plasticity. *Vitam Horm* **104**, 153-195.

[44] Myers KA, McGlade A, Neubauer BA, Lal D, Berkovic SF, Scheffer IE, Hildebrand MS (2018) KANSL1 variation is not a major contributing factor in self-limited focal epilepsy syndromes of childhood. *PLoS One* **13**, e0191546.

[45] Soliman ML, Smith MD, Houdek HM, Rosenberger TA (2012) Acetate supplementation modulates brain histone acetylation and decreases interleukin-1β expression in a rat model of neuroinflammation. *J Neuroinflammation* **9**, 51.

[46] Vadnal J, Houston S, Bhatta S, Freeman E, McDonough J (2012) Transcriptional signatures mediated by acetylation overlap with early-stage Alzheimer's disease. *Exp Brain Res* **221**, 287-297.

[47] Kim S, Kaang BK (2017) Epigenetic regulation and chromatin remodeling in learning and memory. *Exp Mol Med* **49**, e281.

[48] Moreno-Igoa M, Hernández-Charro B, Bengoa-Alonso A, Pérez-Juana-del-Casal A, Romero-Ibarra C, Nieva-Echebarria B, Ramos-Arroyo MA (2015) KANSL1 gene disruption associated with the full clinical spectrum of 17q21.31 microdeletion syndrome. *BMC Med Genet* **16**, 68.

[49] Rouault JP, Falette N, Guéhenneux F, Guillot C, Rimokh R, Wang Q, Berthet C, Moyret-Lalle C, Savatier P, Pain B, Shaw P, Berger R, Samarut J, Magaud JP, Ozturk M, Samarut C, Puisieux A (1996) Identification of BTG2, an antiproliferative p53-dependent component of the DNA damage cellular response pathway. *Nat Genet* **14**, 482-486.

[50] Rouault JP, Prévôt D, Berthet C, Birot AM, Billaud M, Magaud JP, Corbo L (1998) Interaction of BTG1 and p53-regulated BTG2 gene products with mCaf1, the murine homolog of a component of the yeast CCR4 transcriptional regulatory complex. *J Biol Chem* **273**, 22563-22569.

[51] Terra R, Luo H, Qiao X, Wu J (2008) Tissue-specific expression of B-cell translocation gene 2 (BTG2) and its function in T-cell immune responses in a transgenic mouse model. *Int Immunol* **20**, 317-326.

[52] Rossi G, Dalprá L, Crosti F, Lissoni S, Sciacca FL, Catania M, Di Fede G, Mangieri M, Giaccone G, Croci D, Tagliavini

F (2008) A new function of microtubule-associated protein tau: Involvement in chromosome stability. *Cell Cycle* **7**, 1788-1794.

[53] Jordán J, Galindo MF, Prehn JH, Weichselbaum RR, Beckett M, Ghadge GD, Roos RP, Leiden JM, Miller RJ (1997) p53 expression induces apoptosis in hippocampal pyramidal neuron cultures. *J Neurosci* **17**, 1397-1405.

[54] Mattson MP, Liu D (2002) Energetics and oxidative stress in synaptic plasticity and neurodegenerative disorders. *Neuromolecular Med* **2**, 215-231.

[55] Hroudová J, Fišar Z (2013) Control mechanisms in mitochondrial oxidative phosphorylation. *Neural Regen Res* **8**, 363-375.

[56] Chatterjee A, Seyfferth J, Lucci J, Gilsbach R, Preissl S, Böttinger L, Mårtensson CU, Panhale A, Stehle T, Kretz O, Sahyoun AH, Avilov S, Eimer S, Hein L, Pfanner N, Becker T, Akhtar A (2016) MOF acetyl transferase regulates transcription and respiration in mitochondria. *Cell* **167**, 722-738.

[57] Stryer L, L Tymoczko J, M Berg J (2002) *Biochemistry, Fifth Edition*. W.H. Freeman and Company, ISBN-10:0-7167-3051-0.

[58] Johns P (2016) *Clinical Neuroscience*. Churchill Livingstone, eBook ISBN: 9780702057137

[59] Stepien KM, Heaton R, Rankin S, Murphy A, Bentley J, Sexton D, Hargreaves IP (2017) Evidence of oxidative stress and secondary mitochondrial dysfunction in metabolic and non-metabolic disorders. *J Clin Med* **6**, 71.

[60] Koudinov AR, Koudinova NV (2001) Essential role for cholesterol in synaptic plasticity and neuronal degeneration. *FASEB J* **15**, 1858-1860.

[61] Wang D, Zheng W (2015) Dietary cholesterol concentration affects synaptic plasticity and dendrite spine morphology of rabbit hippocampal neurons. *Brain Res* **1622**, 350-360.

[62] Koudinov AR, Koudinova NV (2003) Cholesterol, synaptic function and Alzheimer's disease. *Pharmacopsychiatry* **36**(Suppl 2), S107-112.

[63] Cutler RG, Kelly J, Storie K, Pedersen WA, Tammara A, Hatanpaa K, Troncoso JC, Mattson MP (2004) Involvement of oxidative stress-induced abnormalities in ceramide and cholesterol metabolism in brain aging and Alzheimer's disease. *Proc Natl Acad Sci U S A* **101**, 2070-2075.

[64] Zhou Q, Homma KJ, Poo MM (2004) Shrinkage of dendritic spines associated with long-term depression of hippocampal synapses. *Neuron* **44**, 749-57.

[65] Whalley K (2007) Balancing LTP and LTD. *Neuroscience* **8**, 249-249.

[66] Edelmann E, Lessmann V, Brigadski T (2014) Pre- and post-synaptic twists in BDNF secretion and action in synaptic plasticity. *Neuropharmacology* **76**, 610-627.

[67] Bagheri A, Habibzadeh P, Razavipour SF, Volmar CH, Chee NT, Brothers SP, Wahlestedt C, Mowla SJ, Faghihi MA (2019) HDAC inhibitors induce BDNF expression and promote neurite outgrowth in human neural progenitor cells-derived neurons. *Int J Mol Sci* **20**, 1109.

[68] Malenka RC, Bear MF (2004) LTP and LTD: An embarrassment of riches. *Neuron* **44**, 5-21.

[69] Geetha T, Zheng C, McGregor WC, White BD, Diaz-Meco MT, Moscat J, Babu JR (2012) TRAF6 and p62 inhibit amyloid β-induced neuronal death through p75 neurotrophin receptor. *Neurochem Int* **61**, 1289-1293.

[70] Koss DJ, Robinson L, Drever BD, Plucińska K, Stoppelkamp S, Veselcic P, Riedel G, Platt B (2016) Mutant Tau knock-in mice display frontotemporal dementia relevant behavior and histopathology. *Neurobiol Dis* **91**, 105-123.

[71] Rodríguez-Martín T, Pooler AM, Lau DHW, Mórotz GM, De Vos KJ, Gilley J, Coleman MP, Hanger DP (2016) Reduced number of axonal mitochondria and tau hypophosphorylation in mouse P301L tau knockin neurons. *Neurobiol Dis* **85**, 1-10.

[72] Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-1797.

[73] Dong S, Boyle AP (2019) Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum Mutat* **40**, 1292-1298.

[74] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005) The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am* **15**, 869-877.

[75] Lovestone S, Francis P, Kloszewska I, Mecocci P, Simmons A, Soininen H, Spenger C,Tsolaki M, Vellas B, Wahlund LO, Ward M; AddNeuroMed Consortium (2009) AddNeuroMed — the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Ann N Y Acad Sci* **1180**, 36-46.

[76] Emon MA, Domingo-Fernández D, Hoyt CT, Hofmann-Apitius M (2020) PS4DR: A multimodal workflow for identification and prioritization of drugs based on pathway signatures. *BMC Bioinformatics* **21**, 1-21.

[77] Peyvandipour A, Saberian N, Shafi A, Donato M, Draghici S (2018) A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **34**, 2817-2825.

7.4 Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures

Reprinted with permission from "Golriz Khatami, S. *et al.*. "Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures", *npj Syst Biol Appl* 7(1), 40, (2021).

npj Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

Check for updates

Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures

Sepehr Golriz Khatami[1,2✉], Sarah Mubeen[1,2,3], Vinay Srinivas Bharadhwaj[1,2], Alpha Tom Kodamullil[1], Martin Hofmann-Apitius[1,2] and Daniel Domingo-Fernández[1,3,4✉]

The utility of pathway signatures lies in their capability to determine whether a specific pathway or biological process is dysregulated in a given patient. These signatures have been widely used in machine learning (ML) methods for a variety of applications including precision medicine, drug repurposing, and drug discovery. In this work, we leverage highly predictive ML models for drug response simulation in individual patients by calibrating the pathway activity scores of disease samples. Using these ML models and an intuitive scoring algorithm to modify the signatures of patients, we evaluate whether a given sample that was formerly classified as diseased, could be predicted as normal following drug treatment simulation. We then use this technique as a proxy for the identification of potential drug candidates. Furthermore, we demonstrate the ability of our methodology to successfully identify approved and clinically investigated drugs for four different cancers, outperforming six comparable state-of-the-art methods. We also show how this approach can deconvolute a drugs' mechanism of action and propose combination therapies. Taken together, our methodology could be promising to support clinical decision-making in personalized medicine by simulating a drugs' effect on a given patient.

npj Systems Biology and Applications (2021)7:40 ; https://doi.org/10.1038/s41540-021-00199-1

INTRODUCTION

Applying machine learning (ML) methods to biomedical data has enormous potential for the development of personalized therapies,[1] drug repurposing,[2] and drug discovery.[3] The data exploited by these methods can comprise multiple modalities including imaging data,[4] chemical structure information,[5] and natural language data.[6] However, the widespread availability of transcriptomics data (e.g., RNA-Sequencing (RNA-Seq), microarrays, etc.) along with its capacity to provide a comprehensive overview of biological systems have made this particular modality a popular choice for various computational methods. Although this modality can reveal both molecular signatures as well as phenotypic changes that occur in altered states, pathway analyses are often performed to map measured transcripts to the pathway level due to high dimensionality and correlations present in transcriptomics datasets.[7,8] This transformation facilitates the training of ML/AI models by reducing dimensional complexity whilst enhancing interpretive power.[9] However, such a transformation implicates the use of prior pathway knowledge[10] from databases such as KEGG[11] and Reactome.[12,13]

The transformation of data from the transcriptomics to the pathway level can be used to generate pathway features (i.e., sets of genes involved in a given pathway that are coordinately up or down-regulated), the latter of which have broad applications in drug discovery and drug response prediction.[14] For instance,[15–17] exploited the concept of anti-similarity between drugs and disease-specific pathway signatures to identify therapeutic candidate drugs that can potentially revert disease pathophysiology. Furthermore,[18] shows how pathway signatures derived from cell lines using kernelized Bayesian matrix factorization can be used for drug response prediction.

Alternatively, other methods can generate individualized pathway features from a population of patients or cell lines.[19] These features, or pathway activity scores, can subsequently be used for several downstream ML applications including classification tasks and survival prediction.[8,20] In addition,[21] showed how ML models can be used to predict drug response using pathway activity scores derived from cell lines. Furthermore, another example from[22] demonstrated how modeling individualized pathway activity scores from Fanconi anemia patients can reveal potential targets for therapeutic interventions. Finally, similar approaches have been used to prioritize drug treatments in the cancer context.[23,24]

While these methods have shown how pathway signatures can be used for drug discovery and drug response prediction, existing methods thus far fail to account for two important factors. First, as the response triggered by a drug in a given patient may differ if administered in another, these methods should account for patient heterogeneity which is crucial in designing individualized therapies. Second, specific indications may be improved or corrected by a drug combination approach or through the administration of multi-target drugs.

In this work, we present an intuitive methodology that exploits the predictive power of ML models to simulate drug response by calibrating pathway signatures of patients. We first trained an ML model (i.e., elastic net penalized logistic regression model) to discriminate between disease samples and controls based on sample-specific pathway activity scores. Next, we simulate drug responses in patients using a scoring algorithm that modifies a patient's pathway signatures using existing knowledge on drug-target interactions. We hypothesize that promising drug candidates for a given condition would modify pathway activity scores of patients in such a way that they closely resemble scores of

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53757, Germany. [2]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115 Bonn, Germany. [3]Fraunhofer Center for Machine Learning, Sankt Augustin, Germany. [4]Enveda Biosciences, Boulder, CO 80301, USA. ✉email: sepehr.golriz.khatami@scai.fraunhofer.de; daniel.domingo.fernandez@scai.fraunhofer.de

