Building Biomarkers

to

Bridge the Brain, Behavior, and Brain Diseases



Oliver Y. Chén

Supervised by

Professor Maarten de Vos

and

Professor Lionel Tarassenko

This thesis is submitted to the Department of Engineering Science,

University of Oxford, in fulfillment of the requirements for the degree of

Doctor of Philosophy

St. Edmund Hall

Friday 22nd January, 2021

Acknowledgements

I would like to thank Maarten de Vos and Lionel Tarassenko for their guidance, encouragement, and, above all, trust.

I would like to acknowledge funding from the University of Oxford and F. Hoffmann - La Roche.

I would like to thank Huy Phan, Guy Nagels, Tianchen Qian, Luke Gustafson, Serena Ly, and Chris Hays for helpful comments on earlier versions of the thesis.

I would like to thank my friends and colleagues from the Engineering Department and the St. Edmund Hall for their conversations, time, and support.

I would like to thank the assessors during the *Transfer of Status* examination, *Confirmation of Status* examination, and the final *Viva*. Their suggestions, criticisms, and encouragements have improved the rigor and readability of the work.

I am indebted to my mentors throughout my career who have taught me knowledge, and, more importantly, humanity. Their character I will remember for evermore.

I would like to thank my friends who accompanied me during good times and supported me during bad times.

Finally, I would like to thank my parents, maternal grandparents, and Godparents for their love and support. I attribute my minor achievement in this thesis, if any, to my late paternal grandparents, who taught me to be honest, to be diligent, and to serve people and society.

To God

for His grace.

Abstract

The brain is one of the most important and complex organs in humans. On the one hand, it dictates our behavioral outcomes, such as seeing, speaking, and walking, and, through these behaviors, it determines how we interact with and perceive the world. On the other hand, its pathological outcomes, such as the development of neurodegenerative diseases, affect the longevity and quality of millions of people's lives. Today, the brain's neural activities can be captured by neuroimaging modalities such as functional magnetic resonance imaging (fMRI), and its behavioral markers can be recorded using smartphone sensors. Through uncovering how patterns of these large-scale data are related to the brain's behavioral and pathological outputs, one contributes to our understanding of biological underpinnings of knowledge, consciousness, cognition, and behavior. In parallel, by developing targeted statistical frameworks, one begins to unveil the intrinsic spatial and temporal relationships between neural signatures, behavioral features, and the brain's pathological outcomes, and to assess and monitor neurodegenerative diseases more effectively and accurately.

This thesis consists of four inter-correlated topics aimed at building targeted and efficient biomarkers using large-scale data to inquire into human behavior and brain disorders. The four topics are arranged according to their logic flow (from relatively simple to reasonably complex) and will be present in the following order: assessing outcomes of the brain using (a) **static**, (b) **dynamic**, (c) **intermediating**, and (d) **longitudinal** biomarkers.

1. **Building static biomarkers.** Methodological approaches are developed to identify time-averaged biomarkers to assess brain disease status and severity. This is supported by an article on developing static biomarkers to assess the status and severity of Parkinson's disease.

2. **Building dynamic biomarkers.** Methodological approaches are developed to identify and isolate dynamic markers or directed information flows in the whole brain. This is supported by an article on extracting dynamic neural information flows and using them to predict cognitive flexibility in humans.

3. **Building intermediating biomarkers.** In a complex regulative biological system, some markers are on the one hand associated with behavior (*e.g.*, prodromal symptoms), and on the other hand give rise to brain disorders (*e.g.*, psychosis). Methodological approaches are developed to identify, isolate, and quantify these intermediating biomarkers or neural mediators. This is supported by an article on brain-wide functional mediation analysis.

4. **Building longitudinal biomarkers.** The patterns of biomarkers fluctuate over time. Identifying biomarkers whose temporal patterns are coupled with neurodegenerative disorders longitudinally is useful for assessing the development and trajectory of the diseases. This is supported by an article on developing longitudinal biomarkers to assess multiple sclerosis over time.

Taken together, the methodological explorations and neurobiological discoveries in this thesis suggest that biomarker-based predictive models are promising for advancing our understanding of the human brain and behavior, and for facilitating brain illness assessment and monitoring in an automated way.

Table of Contents

1 Introduction 1

2 Building static biomarkers 72

3 Building dynamic biomarkers 104

4 Building intermediate biomarkers 138

5 Building longitudinal biomarkers 191

6 Conclusions and discussions 219

Chapter 1

Introduction

Prologue

All behavior is the result of brain function; all behavioral disorders that characterize mental and psychiatric illnesses are, in one way or another, consequences of disturbances of brain function (see p. 5 in Kandel *et al.*[1]). One therefore cannot paint a complete picture of human behavior without referring to brain functions that dictate it. Nor can one fully unfold the neural basis of brain disorders without considering their corresponding behavioral symptoms.

Attempts aiming at charting the relationships between the brain, mind, body, and behavior have run through the history of scientific and philosophical enquiries from their earliest days to the present time. During the Roman Empire, Greek physician Aelius Galenus (Galen) argued in *On the Doctrines of Hippocrates and Plato* that the brain is the location of the ruling or directive part of psyche[2,3]. For centuries, Western medicine followed Galen's view (see p. 6 in Kandel *et al.*[1]). The 17th century saw a rise of the mind-brain argument, signified by opposing views from two towering philosophers René Descartes and Baruch Spinoza. Descartes gave a proof in *Meditations on First Philosophy* that "mind is a reality distinct from body"[4], whereas Spinoza argued in his *Ethics* that "the mind and the body are a single thing that is being thought of in two different ways"[5]. Subsequent thinkers extended the discussion to include emotion, behavior, and perception. Immanuel Kant wrote in *The Critique of Pure Reason* that, "intuitions [perceptions] without concepts are blind," arguing that all inputs into the mind must be somehow organized by being based on, or interfaced through, concepts[6]. Charles Darwin connected the brain with joy, despair, depression, and pain, and gave examples in *The Expression of the Emotions in Man and Animals*, for example, that "[j]oy quickens the

circulation, and this stimulates the brain, which again reacts on the whole body". Darwin also inquired into the neural origin of irregular behavior, stating that "[i]nflammation of the membranes of the brain leads to the engorgement of the face, ears, and eyes with blood"[7].

But it was not until the neurobiologists who had studied the anatomy of the brain and on patients with brain disorders *postmortem* did the myths about how functional organization and anatomical structure of the brain underpin behavior, and how brain lesions and disrupting neural functions give rise to brain diseases, begin to unfold.

Among the contributions made by neurobiologists are two fundamental theories about the brain: functional specialization and functional integration.

Functional specialization of the brain. A central principle underlying the relationship between brain and behavior is functional specialization, introduced in 1861 in Paris. It states that different brain areas process different perceptual or motor tasks. Previously, German neuroanatomist Franz Joseph Gall had hypothesized, without empirical evidence, that each region of the cerebral cortex controlled a specific function. By studying patients *postmortem* who were unable to formulate language but able to comprehend it – now referred to as Broca's aphasia - French neurologist Pierre Paul Broca located an area in the posterior region of the frontal lobe specialized in language production. Using this evidence as a gateway, Broca delivered his new-born idea of functional specialization at a meeting of the Société d'Anthropologie in 1861. Today, this language area is named in honor of him as Broca's area. The localization of a special language area opened the door for pinpointing other brain areas specialized for a broad range of functions. For example, Gustav Fritsch and Eduard Hitzig discovered the motor cortex, via studying dogs' brains, in Germany in 1870; Karl Wernicke

discovered the brain (or Wernicke's) area specialized for comprehending words but not for forming them in 1876. In the late 1930s, Edgar Bard and Wade Marshall charted a precise brain map with each anatomical area linked to a part of the body. In 1991, British neurologist Semir Zeki directly demonstrated functional specialization within the visual cortex related to dynamic form (in V3), color and form-with-color (V4), and motion (V5)[8]. Today, it is generally agreed that the frontal lobe of the brain is mainly specialized in short-term memory, future planning, and movement control; the parietal lobe in somatic sensation and in forming a body image and relating it to extrapersonal space; the occipital lobe in vision; the temporal lobe in hearing; the hippocampus and amygdaloid nuclei in learning, memory, and emotion[1].

Functional integration of the brain. Another fundamental principle underpinning the brain-behavior interface is functional integration. Functional integration states that the brain brings together signals from multiple physically distinct but interconnect cortical and sub-cortical areas to construct complex perceptual and motor activities. The idea of functional integration was first formed by Wernicke. He combined the two then predominant theories about the brain: the holistic view wherein every mental function involved the entire cerebral cortex, and the cellular connectionism view that the cortex was a mosaic of functionally specific, separated areas. He argued that complex cognitive functions were products of interconnections between several potentially distinctive and separate functional sites, and that different components of a single behavior were likely to be processed in multiple brain regions (see p. 12 in Kandel *et al.*[1]).

Evidence of functional integration was found in language and vision processing. Wernicke postulated that language involved the integration of separate motor and sensory programs: the former governed the mouth movements for speech (speech production) and was located in Broca's area and the latter dictated word perception (speech comprehension) and was located

in the temporal lobe (Wernicke's area). Based on the fact that these two areas were located in distinctive sites connected by the arcuate fasciculus, Wernicke predicted a third type of aphasia, *paraphasia* (or *conduction aphasia*), as a result of disconnections between the two areas (which disrupted the auditory-motor integration). Patients with paraphasia can speak fluently but not coherently (*i.e.*, paraphasic) and have difficulty repeating phrases; yet they understand the words that they hear and read (see p. 12 in Kandel *et al.*[1]). Another set of evidence was found in visual integration. The visual brain collects signals from distinct visual areas each of which specialized in processing color, form, and motion information; integrating these visual components one produces a unified perception of the vision world[8,9].

1. The need to build biomarkers to study and manage brain disorders

The facts of functional specialization and integration of the brain suggest the possibility to isolate and pinpoint behavior- and disease-related brain regions. Additionally, they highlight the importance of combining neural and behavioral observations to assess and predict human behavior and brain disorders.

1.1 An overview of brain illnesses studied in this thesis

Before I introduce the definition and categorization of biomarkers used in this thesis, here I first give an overview of three common brain diseases studied in this thesis: Parkinson's disease, psychosis, and multiple sclerosis.

1.1.1 Parkinson's disease

Parkinson's disease (PD) is a neurodegenerative disorder that affects multiple neurological systems. It disturbs seven million people worldwide; the prevalence increases from 1% of the population for those over 60 years of age to 4% over 80[10]. The fundamental neural basis of

Parkinson's disease (PD) is the progressive loss of dopaminergic neurons, largely in the substantia nigra. The behavioral hallmarks of the PD are resting tremor, bradykinesia, rigidity, and impairment to initiate and sustain movements. Additionally, patients may have shuffling gait, experience difficulty to maintain balance, and present diminished facial movement with little expression, as if wearing a mask. In some forms of PD, there are also genetic causes. Several genetic loci have been identified (*e.g.*, PARK1-PARK8, PARK 10, and PARK 11).

PD diagnosis. At present, a definitive diagnosis of PD can be only made via autopsy *postmortem*, thereby making behavior feature identification paramount in early diagnosis of PD. The severity of the disease can be quantified using the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[11]. MDS-UPDRS scores measure the progression of an individual's Parkinson's disease, and serve as the gold standard for validation[12]. The total score equals the summation of subscores obtained from 42 items covering four subscales: Part I: mentation, behavior, and mood; Part II: activities of daily living; Part III: motor examination; Part IV: complications of therapy. The total score ranges from 0 and 199 points, where a patient with a higher score would be considered to have more severed PD[13].

PD treatment. Whereas there is at present no cure for PD[14], there are several ways to potentially manage the disease. For patients with mild motor symptoms, a monoamine oxidase type B inhibitor (MAOBI; selegiline or rasagiline) may be tried before using a more potent treatment such as levodopa or a dopamine agonist[15]. As the disease progresses, the medication should be adjusted to obtain optimal symptom control[16]. For patients whose daily life is severely impaired by the disease, levodopa or a dopamine agonist is usually initiated. The former is the most employed PD treatment to reduce motor symptoms[17], and is more

favored (compared to dopamine agonists and alternative therapies) to treat older patients, especially those with preexisting cognitive dysfunction, who are at greater risk of developing psychiatric adverse effects[16]. If patients' dyskinesia and motor fluctuations are refractory to medical therapy and postural instability and/or gait impairment remains responsive to levodopa, one could consider deep brain stimulation; but this needs to generally only be done to patients younger than 70 years[18–20].

PD data used in the thesis. In this thesis, I used data collected from two independent smartphone-based remote monitoring studies[21]. The first study was a six-month-long phase 1b clinical drug trial of PRX002/RG7935 (now known as prasinezumab) conducted by Prothena and Roche, which consisted of 44 PD participants (NCT02157714). The second study was a six-week-long observational study of 35 age- and sex-matched healthy controls (HCs). All controls scored $\geq$ 26 points on the Montreal Cognitive Assessment (MoCA)[11] and were free of a cardiovascular, neurological, or psychiatric condition, and had no first-degree relative with PD. The study also included the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). MDS-UPDRS scores measure the progression of an individual's Parkinson's disease, and serve as the gold standard for validation[12].

1.1.2 Psychosis

Psychosis is a brain disorder characterized by three chief symptoms: hallucinations, delusions, and confused and disturbed thoughts. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) characterizes disorders as psychotic or on the schizophrenia spectrum if they involve hallucinations, delusions, disorganized thinking, grossly disorganized motor behavior, or negative symptoms[22]. The pooled median global prevalence of psychotic disorders was 4.6 per 1000 persons and the median lifetime prevalence was 7.49 per 1000

persons[23]. The functional deficits (for example, inability to work or be an independent member of the community, social isolation, and lack of self-care) are long-established hallmarks for chronic psychosis[24–26]; they are reported to be nonresponsive to antipsychotic treatment, even when positive symptoms show marked improvement[27–29]. Although the neural basis of psychosis is not fully charted, there is converging evidence suggesting that the limbic structures play a central role in the psychopathology of psychosis; but the frontal and parietal structures may also play a role, for example, in the delusional misidentification syndromes (in which impairments of face recognition memory are present)[30].