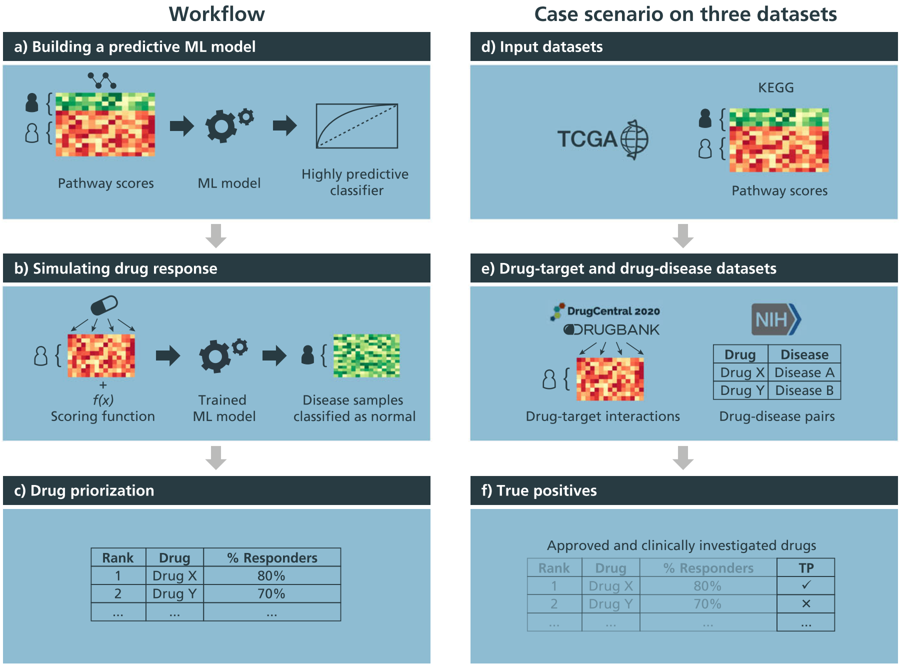


Fig. 1 Conceptual overview of the drug simulation workflow and case scenario on multiple datasets. (**a**) Pathway activity scores are used to train a highly predictive ML model that differentiates between normal and disease samples, labeled green and red on the heatmap, respectively. (**b**) Next, pathway scores of disease samples are modified by using drug-target information and applying a scoring algorithm that simulates the effect of a given drug at the pathway-level. Using the modified pathway scores of disease samples, the trained ML classifier is then used to evaluate whether these modified disease samples that were previously classified as "diseased" could now be classified as "normal". (**c**) Finally, we use the proportion of disease samples now classified as normal (i.e., % responders) as a proxy to identify candidate drugs and propose combination therapies. (**d**) To demonstrate the methodology in a case scenario, we first performed ssGSEA using pathways from KEGG and the BRCA, LIHC, PRAD, and KIRC TCGA datasets to acquire sample-wise pathway activity scores. (**e**) Next, we obtained known drug-target interactions from DrugBank and DrugCentral and drug-disease pairs (i.e., FDA-approved drugs and drugs under clinical trials for a given condition) from Clinicaltrials.gov and FDA-approved drugs, of which, the latter two were used as a ground-truth list of true positives (TP). (**f**) To simulate drug treatments of patients from the aforementioned TCGA datasets using their pathway activity scores (i.e., Fig. 1d), we applied the methodology described in Fig. 1a–c to acquire a ranking of drugs based on the proportion of disease samples that were treated. Finally, we identified the proportion of drugs ranked by our methodology that were true positives for the four TCGA datasets and compared this proportion to random chance.

controls. Thus, using the previously trained ML model, we then evaluate whether patients with modified pathway scores are now classified as normal as a proxy for promising drug candidates. We demonstrate the scalability and generalizability of our methodology by simulating over one thousand drugs from two independent drug-target datasets on four cancer indications. Furthermore, we show how our methodology is able to recover a large proportion of clinically investigated drugs on these four indications, outperforming six comparable state-of-the-art methods. Finally, we show how the most relevant pathways identified by our methodology can be used to better understand the biology pertaining to a given condition.

RESULTS

We present a workflow designed to approximate a drug's effect on a patient by intentional modifications to patient-specific features, specifically, pathway activity scores, by employing highly predictive ML models trained to differentiate between normal and disease samples (Fig. 1). In the first subsection, we validate our approach by (i) evaluating its capability in retrieving FDA-approved drugs and those in clinical trials for multiple cancer datasets and, (ii) comparing the results yielded by our approach against several equivalent methods. Then, in the following two subsections, we investigate the drug candidates prioritized by our approach and the specific pathways targeted by these prioritized drugs, respectively. Finally, we show the utility of our approach in predicting the effects of a combination of drugs for applications in combination therapy and for the identification of potential adverse events associated with drug combinations.

Validation of the methodology and comparison against equivalent approaches

In this subsection, we investigate the drug candidates prioritized by our methodology in four different cancers and evaluate the ability of our approach to identify approved and clinically investigated drugs (i.e., true positives). Table 1 shows that only a minority of the drugs present in both drug-target datasets were prioritized by our methodology given that a stringent threshold was employed which required that prioritized drugs change the predictions of at least 80% of the patients (see "Materials and Methods" and Supplementary Figs. 7, and 8 for details on the selection of this threshold). Overall, our methodology is able to recover a large proportion of true positives (ranging from 13% to 32%) in all four cancers as well as in both drug-target datasets (Table 1). This wide range may be attributable to a disproportion in the number of true positives that exist for each of the cancer datasets (e.g., BRCA has more than twice as many FDA-approved drugs and drugs in clinical trials than LIHC) as well as to the size of the drug-target datasets (i.e., DrugBank contains twice as many drugs as DrugCentral).

As a comparison, the methodology proposed by[25] reported lower proportions of true positives than our approach for the BRCA and PRAD datasets with 21.42% and 15.94%, respectively (Supplementary Table 1). Furthermore, four additional methods present that were benchmarked by[25] yielded even lower results

Table 1. Number of FDA-approved and clinically tested drugs recovered for both drug-target datasets (i.e., DrugBank (DB) and DrugCentral (DC)) across the four investigated cancers.

Dataset	DB Prioritized	DB Approved (total)	DB Clinical trials (total)	DB Proportion of true positives (%)	DC Prioritized	DC Approved (total)	DC Clinical trials (total)	DC Proportion of true positives (%)
BRCA	129	8 (26)	23 (182)	31/129 (24.03%)	19	2 (14)	4 (115)	6/19 (31.57%)
LIHC	74	2 (5)	11 (50)	13/74 (17.56%)	19	1 (1)	2 (35)	3/19 (15.78%)
PRAD	68	2 (13)	18 (134)	20/68 (29.41%)	19	1 (7)	3 (84)	4/19 (21.05%)
KIRC	88	2 (8)	10 (44)	12/88 (13.63%)	26	3 (3)	2 (25)	5/26 (19.2%)

In the first column for each drug-target dataset ("Prioritized"), we report the number of drugs that changed the predictions for at least 80% of the patients for each cancer type. The second column ("Approved") reports the number of FDA-approved drugs among these prioritized drugs as well as the total number of FDA-approved/clinically tested drugs present in each dataset between parentheses. Similarly, the third column ("Clinical trials") reports the number of drugs tested in clinical trials among the prioritized drugs and the total number of FDA-approved/clinically tested drugs between parentheses. Finally, the last column ("Proportion of true positives") reports the proportion of true positives (both FDA-approved and clinically tested drugs) among the prioritized drugs.

on the same two cancer datasets (Supplementary Tables 2–8). Similarly,[26] also reported a lower proportion of true positives than our approach for the BRCA and PRAD datasets with 0.8% and 0.4%, respectively (Supplementary Table 9). Overall, the performance across all six methods varied from 0% to 11.53% for BRCA, and from 0.50% to 22.22% for PRAD and is summarized in Supplementary Table 10.

In addition, the proportion of true positives yielded by our methodology is significantly higher than what one would expect by chance (see "Materials and Methods"). Furthermore, we compared the number of prioritized drugs found in the original DrugBank and DrugCentral datasets to the number of prioritized drugs obtained in the robustness experiments in which we applied our methodology to drugs with randomly generated targets and target interactions (Supplementary Fig. 1). We found that all permutation experiments yielded a significantly lower number of prioritized drugs. Because our methodology can capture a much greater number of prioritized drugs on a real dataset, this validation highlights the capability of our approach to prioritize drugs with targets in relevant pathways that are key to change the predictions of patients.

As a final remark, we explored the performance of our methodology when varying one of the weights while keeping the other two constant to better understand how sensitive the results are to the selected weights **(**Supplementary Tables 11, 12). We have observed that the proportions of true positives recovered mainly vary between 15% and 35% in the three test disease datasets for both drug-target datasets when W_1 (i.e., the weight assigned to the quartile that contains the most dysregulated pathways) is in the range of 10–20. There are multiple cases where we found sets of weights yielding better results than the ones presented in Table 1 if exclusively looking at a single or two specific disease datasets (Supplementary Table 13). In contrast, we observed that when weights are low (e.g., $W_1 = 1$), our approach often does not yield any prioritized drugs (Supplementary Table 14), as in these cases, the modified pathway activity scores are not sufficient enough to change the predictions of the ML model.

In-depth investigation of the prioritized candidate drugs

Apart from the previous quantitative evaluation of our methodology, we conducted an in-depth analysis of the prioritized drugs to better understand the predictions made by our approach. Below, we focus on drugs prioritized using the DrugCentral dataset as this dataset contains a fewer number of prioritized drugs than DrugBank.

In the breast cancer dataset (BRCA), we identified a major class of drugs based on their mechanisms of action (Fig. 2a). This class targeted DNA and RNA metabolism and included commonly used anti-tumor drugs. One example of this group of drugs is fluorouracil, which targets thymidylate synthase, thereby inhibiting the formation of thymidylate from uracil.[27] This drug is a chemotherapy medication commonly used to treat several cancers.

In the prostate cancer dataset (PRAD), we found that the majority of drugs were related to hormone metabolism and regulation (Fig. 2c). Due to the key role of sex steroid hormones in its initiation and progression,[26] this cancer is classified as hormone-dependent. Thus, current treatments are often directly targeted towards these hormones, such as androgen deprivation therapy, which represents the major therapeutic option for treatment of advanced stages of this cancer.[28–30]

The third dataset, LIHC, corresponds to hepatocarcinoma. Interestingly, the vast majority of the candidate drugs in this dataset (14/19) are tyrosine kinase inhibitors (TKI) corresponding to anti-tumor drugs already FDA-approved for other cancers[31] (Fig. 2b). Since these kinases act as regulatory players in several cancer signaling pathways that can be hyperactivated, TKIs are used to "switch-off" these pathways, indirectly inhibiting cell growth.[32] One of the predicted drugs is sorafenib, which was the first TKI to be approved for the treatment of liver carcinoma and still remains as a first-line therapy. Similarly, another predicted drug, trametinib, is a dual-kinase inhibitor that is used in the treatment of advanced liver cancer. Finally, two of the remaining non-TKIs are also employed as chemotherapy drugs as they inhibit the synthesis of nucleotides.

Investigation of pathways targeted by the prioritized drugs

Here, we interpret and analyze the results yielded by our methodology for multiple datasets by investigating the pathways targeted by the drugs prioritized through our approach. We identified clusters of pathways belonging to several distinct classes (Fig. 2). Not surprisingly, we found that various metabolic pathways appeared in all three test datasets as the regulation of metabolism plays an important role in numerous cancers. Given that each of the three test datasets were cancer subtypes, intuitively, we also observed several disease-relevant pathways targeted by the prioritized drugs, among which were ~30 cancer-related pathways from KEGG (e.g., prostate cancer, pancreatic cancer, bladder cancer, and breast cancer).

Drugs that were prioritized by our approach (Fig. 2) were likewise clustered based on the pathways they targeted to assess whether drugs that targeted the same pathway fell within the same class of drugs. Prioritized drugs for liver cancer could be clustered into four different classes of tyrosine kinase inhibitors: (i) JAK inhibitors (i.e., sorafenib, vandetanib, erlotinib, and lapatinib), (ii) ALK inhibitors (i.e., lorlatinib), (iii) BCR–Abl (i.e., nilotinib, dasatinib, and imatinib), and (iv) and EGFR inhibitors (i.e., afatinib).[33] In addition, we found MEK kinase inhibitors, specifically

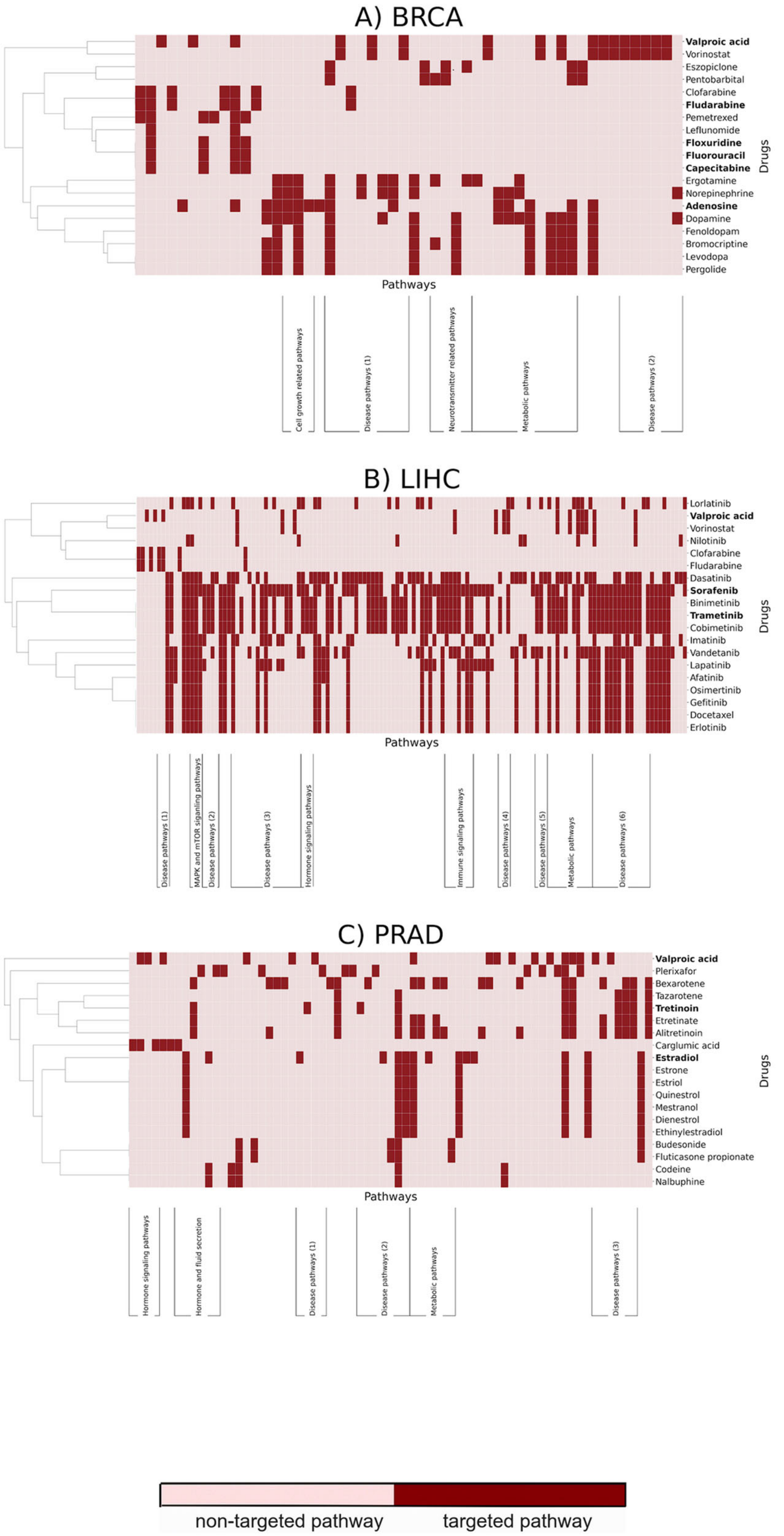


Fig. 2 Pathways targeted by prioritized drugs in DrugCentral for each of the three cancer test datasets. The X axis corresponds to pathways targeted by any of the prioritized drugs (i.e., pathways not targeted by any prioritized drug are omitted for better visualization). Prioritized drugs for each cancer dataset have been clustered based on the pathways they target and are reported on the Y axis. Of the prioritized drugs, those that correspond to true positives are highlighted in bold. If a set of three or more similar pathways was clustered together, we manually assigned these pathways into distinct classes (Y axis) Pathway names and cluster information are available as a Supplementary File and the equivalent figures for DrugBank are available as Supplementary Figs. 2–4.

Table 2. Examples of predicted combination therapies.

Cancer type	Drug 1	Drug 2	Proportion of responders (%)	Reference
Liver cancer	Sorafenib	Trametinib	87%	[53]
Liver cancer	Erlotinib	Sorafenib	87%	[54]
Breast cancer	Vorinostat	Capecitabine	88%	[55]

trametinib and cobimetinib. Finally, we found that while some drugs were able to change the predictions by targeting only a limited number of pathways (e.g., fludarabine in breast cancer and liver cancer), other drugs could change predictions by targeting several pathways (e.g., tretinoin in prostate cancer and trametinib in liver cancer).

Among the most commonly targeted pathways by the prioritized drugs in liver carcinoma, we found Ras/Raf/MAPK and PI3K/AKT/mTOR signaling, both of which have been reported to play important roles in the development of this type of cancer.[34] One of the prioritized drugs, sorafenib, is a multi-kinase inhibitor that targets several kinases including RFA1, PDGFR, and FLT3, which are involved in both tumor proliferation and angiogenesis.[35,36] Sorafenib has been shown to inhibit tumor cell proliferation by blocking the Ras/Raf/MAPK pathway and to inhibit angiogenesis by blocking PDGFR signaling[37] (Supplementary Table 15).

Prioritizing combination therapies

Combination therapies are widely used for treating indications like cancer as they can often lead to the inhibition of the compensatory signaling pathways that maintain the growth and survival of tumor cells. Here, we demonstrate how our methodology can be extended to predict the effects of a combination of drugs. To reduce the computational complexity associated with running our methodology on all possible combinations of drug pairs from both drug-target datasets (i.e., DrugBank and DrugCentral), we exclusively applied our method on all possible pairs from the set of prioritized drugs. Table 2 lists a subset of combinations of prioritized drugs, alongside the proportion of patients that they reclassify as normal (i.e., proportion of treated patients).

For two of the three test datasets (i.e, LIHC and PRAD), nearly all drug pairs yielded better results (i.e., larger proportion of disease samples predicted as normal) than the use of a single drug alone. In the BRCA dataset, however, multiple combinations yielded worse results than those observed with single drug therapy. For example, the combination of bromocriptine with valproic acid decreased the proportion of treated patients from 80% to <10%. Specifically, bromocriptine is an adrenergic receptor agonist that stimulates the beta-adrenergic signaling pathway, which in turn prompts tumor angiogenesis and cancer development.[38] Similarly, valproic acid is a histone deacetylase which also induces beta-adrenergic signaling, thus promoting cancer progression.[39] Therefore, the combination of these two drugs not only fails to treat the cancer, but may in fact lead to the worsening of the condition.

DISCUSSION

Here, we have presented a powerful machine learning framework to simulate drug responses for applications in drug discovery and precision medicine. We demonstrate our methodology on four different cancer datasets and two independent drug-target datasets by using patient-specific pathway signatures to train highly predictive models which we use as a proxy for drug candidate identification. Across all datasets, our results yielded a larger proportion of FDA-approved drugs as well as drugs investigated in clinical trials than six comparable approaches for the indications we studied, suggesting that other drugs prioritized by our methodology may also represent promising candidates for repurposing. In addition, in contrast to the other methodologies, our approach is able to prioritize drugs for individual patients, making it suitable for precision medicine applications. Finally, we also show how our methodology can be applied to propose drug combinations as well as to reveal sets of dysregulated pathways that could be used as possible targets.

Currently, there exist several limitations to this study; first, although our scoring algorithm used to simulate drug response has been shown to yield promising results in the four datasets analyzed, other scoring algorithms may be better suited for different datasets and/or applications. For instance, we could tailor the current scoring algorithm for drug discovery to learn pathway signatures from approved drugs and use these drugs to prioritize candidates that exhibit similar patterns of activity. Second, although we recommend the selection of weights following a similar logic to the one we have presented here (i.e., assigning larger weights to the quartile containing the most dysregulated pathways and lower weights for others), it may be the case that weights must be tuned for other datasets to yield promising candidates. Third, since our methodology relies on pathway signatures derived from transcriptomics data, it is inherently limited to indications where this modality is highly predictive. In other words, pathway activity scores must be readily separable between disease and normal samples in the disease we investigate as we require highly predictive models that can guarantee the change in the predicted class label is exclusively caused by the drug simulation step and not by the lack of accuracy of the model. Thus, it would be less effective in indications where transcriptomics have limited prediction power to discriminate between normal and disease samples, such as Parkinson's disease.[40] Finally, while we have demonstrated our approach with a commonly used sample-wise enrichment method, ssGSEA does not take network topology into consideration. Thus, in the future, other enrichment methods that leverage the topological information of pathways can be used to generate the pathway activity scores used by our algorithm.

Beyond this proof-of-concept, our methodology can be extended to include several additional functionalities. For instance, drug administration could be simulated in an ML model that takes into consideration temporal dimensions (e.g., event-based models,[41] survival analysis[42]). Furthermore, in this paper we trained a simple ML model, nonetheless, the same strategy could be applied to more complex ML or AI models. Since the elastic net penalty encourages sparsity, one may also use the coefficients of an ML model as a preliminary method of filtering for significant features. To save time, the total set of drug candidates can be subset to only those which directly affect the features that significantly affect the prediction capabilities of the model. In addition, we restricted our analysis to a single pathway database as it was sufficient to deploy a predictive ML model for the specific classification task we presented. However, by incorporating pathway information from other databases into the ML model, we can increase the total number and coverage of pathways to potentially reveal additional pathway targets. Similarly, the use of different drug-target databases such as ExCAPE-DB[43] could broaden the chemical space and lead to the identification of new candidates. By combining brute-force and reverse engineering approaches, one can also identify the most effective pathway scores a drug should target for any given indication; thus, tailoring the presented methodology towards drug discovery. Finally, due to limited data for all possible responses a given patient could have to a particular drug in large cohorts, we relied upon classic drug repurposing validation strategies to demonstrate the efficacy of our approach. However, with enough training data, our methodology could be deployed to support clinical decision-making in personalized medicine by simulating the effect of drugs on individual patients.

MATERIALS AND METHODS

The initial step of our methodology consists of generating patient-specific features that can be used for model training. Although in this work, we employed pathway activity scores (see subsection "Calculating individualized pathway activity scores"), other features could also be used for the same purpose. Using these scores, we trained an ML model (subsection "Building a predictive classifier") that can accurately discriminate between sample classes (e.g., disease vs normal). Next, we developed a scoring algorithm aimed to simulate the effect of a drug intervention at the pathway-level by modifying the pathway activity scores of disease samples (subsection "Scoring algorithm"). Then, the method uses the modified pathway activity scores as an input in the trained model to assess whether samples that were previously classified as "diseased" could now be classified as "normal" as a proxy for drug candidates (Subsection "Drug response prediction and prioritization"). Then we validate and evaluate our approach by presenting the datasets used for our case scenario and comparing our methodology against six equivalent approaches. Finally, we provide details on the implementation.

Datasets

Datasets from The Cancer Genome Atlas (TCGA)[44] were retrieved from the Genomic Data Commons (GDC; https://gdc.cancer.gov) portal through the R/Bioconductor package, TCGAbiolinks (version 2.16.3;[45]) on 04-08-2020 (Fig. 1d). Gene expression data from RNA-Seq was quantified using the HTSeq and raw read counts were normalized using Fragments Per Kilobase of transcript per Million mapped reads upper quartile (FPKM-UQ). Gene identifiers were mapped to HUGO Gene Nomenclature Committee (HGNC) symbols where possible. The datasets downloaded include The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), The Cancer Genome Atlas Prostate Adenocarcinoma (TCGA-PRAD), The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC), and The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) (Supplementary Table 16). We would like to note that due to the design of our methodology, we required the datasets to have a large sample size to conduct the hyperparameter optimization of the ML model and the cross validation strategy described below.

Calculating individualized pathway activity scores

We used single-sample GSEA (ssGSEA),[46] a commonly used tool to generate patient-specific pathway activity scores. Normalized gene expression (FPKM-UQ) and pathway definitions (i.e., gene sets) were provided as input and were converted to scores through ssGSEA (Supplementary Table 17; Supplementary Fig. 5). As a reference database, we used 337 pathways from KEGG (downloaded on 01-04-2020) as it is the most widely used pathway database and a standard for the most commonly used pathway activity scoring methods[18] (Fig. 1d).

Building a predictive classifier

Patient-specific pathway activity scores generated by ssGSEA were used to generate a ML classifier to distinguish between normal and tumor sample labels for each of the four datasets. The classification was conducted using an elastic net penalized logistic regression model[47] as regularized models have been shown to be generally well suited for -omics data which typically contains a disproportionate number of features to samples, and specifically well suited for these datasets.[21] Furthermore, we previously used this ML model on the same TCGA datasets,[19] yielding AUC-ROC and AUC-PR values close to 1 (Supplementary Fig. 6), in line with Mubeen et al. (2019). Prediction performance was evaluated via 10 times repeated 10-fold stratified cross-validation and tuning of elastic net hyper-parameters (i.e., l_1, l_2 regularization parameters) via grid search was performed within the cross-validation loop to avoid over-optimism.[48]

Scoring algorithm

To modify the pathway activity scores for disease samples, we developed a scoring algorithm to replicate the effect of a drug at the pathway-level. The scoring algorithm exploits interactions from drug-target datasets to modify the activity scores of pathways containing the target(s) of a drug (see example in Supplementary Fig. 10). We describe the scoring algorithm in Box 1.

Box 1 Scoring algorithm pseudocode. The pseudocode outlines the scoring algorithm used to modify the pathway activity scores of a given patient

```
Scoring Algorithm
Require:
        Set of pathways containing the protein target(s) of the drug, $\{P|p \in P\}$
        Set of samples, $\{S|s \in S\}$
        Set of healthy and disease samples, $\{H, D|H, D \in S|\forall h \in H, d \in D\}$
        Set of target labels, $\{T|t \in T\}$
        Array consisting of effect scores for all pathways,

            $\{\mathrm{ES}|\mathrm{ES}(p) \in \mathrm{ES}, \mathrm{ES}(p) = \frac{1}{N}\sum_{j=0}^{N} t_j(p)\}$

        Where, $N$ is the number of targets that are affected by a drug in pathway $p$
        Matrix consisting of original pathway activity scores for disease samples, PAS
        Array consisting of the absolute values of mean differences between
        sample groups for each $p$, $\mu_{\mathrm{H-D}} = |\mu_{\mathrm{H}} - \mu_{\mathrm{D}}|$

 1:  function SCORING_FUNCTION ($D, P, \mathrm{ES}, \mathrm{PAS}, \mu_{\mathrm{H-D}}$)
 2:      Compute quartiles, $Q_1, Q_2, Q_3$, for all values of $\mu_{\mathrm{H-D}}$
 3:      for all $d \in D$ do
 4:          for all $p \in P$ do
 5:              $sgn(p) := \begin{cases} -1 & \text{if } \mathrm{ES}(p) < 0, \\ 0 & \text{if } \mathrm{ES}(p) = 0, \\ 1 & \text{if } \mathrm{ES}(p) > 0. \end{cases}$
 6:              if $ES \neq 0$ then
 7:                  if $\mu_{H-D}(p) \in (Q_3, +\infty)$ then
 8:                      $\mathrm{CS}(p, d) = |\mathrm{PAS}(p, d)| * (w_1 * sgn(p))$
 9:                  else if $\mu_{H-D}(p) \in |Q_2, Q_3|$ then
10:                      $\mathrm{CS}(p, d) = |\mathrm{PAS}(p, d)| * (w_2 * sgn(p))$
11:                  else
12:                      $\mathrm{CS}(p, d) = |\mathrm{PAS}(p, d)| * (w_3 * sgn(p))$
13:              else
14:                  $\mathrm{CS}(p, d) = \mathrm{PAS}(p, d)$
        $\Rightarrow$ CS, Matrix consisting of calibrated pathway scores after drug treatment
15:  return CS
```

For each drug-pathway association, the pathway is assigned an effect score *ES* which is equivalent to a drug's effect on a protein target coming from drug-target datasets (i.e., activation and inhibition relationships given +1 and −1 labels, respectively). For pathways that contain multiple protein targets, the ES is equivalent to the mean of these effects (e.g., if a drug activates a protein in a pathway but also inhibits a second protein in the same pathway, the overall effect of the drug on the pathway (ES) would be 0). The absolute values of the mean differences between healthy and disease groups are calculated for each pathway $\mu_{\mathrm{H\text{-}D}}(p)$ while their quartiles are then computed on line 2. Then, from lines 3–12, for each disease sample, if the *ES* of a pathway *p* is less than or greater than 0, the scoring algorithm calculates a calibration score *CS* as the product of the absolute value of the original pathway activity score *PAS*, the weight *w*, and the effect of the drug on the pathway *sgn(p)* (i.e., −1, 0 or 1). We assign *w* based on the quartile $\mu_{\mathrm{H\text{-}D}}(p)$ pathway *p* falls into. For pathways with larger mean differences between groups, weights are assigned greater values, while pathways with smaller differences are weighted less (see example in Supplementary Text 1). On lines 13–14, if the *ES* of a pathway *p* is 0, the *CS* is assigned the value of the original *PAS*. Finally, on line 15, the *CS* is returned as a score that simulates the effect of a drug on a pathway for a disease sample.