Psychosis diagnosis. The disease can be formally examined using The Brief Psychiatric Rating Scale (BPRS)[31]; it assesses the level of 18-24 (depending on the version) symptom constructs of psychosis such as depression, anxiety, hallucinations, and unusual behavior, and each symptom is rated 1-7.

Psychosis treatment. For individuals at increased risk of developing psychosis, they could be offered individual cognitive behavioural therapy (CBT) with or without family intervention, and be offered interventions recommended in National Institute for Health and Care Excellence (NICE) guidance for people with any of the anxiety disorders, depression, emerging personality disorder or substance misuse[32]. For individuals with first-episode psychosis, or those with an acute exacerbation or recurrence of psychosis or schizophrenia, treatments include oral antipsychotic medication in conjunction with psychological interventions (family intervention and individual CBT)[32].

Psychosis data used in the thesis. In this thesis, I used fMRI data drawn from the second phase of the North American Prodrome Longitudinal Study (NAPLS-2) consortium[33], which

included 263 subjects recruited from eight study sites across the United States and Canada. All subjects met the criteria for the prodromal syndromes at the point of recruitment according to the Structured Interview for Prodromal Syndromes (SIPS)[34] and were clinically followed-up for two years. During follow-up, 25 subjects developed one type of Axis-I psychotic disorders (CHR convertors, age 18.52 ± 4.08 years, 17 males) and 238 did not (CHR non-convertors, 19.07 ± 4.16 years, 136 males). The conversion was defined as the individual either met the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV)[35] criteria for an Axis-I psychotic disorder or had at least one fully psychotic symptom assessed by the SIPS at follow-up[36,37].

1.1.3 Multiple Sclerosis

Multiple Sclerosis (MS) is a neuro-degenerative and -inflammatory disease leading to physical and cognitive impairment, pathological fatigue, and depression[38]. Myelin (insulating covers) of nerve cells in MS patients' brain and spinal cord are damaged; the degeneration of myelin results in deteriorated communication within and between nervous subsystems, causing physical, mental, and sometimes psychiatric problems[39,40]. At the initial stage, patients may have blurred or double vision and red-green color distortion; some patients may even experience blindness in one eye. During later stages, most MS patients experience muscle weakness and have difficulties with coordination and balance; additionally, most patients also exhibit paresthesia, transitory abnormal sensory feelings such as numbness, prickling, or "pins and needles" sensations. Although the definitive neural basis for MS is unknown, evidence suggests that neurodegeneration occurs from the early stages of MS[41]. Critically, brain atrophy and abnormal microstructural integrity are found among MS patients. For the former, MS-related cognitive impairment has been associated with both cortical[42–44] and subcortical[45–50] gray matter atrophy; cortical lesions[51] explain between 20% and 60% of the variance of a

variety of cognitive tests assessing the most commonly affected cognitive domains using multilinear models. For the latter, MS leads to demyelination entailing reduced structural connectivity; several studies have found microstructural abnormalities in specific tracts such as the fornix[52,53], the cingulum[54], and the uncinate fasciculus[55].

MS diagnosis. At present, MS is diagnosed based on a combination of in-clinic assessments, including checking for MS-related signs and symptoms, neuroimaging studies, and laboratory testing[56]. A useful way to quantify MS severity is to use the Multiple Sclerosis Impact Scale (MSIS-29), a patient-based rating scale measuring the physical (20 items) and psychological (nine items) impact of the disease[57–59].

MS treatment. There is at present no cure for MS, but there are a few therapies that are used to manage the disease. One may begin treatment after the first attack, when more than two lesions are seen on MRI[60]. For patients during symptomatic attacks, one could administer high doses of intravenous corticosteroids, such as methylprednisolone[39]; alternatively, one could administer oral corticosteroids, which have a similar efficacy and safety profile[61]. Corticosteroid treatments are effective in the short term for relieving symptoms, but they seem to not have a significant effect on long term recovery[62,63]. For patients whose symptoms do not respond to corticosteroids, one could consider applying plasmapheresis[39]. Several disease-modifying treatments are currently available to manage relapsing-remitting MS and progressive MS. To treat relapsing-remitting MS, one could use interferon beta-1a and beta-1b[64], glatiramer acetate, mitoxantrone, natalizumab[65], fingolimod[66], teriflunomide[67], dimethyl fumarate[68], alemtuzumab[69], ocrelizumab[70], siponimod[70,71], cladribine[70], and ozanimod[72]. One could consider ocrelizumab to treat primary progressive MS (it is also used to treat clinically isolated MS syndrome,

relapsing-remitting MS, and active secondary progressive MS)[70,73], or siponimod and cladribine to treat secondary progressive MS[70].

MS data used in the thesis. In this thesis, I used data from a smartphone-based remote monitoring study (NIH Clinical Trial Identifier: NCT02952911). The study consisted of 76 MS patients (18–55 years; EDSS score 0–5.5). Smartphone data were collected using the Floodlight application, developed by Roche and Genentech[74]. The application contained a range of remote monitoring tests that measure a person's ability to perform simple tasks on smartphones. Each participant received a preconfigured smartphone that prompts the user to perform the Floodlight test battery, comprising 'active tests' and 'passive monitoring', for 24 weeks[75]. The data considered in this thesis are objective measurements obtained from a Two-Minute Walk (TMW) test, which contains gait features, and a Draw a Shape (DaS) test, which contains hand function features. Specifically, the gait features assess difficulties, fatigability, or unusual patterns during a two-minute walk test. The DaS test consists of drawing five shapes with circular and linear features to assess fine finger control and stroke sequencing. The study also included the clinician-rated Expanded Disability Status Scale (EDSS) scores, and the self-reported MS Impact Scale (29-item scale) (MSIS-29) scores.

1.2 The importance of developing biomarkers

Despite advances, the etiological and pathogenetic origins of most mental disorders have not been fully charted. For example, disease assessment and diagnosis at present are primarily based upon patients' description, clinic observation, recorded natural history, and the response to treatment (*e.g.*, see p. 1389 in Kandel *et al.*[1]). The search for and development of biomarkers, therefore, are crucial for early diagnosing brain illnesses and therefore early managing and treating them. But their search and development must take a parallel approach. On the one

hand, one shall continue to pursue (etiological and pathogenetic) biomarkers at the genetic, molecular, and cell levels that directly underpin behavioral outputs and brain disorders. On the other hand, when a definitive etiological and pathogenetic cause is unknown, enquiry expensive (*e.g.*, taking fMRI scans of a large number of individuals' brain), and measurements inconvenient (*e.g.*, frequently obtaining diagnostic markers and clinical scores during in-clinic visits), one shall explore alternative markers, such as neuroimaging, sensor, and behavior features, that may be associated with or predictive of behavior and brain disorders.

1.3 The definition and division of biomarkers

1.3.1 The definition of biomarkers

There are several ways to define biomarkers in scientific studies. The classic definition of biomarkers considers only biological entities, namely *cellular, biochemical, or molecular alterations that are measurable via biological media such as human tissues, cells, or fluids*[81]. Another, broader, definition considers a biomarker to be any biological characteristic that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention[82].

In this thesis, I define biomarkers as measurements, tools, and technologies that are useful to study brain diseases, brain functions, and human behavior. They include etiological factors (genetic, biochemical, toxin, and microbe) and pathogenesis (the mechanism by which etiological factors give rise to disease) that are directly linked with behavior or disease, and biological features—obtained from advanced tools and technologies—including imaging data (such as functional magnetic resonance imaging or fMRI data) and digital measurements (such as the severity of bradykinesia measured by a smartphone sensor) that are associated with and predictive of brain activities, human behavior, and mental disorders.

Because of the diverse and complex nature of human behavior and the brain, future methodological and neurobiological studies may spawn new, broader definitions and finer classifications for biomarkers. I, however, hope that the definition and classification made in this thesis (see **Section** 1.3.2 and **Chapters** 2-5) may provide some useful insights both conceptually and practically to produce targeted and efficient biomarkers.

Naturally, the introduction of biomarkers raises the question of causation. Probing into the relationships between biomarkers, brain functions, and human behavior is not merely a problem of neurobiological or statistical interest. It is an important problem in the causal inference that has fascinated scholars for ages—namely, if a selected feature is associated with or can be used to predict outcomes (even) in novel subjects, is this feature causally linked to the outcome? The short answer is not necessarily. After presenting a collection of works in building different types of biomarkers related to the brain and behavior, I shall discuss some aspects regarding the biological basis for association, explanation, prediction, and causation in **Chapter** 6 of the thesis.

1.3.2 The division of biomarkers

Here I shall give a brief account of the nature of biomarkers in behavioral and brain studies. This will help in understanding not only the importance and convenience of classifying different biomarkers, but also many of the descriptions given later in this thesis.

Traditionally, biomarkers in brain studies have been divided into two classes: biomarkers of exposure, and biomarkers of disease[83]. The former class is useful in risk prediction, and the latter is used in disease screening as well as diagnosis and monitoring of disease progression. Although such a division has many useful implications in the clinic and neurological studies, it

does not provide detailed descriptions of the functional or structural properties of neural biomarkers. To better understand the very contribution biomarkers make to chart neural and behavioral pathways and assess, predict, and monitor brain functions and human behavior in health and disease, here I suggest a modified version of biomarker classification. Chiefly, I divide biomarkers into four categories: static, dynamic, intermediate, and longitudinal biomarkers (see **Fig.** 1).

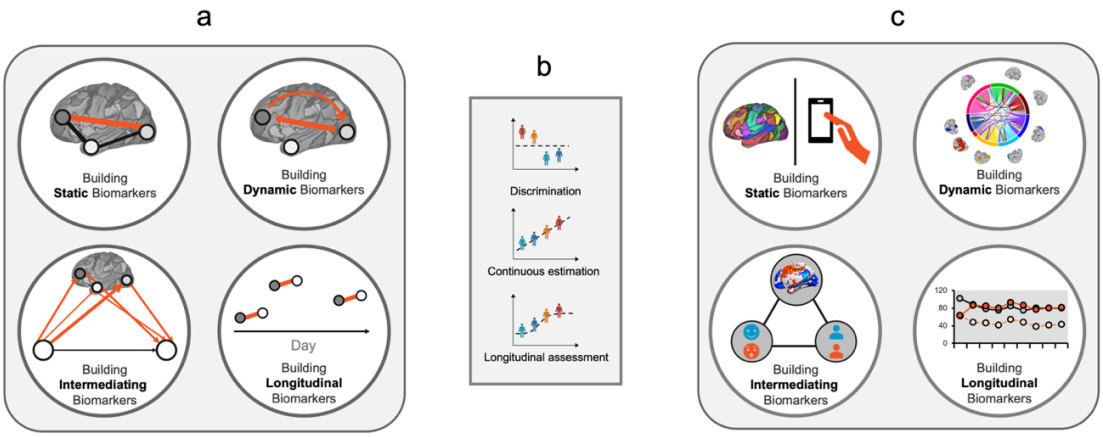


Fig. 1. Four types of biomarkers. (a) A schematic representation of static, dynamic, intermediating, and longitudinal biomarkers. **(b)** Three key goals of predictive modeling using biomarkers: discrimination analysis, continuous outcome estimation, and longitudinal assessment. In brain disease studies, these three goals correspond to, for example, classifying subjects into patients and healthy controls, estimating patients' disease scores (such as the UPDRS scores of Parkinson's patients), and forecast the disease course over time (such as the fluctuation of EDSS scores of multiple sclerosis patients over time). In behavioral studies, these three goals correspond to classifying categorical behavior outputs (such as normal IQ and high IQ), estimating behavior measurements (such as how flexible one's cognitive ability is), and monitor the longitudinal trajectory of behavior measurement (such as one's physical activity over time). **(c)** Schematic representation of studies included in this thesis. Static features obtained from the smartphone are used to identify Parkinson's patients and to estimate their UPDRS scores. Dynamic information flows from the whole brain are mapped; they are subsequently used to predict cognitive flexibility in humans. Intermediate brain patterns are uncovered; they are on the one hand associated with prodromal symptoms and on the other hand potentially linked to later conversion into psychosis. Longitudinal features recorded from smartphones are employed to monitor the trajectory of disease scores among patients with multiple sclerosis.

Static biomarkers

A static biomarker summarizes the patterns of each behavioral or neural feature measured during a period into a point. A potential static biomarker can be a feature obtained cross-sectionally, a single value averaged from longitudinal measurements of one variable (*e.g.*, the arithmetic mean of a time series), or a single metric summarized from multiple longitudinal variables (*e.g.*, the correlation between two time series). In other words, a static feature is a point (estimate) of the feature(s). That to summarize a feature into a point does not mean that the problem becomes univariate; rather, it can still be a multivariate and potentially multilevel problem, where a set of multivariate features and multiple sub-groups - each consisting of a single point - are concerned.

The usefulness of static biomarkers is threefold. First, it is handy to map static features onto an outcome variable that is also a point (such as an individual's disease status) to estimate their relationship, and to perform classification or continuous estimation. Second, it is a useful de-noising approach (see **Chapter** 2 for a detailed explanation and discussion) and provides a succinct summary of the features. Third, it is simple, and as a result, it is convenient to interpret the associations between the features and the outcome. Examples of static biomarkers are averaged health data (such as physical activities averaged over a period) and brain connectivity (correlations between brain activations during a period).