Drug response prediction and prioritization

The methodology then aims at identifying drug candidates based on the predicted response of a patient to the simulated drug treatment. To do so, we input the modified features generated by the scoring algorithm in the trained ML model and re-evaluate the new class assignment of the patient.

Since the ML model has learnt to accurately differentiate between normal and disease samples, we expect that if a drug fails to affect a set of relevant pathways, the labels of the disease samples would remain unchanged. However, if the drug were to target a set of pathways dysregulated in a disease, we expect that the scoring algorithm could modify the scores so that they resemble those observed in control

Table 3. Number of FDA-approved and clinically tested drugs present in both drug-target datasets across the four investigated cancers.

Dataset	DrugBank Approved	DrugBank Clinical trials	DrugCentral Approved	DrugCentral Clinical trials
BRCA	26/1346 (1.93%)	182/1346 (13.52%)	14/638 (2.19%)	115/638 (18.02%)
LIHC	5/1346 (0.37%)	50/1346 (3.71%)	1/638 (0.16%)	35/638 (5.49%)
PRAD	13/1346 (0.97%)	134/1346 (9.96%)	7/638 (1.10%)	84/638 (13.17%)
KIRC	8/1346 (0.60%)	44/1346 (3.26%)	3/638 (0.47%)	25/638 (3.91%)

The percentage for the number of FDA-approved/clinically investigated drugs for each cancer type over the total number of drugs present in the drug-target dataset is reported between parentheses.

samples. Thus, by inputting these modified scores into the trained ML model, we can assess whether disease samples can now be classified as normal. Finally, after re-evaluating the predictions made by the ML model, we can rank promising drugs by the proportion of disease samples that are classified as normal as a proxy of the effectiveness of the drug.

Validation and robustness analysis

Here, we outline the robustness experiments conducted to assess the ability of our methodology to identify drugs which are already FDA-approved or have been tested in clinical trials for each of the four cancer types (i.e., TCGA datasets).

First, to simulate drug treatment using the scoring algorithm described in Box 1, we used two different drug-target datasets: DrugBank (version 5.1.6)[49] and DrugCentral (version 9.18.2020).[50] For each of the datasets, we mapped drugs to DrugBank identifiers and protein targets to HGNC symbols. In total, we retrieved 1346 unique drugs and 4673 drug-target interactions from DrugBank and 638 unique drugs and 1481 drug-target interactions from DrugCentral. Here, we would like to note that both datasets are largely overlapping (Supplementary Fig. 11). We then used these drug-target interactions as the input to our methodology to simulate patient treatments (Fig. 1e).

For validation purposes, we used two ground-truth lists containing drug-disease pairs as true positives to verify the predictions made by our methodology (Fig. 1f). The first ground-truth list contained FDA-approved drugs for the four cancer types manually retrieved from the National Cancer Institute (https://www.cancer.gov/about-cancer/treatment/drugs/cancer-type) which we mapped to the two drug-target datasets previously described. The second ground-truth list contained drugs investigated in clinical trials for the four cancer datasets retrieved from the ClinicalTrials.gov website (downloaded on 16.04.2020). Table 3 lists the number of approved and clinically tested drugs present in both drug-target datasets across the four investigated cancers.

As validation, both ground-truth lists were compared against the list of prioritized drugs that, according to our methodology, changed the predictions of 80% of the patients and subsequently classified them as normal. This threshold was selected as there were no drugs that changed the prediction for 90% or more of the patients with the parameters used by our scoring algorithm (Supplementary Figs. 7, 8). In addition, we would like to note that the vast majority of the drugs do not change the predictions for most patients. Thus, we were exclusively interested in assessing the ability of our approach to recover true positives (i.e., positive predictive value) from the list of prioritized drugs. However, since our methodology aims to prioritize drug candidates, it suffers from an early retrieval problem.[51] Furthermore, only a small minority of drugs from the drug-target datasets can be used as positive labels for each of the indications, while the majority of drugs are not known to have therapeutic benefits for them, thus, creating a large imbalance between positive and negative labels. Due to these reasons, we maintain that the evaluation strategy we present is more suitable than other conventional metrics such as the receiver operating characteristic (ROC) curves.

To identify a set of weights for the three quartiles (i.e., Q_1, Q_2 *and* Q_3 (see Box 1)) that perform well in three cancer test datasets, we followed a similar strategy to[26] where we tested different weight combinations with the intention of assigning larger weights to pathways with significantly higher dysregulations between disease and normal samples. We would like to note that the purpose of using weights in the algorithm was to modify the pathway activity scores of the few but relevant pathways targeted by the drug while maintaining the underlying distribution of pathway scores (Supplementary Fig. 9). We performed the drug simulation and conducted this parameter optimization independently on the three cancer test datasets on DrugBank, the first of two drug-target datasets. Consequently, we found a set of weights (i.e., $W_1 = 20$, $W_2 = 5$, and $W_3 = 10$ for Q_3 (the upper quartile representing the most dysregulated pathways), Q_2 (middle quartile), and Q_1 (lower quartile), respectively), that yielded both a large proportion of true positives among the prioritized drugs and also performed better than any of the six methods we compared our methodology against, as described below. Finally, we validated whether this same set of weights could also yield a large proportion of true positives on the second drug-target dataset (i.e., DrugCentral) as well as the fourth cancer dataset (i.e, KIRC).

To test the robustness of our methodology, we replicated our experiments by generating one hundred sets of 1346 drugs (the size of the DrugBank dataset) where each drug was assigned to a randomly selected protein target (from the set of all HGNC symbols) with a random causal effect following the same distribution as the original dataset (i.e., activation or inhibition). Next, we compared the number of drugs prioritized by these permutation experiments against the number of drugs prioritized by our methodology for the DrugBank dataset in the three cancer test datasets. Since we use a method to generate pathway activity scores that ignores network topology (i.e., ssGSEA), we did not conduct a robustness analysis that focused on perturbing pathway networks.

Performance comparison against equivalent drug-repurposing approaches

To evaluate our methodology, we compared it to six similar approaches that also employ transcriptomics data and pathway information to repurpose drugs on the BRCA and PRAD datasets[25,26] (note that the LIHC dataset is not included in their analyses). In the first of the two studies,[25] evaluated the ability of their methodology and four additional approaches to predict known drugs (i.e., FDA-approved or in advanced clinical trials) for breast and prostate cancer. Similarly,[26] reported the ability of their approach to identify FDA-approved drugs on the same datasets. We were thus able to directly compare the proportion of true positives that were recovered by other approaches as reported in the aforementioned studies against the proportion recovered by our approach.

Implementation

We performed ssGSEA with the Python package, GSEApy (version 0.9.12; https://github.com/zqfang/gseapy) and generated the ML models using scikit-learn.[52] We would like to note that ssGSEA does not take the topology of the pathways into account.

DATA AVAILABILITY

Data used in this manuscript are available at https://github.com/sepehrgolriz/simdrugs under the Apache 2.0 License.

CODE AVAILABILITY

Source code used in this manuscript is available at https://github.com/sepehrgolriz/simdrugs under the Apache 2.0 License.

Received: 2 April 2021; Accepted: 21 September 2021;
Published online: 27 October 2021

REFERENCES

1. Pai, S. et al. netDx: Interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.* **15**, e8497 (2019).
2. Zhao, K. & So, H. C. Using drug expression profiles and machine learning approach for drug repurposing. *Computational methods for drug repurposing*, 219–237. Humana Press, New York, NY (2019).
3. Réda, C. et al. Machine learning applications in drug development. *Computational Struct. Biotechnol. J.* **18**, 241–252 (2020).
4. Liu, S. et al. Early diagnosis of Alzheimer's disease with deep learning. *IEEE 11th international symposium on biomedical imaging (ISBI)* 1015–1018 (2014).
5. Hirohara, M. et al. Convolutional neural network based on SMILES representation of compounds for detecting chemical motifs. *BMC Bioinforma.* **19**, 526 (2018).
6. Castro, V. M. et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology* **88**, 164–168 (2017).
7. Su, J., Yoon, B. J. & Dougherty, E. R. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PloS ONE* **4**, e8161 (2009).
8. Lim, S. et al. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Brief. Bioinforma.* **21**, 36–46 (2020).
9. Reimand, J. et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517 (2019).
10. Perscheid, C. Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches. *Brief. Bioinforma.* **22**, bbaa151 (2020).
11. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
12. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
13. Nguyen, T. M., Shafi, A., Nguyen, T. & Draghici, S. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.* **20**, 1–15 (2019).
14. Adam, G. et al. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis. Oncol.* **4**, 1–10 (2020).
15. Peyvandipour, A., Saberian, N., Shafi, A., Donato, M. & Draghici, S. A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **34**, 2817–2825 (2018).
16. Saberian, N., Peyvandipour, A., Donato, M., Ansari, S. & Draghici, S. A new computational drug repurposing method using established disease–drug pair knowledge. *Bioinformatics* **35**, 3672–3678 (2019).
17. Emon, M. A. et al. PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures. *BMC Bioinforma.* **21**, 1–21 (2020).
18. Ammad-ud-din, M. et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* **32**, i455–i463 (2016).
19. Amadoz, A. et al. A comparison of mechanistic signaling pathway activity analysis methods. *Brief. Bioinforma.* **20**, 1655–1668 (2019).
20. Mubeen, S. et al. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front. Genet.* **10**, 1203 (2019).
21. Smith, A. M. et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinforma.* **21**, 119 (2020).
22. Esteban-Medina, M. et al. Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. *BMC Bioinforma.* **20**, 370 (2019).
23. Cubuk, C. et al. Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. *Cancer Res.* **78**, 6059–6072 (2018).
24. Çubuk, C. et al. Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. *NPJ Syst. Biol. Appl.* **5**, 1–11 (2019).
25. Chan, J., Wang, X., Turner, J. A., Baldwin, N. E. & Gu, J. Breaking the paradigm: Dr Insight empowers signature-free, enhanced drug repurposing. *Bioinformatics* **35**, 2818–2826 (2019).
26. Chen, H. R., Sherr, D. H., Hu, Z. & DeLisi, C. A network based approach to drug repositioning identifies plausible candidates for breast cancer and prostate cancer. *BMC Med. Genomics* **9**, 1–11 (2016).
27. Zhang, N. et al. 5-Fluorouracil: mechanisms of resistance and reversal strategies. *Molecules* **13**, 1551–1569 (2008).
28. Snaterse, G. et al. Circulating steroid hormone variations throughout different stages of prostate cancer. *Endocr.-Relat. Cancer* **24**, R403–R420 (2017).
29. Harris, W. P. et al. Androgen deprivation therapy: progress in understanding mechanisms of resistance and optimizing androgen depletion. *Nat. Clin. Pract. Urol.* **6**, 76–85 (2009).
30. Karantanos, T., Corn, P. G. & Thompson, T. C. Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene* **32**, 5501–5511 (2013).
31. Huynh, H. Tyrosine kinase inhibitors to treat liver cancer. *Expert Opin. Emerg. Drugs* **15**, 13–26 (2010).
32. Khoo T. S. W. L., Rehman A. & Olynyk J. K. Tyrosine kinase inhibitors in the treatment of hepatocellular carcinoma. *Exon Publications*. 127–139 (2019).
33. Bhullar, K. S. et al. Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol. Cancer* **17**, 1–20 (2018).
34. Gedaly, R. et al. PI-103 and sorafenib inhibit hepatocellular carcinoma cell proliferation by blocking Ras/Raf/MAPK and PI3K/AKT/mTOR pathways. *Anticancer Res.* **30**, 4951–4958 (2010).
35. Mousa, A. B. Sorafenib in the treatment of advanced hepatocellular carcinoma. *Saudi J. Gastroenterol.: Off. J. Saudi Gastroenterol. Assoc.* **14**, 40 (2008).
36. Zhu, Y. J., Zheng, B., Wang, H. Y. & Chen, L. New knowledge of the mechanisms of sorafenib resistance in liver cancer. *Acta Pharmacologica Sin.* **38**, 614–622 (2017).
37. Llovet, J. M. et al. Sorafenib in advanced hepatocellular carcinoma. *N. Engl. J. Med.* **359**, 378–390 (2008).
38. Chen, H. et al. Adrenergic signaling promotes angiogenesis through endothelial cell-tumor cell crosstalk. *Endocr.-Relat. Cancer* **21**, 783–795 (2014).
39. Hulsurkar, M. et al. Beta-adrenergic signaling promotes tumor angiogenesis and prostate cancer progression through HDAC2-mediated suppression of thrombospondin-1. *Oncogene* **36**, 1525–1536 (2017).
40. Chen-Plotkin, A. S. Blood transcriptomics for Parkinson disease? *Nat. Rev. Neurol.* **14**, 5–6 (2018).
41. 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 (2012).
42. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).
43. Sun, J. et al. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminformatics* **9**, 17 (2017).
44. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113 (2013).
45. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71–e71 (2015).
46. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108 (2009).
47. Zou, H. & Trevor, H. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67**, 301–320 (2005).
48. Molinaro, A. M., Simon, R. & Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307 (2005).
49. Knox, C. et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **39**, D1035–D1041 (2010).
50. Ursu, O. et al. DrugCentral: online drug compendium. *Nucleic Acids Res.* **45**, D932–D939 (2016).
51. Berrar, D. & Flach, P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief. Bioinforma.* **13**, 83–97 (2012).
52. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. Kim, R. et al. A Phase I trial of trametinib in combination with sorafenib in patients with advanced hepatocellular cancer. *Oncologist* **25**, e1893–e1899 (2020).
54. Zhang, J. et al. Erlotinib for advanced hepatocellular carcinoma: a systematic review of phase II/III clinical trials. *Saudi Med. J.* **37**, 1184 (2016).
55. Di Gennaro, E. et al. Vorinostat synergises with capecitabine through upregulation of thymidine phosphorylase. *Br. J. Cancer* **103**, 1680–1691 (2010).

ACKNOWLEDGEMENTS

The authors would like to thank Colin Birkenbihl and Lauren DeLong for their valuable feedback. The authors would like to thank the reviewers for their valuable feedback and suggestions. This work was developed in the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies".