Dynamic biomarkers

Static features, such as brain connectivity and averaged behavior measurements, can only describe a limited aspect of the functions, functioning, and functional integration of the brain and behavior. For how could one chart, for example, the visual pathways in the cerebral cortex using a directionless metric such as correlation or average activation? One could of course

describe the neural pathways using undirected connectivity (*i.e.*, links between brain regions that do not indicate directed influence; see **Chapter** 3 for details). But there is a problem. For example, it makes sense to say the retina is connected with the primary visual cortex (V1) via the lateral geniculate nucleus (LGN); it also makes sense to say that the retina sends visual information to the V1 for processing via LGN; it does not make sense when the order is reversed. Indeed, it is the directed intra- and inter-region connectivity that gives rise to the specialized visual perception; and each type of visual perception is achieved, and can only be achieved, via the underlying directed pathways, not *vice versa*.

Here, I define dynamic biomarkers as directed edges between brain areas. Suppose we use two circles to denote two brain areas A and B. A directed edge between areas A and B could be either a link between A and B with an arrow pointing towards B (which indicates the direction is from A to B), or a link with an arrow pointing towards A (which indicates the direction is from B to A). Thus, there are four possible scenarios regarding directed edges between A and B: (i) there is not any link between A and B, (ii) there is a directed link from A to B, (iii) there is a directed link from B to A, and (iv) there is a directed link from A to B and another from B to A. In this thesis, they are used to describe the cumulative directed influence one brain area has over another during a period. Their dynamic property is hence highlighted by both the directions of the information transmission between every pair of brain regions and the potentially heterogeneous flow strength corresponding to each direction. The dynamic biomarkers should not be confused with biomarkers that change from moment to moment over time, which fall into the class of longitudinal biomarkers (see **Section** 3.4).

Chief to directed edges in brain studies is the concept called effective connectivity, “a circuit diagram that would replicate the observed timing relationships between the recorded

neurons"[87]. Effective connectivity is dynamic (activity-dependent), and depends on a model of interactions or coupling[88]. The idea of effective connectivity originates from the analysis of individual spike trains obtained from multiunit electrode recordings[87,89–91] but has since extended to study general neural systems. Compared to directionless functional connectivity, effective connectivity *mediates the influence that one neuronal system (either at a synaptic or cortical level) exerts on another and, therefore, discounts other influences*[88].

An enquiry into the dynamic biomarkers underpinning brain function and human behavior can be made in two steps. First, one charts a diagram formed by directed edges between brain areas. Subsequently, one associates each directed edge with human behavior or brain diseases to reduce the likelihood that the directed edge is spurious and to make further neurobiological enquires about their association.

Yet it remains possible that some of the identified directed edges are a mere numerical coincidence. A simple example can be found in Granger-analysis of the brain (see **Chapter** 3 for a detailed discussion on Granger analysis), where signals recorded from a brain region A that Granger-cause (causally antecede in the Granger-sense) signals recorded from another region B does not endorse that area A sends signals to area B to produce a particular motion or perception. A useful way to guard a method against yielding spurious claims on effective connectivity is to perform out-of-sample predictive modeling using directed edges. For example, by linking directed edges with brain disease status or behavior metric, one can select directed edges that are associated with the disease or behavior; by validating selected edges on novel subjects, one can further verify if the directed edges can be extrapolated (see **Chapter** 3). After presenting a study of dynamic biomarkers in the whole brain related to cognition, I shall return to discuss some aspects related to biomarkers and causation in **Chapter** 6 of the thesis.

Intermediate biomarkers

Both static and dynamic biomarkers could advance the studies of brain functions in relation to human behaviour and brain disease outcomes. Oftentimes, however, one may find that the brain is interposed between an input variable and an outcome variable. How can one identify and isolate brain areas that are potentially intermediating the input and outcome? Here, I define intermediate biomarkers as a subset of brain areas whose patterns are potentially mediating the input variable (*e.g.*, prodromal symptoms) and the outcome variable (*e.g.*, later conversion into psychosis).

Mediation analysis is useful to uncover intermediate biomarkers. But it needs to confront a few challenges. Although existing mediation models have made the search for mediators fruitful, they are not suitable for studying high-dimensional mediation analysis with binary disease outcomes. For example, existing multi-level mediation models assume that the outcomes are continuously distributed[4–7]; mediation frameworks concerning binary outcomes are at present restricted to a relatively small number of mediators[6,8]; high-dimensional mediation models whose outcomes are not normally distributed do not have a closed-form solution (therefore it is difficult to estimate parameters analytically, as, for example, in Chén *et al.*[4]). Second, although functional mediation analysis[9] has considerably advanced knowledge about the functional signal organization of the brain in relation to explanatory and outcome variables, it remains unclear whether it is suitable for analyzing high-dimensional brain data, and if so, how the underlying data configuration, such as the sample size and noise level, would affect parameter estimation. In parallel, its efficacy needs to be evaluated for brain disease studies.

In this thesis, I further extend mediation analysis to general cases, where the intermediator variable is high-dimensional and functional and the outcome variable has a relaxed (*i.e.*, non-

Gaussian) distribution. I will demonstrate how to use this framework to identify brain regions that are potentially mediating prodromal behavioral symptoms and later conversion into neurological diseases.

Longitudinal biomarkers

The hallmark of neural degenerative diseases is longitudinal progression. The course of degeneration, however, differs significantly between disease types, across individuals, and over time.

Let's take multiple sclerosis (MS) as an example. There are four subtypes of MS: clinically isolated syndrome (CIS), relapsing-remitting MS (RRMS), primary progressive MS (PPMS), and secondary progressive MS (SPMS). Each type is characterized by a unique longitudinal disease property. A CIS patient has an attack suggestive of demyelination but does not fulfill the criteria for multiple sclerosis[39,56]; the patient, however, has a 30% to 70% chance of developing MS subsequently[96]. An RRMS patient presents unpredictable relapses followed by periods of months to years of relative quiet (remission) with no new signs of disease activity[39,56]. A PPMS patient shows a progression of disability from onset, with no, or only occasional and minor, remissions and improvements[96,97]. Finally, an SPMS patient displays an initial relapsing-remitting MS, and eventually progresses between acute attacks without any definite periods of remission (although occasional relapses and minor remissions may appear)[39,97].

How, then, can one differentiate subjects with different types of longitudinal MS development? Can we forecast the next onset of relapses for MS patients with RRMS or SPMS? When is the optimal time to assign a drug treatment to the patients? Central to addressing these questions

is to find longitudinal neural and behavioral biomarkers that can be used to assess, monitor, and predict the disease profile over time.

To qualify for a longitudinal biomarker, a feature needs the following three properties. First, it can be measured repetitively over time. Second, its temporal course is associated with either a behavioral variable or disease outcome that is also measured over time. Finally, its longitudinal pattern can be generalized to a broad population and thus is suitable for assessing, monitoring, and predicting outcomes over time in novel subjects. Although it is possible to establish the longitudinal relationship between one feature recorded over time (*i.e.*, one time series) and one (disease) outcome recorded over time (*i.e.*, another time series). The longitudinal biomarkers we discuss in this thesis are multivariate. In other words, we aim to find multiple (longitudinal) biomarkers whose longitudinally recorded measurements are associated with and predictive of (disease) outcomes over time.

Let us return to the example of MS. A well-established feature of the MS is one's walking ability[98]. For example, an MS subject with an EDSS (Expanded Disability Status Scale, ranging from 0 to 10, where the larger the EDSS score, the more severe MS is) of 4 is able to walk without aid or rest for 500 meters; a subject with an EDSS score of 6 requires a walking aid – cane, crutch, *etc*. and can walk about 100 meters with or without resting; a subject with an EDSS score of 9, although can still communicate and eat, is confined to bed. Naturally, if one records features related to walking ability regularly, one should be able to use it to paint a longitudinal picture of MS progression.

MS symptoms, however, span across multiple functional systems, including pyramidal (muscle weakness or difficulty moving limbs), cerebellar (ataxia, loss of balance, coordination, or

tremor), brainstem (problems with speech, swallowing, and nystagmus), sensory (numbness or loss of sensations), bowel and bladder function, visual function (problems with sight), cerebral functions (problems with thinking and memory). One feature, such as walking ability, naturally only provides one aspect of the disease. A multivariate longitudinal approach that combines features from multiple feature modalities, therefore, is chief to provide a fuller, and perhaps more accurate, description, assessment, and prediction of the disease outcome over time. In this thesis, I develop a longitudinal statistical model combining the generalized estimation equations and multiple imputation techniques. I will demonstrate how to use this framework to select and combine longitudinal walking and dexterity features to assess and predict MS development over time.

Although I focus on building biomarkers associated with and predictive of neurodegenerative diseases, a probe into the longitudinal brain and behavior trajectory during infancy and adolescence is important to study the developmental brain. After presenting a longitudinal biomarker study with regards to MS patients, I shall provide a discussion regarding longitudinal studies of the developmental brain in **Chapter** 6.

2. The aim and outline of the thesis

2.1 The aim of the thesis

The central theme of this thesis forms is to uncover biomarkers from neuroimaging, sensor, and behavior data and use them to distinguish different types of biomarkers and to develop suitable methods to identify them and use them to analyze brain functions, brain illnesses, and behavior. More specifically, to achieve the first goal, I divide the biomarkers into four categories: static, dynamic, intermediate, and longitudinal biomarkers (see **Section** 1). To achieve the second goal, I first overview statistical frameworks that are useful to study biomarkers from

each category, and then give four examples in **Chapters** 2-5 to showcase their usages in real-world applications.

I summarize biomarkers into four general categories that most would agree are distinctive and important for brain and behavioral studies, and develop methodological frameworks that are, to the best of my knowledge, suitable to handle them at present. I am informed that the future of brain studies, statistical science, and engineering science will witness a more refined categorization of biomarkers and a line of better methodological avenues to deal with them.

2.2 The outline of the thesis and the contributions made in each chapter

Chapter 2 (Building static biomarkers). In this chapter, I introduce a statistical framework utilizing elastic-net regularization and a two-step feature selection procedure to develop statistic biomarkers useful for remotely assessing Parkinson's disease. Using 437 behavioral features obtained from 72 subjects (37 PD patients and 35 HCs) sampled unevenly from 17 noncontiguous days during a period of up to 6 months, the model identifies PD-specific static biomarkers and shows promise to (a) discriminate PD participants from HCs, and (b) estimate the disease severity of individuals with PD.

Chapter 3 (Building dynamic biomarkers). In this chapter, I introduce a framework that includes a modified Granger-Geweke causality model that accounts for the subject- and edge-specific lags in brain effective connectivity to develop dynamic biomarkers that are potentially associated with and predictive of human cognition. Using this model, I map the whole brain's directed information flow architecture using resting-state fMRI data from 783 healthy adults in the Human Connectome Project (HCP) dataset. Additionally, I show that the identified directed brain information flows potentially underpin human cognition.

Chapter 4 (Building intermediating biomarkers). In this chapter, I present a statistical framework to identify intermediate biomarkers from the brain that are potentially mediating prodromal symptoms and later conversion to psychosis. The framework introduces a functional principal component regression and a generalized linear model (GLM) into classic mediation models. I use this method to analyze functional magnetic resonance imaging (fMRI) data of 263 subjects at clinical high risk (CHR) for psychosis, 25 of whom later developed a full-blown psychotic disorder (CHR converters); I identify a subset of brain areas (from a total of 130,992 areas) that are potentially meditating prodromal symptoms and psychosis onset.

Chapter 5 (Building longitudinal biomarkers). In this chapter, I introduce a longitudinal statistical model suitable for studying multiple sclerosis (MS) over time using smartphone data. In face of missing data, the model first imputes multiple sets of complete data using *multiple imputation*. Subsequently, it discovers the potential longitudinal biomarkers of MS by employing a generalized estimation equation (GEE). Next, parameters learned from multiple training datasets are ensembled to form a simple, unified longitudinal predictive model to forecast MS over time in previously unseen subjects. Lastly, to mitigate potential underestimation for individuals with severe disease scores, the final model incorporates an additional subject-specific fine-tuning using data from the first day. The results show that the proposed model is promising to achieve personalized longitudinal MS assessment and monitoring; they also suggest that hand and gait features collected remotely from smartphones may be useful digital markers for predicting MS over time.

Chapter 6 (Conclusions and discussions). This chapter concludes the thesis and gives discussions that are related to, but may extend, the topics discussed in this thesis, including the projections between neural, behavior, and disease data; the potential biological basis for

association, explanation, prediction, and causation; statistical methods for dealing with high-dimensional data; statistical methods for dealing with missing data; and longitudinal methods for studying the developmental brain.

3. Statistical methods for building biomarkers

This section overviews methodological frameworks that in my view are useful to study different types of biomarkers outlined in **Section** 1. Specifically, **Section** 3.1 overviews methods that can be used to build statics biomarker to classify categorical outcomes such as disease status (**Section** 3.1.1) and to estimate continuous outcomes such as disease severity (**Section** 3.1.2). **Section** 3.2 introduces three methods that can be used to build dynamic biomarkers: dynamic causal modeling (DCM) (**Section** 3.2.1), Granger-Geweke causality (**Section** 3.2.2), and structural equation modeling (**Section** 3.2.3). **Section** 3.3 discusses methods for mediation analysis, including univariate (**Section** 3.3.1), multivariate (**Section** 3.3.2), and high-dimensional (**Section** 3.3.3) mediation analyses. **Section** 3.4 discusses parametric (**Section** 3.4.1), semi-parametric (**Section** 3.4.2), and non-parametric (**Section** 3.4.3) models useful for building longitudinal biomarkers to assess and predict outcomes over time.

Notations. Let's first define a few key notations used throughout this section. I shall use the uppercase Roman bold (*e.g.*, $\boldsymbol{X}$) to denote a matrix, lowercase Roman bold (*e.g.*, $\boldsymbol{x}$) to denote a (column) vector, and lowercase Roman letter (*e.g.*, x) to denote a scalar. When random variables and their realizations are concerned, I use the uppercase Roman letter (*e.g.*, X) to denote a random variable and use lowercase (*e.g.*, x) for its realization. I shall use $X = x$ whenever possible to denote that x is a realization of X. I use superscript T (*e.g.*, $\boldsymbol{x}^T$) to indicate transpose. Unless otherwise specified, I use Latin alphabet (*e.g.*, X, Y, Z, and M) to denote data, Greek letters (*e.g.*, α, β, γ, and τ) to designate parameters, and lowercase letters such

as i, j, k, and t to indicate repeated measurements. For example, $\boldsymbol{X}_i$, $\boldsymbol{x}_i$, and x_i represent a data matrix, a vector, and a scalar, respectively, that are related to a specific subject i, and $\boldsymbol{X}_{ij}$, $\boldsymbol{x}_{ij}$, and x_{ij} stand for their j^{th} repeated measurements. I use $\boldsymbol{\beta}(k)$ or $\boldsymbol{\beta}_k$ to denote the k^{th} entry of a vector parameter $\boldsymbol{\beta}$. Finally, to avoid confusion, in the following, we use feature vector (or matrix) to represent one (or all) candidate biomarker(s). Each method then aims at finding a subset of features from all candidate features that may best describe the relationship between the features and outcome (regarding building static, dynamic, and longitudinal outcomes) or the relationship between the features, the input, and the outcome (regarding building intermediate biomarkers). The selected, subset of features are potential (static, dynamic, intermediate, or longitudinal) biomarkers.

3.1 Statistical methods for building static biomarkers

Statistical models for developing static biomarkers need to both address the scientific question of interest and accommodate the nature of the (univariate) outcome variable. For the former, one has to distinguish whether the scientific goal is to find biomarkers that can explain the relationship between the features and the outcome, or to develop biomarkers that are useful to predict an outcome[99]. We shall return to this for a more detailed discussion, namely, how to develop biomarkers to address association, explanation, prediction, and causation, in **Chapter 6**. For the latter, in the terrain of statistical methods, the nature of the outcome is largely governed by its (probabilistic) distribution. In general, one can separate the outcome distributions into discrete and continuous ones. If the outcome variable is discrete, the statistical problem is inherently classification, and an optimal model is one that can first handle the given discrete distribution, then select static features from the individuals, and finally allocate individuals accurately into each outcome category. If the outcome is continuous, the

statistical problem is essentially continuous regression, and the optimal model is one that can first deal with the given continuous distribution, and subsequently select the static features that, via the model, yield estimated outcomes closest to their observed values.

3.1.1 Statistical methods for building static biomarkers for classifying categorical outcomes

Let us begin by defining what a classification problem is. Consider a multivariate feature matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^T$ of size $n \times p$, where $\boldsymbol{x}_i$ is a column vector of length p which contains p features for subject i, for $1 \leq i \leq n$. In general, a classification model aims at mapping $\boldsymbol{x}$ over the feature domain $\mathbb{R}^p$ to a category $f(\boldsymbol{x}) \in \mathbb{R}$ via a function f. The chief goal in classification is to obtain an estimated function $\hat{f}(\cdot)$ that can allocate every query feature $\boldsymbol{x}_i$ into a category $\hat{f}(\boldsymbol{x}_i)$, for $i = 1, 2, \dots, n$. The performance of the classification can be evaluated by comparing the estimated category $\hat{f}(\boldsymbol{x}_i)$ with the observed category y_i contained in a discrete set $\mathfrak{S}$. If the estimated $\hat{f}(\boldsymbol{x}_i)$'s are not discrete, one can always partition the space into collection of regions labeled so the output categories match those in $\mathfrak{S}$. One can then quantify and assess the classification performance using a metric such as the area under the curve (AUC), Cohen's kappa, sensitivity, and specificity (see Sokolova and Lapalme[100] for a systematic review of performance measures for classification tasks).

Linear classification models. Perhaps the simplest methods for performing classification tasks are linear models. Common linear classification models include the logistic regression model and linear discriminant analysis (LDA). Much has been discussed about their modeling and estimation. Parameter estimation of logistic regression model aims at maximizing the conditional likelihood and can be performed via Newton-Raphson algorithm for the binary case,

and iteratively reweighted least-squares algorithm or coordinate-descent methods for cases with more than two classes. Parameter estimation of LDA is done via eigen-decomposition of the covariance to find the linear combination of the data that maximizes the between-class variance relative to the within-class variance. Here I shall highlight their major similarities and dissimilarities.

The similarity between logistic regression and LDA lies in the form. Specifically, for classification between class i and (reference) class j, both *can be* written as a log-linear model:

$$\log\frac{P(Y=i|X=\boldsymbol{x})}{P(Y=j|X=\boldsymbol{x})}=\beta_{i0}+\boldsymbol{x}^T\boldsymbol{\beta}_i$$

where the right-hand side is the classic linear logits in logistic regression and the uppercase X and Y indicate random variables. To see the resemblance of LDA to logistic regression, suppose that each class density follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i$ and a common covariance structure $\boldsymbol{\Sigma}$. Let π_i denote the prior probability of classes i, $\forall i$. Then LDA bears the above logistic form with $\beta_{i0}=\log\frac{\pi_i}{\pi_j}-\frac{1}{2}\left(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j\right)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$ and $\boldsymbol{\beta}_i=\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)$ (see p. 108 in Friedman *et al.*[101]).

The difference between logistic regression and LDA is twofold. First, LDA assumes that the class density is Gaussian; the logistic regression does not. Second, the logistics regression finds its parameters by maximizing the conditional likelihood, whereas LDA aims at maximizing the full log-likelihood. Having an additional normal assumption, LDA has more information about the parameters, thereby yielding more efficient estimates (in terms of having lower variance). On the other hand, with fewer assumptions, the logistic regression is more general. Overall, logistic regression is safer and more robust than LDA; but both typically yield similar results, even when LDA is used inappropriately when the input contains categorical variables

(see p. 128 in Friedman *et al.*[101]). When data is not linearly separable, generalized LDA (*e.g.*, flexible discriminant analysis and mixture discriminant analysis) has been proposed to address this issue[102]. When the feature is high-dimensional, feature selection and regularized approaches can be used, which I shall give a discussion in **Chapter** 6 of the thesis.

Another approach for linear classification is *separating hyperplane classifiers*. This approach designs linear decision boundaries aiming at separating the data into different classes, meanwhile optimizing some criterion (*e.g.*, minimizing the distance of misclassified points to the decision boundary or maximizing the distance to the closest point from either class).

The simplest case of the separating hyperplane classifier perhaps is Rosenblatt's *Perceptron* learning algorithm[103], the predecessor of the modern neural networks. Consider a decision boundary (for two classes) defined by $\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0$. If the outcome $y_i = 1$ is misclassified, then $\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 < 0$; if the outcome $y_i = -1$ is misclassified, then $\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 > 0$. The *Perceptron* aims at minimizing the following distance function:

$$D(\boldsymbol{\beta}, \beta_0) = -\sum_{i \in \mathcal{S}} y_i (x_i^T \boldsymbol{\beta} + \beta_0) \qquad (1)$$

where $\mathcal{S}$ contains all misclassified values.

The solution for **Eq.** (1) can be found via stochastic gradient descent (by taking derivatives with respect to β and β_0, respectively) and the parameters are iteratively updated:

$$\begin{pmatrix} \boldsymbol{\beta} \\ \beta_0 \end{pmatrix} \leftarrow \begin{pmatrix} \boldsymbol{\beta} \\ \beta_0 \end{pmatrix} + \rho \begin{pmatrix} y_i \boldsymbol{x}_i \\ y_i \end{pmatrix}$$

where ρ represents the learning rate.

A general form of *Perceptron* can be stated as the following optimization problem[104]:

$$\max_{\boldsymbol{\beta},\beta_0,\|\boldsymbol{\beta}\|=1} M \qquad (2)$$

$$\text{such that} \quad y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta}+\beta_0\right) \geq M, i = 1, \ldots, N.$$

In words, this optimization problem put the points at least a signed distance (margin) M away from the decision boundary defined by parameters $\boldsymbol{\beta}$ and β_0.

Without loss of generality, let $\|\boldsymbol{\beta}\| = \frac{1}{M}$. The optimization problem in **Eq.** (2) is equivalent to:

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 \qquad (3)$$

$$\text{such that} \quad y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta}+\beta_0\right) \geq 1, i = 1, \ldots, N.$$

In words, the optimization problem translates to one that looks for parameters $\boldsymbol{\beta}$ and β_0 that maximize the margin of width $\frac{1}{\|\boldsymbol{\beta}\|}$ around the linear decision boundary. Since this is a convex optimization, the problem in **Eq.** (3) can be solved using the Lagrange method by taking derivatives of:

$$f(\boldsymbol{\beta},\beta_0) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{N}\alpha_i[y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta}+\beta_0\right)-1]$$

with respect to $\boldsymbol{\beta}$ and β_0, respectively, and set the derivatives to 0; α_i is used so that the supporting points $\boldsymbol{x}_i$'s are defined to be on the boundary of the slab when $\alpha_i > 0$. The solution is then solved using standard software (see Friedman *et al.*[101] for details).

The methods we have so far discussed assume that the outcome classes can be separated by a linear boundary. When there is not a clear linear boundary, one can use SVMs and neural networks. We discuss them below in sequence.

Support vector machines. When groups are not linearly separable, imposing any linear separating hyperplanes will end up with some points falling into the wrong group. Support vector machines (SVMs) can address this issue by adding slack variables $\xi_i \ \forall i$ to the optimization problem defined by the separating hyperplanes. Specifically, SVM solves:

$$\min \frac{1}{2}\|\boldsymbol{\beta}\|^2 \qquad \text{such that} \qquad \begin{cases} y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta} + \beta_0\right) \geq 1 - \xi_i \ \forall i \\ \xi_i \geq 0 \ \forall i, \sum_{i=1}^{N} \xi_i \leq C \end{cases} \qquad (4)$$

where C determines the amount of error one would allow the classifier to make.

Similar to a separating hyperplane problem, this is a convex optimization, and is solved using the Lagrange method (see pp. 420-421 in Friedman *et al.*[101]), and the solution can be estimated using standard techniques[105].

There are two points I wish to make about SVMs: (1) the origin of the name "support vector" machines; and (2) the kernel (trick) they employ.

For the first point, the origin of the name for SVM springs from the solution of **Eq.** (4), which bears the form:

$$\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \boldsymbol{x}_i$$

where $\hat{\alpha}_i$'s are non-zero for observations that meet the constraint stated in the optimization. These observations are called *support vectors* (as $\widehat{\boldsymbol{\beta}}$ is only built based upon them), hence the name *support vector machines*.

For the second point, the reason for using transformed features (via a kernel) than raw features in SVMs is as follows: behind a kernel is a mapping that maps the feature data to another

space so that they are, hopefully, linearly separable; adding slack variables alone cannot do that job. Specifically, it is made by obtaining the inner products $\langle h(\boldsymbol{x}), h(\boldsymbol{x}')\rangle$ of the transformed features $h(\boldsymbol{x}_i)$. The solution function is:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{\beta}_0.$$

where $K(\boldsymbol{x}, \boldsymbol{x}') = \langle h(\boldsymbol{x}), h(\boldsymbol{x}_i)\rangle$ represents the kernel. Additionally, as long as we know the (symmetric positive semi-definite) kernel function K, we do not need to know the explicit transform function h.

Neural networks. We begin with an overview on classification using a single hidden layer fully connected neural network (one can naturally extend it to multilayer neural networks as, for example, in LeCun *et al.*[106] and Goodfellow *et al.*[107]). First, let's suppose that the input features $\boldsymbol{x}$ are in $\mathbb{R}^p$ represented by p nodes, the single hidden layer contains M transformed features represented by M nodes (hidden units), and there are K possible outcome classes label by K nodes. A single layer neural network can be expressed as follows:

1. From input layer to hidden layer: $\boldsymbol{z}_m = \sigma(\alpha_{0m} + \boldsymbol{x}^T \boldsymbol{\alpha}_m),\ m = 1, \dots, M$

where σ is an activation function, typically chosen as a sigmoid function $\sigma(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x})}$.

2. From hidden layer to the output layer: $t_k = \beta_{0m} + \boldsymbol{x}^T \boldsymbol{\beta}_K,\ k = 1, \dots, K.$
3. From output layer to classification: $f_k(\boldsymbol{x}) = g_k(\boldsymbol{t}),\ k = 1, \dots, K$

where g_k is an output function and $\boldsymbol{t} = (t_1, t_2, \dots, t_K)^T$. For classification, it is typically chosen as a softmax function: $g_k(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_{i=1}^{K} \exp(\boldsymbol{x})}$.

The parameter estimation for hidden layers is performed via back-propagation (namely gradient descent) to minimize a loss function (*e.g.*, the sum of squared errors) (see pp. 395-397 in Friedman *et al.*[101] and Goodfellow *et al.*[107]).

There are a few points I wish to make about using a neural network for classification. First, thanks to the softmax function, the sum of the output equals 1; therefore, one can interpret the outcomes as probabilities over K categories. Second, the one-layer neural network is closely related to the Projection Pursuit Regression (PPR)[108–110]; the difference is that the neural network uses a sigmoid function and PPR uses a nonparametric function. Third, the last layer of a neural network used for classification is similar to a multi-logit model, both of which use a softmax function that sums the outputs into 1 and can be interpreted as the probability each node falls into a particular class.