AUTHOR CONTRIBUTIONS

D.D.F. conceived and designed the study. S.G.K. implemented the scoring algorithm and ran the validation experiments with assistance from D.D.F. S.G.K. analyzed the case scenario and M.H.A. and D.D.F. assisted in the interpretation of the results. S.M. processed and prepared the datasets. S.M. and S.G.K. ran the datasets with the pathway enrichment method to generate the pathway activity scores. S.M. and V.S.B. trained the ML models. A.T.K., M.H.A., and D.D.F. acquired the funding. S.G.K., S.M., and D.D.F. wrote the paper. All authors have read and approved the final paper.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

D.D.F. received salary from Enveda Biosciences.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-021-00199-1.

Correspondence and requests for materials should be addressed to Sepehr Golriz Khatami or Daniel Domingo-Fernández.

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.

Supplementary File

Supplementary Tables

Dr. Insight Performance		
Dataset	**TCGA BRCA**	**TCGA PRAD**
# of proposed drug treatments	70	69
# identified ground-truth drug treatments	15	11
Proportion of true positives (%)	**21.42**	**15.94**

Supplementary Table 1. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using Dr. Insight on the BRCA and PRAD datasets. See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported in Chan *et al*. (2019) (Table S5.11) https://doi.org/10.1093/bioinformatics/btz006.

CMap performance on BRCA								
Gene signature sizes (threshold)	50	100	200	300	400	600	800	1000
# of proposed drug treatments	56	52	55	67	74	74	71	56
# identified ground-truth drug treatments	5	6	3	6	5	6	6	5
Proportion of true positives (%)	8.92	**11.53**	5.45	8.92	6.75	8.1	8.45	8.92

Supplementary Table 2. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using CMap on the BRCA dataset. The proportion of true positives highlighted in bold indicates the highest percentage amongst all parameters (i.e., threshold gene set size). See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.1) https://doi.org/10.1093/bioinformatics/btz006.

CMap performance on PRAD								
Gene signature sizes (threshold)	50	100	200	300	400	600	800	1000
# of proposed drug treatments	52	83	65	66	80	73	76	71
# identified ground-truth drug treatments	8	11	7	7	9	10	8	9
Proportion of true positives (%)	**15.38**	13.25	10.76	10.60	11.25	13.69	10.52	12.67

Supplementary Table 3. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using CMap on the PRAD dataset. The proportion of true positives highlighted in bold indicates the highest percentage amongst all parameters (i.e., threshold gene set size). See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.2) https://doi.org/10.1093/bioinformatics/btz006.

sscMap performance on BRCA								
Gene signature sizes (threshold)	50	100	200	300	400	600	800	1000
# of proposed drug treatments	6	19	659	998	1185	1316	1404	1436
# identified ground-truth drug treatments	0	1	31	45	61	63	72	77
Proportion of true positives (%)	0	5.26	4.70	4.50	5.14	4.78	5.12	**5.36**

Supplementary Table 4. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using sscMap on the BRCA dataset. The proportion of true positives highlighted in bold indicates the highest percentage amongst all parameters (i.e., threshold gene set size). See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.4) https://doi.org/10.1093/bioinformatics/btz006.

sscMap performance on PRAD								
Gene signature sizes (threshold)	50	100	200	300	400	600	800	1000
# of proposed drug treatments	8	18	100	177	202	381	653	810
# identified ground-truth drug treatments	1	4	9	11	14	17	21	25
Proportion of true positives (%)	12.5	**22.22**	9	6.21	6.93	4.46	3.21	3.08

Supplementary Table 5. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using sscMap on the PRAD dataset. The proportion of true positives highlighted in bold indicates the highest percentage amongst all parameters (i.e., threshold gene set size). See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.5) https://doi.org/10.1093/bioinformatics/btz006.

NFFinder performance on BRCA								
Gene signature sizes (threshold)	50	100	200	300	400	600	800	1000
# of proposed drug treatments	329	285	623	782	651	854	1007	1069
# identified ground-truth drug treatments	26	24	40	45	43	54	66	62
Proportion of true positives (%)	7.90	**8.42**	6.42	5.75	6.60	6.32	6.65	5.59

Supplementary Table 6. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using NFFinder on the BRCA dataset. The proportion of true positives highlighted in bold indicates the highest percentage amongst all parameters (i.e., threshold gene set size). See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.7) https://doi.org/10.1093/bioinformatics/btz006.

NFFinder performance on PRAD								
Gene signature sizes (threshold)	50	100	200	300	400	600	800	1000
# of proposed drug treatments	347	592	548	717	529	719	842	783
# identified ground-truth drug treatments	18	28	32	34	32	38	42	38
Proportion of true positives (%)	5.18	4.72	**5.83**	4.50	4.74	5.28	4.98	4.85

Supplementary Table 7. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using NFFinder on the PRAD dataset. The proportion of true positives highlighted in bold indicates the highest percentage amongst all parameters (i.e., threshold gene set size). See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.8) https://doi.org/10.1093/bioinformatics/btz006.

Cogena		
Cogena	TCGA BRCA	TCGA PRAD
# of proposed drug treatments	335	982
# identified ground-truth drug treatments	30	5
Proportion of true positives (%)	**8.95**	**0.50**

Supplementary Table 8. Number of approved drugs or drugs in clinical trials (i.e., ground-truth drug treatments) recovered using Cogena on the BRCA and PRAD datasets. See the **Supplementary Text** for a detailed description of the information reported in each of the rows. The results of this table are reported by Chan *et al*. (2019) (Table S5.10) https://doi.org/10.1093/bioinformatics/btz006.

Chen *et al.* (2016) study performance			
Dataset	**Prioritized**	**Approved (total)**	**Proportion of true positives (%)**
BRCA	2435	20 (20)	20/2435 **(0.81%)**
PRAD	2500	10 (11)	10/2500 **(0.40%)**

Supplementary Table 9. Number of approved drugs recovered reported by Chen *et al.* (2016) on the BRCA and PRAD datasets. The results from this table are reported in Table 1 of the paper (https://doi.org/10.1186/s12920-016-0212-7).

Method	**BRCA**	**PRAD**
Dr. Insight	21.42 (%)	15.94 (%)
CMap	6.75 - 11.53 (%)	10.52 - 15.38 (%)
sscMap	0 - 5.36 (%)	3.08 - 22.22 (%)
NFFinder	5.59 - 8.42 (%)	4.5 - 5.83 (%)
Cogena	8.95 (%)	0.50 (%)
Chen *et al.* (2016)	0.81 (%)	0.40 (%)

Supplementary Table 10. Summary of performance for all methods benchmarked on the BRCA and PRAD datasets. Performances are measured as % of approved drugs or drugs in clinical trials recovered and are taken from Supplementary Tables 1-9.

DrugCentral (638 drugs)						
TCGA dataset	**Replaced weight**	**Weight sets (W1_W2_W3)**	**# Prioritized**	**# Clinical trials (total)**	**# Approved (total)**	**Proportion of true positives (%)**
BRCA	**W1**	1_5_10	0	-	-	-
		5_5_10	0	-	-	-
		10_5_10	3	1(115)	-	1/3(33%)
		15_5_10	7	2(115)	-	2/7(28%)
	W2	20_1_10	17	4(115)	1(14)	5/17(29%)
		20_10_10	29	5(115)	1(14)	6/29(20%)
		20_15_10	31	5(115)	1(14)	6/31(19%)
		20_20_10	46	6(115)	1(14)	7/46(15%)
	W3	20_5_1	16	4(115)	1(14)	5/16(31%)
		20_5_5	17	4(115)	1(14)	5/17(29%)
		20_5_15	18	4(115)	1(14)	5/18(27%)
		20_5_20	19	4(115)	1(14)	5/18(27%)
LIHC	**W1**	1_5_10	14	1(35)	1(1)	2/14(14%)
		5_5_10	16	1(35)	1(1)	2/16(12%)
		10_5_10	19	2(35)	1(1)	3/19(15%)
		15_5_10	19	2(35)	1(1)	3/19(15%)
	W2	20_1_10	19	2(35)	1(1)	3/19(15%)
		20_10_10	20	2(35)	1(1)	3/20(15%)

		20_15_10	21	2(35)	1(1)	3/21(14%)
		20_20_10	22	2(35)	1(1)	3/22(13%)
	W3	20_5_1	9	1(35)	-	1/9(11%)
		20_5_5	10	1(35)	1(1)	2/10(20%)
		20_5_15	20	2(35)	1(1)	3/20(15%)
		20_5_20	20	2(35)	1(1)	3/20(15%)
PRAD	W1	1_5_10	1	1(84)	-	1/1(100%)
		5_5_10	1	1(84)	-	1/1(100%)
		10_5_10	1	1(84)	-	1/1(100%)
		15_5_10	9	3(84)	-	3/9(33%)
	W2	20_1_10	11	2(84)	-	2/11(18%)
		20_10_10	19	3(84)	1(7)	4/19(21%)
		20_15_10	40	7(84)	1(7)	8/40(20%)
		20_20_10	41	8(84)	1(7)	9/41(21%)
	W3	20_5_1	16	3(84)	-	3/16(18%)
		20_5_5	16	3(84)	-	3/16(18%)
		20_5_15	32	7(84)	1(7)	8/32(25%)
		20_5_20	32	7(84)	1(7)	8/32(25%)

Supplementary Table 11. Number of FDA-approved and clinically tested drugs recovered across the three investigated cancers using different weights in the DrugCentral dataset. In the fourth column (i.e., # Prioritized), we report the number of drugs that changed the predictions for at least 80% of the patients for each cancer type.

DrugBank (1346 drugs)						
TCGA dataset	Replaced weight	Weight sets (W1_W2_W3)	# Prioritized	# Clinical trials (total)	# Approved (total)	Proportion of true positives (%)
BRCA	W1	1_5_10	22	4(182)	0(26)	4/22(18%)
		5_5_10	55	8(182)	2(26)	10/55(18%)
		10_5_10	81	12(182)	2(26)	14/81(17%)
		15_5_10	92	14(182)	2(26)	16/92(17%)
	W2	20_1_10	124	21(182)	3(26)	24/124(19%)
		20_10_10	142	24(182)	3(26)	27/142(19%)
		20_15_10	148	24(182)	3(26)	27/148(18%)
		20_20_10	172	27(182)	3(26)	30/172(14%)
	W3	20_5_1	85	16(182)	3(26)	19/85(22%)
		20_5_5	89	18(182)	3(26)	21/89(23%)
		20_5_15	142	25(182)	3(26)	28/142(19%)
		20_5_20	143	25(182)	3(26)	28/142(19%)
LIHC	W1	1_5_10	67	9(50)	2(5)	11/67(16%)
		5_5_10	70	10(50)	2(5)	12/70(17%)
		10_5_10	70	9(50)	2(5)	11/70(16%)
		15_5_10	73	10(50)	2(5)	12/73(16%)
	W2	20_1_10	71	11(50)	2(5)	13/71(18%)
		20_10_10	81	13(50)	2(5)	15/81(18%)
		20_15_10	89	14(50)	2(5)	16/89(18%)
		20_20_10	93	14(50)	2(5)	16/93(18%)
	W3	20_5_1	36	8(50)	1(5)	9/36(25%)
		20_5_5	46	8(50)	2(5)	10/46(21%)
		20_5_15	101	13(50)	3(5)	16/101(19%)
		20_5_20	117	13(50)	3(5)	16/101(19%)
PRAD	W1	1_5_10	3	-	-	-
		5_5_10	7	-	-	-
		10_5_10	19	1(134)	-	1/19(5%)

		15_5_10	27	2(134)	-	2/27(7%)
	W2	20_1_10	46	10(134)	-	10/46(22%)
		20_10_10	82	19(134)	2(13)	21/82(25%)
		20_15_10	59	8(134)	2(13)	10/59(16%)
		20_20_10	94	19(134)	2(13)	21/94(22%)
	W3	20_5_1	66	16(134)	-	16/66(24%)
		20_5_5	64	15(134)	-	15/64(23%)
		20_5_15	59	8(134)	2(13)	10/59(17%)
		20_5_20	63	8(134)	0(13)	8/59(14%)

Supplementary Table 12. Number of FDA-approved and clinically tested drugs recovered across the three investigated cancers using different weights in the DrugBank dataset. In the fourth column (i.e., # Prioritized), we report the number of drugs that changed the predictions for at least 80% of the patients for each cancer type.

	Weights 10(Q3), 5(Q2), 2(Q1)							
-	**DrugBank**				**DrugCentral**			
Dataset	**Prioritized**	**Approved (total)**	**Clinical trials (total)**	**Proportion of true positives (%)**	**Prioritized**	**Approved (total)**	**Clinical trials (total)**	**Proportion of true positives (%)**
BRCA	20	3(26)	2(182)	5/20**(25%)**	2	0(14)	1(115)	1/2**(50%)**
LIHC	27	1(5)	6(50)	7/27**(25.95%)**	4	0(1)	1(35)	1/4**(25%)**
PRAD	17	0(13)	0(134)	0/17**(0%)**	0	0(7)	0(84)	**0**

Supplementary Table 13. Number of FDA-approved and clinically tested drugs recovered for both drug-target datasets across the three investigated cancers. In the columns labelled “Prioritized”, we report the number of drugs that changed the predictions for at least 80% of the patients for each cancer type. A different set of weights were used than the ones used to generate the results of Table 2, leading to comparatively better results for the BRCA and LIHC datasets, but resulting in no true positives recovered for PRAD.

-	**Weights 1(Q3), 1(Q2), 1(Q1)**							
-	**DrugBank**				**DrugCentral**			
Dataset	**Prioritized**	**Approved (total)**	**Clinical trials (total**	**Proportion of true positives (%)**	**Prioritized**	**Approved (total)**	**Clinical trials (total**	**Proportion of true positives (%)**
BRCA	0	0(26)	0(182)	**0**	0	0(14)	0(115)	**0**
LIHC	0	0(5)	0(50)	**0**	0	0(1)	0(35)	**0**
PRAD	0	0(13)	0(134)	**0**	0	0(7)	0(84)	**0**

Supplementary Table 14. Number of FDA-approved and clinically tested drugs recovered for both drug-target datasets across the three investigated cancers. In the columns labelled “Prioritized”, we report the number of drugs that changed the predictions for at least 80% of the patients for each cancer type. Here, we set all weights equal to one and find that there are no prioritized drugs for any of the three cancer datasets.