Kernel smoothing methods. Let us return to the problem of mapping p-dimensional multivariate feature $\boldsymbol{x}$ over the domain $\mathbb{R}^p$ to a category vector $f(\boldsymbol{x}) \in \mathbb{R}$ via a regression function f. The central theme around *kernel smoothing methods* is to obtain an estimated function $\hat{f}(\cdot)$ that is smooth in $\mathbb{R}^p$ for every query point $\boldsymbol{x}_0$ using observations close to $\boldsymbol{x}_0$. The smoothing is achieved by adopting a kernel function $K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)$, which assigs a weight to $\boldsymbol{x}_i$ depending on its distance to $\boldsymbol{x}_0$ and λ determines the size of neighborhood. *Note that the kernel here is different from the kernel used for nonlinear modeling where it is defined as an inner product in a high-dimensional (implicit) feature space.*

Kernel smoothing methods can be used for nonparametric classification using Bayes's theorem. Suppose we have already obtained the estimated function $\hat{f}(\cdot)$ and we are interested in grouping observations into J classes:

$$P(Y = j|X = \boldsymbol{x}) = \frac{\hat{\pi}_j \hat{f}_j(\boldsymbol{x})}{\sum_{k=1}^{J} \hat{\pi}_k \hat{f}_k(\boldsymbol{x})}$$

where $\hat{\pi}_j$ is the estimated class prior (sample proportion) for class $j = 1, \ldots, J$ and the uppercase X and Y indicate random variables.

When features are high-dimensional, density estimation becomes challenging. Another classification approach using kernel smoothing, *the naïve Bayes classifier*, is useful for such scenarios. The naïve Bayes classifiers assume that features are independent given a class $Y = j$: $f_j(\boldsymbol{x}) = \prod_{k=1}^{p} f_{jk}(x_k)$. Although a strong assumption, they often outperform far more sophisticated alternatives, because even though the individual class density estimates may be biased, this bias might not affect the posterior probabilities as much, especially near the decision regions[101]. The estimation of the naïve Bayes classifier begins with estimating each class-conditional marginal density $f_{jk}(x_k)$ using one-dimensional kernel density estimates; when a feature is discrete, one can use a histogram approximate. Assuming independence among features, the classification is achieved using a logit-transformation, namely:

$$\begin{aligned}
\log \frac{P(Y = j|X = \boldsymbol{x})}{P(Y = i|X = \boldsymbol{x})} &= \log \frac{\pi_j f_j(\boldsymbol{x})}{\pi_i f_i(\boldsymbol{x})} \\
&= \log \frac{\pi_j}{\pi_i} + \sum_{k=1}^{p} \log \frac{f_{jk}(x_k)}{f_{ik}(x_k)} \\
&= \alpha_j + \sum_{k=1}^{p} g_{jk}(x_k)
\end{aligned}$$

where group i is used as a reference, and the final model bears the form of a generalized additive model.

A special case of kernel method is the *mixture model*, written as a combination of densities from different distributions:

$$f(\boldsymbol{x}) = \sum_{j=1}^{J} \alpha_j \phi(\boldsymbol{x};\ \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where ϕ is a density function (here we use a Gaussian density ϕ with mean and variance-covariance evaluated at x as an example; each ϕ, in general, could be replaced with any arbitrary density), α_j is a mixing parameter determining the proportion of each density ϕ and $\sum_{j=1}^{J} \alpha_j = 1$.

It is kernel smoothing because the output $f(\boldsymbol{x})$ can be viewed as weighted average of distributions evaluated at $\boldsymbol{x}$. The classification can subsequently be performed as follows. For K target groups, the probability of subject i belongs to group j is:

$$\hat{r}_{ij} = \frac{\hat{\alpha}_j \phi(\boldsymbol{x}_i;\ \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}{\sum_{k=1}^{K} \hat{\alpha}_k \phi(\boldsymbol{x}_i;\ \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}$$

and subsequently, one can apply a threshold τ to $\hat{r}_{ij}$, say $\hat{\delta}_i = I(\hat{r}_{ij} > \tau)$, where I is an indicator function, to determine which group i falls into.

Classification trees. The key point of the tree-based methods (for both classification tree and regression tree (discussed below)) is to partition the feature space into discrete regions that ensure: (1) the partition allows for binary search (therefore it is convenient) and (2) a simple model can be fit for each partition.

Let's consider a partition of the feature space $\mathbb{R}^p$ into M regions $R_1, \dots, R_M$. We grow a tree T_0 and stop until some minimum node size is reached. Let $T \subset T_0$ be any tree that can be obtained by pruning T_0, $|T|$ be the number of terminal nodes in T, and R_m be the m^{th} terminal node. Further, let:

1. $N_m = \{\boldsymbol{x}_i \in R_m\}$.
2. $\hat{c}_m$ be an estimated constant for each region.
3. $Q_m(T)$ be the node impurity measure (a cost function).
4. $C_\alpha(T)$ be the cost complexity criteria, which balances the cost function $Q_m(T)$ and the number terminal nodes $|T|$, where:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m \, Q_m(T) + \alpha|T|.$$

The key then is to find the subtree $T_\alpha \subseteq T_0$ that minimizes $C_\alpha(T)$ for each α (see Breiman, L. and Ihaka[111] and Ripley[112] for details). The major difference between classification trees and regression trees (the latter of which we shall introduce below when we discuss models for dealing with continuous variables) are the constant $\hat{c}_m$ the node impurity measure $Q_m(T)$. For classification, the estimated constant $\hat{c}_m \equiv \hat{p}_{mk} = \frac{1}{N_m}\sum_{\boldsymbol{x}_i \in R_m} I(y_i = k)$, which can be interpreted as the probability observation i belonging to class k at node m. The three common cost complexity criterions for classifications are: misclassification error, where $Q_m(T) = \frac{1}{N_m}\sum_{i \in R_m} I\big(y_i \neq k(m)\big) = 1 - \hat{p}_{mk}(m)$; Gini index, where $Q_m(T) = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$; and cross-entropy or deviance, where $Q_m(T) = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$. The formation of the cost complexity criterions determines their classification performance. On the one hand, the cross-entropy and the Gini index are differentiable, therefore they are more suitable for optimization. On the other hand, the Gini index and cross-entropy are more

sensitive to changes in the node probabilities than the misclassification rate (see p. 309 in Friedman *et al.*[101]).

Ensemble learning: bagging, boosting, and random forest. I introduce these three classes of methods together for one simple reason, that they all form a "committee" of leaners who each would form one's own modeling opinion; by assembling opinions from the committee, the final aggregated result is drastically improved. For the classification problem, each learner would cast a vote; the final outcome would be some weighted version of the votes from the committee members. For continuous estimation, each learner would offer an estimate, the final outcome would be some weighted version of the estimates from all committee members. Because the statistical method each learner uses is typically a standard classification or regression method, we will not discuss separately methods for dealing with continuous outcomes below. Rather, we overview the committee methods for classification and continuous estimation together and highlight the key steps and differences.

Bagging, or bootstrap aggregation[113], first obtains a committee of multiple bootstrap data sets, then fits a (classification or regression) model on each bootstrap sample, and lastly averages the outcomes from each bootstrap to obtain a final estimation. It works well for high-variance and low-bias problems.

Boosting methods form a committee of multiple "weak" classifiers (one whose error rate is only slightly better than one from a random guess) and combine their outcomes to produce a weighted outcome that is significantly better than the outcome of each classifier. The classification trees have been widely used as the "base learner". AdaBoost[114], which combined a weighted majority vote (with higher weights assigned to classifiers with better performance)

to produce a final prediction, is arguable the most popular[101] and “best off-the-shelf classifier in the world”[115].

A random forest[116] is, in essence, bagging. The key improvement in a random forest is that it attempts to make trees less correlated by randomly selecting the input variables during the tree-growing step.

Overall, boosting methods outperform bagging in most problems; the performance of random forests is similar to boosting, but they are simpler to train and tune (see p. 587 in Friedman *et al.*[101]).

3.1.2 Statistical methods for building static biomarkers for estimating continuous outcomes

Common methodological frameworks for dealing with continuous outcomes include linear regression models, regression trees, neural networks, and kernel smoothing methods. Here, I shall introduce methods for dealing with continuous outcomes that are distinctive from those developed to treat classification problems; for methods that can deal with both classification and regression problems, I shall only highlight the steps that are different. Ensemble learning, such as bagging, boosting, and random forest, are powerful committee methods for estimating and predicting continuous outcome; the technical concept regarding how they treat continuous outcomes, however, do not depart much from how they manage classification problem (see **Section** 3.1.1): the former aggregate estimates from multiple committee members, whereas the latter combines votes from committee members. I shall thus avoid repeating them here again; readers could refer to technic notes for bagging[113], boosting[117], and random forests[116].

Let us begin by defining what a regression problem is. In general, a regression model aims at mapping a p-dimensional multivariate feature $\boldsymbol{x}$ over the domain $\mathbb{R}^p$to a continuous outcome $f(\boldsymbol{x})$ via a function f. When discussing regression problems, one should not restrict oneself to a multivariate regression bearing the form of $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ that assumes the outcomes are *i.i.d.* (independent and identically distributed) and the errors follow a Gaussian distribution; rather, general regression problems consist of a broad range of cases where the outcomes are non-Gaussian and known (or unknown) correlation exists between outcomes. The chief goal in regression is to obtain an estimated function $\hat{f}(\boldsymbol{x})$ that can map every query feature vector $\boldsymbol{x}_i$ into $\hat{f}(\boldsymbol{x}_i) \in \mathbb{R}$, where $\hat{f}(\boldsymbol{x}_i)$ is as close to the observed continuous variable $y_i \in \mathbb{R}$ as possible, for $i = 1, 2, \ldots, n$. One can then quantify and assess the regression performance using a metric such as the correlation between the estimated and observed outcomes, root mean square errors (RMSE), mean absolute error (MAE), R squared, and adjusted R squared.

Linear regression, general linear regression, and generalized estimation equations. Perhaps the simplest regression model is a multivariate regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is a column vector of length n, $\boldsymbol{X}$ is an n by p feature matrix, $\boldsymbol{\beta}$ is a vector parameter of length p, and $\boldsymbol{\varepsilon}$ denotes the residual term including both noises, errors, and information not explained by the model. When the noise is *i.i.d.*, say $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, then the least square solution for $\boldsymbol{\beta}$ coincides with the maximum likelihood estimators (MLE), namely:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2)$, and $var(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$.

The outcomes can be estimated via:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Gaussian assumptions, however, are not practical to all data. The generalized linear models (GLM) extend linear regressions to the terrains when the errors are non-Gaussian. Via a link function, the GLM transforms the outcome variable such that the transformed outcome is linearly associated with the features $g\big(\mathbb{E}(\boldsymbol{y})\big) = \boldsymbol{X\beta}$, or more commonly:

$$\mathbb{E}(\boldsymbol{y}) = g^{-1}(\boldsymbol{X\beta}).$$

Subsequently, the MLE parameters can be estimated using the iteratively reweighted least squares or the Newton-Raphson method (see p. 40 in McCullagh and Nelder[118]). Note also that the logistic regression discussed in **Section** 3.1.1 is a special case in GLM with a logit link.

When the outcomes are correlated longitudinally – especially within each individual - a similar (conceptually) method called generalized estimation equation (GEE)[119] can be used (see **Section** 3.4.1 below).

Regression trees. The tree methods for treating regression problems and classification problems are in principle the same. The key differences between the two are the constant $\hat{c}_m$ and the node impurity measure $Q_m(T)$.

Resuming the notations and procedures for classification trees in **Section** 3.1.1, for regression trees, one needs to modify the constant as:

$$\hat{c}_m = \frac{1}{N_m} \sum_{\boldsymbol{x}_i \in R_m} y_i$$

which represents the best-estimated response in region R_m.

Additionally, let the node impurity measure be:

$$Q_m(T) = \frac{1}{N_m} \sum_{\boldsymbol{x}_i \in R_m} (y_i - \hat{c}_m)^2$$

which represents the sum of squared errors.

The rest of the methods are identical to classification trees in **Section** 3.1.1, namely one aims at finding the best (estimated) responses that yield both a small cost (in terms of minimizing $Q_m(T)$) and a reasonable number of terminal nodes (so it does take to many trees to finalize an estimate).

Neural networks for regression. The central difference between neural networks for classification and regression is the output function. Recall that the output function g_k maps the outputs from the final layer consisting of K nodes (*e.g.*, $\boldsymbol{t} = (t_1, t_2, \dots, t_K)^T$) to the final outcome $f_k(\boldsymbol{x})$ via $f_k(\boldsymbol{x}) = g_k(\boldsymbol{t}),\ k = 1, \dots, K$. For classification, it is typically chosen as a softmax function: $g_k(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x})}{\sum_{i=1}^{K} \exp(\boldsymbol{x})}$. For regression, one typically uses the identity function $g_k(\boldsymbol{t}) = t_k$.

Kernel smoothing methods for regression. The goal of kernel smoothing for regression is to establish a locally linear or non-linear (*e.g.*, polynomial) relationship between the features and the outcome via a kernel (which can be seen as weights assigned to outcomes), such that the estimated outcome is smooth.

The simplest kernel methods for dealing with the continuous outcome is a *locally weighted average*, with the following general form:

$$\hat{f}(\boldsymbol{x}_0) = \frac{\sum_{i=1}^{N} K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i) y_i}{\sum_{i=1}^{N} K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)}$$

with kernel

$$K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}) = D(\frac{|\boldsymbol{x} - \boldsymbol{x}_0|}{h_\lambda(\boldsymbol{x}_0)})$$

where $D(\cdot)$ is a distance function.