Pathway	**Target**	**Pathway- level drug effect**	**Pathway activity in patients relative to controls**
Gap junction	PDGFRB, RAF1	Inhibition	Upregulated
Fc gamma R mediated phagocytosis	RAF1	Inhibition	Upregulated
Phospholipase D signaling pathway	INSR, PDGFRB, KIT, RAF1	Inhibition	Upregulated
Thyroid hormone signaling pathway	RAF1	Inhibition	Upregulated
Thyroid cancer	BRAF, RET	Inhibition	Upregulated

Hepatitis B	BRAF, RAF1	Inhibition	Upregulated
Human papillomavirus infection	PDGFRB, RAF1	Inhibition	Upregulated
Focal adhesion	BRAF, FLT4, KDR, PDGFRB, RAF1, FLT1	Inhibition	Upregulated
Renal cell carcinoma	BRAF, RAF1	Inhibition	Upregulated
Glioma	BRAF, PDGFRB, RAF1	Inhibition	Upregulated
Axon guidance	RAF1	Inhibition	Upregulated
Endocrine resistance	BRAF, RAF1	Inhibition	Upregulated
Breast cancer	BRAF, FLT4, KIT, RAF1	Inhibition	Upregulated
Sphingolipid signaling pathway	RAF1	Inhibition	Upregulated
Autophagy animal	RAF1	Inhibition	Upregulated
mTOR signaling pathway	INSR, BRAF, RAF1	Inhibition	Upregulated
GnRH signaling pathway	RAF1	Inhibition	Upregulated
Pathways in cancer	FLT3, BRAF, FLT4, PDGFRB, KIT, RAF1, RET	Inhibition	Upregulated
Chronic myeloid leukemia	BRAF, RAF1	Inhibition	Upregulated
Choline metabolism in cancer	PDGFRB, RAF1	Inhibition	Upregulated
Bladder cancer	BRAF, RAF1	Inhibition	Upregulated
Non small cell lung cancer	BRAF, RAF1	Inhibition	Upregulated
Gastric cancer	BRAF, RAF1	Inhibition	Upregulated
Cushing syndrome	BRAF	Inhibition	Upregulated
VEGF signaling pathway	KDR, RAF1	Inhibition	Upregulated
Hepatocellular carcinoma	BRAF, RAF1	Inhibition	Upregulated
MAPK signaling pathway	INSR,FLT3,BRAF,FLT4,KDR,PDGFRB,KIT,RAF1,	Inhibition	Upregulated
Regulation of actin cytoskeleton	BRAF, PDGFRB, RAF1	Inhibition	Upregulated
Human immunodeficiency virus 1 infection	RAF1	Inhibition	Upregulated
Relaxin signaling pathway	RAF1	Inhibition	Upregulated
Estrogen signaling pathway	RAF1	Inhibition	Upregulated
Progesterone mediated oocyte maturation	BRAF, RAF1	Inhibition	Upregulated
MicroRNAs in cancer	PDGFRB, RAF1	Inhibition	Upregulated
Neurotrophin signaling pathway	BRAF, RAF1	Inhibition	Upregulated
Alcoholism	BRAF, RAF1	Inhibition	Upregulated
Fc epsilon RI signaling pathway	RAF1	Inhibition	Upregulated
Apoptosis	RAF1	Inhibition	Upregulated
Cellular senescence	RAF1	Inhibition	Upregulated
Colorectal cancer	BRAF, RAF1	Inhibition	Upregulated
Long term depression	BRAF, RAF1	Inhibition	Upregulated
Melanogenesis	KIT, RAF1	Inhibition	Upregulated

Supplementary Table 15. Effect of Sorafenib on pathway targets in the LIHC dataset. The first column corresponds to the pathways that contain protein targets of Sorafenib while the second column corresponds to the specific protein targets of the drug. The third column corresponds to the effect of the drug on the pathway based on its effect on the target. In this case, all pathways are inhibited as all protein targets are inhibited by Sorafenib. Finally, the last column presents a relative comparison between the pathway activity observed in patients vs. controls: downregulated corresponds to lower pathway activity and upregulated corresponds to the opposite.

Dataset	Normal samples	Tumor samples	Reference	DOI
BRCA	113	1102	(The Cancer Genome Atlas Network, 2012)	https://doi.org/10.1038/nature11412
LIHC	50	371	(The Cancer Genome Atlas Research Network, 2017)	https://doi.org/10.1016/j.cell.2017.05.046
PRAD	52	498	(The Cancer Genome Atlas Research Network, 2015)	https://doi.org/10.1016/j.cell.2015.10.025
KIRC	72	538	(The Cancer Genome Atlas Research Network, 2013)	https://doi.org/10.1038/nature12222

Supplementary Table 16. Number of normal and tumor samples in the TCGA datasets used in this work.

Parameter	Configuration
Method	rank
Minimum size of gene set	15
Maximum size of gene set	3000

Supplementary Table 17. Parameter configuration settings for running ssGSEA with GSEApy (version 0.9.12).

Dr. Insight (Chan *et al.* (2019))				
Dataset	**Normal samples**	**Tumor samples**	**Reference**	**DOI**
BRCA	111	1099	(The Cancer Genome Atlas Network, 2012)	https://doi.org/10.1038/nature11412
PRAD	52	498	(The Cancer Genome Atlas Research Network, 2015)	https://doi.org/10.1016/j.cell.2015.10.025

Supplementary Table 18. Number of normal and tumor samples in the TCGA datasets used in the Chan *et al.* (2019) study. Datasets were retrieved through the Genomic Data Commons (GDC; https://gdc.cancer.gov) by Chan *et al.* (2019). Log transformed TCGA level-3 normalized count data was used in their study. Study details can be found at https://doi.org/10.1093/bioinformatics/btz006.

Chen *et al.* (2016)		
	Breast Cancer	**Prostate Cancer**
Total compounds	3678	4228
Compounds that are FDA-approved drugs	632	676
Compounds that are FDA-approved drugs for target disease	20	11
Compounds that are in clinical trial for target disease	154	106
Total number of pathways	287	

Supplementary Table 19. Information about the chemicals used by Chen *et al.* (2016). Details of the approach can be found at https://doi.org/10.1186/s12920-016-0212-7.

Dr. Insight (Chan *et al.* (2019))		
	PRAD	**BRCA**
FDA-approved drugs	7	9
Clinical trials drugs	47	63
Total number of drugs	54	72
Total number of pathways	222	

Supplementary Table 20. Information about the chemicals used by Chan *et al.* (2019) study. Study details can be found at https://doi.org/10.1093/bioinformatics/btz006.

Supplementary Text

1. Drug simulation scenario

Suppose you have a score of 0.2 for patient *A* on pathway *X*. If a drug is activating the pathway and the mean difference between healthy and disease groups is large, this pathway will be in the first quartile and the initial pathway score will be multiplied by a higher weight, (e.g., 3). Thus, the modified score for patient *A* on pathway *X* (originally 0.2) will be 0.6. However, if the mean difference between healthy and disease groups is not large, the weight will be smaller (e.g., 2) and the modified score for the patient on pathway *X* will be 0.4. These steps are repeated for all pathways which contain protein targets of a particular drug for a given patient. Finally, all modified scores are then passed to the classifier to determine whether the patient is subsequently classified as normal.

2. Measurements reported by equivalent approaches

Below, we describe each of the measurements reported by studies on similar drug-repurposing approaches that can be found in **Supplementary Tables 1-8**.

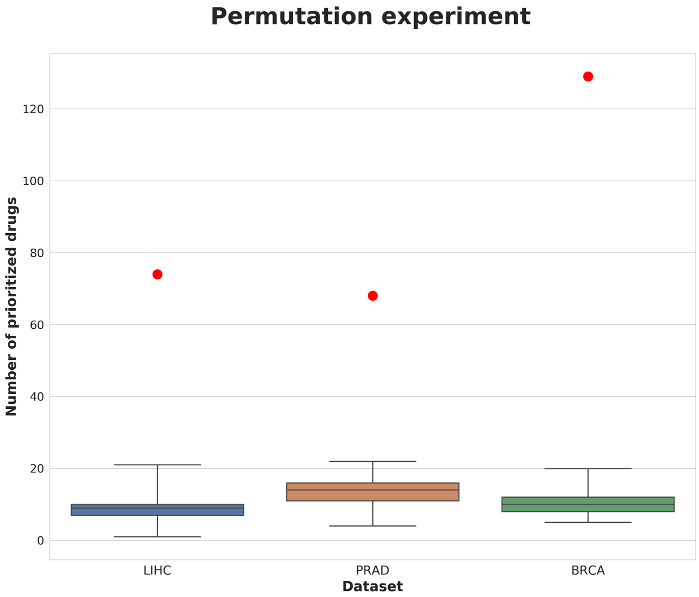
Gene signature sizes (threshold): This refers to the size of the list of query gene signatures used to evaluate the drug repurposing performance of CMap, sscMap and NFFinder. Specifically, the gene signatures were composed of the top- and bottom-ranked most differentially expressed genes of varying sizes. As the CMap, sscMap and NFFinder methods did not provide specific recommendations for the size of query gene signatures in their original work, the developers of the Dr. Insight method used gene lists of varying sizes (50, 100, 200, 300, 400, 600, 800 and 1000 Affymetrix probes) to evaluate the drug repurposing performance of CMap, sscMap and NFFinder and compare them with their method.

Number of identified drug treatments: This number refers to the drugs which were prioritized by different methods (i.e., NFFinder, CMap, sscMap, cogena, Dr. Insight).

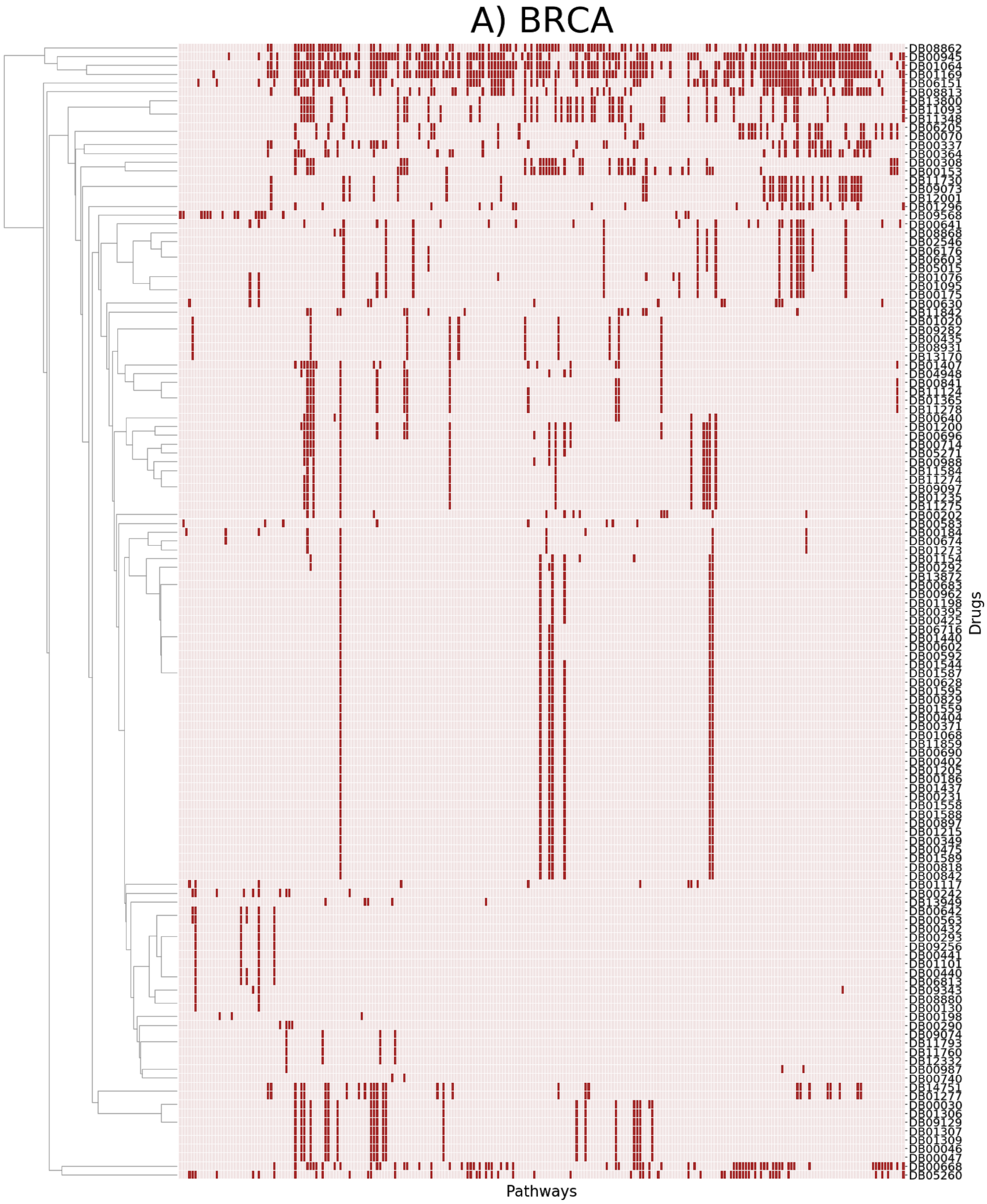
Number of identified ground-truth drug treatments: This number refers to the number of FDA-approved drugs and clinical-trial drugs from the “# Identified drug treatments”.

Proportion of true positives (%): This proportion refers to the "# identified ground-truth drug treatments" over the "# Identified drug treatments".

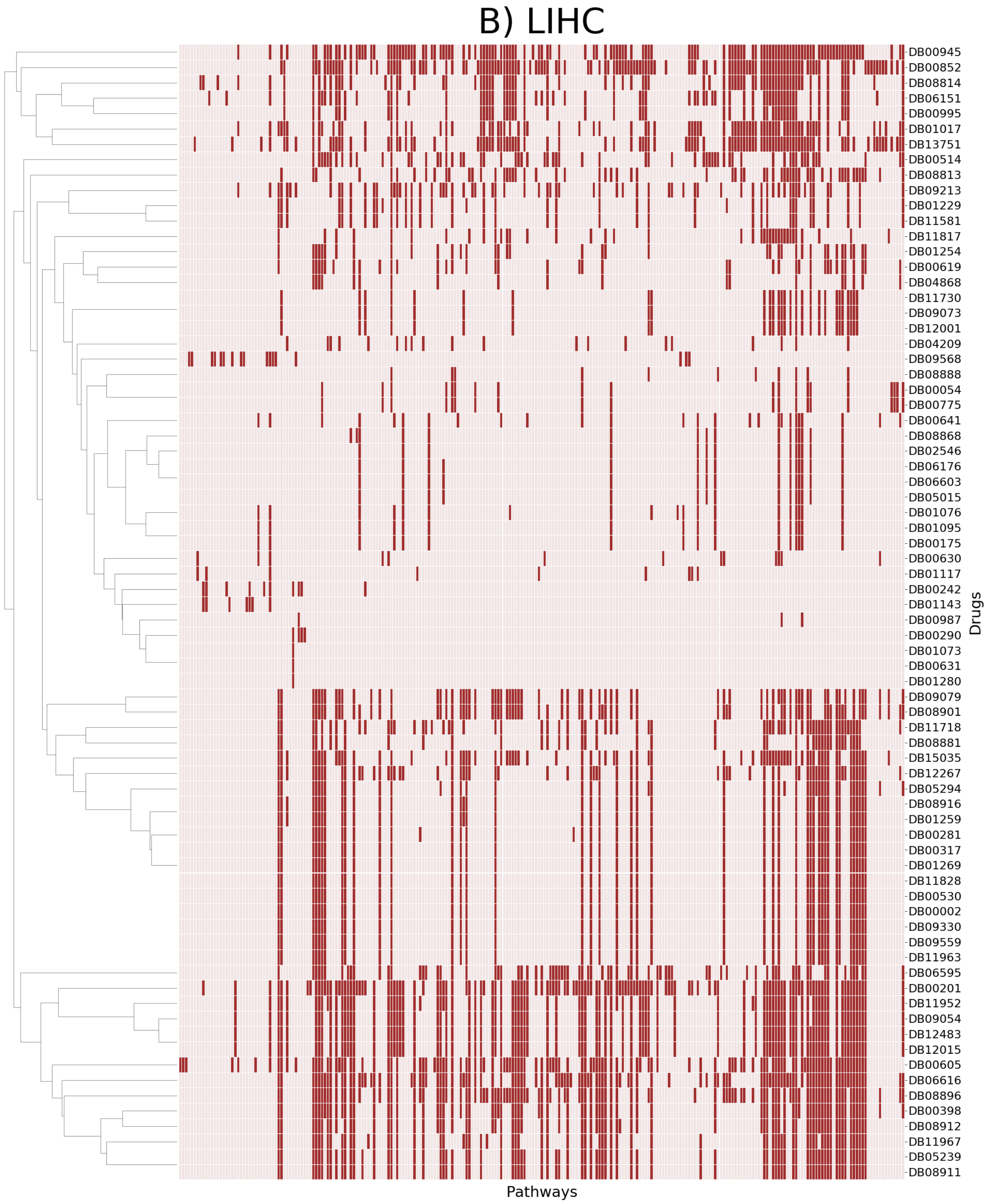
Supplementary Figures



Supplementary Figure 1. Comparison of the results of the permutation experiments against the number of prioritized drugs in DrugBank for the three cancer test datasets. While the three boxplots correspond to the number of prioritized drugs in the 100 permutation experiments for each of the three datasets, the number of prioritized drugs in the original DrugBank dataset has been indicated with a red circle. The number of prioritized drugs from DrugBank is significantly higher than for any of the permutations experiments. *p*-values have been omitted as all permutation experiments yielded a lower number of prioritized drugs compared to the original dataset and thus, *p*-values would be dominated by the number of experiments (i.e., 100 experiments would yield a *p*-value of 0.01, and 1,000 experiments would yield a *p*-value of 0.001). We would like to note that we compare the permutation experiments against DrugBank as the number of simulated drugs is equal to the size of this dataset (1,346 drugs). Furthermore, a comparison to the DrugCentral dataset would yield an even greater difference in the number of prioritized drugs as DrugCentral is smaller in size (638 drugs).

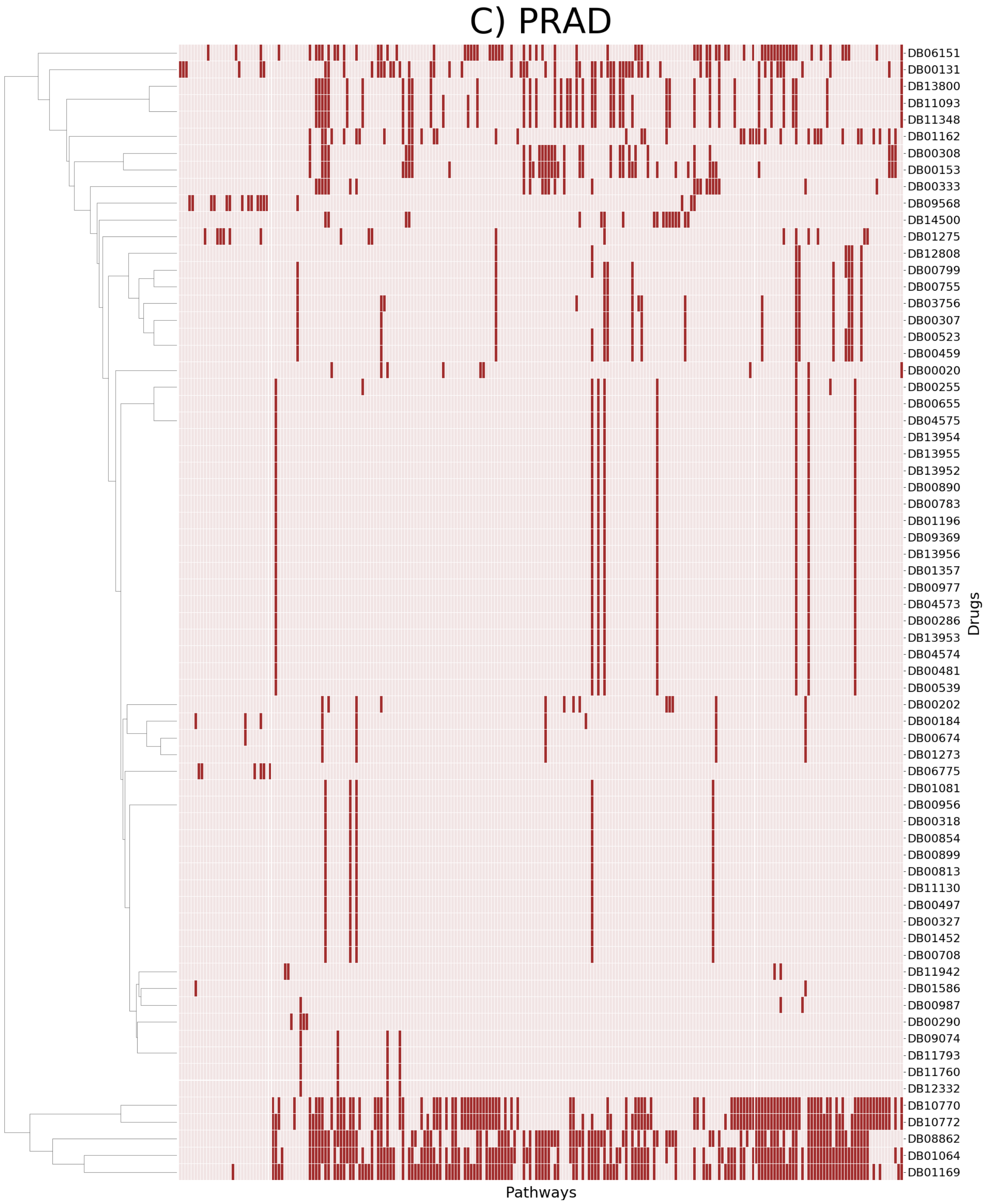


Supplementary Figure 2. Pathways targeted by the prioritized drugs in DrugBank for BRCA. The X-axis corresponds to the pathways targeted by any of the prioritized drugs (KEGG pathways not targeted by any prioritized drug have been omitted for better visualization). Drugs (Y-axis) have been clustered based on the pathways they target. Due to the large number of pathways, we have clustered pathway groups together for visualization purposes. Black cells correspond to pathways targeted for each drug. Details about each pathway are displayed in the following Jupyter notebook https://github.com/sepehrgolriz/simdrugs/blob/main/scripts_and_notebooks/heatmaps.ipynb.

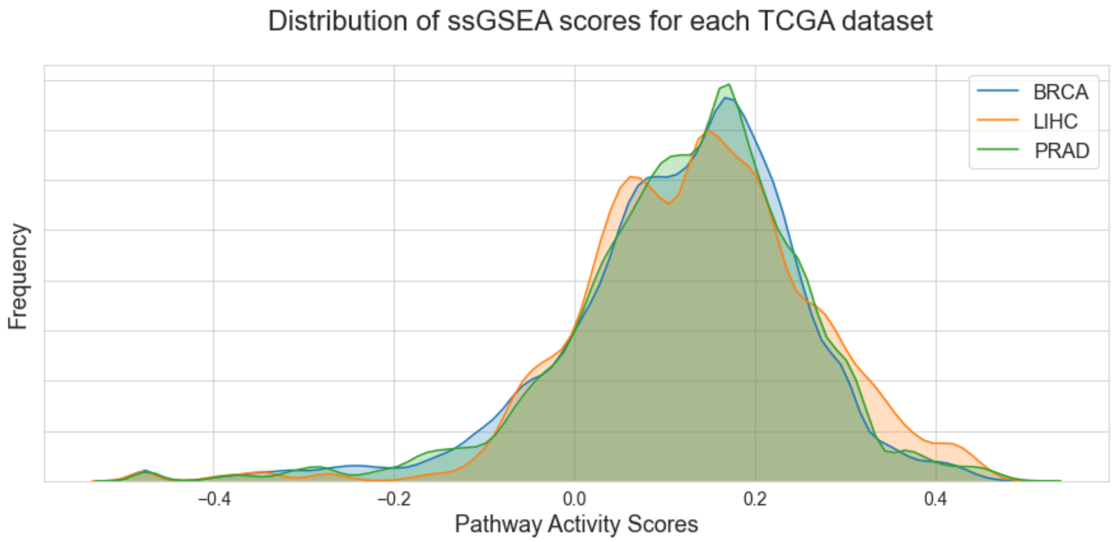


B) LIHC

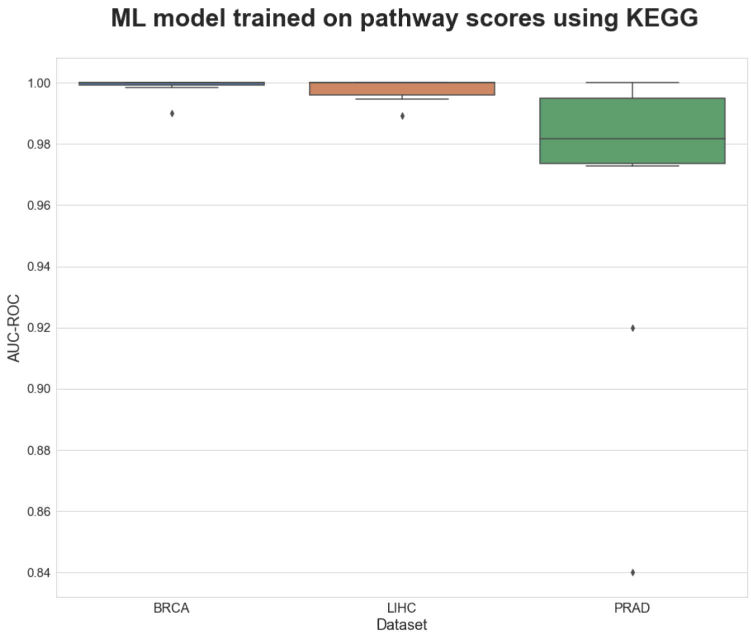
Supplementary Figure 3. Pathways targeted by the prioritized drugs in DrugBank for LIHC. The X-axis corresponds to the pathways targeted by any of the prioritized drugs (KEGG pathways not targeted by any prioritized drug have been omitted for better visualization). Drugs (Y-axis) have been clustered based on the pathways they target. Due to the large number of pathways, we have clustered pathway groups together for visualization purposes. Black cells correspond to pathways targeted for each drug. Details about each pathway are displayed in the following Jupyter notebook https://github.com/sepehrgolriz/simdrugs/blob/main/scripts_and_notebooks/heatmaps.ipynb.



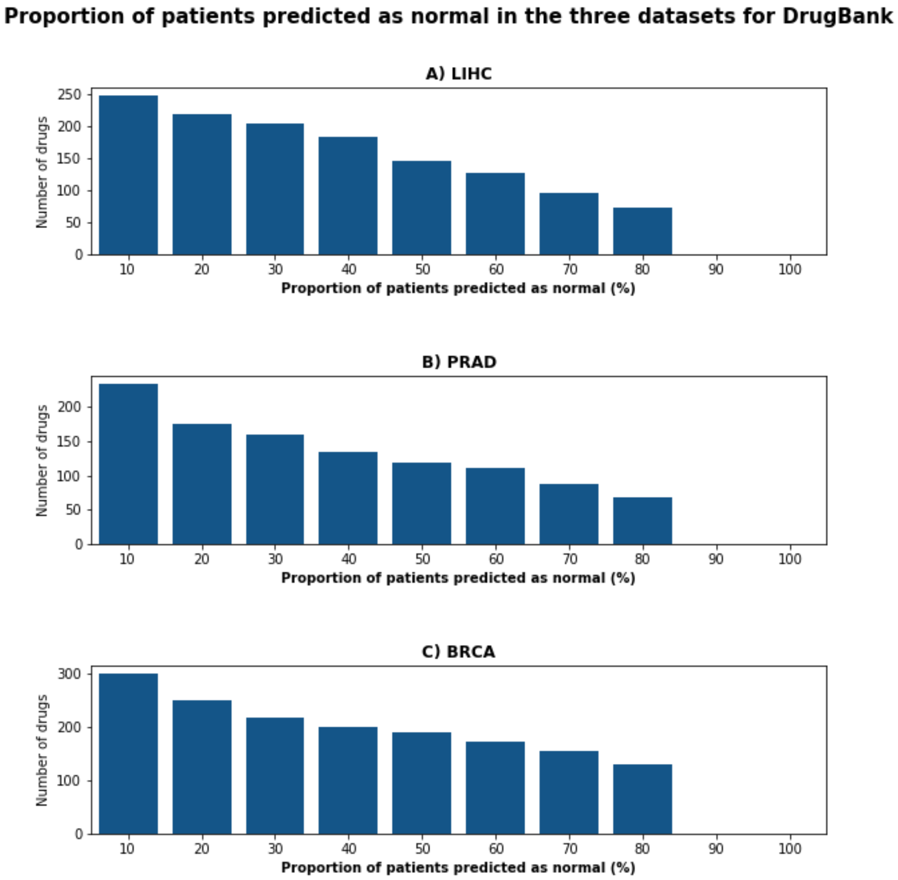
Supplementary Figure 4. Pathways targeted by the prioritized drugs in DrugBank for PRAD. The X-axis corresponds to the pathways targeted by any of the prioritized drugs (KEGG pathways not targeted by any prioritized drug have been omitted for better visualization). Drugs (Y-axis) have been clustered based on the pathways they target. Due to the large number of pathways, we have clustered pathway groups together for visualization purposes. Black cells correspond to pathways targeted for each drug. Details about each pathway are displayed in the following Jupyter notebook https://github.com/sepehrgolriz/simdrugs/blob/main/scripts_and_notebooks/heatmaps.ipynb.