The smoothing parameter or $h_\lambda(\boldsymbol{x}_0)$ determines the size of the neighborhood centering at $\boldsymbol{x}_0$. When $h_\lambda(\boldsymbol{x}_0) = \lambda$ is constant, the kernel $K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}) = D(\frac{|\boldsymbol{x}-\boldsymbol{x}_0|}{\lambda})$ is the so-called Epanechnikov quadratic kernel; when one replaces λ with a neighborhood size k, namely $h_k(\boldsymbol{x}_0) = |\boldsymbol{x}_0 - \boldsymbol{x}_{[k]}|$, where $\boldsymbol{x}_{[k]}$ is the k^{th} closest point to $\boldsymbol{x}_0$, then the kernel is the k-nearest neighborhood. Popular distance functions are the tri-cube function where $D(\boldsymbol{x}) = \begin{cases}(1 - |\boldsymbol{x}|^3)^3, & \text{when } |\boldsymbol{x}| \leq 1; \\ 0, & \text{otherwise}\end{cases}$ and Gaussian kernel where $D(\boldsymbol{x}) = \phi(\boldsymbol{x})$.

There is a major issue with the *locally weighted average* method: the estimates are biased on the boundaries of the domain. To deal with this, one could use local linear regression, which performs a regression on observations in each neighborhood. There are two classes of local regression: local linear regression and local polynomial regression. The locally weighted linear regression aims at solving an optimization problem with respect to a weighted (by the kernel) least square at each query point $\boldsymbol{x}_0$:

$$\min_{\alpha(\boldsymbol{x}_0), \beta(\boldsymbol{x}_0)} \sum_{i=1}^{N} K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)[y_i - \alpha(\boldsymbol{x}_0) - \beta(\boldsymbol{x}_0)\boldsymbol{x}_i]^2$$

where the estimate is $\hat{f}(\boldsymbol{x}_0) = \hat{\alpha}(\boldsymbol{x}_0) + \hat{\beta}(\boldsymbol{x}_0)\boldsymbol{x}_0$ for every point $\boldsymbol{x}_0$.

Similarly, the locally weighted polynomial regression aims at solving an optimization problem involving a weighted (by the kernel) least square polynomial fit at each query point $\boldsymbol{x}_0$:

$$\min_{\alpha(\boldsymbol{x}_0), \beta(\boldsymbol{x}_0), j=1,\ldots,d} \sum_{i=1}^{N} K_\lambda(\boldsymbol{x}_0, \boldsymbol{x}_i)[y_i - \alpha(\boldsymbol{x}_0) - \sum_{j=1}^{d} \beta_j(\boldsymbol{x}_0)\boldsymbol{x}_i^j]^2$$

where the estimate is $\hat{f}(\boldsymbol{x}_0) = \hat{\alpha}(\boldsymbol{x}_0) + \sum_{j=1}^{d} \hat{\beta}_j(\boldsymbol{x}_0)\boldsymbol{x}_0^j$ for every point $\boldsymbol{x}_0$.

Related to the locally weighted regression via kernel smoothing are the basis expansion and its regularized (smooth) version, where the outcome variable is expressed as a weighted sum of basis functions of the feature $f(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{x})$, where $h_m(\cdot)$ is a basis function and β_m its weight, or a weighted sum of basis functions of the feature plus a penalty term (which is used to regularize – or smooth – the estimated curve) (see Chapter 5 of Friedman *et al.*[101] and Ramsay and Silverman[120]). Common applications of splines and their functional versions are seen in analyzing functional data observed over time; we will drive into this direction in more detail in **Chapter** 4 with a discussion on longitudinal functional mediation.

3.2 Statistical methods for building dynamic biomarkers

I define the concept of dynamic biomarkers as directed edges strength between different brain areas (say areas A and B) that imply the amount of information that is passing from one brain area to another, and *vice versa*. More concretely, such a concept is useful to unveil three pieces of information and can be visualized as follows: (1) whether there exists information passing between A and B; (2) if so, whether the information is only passed from A to B (indicated by a directed arrow from A to B: ($A \rightarrow B$), only from B to A ($B \rightarrow A$), or that there is information passing from A to B and from B to A ($A \rightleftarrows B$); (c) how much information is passed (this can be visualized by the width of the arrows).

Dynamic causal modeling (DCM), Granger-Geweke causality, and structural equation modeling are common methodological frameworks to study dynamic biomarkers to unveil

directed functional pathways of the human brain. Here, it is desirable to familiarize oneself with the basic structure of these methods.

3.2.1 Dynamic causal modeling

The dynamic causal modeling (DCM) is stated in terms of differential equations. In its essence, it is a state-space model using bilinear approximation (see below) to describe motion that couples neuronal states in different brain areas[121]. Let $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ denote a hidden neural state vector in time and an observed response vector in time, respectively. Let $\boldsymbol{u}(t)$ be exogenous input that could change experimental conditions or contexts that may ultimately affect responses. Let $\boldsymbol{\theta}(t)$ be the effective connectivity which controls how hidden states in one brain area affect the motion of hidden states in other brain areas. The DCM consists of two equations: the first equation models motion, and the second equation outputs observed outcomes. Specifically:

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta}) + \boldsymbol{\omega} \qquad (5)$$

$$\boldsymbol{y} = g(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta}) + \boldsymbol{v} \qquad (6)$$

where ω and v are random fluctuations.

Because directly modeling $\dot{\boldsymbol{x}}$ and $\boldsymbol{y}$ in **Eqs.** (5)-(6) with large-scale hidden states is challenging, one could simplify the time derivative $\dot{\boldsymbol{x}}$ using a local bilinear approximation[121], as follows:

$$\dot{\boldsymbol{x}} = \boldsymbol{\theta}^{x}\boldsymbol{x} + u\boldsymbol{\theta}^{xu}\boldsymbol{x} + \boldsymbol{\theta}^{u}u + \boldsymbol{\omega}$$

$$\boldsymbol{\theta}^{x} = \frac{\partial f}{\partial \boldsymbol{x}};\ \boldsymbol{\theta}^{xu} = \frac{\partial^2 f}{\partial \boldsymbol{x}\partial u};\ \boldsymbol{\theta}^{u} = \left.\frac{\partial f}{\partial u}\right|_{x=0,u=0}$$

where $u \in \{0,1\}$ is a binary exogenous input (the values are chosen per common, *e.g.*, fMRI, practice[88]). When $u = 0$, with $\Delta x = 1$ change, the dynamic biomarkers (effective connectivity)

can be recapitulated by matrix $\boldsymbol{\theta}^{x}$; when $u = 1$, with $\Delta x = 1$ change, the dynamic biomarkers are represented using $\boldsymbol{\theta}^{x} + u\boldsymbol{\theta}^{xu} + \boldsymbol{\theta}^{u}u$.

3.2.2 Granger-Geweke causality

Granger-Geweke causality considers assessing whether there exists a directed link both two nodes X and Y using two steps[122–124]. First, it employs two autoregressive (AR) processes to model signals from the two nodes. Subsequently, it checks if including the past information of node X into the AR process involving Y would help explain more information (in terms of reducing residual variance, see below) of the current states of Y, and if so, then one says that X Granger causes Y and draws a directed edge from X to Y (*i.e.*, a link between two variables with an arrow pointing towards Y), and *vice versa*. In words, a random variable X Granger causes another variable Y, if the prediction of Y is improved using the information of its past and the past of X, compared with when using only the past information of Y.

More specifically, denote Z as a set of integers each of which corresponds to a time point at which two processes $\{X_t\}_{t\in Z}$ and $\{Y_t\}_{t\in Z}$ are measured. Consider two AR processes that model the current state of X using its past information, and the current state of Y using its past information, respectively:

$$\boldsymbol{x}_t = \sum\nolimits_{j=1}^{\infty} \boldsymbol{x}_{t-j}^{T}\,\boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j, \qquad \mathrm{Var}\left(\boldsymbol{\varepsilon}_j\right) = \boldsymbol{\Sigma}_1.$$

$$\boldsymbol{y}_t = \sum\nolimits_{j=1}^{\infty} \boldsymbol{y}_{t-j}^{T}\,\boldsymbol{\beta}_j + \boldsymbol{\eta}_j, \qquad \mathrm{Var}\left(\boldsymbol{\eta}_j\right) = \mathbf{T}_1.$$

Further, consider two AR processes that model the current state of X using both its past information and the past information of Y, and the current state of Y using its past information and the past information of X:

$$\boldsymbol{x}_t = \sum_{j=1}^{\infty} \boldsymbol{x}_{t-j}^T \boldsymbol{a}_j + \sum_{j=1}^{\infty} \boldsymbol{y}_{t-j}^T \boldsymbol{b}_j + \boldsymbol{e}_j, \qquad \mathrm{Var}(\boldsymbol{e}_j) = \boldsymbol{\Sigma}_2.$$

$$\boldsymbol{y}_t = \sum_{j=1}^{\infty} \boldsymbol{y}_{t-j}^T \boldsymbol{c}_j + \sum_{j=1}^{\infty} \boldsymbol{x}_{t-j}^T \boldsymbol{d}_j + \boldsymbol{z}_j, \qquad \mathrm{Var}(\boldsymbol{z}_j) = \mathbf{T}_2.$$

The variance-covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, $\mathbf{T}_1$, and $\mathbf{T}_2$ can be interpreted as the amount of information not explained by each model (plus noises). Critically, the difference between $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ is as follows: the former contains information with respect to the current state of X not explained by the past of X, and the latter contains information with respect to the current state of X not explained by the past of both X and Y. Thus, if $|\boldsymbol{\Sigma}_2|$ is significantly smaller than $|\boldsymbol{\Sigma}_1|$, it implies that the past information of Y can explain a great portion of the information regarding the current state of X. Similar arguments can be made about $|\mathbf{T}_1|$ and $|\mathbf{T}_2|$.

Geweke[125] introduced a formal way to test the directed relationship between X and Y in terms of comparing the amount of variance explained. Denote $F_{X\to Y}$ and $F_{Y\to X}$ as information flows from X to Y and Y to X, respectively. Then:

$$F_{X\to Y} = \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right).$$

$$F_{Y\to X} = \ln\left(\frac{|\mathbf{T}_1|}{|\mathbf{T}_2|}\right).$$

where $|\cdot|$ denoted matrix determinant. If $F_{X\to Y}$ is large, then one concludes X Granger causes Y and if $F_{Y\to X}$ is large, then one concludes Y Granger causes X.

The strength of the dynamic biomarkers can then be quantified using $F_{X\to Y}$ and $F_{Y\to X}$ with the directions indicated by the arrows. In **Chapter** 3, I will demonstrate that a subset of directed neural information flows estimated using the above arguments are potentially associated with and predictive of cognitive flexibility in humans.

3.2.3 Structural equation modeling

The third way to discover directed edges is to use structural equation modeling (SEM). Consider four nodes X, Y, Z, and M. Suppose one knows that X and Y, X and Z, and Y and M are connected (via, *e.g.*, correlation tests) and one wishes to test the directed information flows between X and Y. Using SEM, one could test it as follows:

$$y = x\beta + m\gamma + \varepsilon$$

$$x = y\alpha + z\tau + \eta$$

where α, β, γ and τ are parameters; ε and η are error terms. Judging from the size, significance, and sign of the pathway coefficients α and β, one can determine the direction of the edges between X and Y (see, for example, McLntosh and Gonzalez-Lima[126]).

3.2.4 Other models and some comments

There are a few other methods that may also be useful to discover dynamic biomarkers. For example, probability propagation approaches (*e.g.*, Bayesian network[127–129] and pairwise conditional probability[130]), generative embedding that combines DCMs and support vector machines[131], and the dynamic directional model which applies ordinary differential equations to analyze intracranial electrocorticographic (ECoG) data[132]. Like the three common methods introduced above, all these methods aim at testing the existence of directed links between nodes and, if a link exists, estimating the pathway strength via some metrics.

Once the directed edges between brain areas are identified and their strength estimated, it remains to interpret them neurobiologically. A beginning can be made by linking these directed edges with human behavior, and more specifically, by (1) identifying the subset of all directed edges that are associated with human behavior, and (2) by using the identified subset of directed edges (and their strength) to predict human behavior in novel samples. If the identified directed edges are both associated with a specific type of human behavior (such as cognitive flexibility) and can be used to predict this behavior in novel individuals, then it suggests that these selected edges *may* be associated with the behavior in a general population. In **Chapter** 3, I will give an example where I mapped the directed information pathways across the entire brain and identified directed edges that are predictive of cognitive flexibility in humans.

Another question regarding dynamic biomarkers (or directed edges) is related to causality, namely, if one finds that there is a set of directed edges between brain areas whose edge strengths are associated with and predictive of human behavior, does that imply a causal relationship? I shall expand on this in **Chapter** 6.

3.3 Statistical methods for building intermediate biomarkers

An intermediate biomarker (or a mediator) is a feature that is potentially mediating an input variable and an outcome variable. Said differently, it is a (intermediating) variable that is interposed between, and potentially causally linked with two variables: an input variable and an outcome variable. Mediation analysis refers to the method used to investigate, when there is only one candidate mediator, whether the variable is a mediator, or to identify, from multiple candidate variables, a subset of variables that are mediators; subsequently, the analysis aims to quantify the mediation effect of each mediator. Depending on the number of mediators,

mediation analyses can be categorized into univariate, multivariate, and high-dimensional mediation analyses.

3.3.1 Univariate mediation analysis

There are, broadly speaking, two approaches to conduct univariate mediation analyses: the *product* approach and the *difference* approach[133,134]. Conceptually, both approaches employ a series of regression models to dissect the pathways between an explanatory variable, a mediator, and an outcome. We begin with a review of each case when the intermediate variable is univariate.

The product approach. The product mediation analysis framework involves two steps[137–141]. The first step examines if the explanatory variable X is associated with the mediator M, after controlling for covariates Z. This is done using a conditional regression model:

$$\mathbb{E}(M|X = x, Z = \boldsymbol{z}) = \theta_0 + \alpha x + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{t} \tag{7}$$

where $\mathbb{E}$ refers to the expectation operation, and θ_0, α, and $\boldsymbol{t}$ are coefficients for the intercept, the explanatory variable x, and the covariates $\boldsymbol{z}$. In simpler terms, if the coefficient α in **Eq.** (7) is significantly different from zero, it suggests that there is potentially an effect from the explanatory variable to the mediator.

The second step checks if the mediator M is associated with the outcome Y, after controlling for the explanatory variable X and covariates Z. This is done using another conditional regression model:

$$\mathbb{E}(Y|X = x, M = m, Z = \boldsymbol{z}) = \theta_0' + \gamma x + \beta m + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\tau} \tag{8}$$

where θ_0', γ, β, and $\boldsymbol{\tau}$ are coefficients for the intercept, the explanatory variable x, the mediator m, and the covariates $\boldsymbol{z}$. In simpler terms, if the coefficient β in **Eq.** (8) is significantly different from zero, it suggests that there is potentially an effect from the mediator to the outcome.

Taken together, when both α and β are significantly different from zero, it means that (I) a change of X affects M, and (II) a change in M further affects Y; therefore M is significantly mediating the relationship between X and Y. In simpler terms, the mediation effect exists when the product $\alpha\beta$, which quantifies the mediation effect in this setting, is non-zero. The remaining coefficient γ' in **Eq.** (8) quantifies the direct effect of the explanatory variable X on the outcome Y (that is not mediated by M).

The difference approach. The difference approach has two modeling steps. The first step is a conditional regression model, which investigates if variability in the explanatory variable X is responsible for variability in outcome Y, after controlling for covariates Z:

$$\mathbb{E}(Y|X = x,\ Z = \boldsymbol{z}) = \theta_0 + \gamma x + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{t} \tag{9}$$

where $\mathbb{E}$ stands for expectation, θ_0,γ, and $\boldsymbol{t}$ are coefficients for the intercept, the explanatory variable x, and the covariates $\boldsymbol{z}$.

The second step is a conditional regression model, which checks if variability in the explanatory variable X and the mediator M are both responsible for it in the outcome Y, after controlling for covariates Z:

$$\mathbb{E}(Y|X = x, M = m, Z = \boldsymbol{z}) = \theta_0' + \gamma' x + \beta m + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\tau} \tag{10}$$

where θ_0', γ', β, and are $\boldsymbol{\tau}$ are coefficients for the intercept, the explanatory variable x, the mediator m, and the covariates $\boldsymbol{z}$**.** Note the coefficients are denoted with a prime (*e.g.,* γ' in **Eq.** (10)) to highlight that they are not necessarily the same as their counterparts (*e.g.*, γ in **Eq.** (9)).

If the coefficient (γ') of the explanatory variable on the outcome when a mediator is included is considerably smaller than the coefficient (γ) of the explanatory variable on the outcome when the mediator is not included, then it suggests that the mediator is at least partially responsible for explaining the variance (information) of the outcome Y. To put more concretely, this indicates that including the mediator m, the explanatory variable x explains a less amount of variance (information) of Y. The difference of $\gamma - \gamma'$ thus quantifies the mediation effect in this setting.

The product approach *vs.* the difference approach

When the outcome is continuous, the estimated mediation effects from the difference and the product approaches are the same. When the outcomes are binary (such as being healthy or afflicted with a certain type of disease), and when the (disease) rate is rare ($\leq 10\,\%$), the product approach and difference approach approximately coincide[135]. When the outcomes are binary, but the (disease) rate is common, the product approach and the difference approach diverge. The difference approach becomes particularly challenging because the coefficient of the explanatory variable using this method may not change when there is mediation (see VanderWeele[133] and Chapter 2 of VanderWeele[136] for a discussion).

3.3.2 Multivariate mediation analysis

Mediation analysis concerning a multivariate mediator can be conducted using structural equation models (SEMs)[94,142]. Formally, consider V mediators ($V \geq 2$), denoted as $M(1), M(2), \ldots, M(V)$, an explanatory variable X, and an outcome variable Y. Multivariate mediation analysis considers two conditional models as follows.

First,

$$\mathbb{E}\left(m_j \middle| X = x, Z = \boldsymbol{z}\right) = \theta_0(j) + \alpha(j)x + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{t}(j), \quad j = 1, 2, \dots, V$$

where m_j is the j^{th} mediator; $\theta_0(j)$, $\alpha(j)$, and $\boldsymbol{t}(j)$ are coefficients for the intercept, the observed explanatory variable x, and the observed covariates $\boldsymbol{z}$ that are associated with the j^{th} mediator.

Second,

$$\mathbb{E}(Y|X = x, M = \boldsymbol{m}, Z = \boldsymbol{z}) = \theta_0' + \gamma x + \sum_{j=1}^{V} \beta(j)\, m(j) + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\tau}$$

where $\boldsymbol{m} = (m_1, m_2, \dots, m_V)$ is a vector representing V mediators; θ_0', γ, and $\boldsymbol{\tau}$ are coefficients for the intercept, the observed explanatory variable x, and the observed covariates $\boldsymbol{z}$; $\beta(j)$ is the coefficient associated with the j^{th} mediator.

The j^{th} mediator m_j, for $j = 1, 2, \dots, V$, is said to significantly mediate the relationship between X and Y, if both $\alpha(j)$ and $\beta(j)$ are significantly different from zero after accounting for the covariates. The product, or $\alpha(j)\beta(j)$, quantifies the mediation effect for the j^{th} mediator.

3.3.3 High-dimensional mediation analysis

A special case of multivariate mediation analysis that calls for a unique treatment is when the intermediate variable is high-dimensional. Under such a circumstance, traditional mediation analysis methods would face theoretical and computational challenges. Two common ways to treat the high dimensionality of the intermediate variable are (1) feature selection, and (2) feature transformation. For the former, a subset of mediators is selected (which are more likely to be the true mediators) for further mediation analysis. The feature selection can be done via, for example, sure independence screening (SIS)[143] followed by penalization (see Zhang *et al.*[144]). For the latter, one can project the high-dimensional mediator onto lower-dimensional

space and use the transformed low-dimensional features for final mediation analysis (see Chén *et al.*[92]).

Most existing high-dimensional mediation analytical approaches, however, assume that the outcome variable is continuous (and particularly Gaussian). This assumption offers technical convenience; for example, with Gaussian assumption, one can derive closed-form of the parameter estimates albeit high dimensionality[92]. Yet, in real life, one would come across high-dimension mediation analysis with non-Gaussian or discontinuous outcomes (such as binary disease outcomes). I shall delve into high-dimensional mediation analysis with non-Gaussian outcomes in **Chapter** 4 and discuss general treatment for high-dimensional data analysis in **Chapter** 6.

3.4 Statistical methods for building longitudinal biomarkers

A useful way to view longitudinal biomarkers is to see them as a set of features that are recorded over time, whose measurements are useful to assess, monitor, or predict (*e.g.*, disease status or severity) outcomes longitudinally.

Let's first state a general longitudinal problem. Consider n subjects and p features each of which is measured m times (one can relax it to n_i measurements per person; but for simplicity here we assume every individual has the same number of repeated measurements). Consider y_{ij} as j^{th} repeated (outcome) measurement of the i^{th} subject, where $1 \leq i \leq n$ and $1 \leq j \leq m$. Consider x_{ijk} as the j^{th} repeated (feature) measurement of the i^{th} subject on the k^{th} feature, where $1 \leq k \leq p$.

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{im})^T, \forall i$. Let $\boldsymbol{y} = (y_1^T, y_2^T, \dots, y_n^T)^T$. Let $\boldsymbol{x}_{ij} = \left(x_{ij1}, x_{ij2}, \dots, x_{ijp}\right), \forall i, \forall j$. Let $\boldsymbol{x}_i = \left(\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \dots, \boldsymbol{x}_{im}^T\right)^T, \forall i$. Let $\boldsymbol{X} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \dots, \boldsymbol{x}_n^T)^T$. In words, $\boldsymbol{y}$ is a vector containing all y_{ij}'s, where the first m entries are repeated measurements for the first subject, the second m entries are repeated measures for the second subject, and so on. And $\boldsymbol{X}$ is a matrix containing all x_{ijk}'s, where the first m rows contain features from the first subject, the second m rows contain features from the second subject, and so on; within each $m \times p$ subject-specific feature matrix $\boldsymbol{x}_i$, each column corresponds to a feature.

A standard parametric longitudinal model (see more in **Sections** 3.4.1) can be expressed as follows (see **Sections** 3.4.2 and 3.4.3 for semi- and non-parametric longitudinal models):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (11)$$

where $\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \text{blogdiag}(\boldsymbol{V}_1, \dots, \boldsymbol{V}_n)$, and $\boldsymbol{V}_i$ models the within-subject correlation structure, for $1 \leq i \leq n$. Note that (a) an equivalent expression for the above model is $\boldsymbol{Y} \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$; (b) Although linear, the expression can be modified to be a generalized linear model by adding a link function to recapitulate a non-linear relationship (see **Section** 3.4.1 for details).

If one assumes that longitudinal observations are observed independently and identically (namely with *i.i.d.* noise σ^2), then the parameters in **Eq.** (11) can be estimated straightforwardly using GLM introduced in **Section** 3.1 because $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$.

But the above treatment neglects one important characteristic of longitudinal data, namely there exists a complex variance-covariance structure regarding repeated measurements. In other words, suitable analytical approaches for longitudinal data need to be able to deal with:

Problem 1: Measurements are obtained over time.

Problem 2: There exist complex variance-covariance structures.

Whereas the treatment for **Problem 1** is built-in inherently within our discussion regarding a longitudinal problem, there are three general classes of statistical models to deal with **Problem 2**: parametric, semi-parametric, and nonparametric models.

There are, certainly, other territories of longitudinal models that we do not cover here. For example, combining longitudinal parametric modeling and mediation analysis forms a longitudinal mediation analysis. Uniting features and a hazard function, one has the Cox proportional hazards model. Dynamic systems, in their own right, form a distinctive area of longitudinal analysis. There are many more. Instead of covering all aspects of statistical modeling on longitudinal data, to address the central theme of building longitudinal biomarkers, I here shall focus on longitudinal problems concerning analyzing p features repeatedly measured and how to use them to address the following questions: which features are associated with the observed outcomes over time, and how to use them to assess and predict outcomes longitudinally.

3.4.1 Parametric models for building longitudinal biomarkers

How does a longitudinal parametric model treat variance-covariance structure? At its core, it aims at decomposing the error term $\boldsymbol{\epsilon}$ (as in $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$) into three parts: random effects, serial correlation, and measurement errors. Specifically (see, for example, p. 83 in Diggle *et al.*[145]), for the j^{th} time point of subject i:

$$\epsilon_{ij} = \boldsymbol{d}_{ij}^{T}\boldsymbol{U}_i + W_i(t_{ij}) + Z_{ij}$$

where $\boldsymbol{U}_i$ represents the random effect, which consists of a subject-specific Gaussian random vector with mean **0** and covariance $\boldsymbol{G}$ ($\boldsymbol{d}_{ij}$ are explanatory variables corresponding to individual measurements), $W_i(t_{ij})$ represents serial correlation consisting of a sample from a stationary Gaussian process with mean zero and correlation function $\gamma(u) = \rho(u;\phi)$, and Z_{ij} represents measurement error with Gaussian distribution of mean 0 and variance $\tau^2 > 0$. The correlation function needs to be further specified; popular choices are exponential correlation $\rho(u) = \exp(-\phi u)$ and Gaussian correlation $\rho(u) = \exp(-\phi u^2)$, for some value of $\phi > 0$.

The decomposition of the error term, however, does not always have to yield all three components. Depending on the research interests and assumptions, for example, if one assumes that there are no random individual differences between subjects, then the error term needs not to contain a random subject-specific component. More specifically, models whose error term only contains measurement error are called marginal models; models whose error term contains both random effects and measurement error are called random-effects models; models whose error term contains both serial effect and measurement error are called transition (or Markov) models.

Marginal models. The central focus of marginal models is to estimate population-average effects (in terms of the parameter estimates $\boldsymbol{\beta}$) of features measured over time; in contrast, random-effects models (see below) are interested in estimating subject-specific (random) effects in addition to population-average (fixed) effects.

A marginal model considers modeling the relationship between the marginal expectation of response $\mathbb{E}(Y_{ij}) = \mu_{ij}$ and the features $\boldsymbol{x}_{ij}$, through $\mu_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$. The variance-covariance

depends on $\boldsymbol{\beta}$ and another parameter $\boldsymbol{\alpha}$ that is related to the correlation between observations (*i.e.*, Corr(Y_{ij}, Y_{ik}) $= \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha})$). In other words, $\mathrm{Var}(\boldsymbol{y}_i) = \mathrm{Var}(\boldsymbol{y}_i; \boldsymbol{\beta}, \boldsymbol{\alpha})$. When the errors are non-Gaussian, one can employ a link function h (*i.e.,* let $h(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$) and the population-level parameter β can be estimated using the GEE:

$$s_\beta(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \mathrm{Var}(\boldsymbol{y}_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0$$

where the dependence of the solution $\boldsymbol{\beta}(\boldsymbol{\alpha})$ can be solved by first obtaining a $\sqrt{n}$-consistent estimation of $\boldsymbol{\alpha}$, or $\widehat{\boldsymbol{\alpha}}(\widehat{\boldsymbol{\beta}})$.