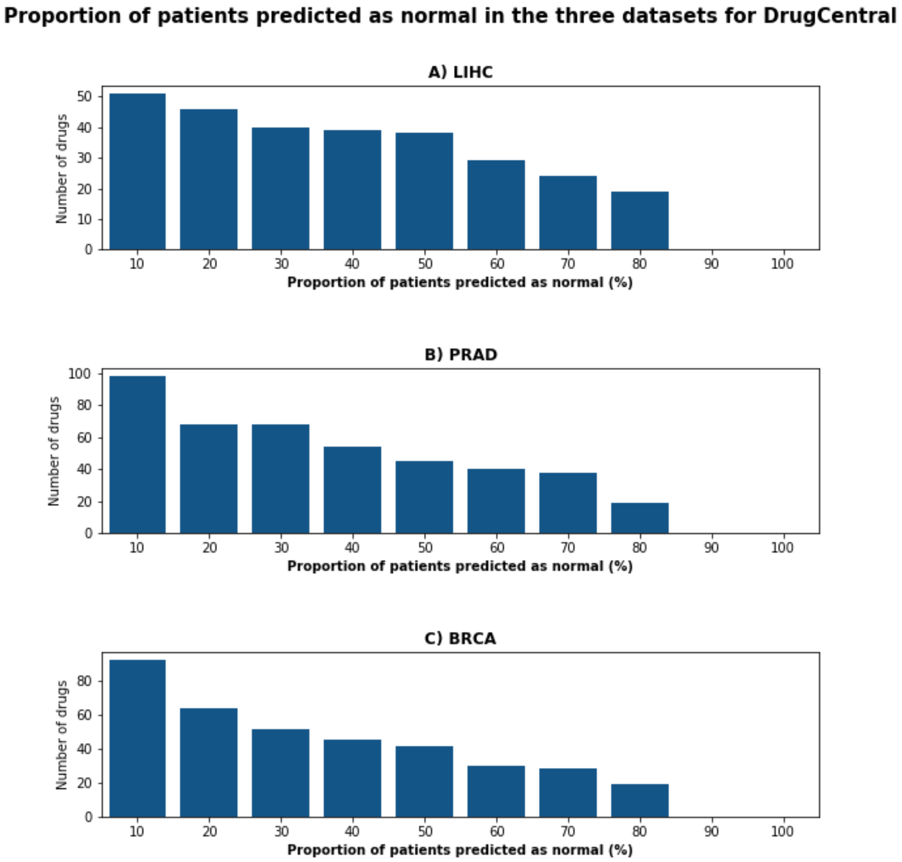
Supplementary Figure 5. Distribution of pathway activity scores for each of the three test datasets.



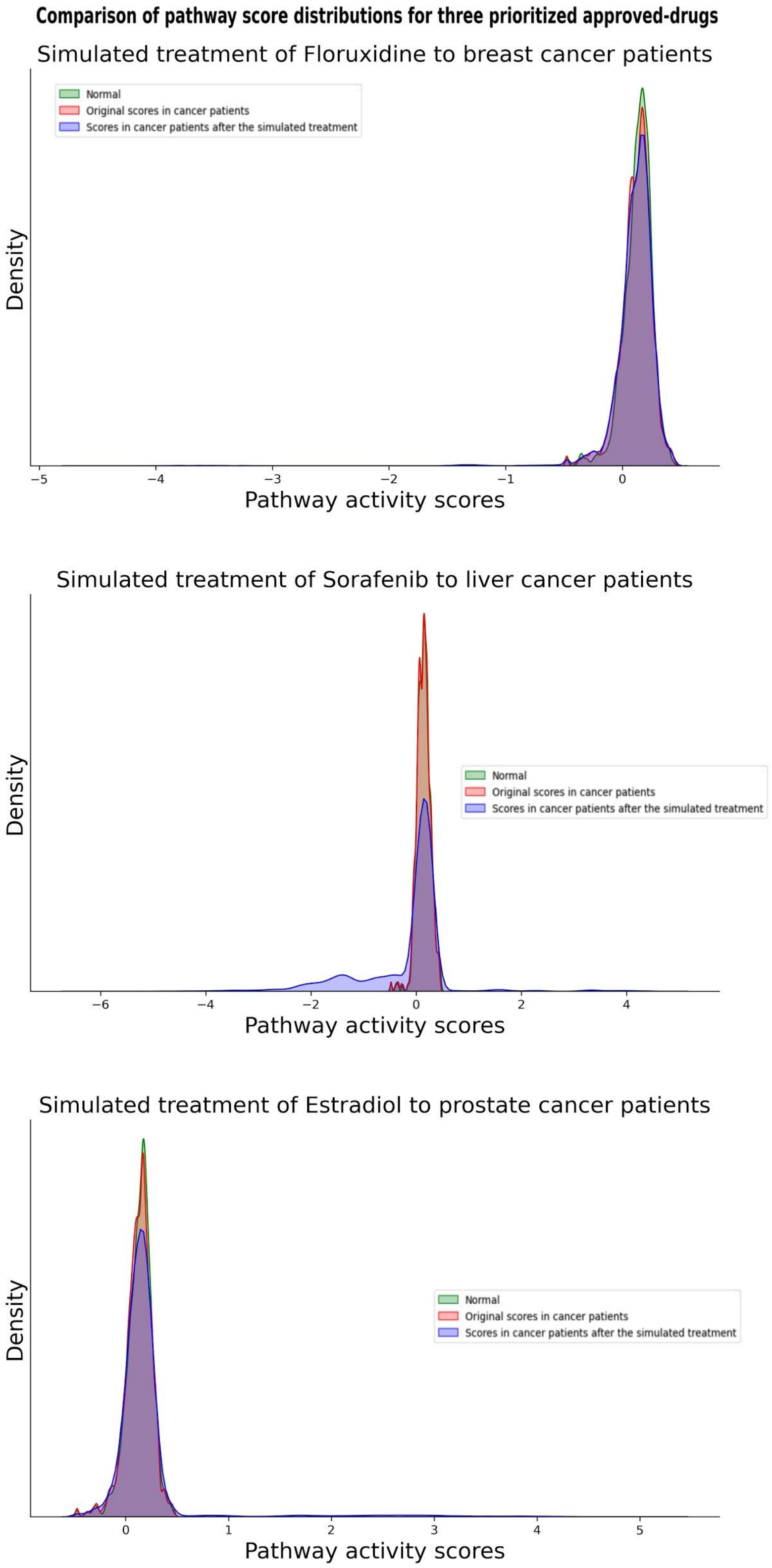
Supplementary Figure 6. Prediction performance measured as AUC-ROC values of an elastic net classifier (tumor vs. normal samples) trained on the three test TCGA datasets using pathway activity scores from ssGSEA run on KEGG. Each boxplot shows the distribution of the AUCs over 10 repeats of the 10-fold cross-validation procedure. The same classifiers yielded equivalent AUC-PR values (data not shown). Similar results were obtained in the KIRC dataset (see https://doi.org/10.3389/fgene.2019.0120 Figure 4).



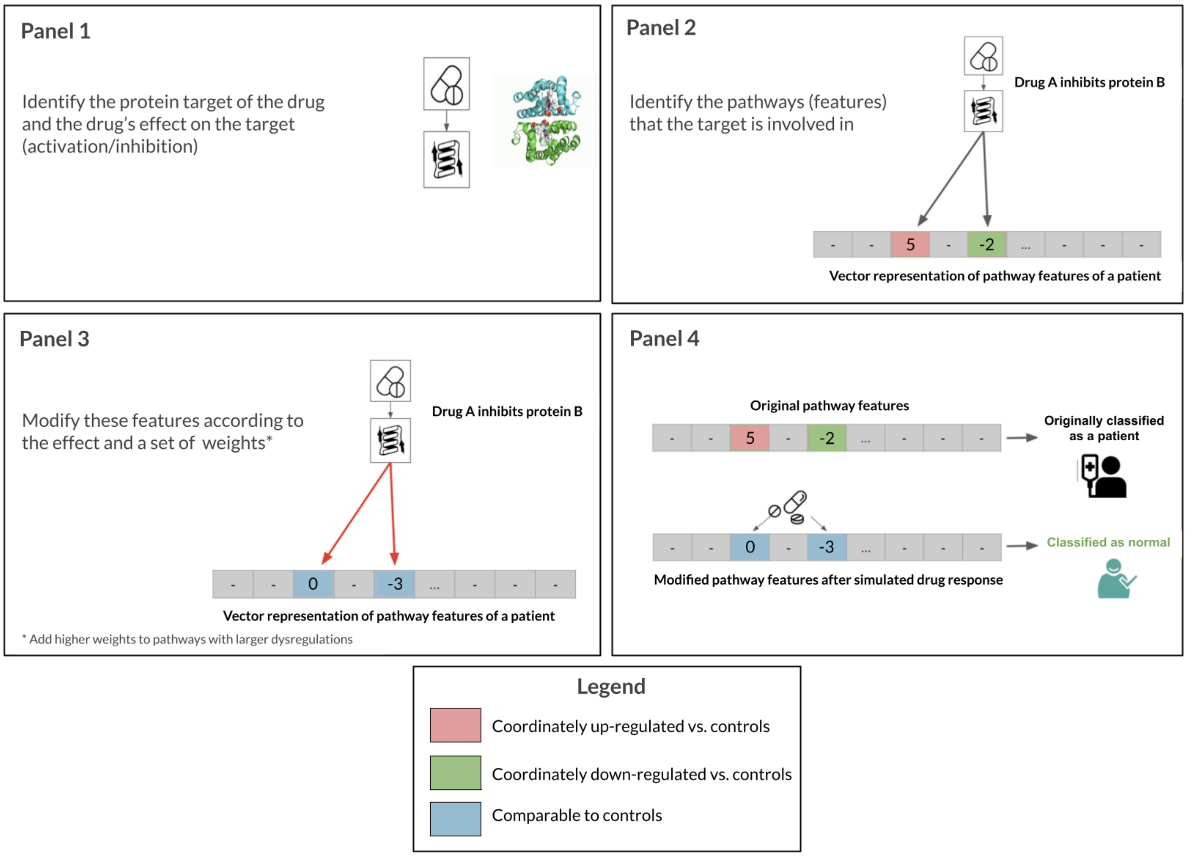
Supplementary Figure 7. Proportion of patients predicted as normal for each cancer test dataset using DrugBank. Only a fraction of all drugs in DrugBank changed the predictions of 10% of the patients to normal. As the proportion of the samples changed increases, the number of prioritized drugs decreases to 19 for all three datasets.



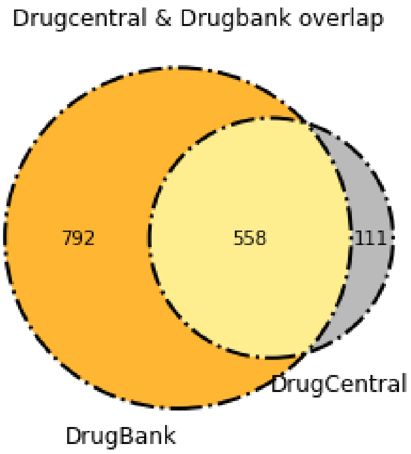
Supplementary Figure 8. Proportion of patients predicted as normal for each cancer test dataset using DrugCentral. Only a fraction of all drugs in DrugCentral changed the predictions of 10% of the patients to normal. As the proportion of the samples changed increases, the number of prioritized drugs decreases to a shortlist of drugs for all three datasets.



Supplementary Figure 9. Comparison of the distributions of pathway activity scores before and after the simulated treatment of three approved drugs prioritized by our approach in each cancer test dataset. In the three cases, we can see that only a minority of the pathway activity scores are modified after the simulated treatment (i.e., outliers that appear on each end of the distribution depending on the effect of the drug).


Panel 1
Identify the protein target of the drug and the drug's effect on the target (activation/inhibition)
Panel 2
Identify the pathways (features) that the target is involved in
Drug A inhibits protein B
- - 5 - -2 ... - - -
Vector representation of pathway features of a patient
Panel 3
Modify these features according to the effect and a set of weights*
Drug A inhibits protein B
- - 0 - -3 ... - - -
Vector representation of pathway features of a patient
* Add higher weights to pathways with larger dysregulations
Panel 4
Original pathway features
- - 5 - -2 ... - - -
Originally classified as a patient
- - 0 - -3 ... - - -
Classified as normal
Modified pathway features after simulated drug response
Legend
Coordinately up-regulated vs. controls
Coordinately down-regulated vs. controls
Comparable to controls

Supplementary Figure 10. Illustration of how the mechanism of action of a drug is simulated by the algorithm.



Supplementary Figure 11. Overlap of drugs present in DrugCental and DrugBank.