Random-effects models. The chief idea behind a random-effects model is that, when modeling real data, there exists a proportion of total variance of the data that cannot be molded or explained by traditional regression approaches which are aimed at estimating the fixed, population-level effects. A useful way to deal with this issue is to consider an approach that can both (a) estimate population-level information using fixed, population-specific parameters; and (b) further model the unexplained variance (for which the population-level parameters cannot account) using heterogeneous (*e.g.*, subject-specific) parameters. The former can be dealt with in the sense of traditional regression models; the latter can be done by considering subject-specific parameters that are randomly sampled from a probability distribution whose variance-covariance structure is estimated from (in other words, controlled to reflect) the data organization (see p. 129 in Diggle *et al.*[145]). Formally, this is modeled using a random-effects model as follows:

$$h(\mathbb{E}(Y_{ij}|\boldsymbol{U}_i)) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{d}_{ij}^T\boldsymbol{U}_i$$

where h is a link function, and $\boldsymbol{x}_{ij}$ and $\boldsymbol{d}_{ij}$ are variables corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{U}_i$, respectively. Additionally, the conditional distribution $Y_{ij}|\boldsymbol{U}_i$ follows a distribution from the

exponential family with density $f(Y_{ij}|\boldsymbol{U}_i;\boldsymbol{\beta})$; given $\boldsymbol{U}_i$, the repeated measurements $Y_{i1},\dots,Y_{im}$ are independent; $\boldsymbol{U}_i$'s are independent and identically distributed with density function $f(\boldsymbol{U}_i;\boldsymbol{G})$ (see p. 171 in Diggle *et al.*[145]).

Thus, the only parameters that need to be estimated are the fixed population-specific effects $\boldsymbol{\beta}$ and the covariance $\boldsymbol{G}$ from which the subject-specific random effects $\boldsymbol{U}_i$ are sampled. This is done using maximum likelihood estimation (MLE) via the EM-algorithm[146]. Specifically, the EM-algorithm will solve the following two score functions (see p. 173 in Diggle *et al.*[145]):

$$S_\beta(\boldsymbol{\beta},\boldsymbol{G}|\boldsymbol{y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\left[y_{ij} - E\{\mu_{ij}(\boldsymbol{U}_i)\big|\boldsymbol{y}_i\}\right] = 0$$

where $\mu_{ij}(\boldsymbol{U}_i) = \mathbb{E}\left(Y_{ij}|\boldsymbol{U}_i\right) = h^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{d}_{ij}^T\boldsymbol{U}_i)$, and

$$S_G(\boldsymbol{\beta},\boldsymbol{G}|\boldsymbol{y}) = \frac{1}{2}G^{-1}\{\textstyle\sum_{i=1}^{n} E\left(\boldsymbol{U}_i\boldsymbol{U}_i^T|\boldsymbol{y}_i\right)\}G^{-1} - \frac{n}{2}\boldsymbol{G}^{-1} = 0.$$

Markov models. Markov models for longitudinal studies consider the conditional distribution of each outcome value y_{ij} for subject i at time j as a function of past responses and covariates. Consider a general Markov model with order q. Here, by using the word general, I mean the conditional distribution falls into an exponential family (hence it is a conditional generalized linear model (GLM)). By order q, I mean the outcome value y_{ij} given its history only depends on the most recent q outcomes and the covariates. Let $\boldsymbol{\mathcal{F}}_{ij}$ denote all past information with regards to subject i at time j. A GLM Markov can be expressed as (see pp. 190-191 in Diggle *et al.*[145]):

$$f\left(y_{ij}\big|\boldsymbol{\mathcal{F}}_{ij}\right) = \exp\left\{\frac{y_{ij}\theta_{ij} - \psi\left(\theta_{ij}\right)}{\phi} + c\left(y_{ij},\phi\right)\right\}$$

with conditional mean $\mu_{ij} = \mathbb{E}\left(Y_{ij}|\boldsymbol{\mathcal{F}}_{ij}\right) = \psi'(\theta_{ij})$ and conditional variance $\sigma_{ij} = \mathbb{E}\left(Y_{ij}|\boldsymbol{\mathcal{F}}_{ij}\right) = \psi''(\theta_{ij})\phi$, where f specifies the conditional distribution and $\psi\left(\theta_{ij}\right)$ and $c\left(y_{ij}, \phi\right)$ are known functions.

One can then model the conditional mean and variance as a function of the most recent q outcomes and the covariance as:

$$h\left(\mu_{ij}\right) = \boldsymbol{x}_{ij}^{T}\beta + \sum_{k=1}^{q} g_k(\boldsymbol{\mathcal{F}}_{ij}; \alpha)$$

and

$$\sigma_{ij} = v(\mu_{ij})\phi$$

where h and v are known link functions.

The linear link (for modeling longitudinal continuous outcomes with Gaussian autoregressive errors) and logit link (for modeling longitudinal binary outcomes) are two common, and useful link functions for treating longitudinal data. A linear Markov model with order q and Gaussian error can be expressed as[147]:

$$y_{ij} = \boldsymbol{x}_{ij}^{T}\beta + \sum_{k=1}^{q} \alpha_k\left(y_{i,j-k} - \boldsymbol{x}_{i,j-k}^{T}\boldsymbol{\beta}\right) + Z_{ij}$$

where Z_{ij} denotes a mean zero Gaussian measurement error. Similarly, a log-linear Markov model with order q can be expressed as[148]:

$$logit\left(P\left(Y_{ij} = y_{ij}\middle|\boldsymbol{\mathcal{F}}_{ij}\right)\right) = \boldsymbol{x}_{ij}^{T}\beta + \textstyle\sum_{k=1}^{q} \alpha_k y_{i,j-k}.$$

Despite the complexity, the essence of Markov models does not depart from the error decomposition (see above):

$$\epsilon_{ij} = \boldsymbol{d}_{ij}^{T}\boldsymbol{U}_i + W_i\left(t_{ij}\right) + Z_{ij}$$

where Markov models only consider ϵ_{ij} to contain serial correlation $W_i(t_{ij})$ and measurement error Z_{ij}. To see this, let's consider a regression model with serial correlation of order q:

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$$

where $\epsilon_{ij} = \sum_{k=1}^{q} \alpha_k \epsilon_{i,j-k} + Z_{ij}$ and $Z_{ij} \sim N(0, \tau^2)$.

Now replace $\epsilon_{ij} = y_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ and $\epsilon_{i,j-k} = y_{i,j-k} - \boldsymbol{x}_{i,j-k}^T\boldsymbol{\beta}$, we have:

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sum_{k=1}^{q} \alpha_k\left(y_{i,j-k} - \boldsymbol{x}_{i,j-k}^T\boldsymbol{\beta}\right) + Z_{ij}$$

which is identical to a linear Markov model with order q and Gaussian error expressed above.

The parameter estimates of Markov models can be done via MLE under Gaussian assumption (namely when $Y_{ij}|\boldsymbol{\mathcal{F}}_{ij}$ is Gaussian) or via maximizing the conditional likelihood, otherwise (see pp. 192-194 in Diggle *et al.*[145]).

3.4.2 Semi-parametric models for building longitudinal biomarkers

The central goals of modeling longitudinal data semi-parametrically are to obtain a smooth (longitudinal) curve and to model the relationship between features and the outcome. The former is done using a nonparametric approach and the latter using a parametric one, hence the name "semi-parametric". Specially, let $\mu(t_{ij})$ denote a curve at time t_{ij} for subject i at time j. Other notations follow the same as parametric models discussed above: y_{ij} and $\boldsymbol{x}_{ij}$ are outcomes and feature data of subject i at time j; $W_i(j)$ represents serial correlation consisting of a sample from a stationary Gaussian process with mean zero and correlation function $\gamma(u) = \rho(u;\theta)$, and Z_{ij} represents measurement error with Gaussian distribution of mean 0 and variance τ^2.

Semi-parametric models for longitudinal data can then be expressed as follows:

$$y_{ij}(t_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \mu(t_{ij}) + W_i(t_{ij}) + Z_{ij}$$

or equivalently:

$$y_{ij}(t) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \mu(t) + W_i(t) + Z_{ij}.$$

Here, $\boldsymbol{\beta}$ denotes the feature parameters, constituting the part that makes the above model "parametric". The term $\mu(t)$ represent an arbitrary smooth function of time and is hence nonparametric (or infinite parametric). Taken together, the full model is semi-parametric.

Both nonparametric and parametric estimates in a semi-parametric model are obtained iteratively. The nonparametric estimate of $\mu(t)$ can be done using kernel estimates (see Zeger and Diggle[149] and Rice and Silverman[150]). One can alternatively use spine estimate and it has been shown that choosing the appropriate bandwidth for spline estimates and kernel for kernel estimates would yield similar results[151]; but the results from kernel approaches are much less sensitive to the kernel chosen than they are to the choice of bandwidth for splines[149]. Additionally, for semiparametric models, the kernels are estimated using the residuals $y_{ij} - \boldsymbol{x}_{ij}^T\widehat{\boldsymbol{\beta}}^{[k]}$ (instead of y_{ij}), where $\widehat{\boldsymbol{\beta}}^{[k]}$ is the estimates for $\boldsymbol{\beta}$ during the k^{th} iteration. The parameter estimates of $\boldsymbol{\beta}$ is done using the backfitting algorithm[152]:

$$\widehat{\boldsymbol{\beta}}^{[k]} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\left(y_{ij} - \mu_{ij}^{[k-1]}\right)$$

where $\boldsymbol{V}_i$ is the covariance matrix of $\boldsymbol{y}_i$. In other words, the algorithm iteratively fits generalized least squares on the residuals or $y_{ij} - \mu_{ij}^{[k-1]}$. Finally, the estimation for W_i is done based upon the conditional expectation of W_i and the properties of the multivariate Gaussian distribution (see Zeger and Diggle[149] for a closed-form solution).

3.4.3 Nonparametric models for building longitudinal biomarkers

Nonparametric models are useful approaches when one is interested in investigating how a baseline covariate (*e.g.*, drug dosage) may affect a longitudinal curve, but does not have (or is not interested in considering) feature variables to perform parametric analysis. For example, consider a baseline covariate, say the amount of dosage, which has $m \geq 1$ levels with the j^{th} level valued at x_j; each dosage may have a different impact on an outcome over time and each may affect individuals differently. Thus, to study the effect of a baseline dosage levelled at x_j on a longitudinal observation $y_{ij}(t)$ for a particular subject i at time t, one could formulate a nonparametric model as follows[153]:

$$y_{ij}(t) = \mu_j(t) + \eta_{ij}(t) + Z_{ij}(t) \qquad (12)$$

where $\mu_j(t) = \mu(x,t)|_{x=x_j}$ which depicts the longitudinal effect from the baseline covariate, $\eta_{ij}(t)$ and $Z_{ij}(t)$ are two independent stochastic processes with zero and covariance functions $V(s,t)$ and $\sigma^2\delta(s,t)$, where $\delta(s,t) = 1$ if $s = t$ and 0 otherwise. Particularly, $\eta_{ij}(t)$ represents longitudinal subject-specific pattern and $Z_{ij}(t)$ denotes the error term. **Eq**. (12) can then be rewritten as follows:

$$y_i(x,t) = \mu(x,t) + \eta_i(x,t) + Z_i(x,t)$$

where $\mu(x,t) = \sum_{k=1}^{\infty} \xi_k(x)\,\psi_k(t) + h(t)$, $\{\psi_k\}_{k=1}^{\infty}$ forms an orthogonal system for the space of square-integrable functions, and x denotes the baseline dosage level.

The primary interest, namely the longitudinal curve that is a function of both baseline covariate and time, can be estimated using kernel or spline estimates, as is done in the nonparametric part of the semi-parametric models (see citations in **Section** 3.4.2); a detailed procedure can be found in Staniswalis and Lee[153].

Nevertheless, there is an important territory of non-parametric models that I do not discuss in this thesis; that is, Bayesian nonparametric models. One, however, can find a thorough overview elsewhere regarding Bayesian models suitable for studying the functional organization and structural architecture of the brain, as well as their relationship to human behavior[154,155,164–166,156–163].

Collectively, I hope this thesis reveals the continuity from static, to dynamic, to intermediate, to longitudinal biomarkers, the fundamental questions we ask about them, and the scientific conclusions we make, through the lenses of statistics and data science, about the brain, behavior, and brain disorders. But the categorization of biomarkers and questions about how to unveil them have been, in one way or another, raised before, and are continued to be studied today. Therefore, I attempt to include in this thesis a few groups of newly identified biomarkers that may bring fresh perspectives about the functions and functioning of the brain, how they give rise to human behavior, and how the brain becomes diseased, as well as a few new or modified methodological frameworks that may provide some insights about building efficient, targeted, and affordable biomarkers. Together, I hope that the collection of papers in this thesis provides some evidence with which future studies will continue to improve the accuracy of biomarker identification and to advance biomarker-based disease and behavior prediction.

Certainly, there are important areas with regards to biomarker development this thesis has not covered. This is in part due to limitations of space and in part due to my ignorance. But, at the minimum, I hope that these papers paint a clear picture about the division of biomarkers and, through them, an exciting look into the future methodological and biological enquiries into the

brain and behavior. Through the glimpse of biomarkers developed from novel statistical methods, I hope that this collection of works could inject useful insights towards a unified view of the studies of human behavior, mental disorders, and, above all, the functional organization of the human brain.

References

1. Kandel, E. R., Schwartz, J. H. & Jessell, T. M. *Principles of neural science*. McGraw-Hill, New York, NY, USA (2000).
2. Galen, A. *On the doctrines of Hippocrates and Plato*. Akademie-Verlag, Berlin, Germany (1981).
3. Gill, C. Galen and the stoics: Mortal enemies or blood brothers? *Phronesis* **52**, pp. 88–120 (2007).
4. Descartes, R. *Meditationes de prima philosophia* (Meditations on first philosophy). Printed by Michaelem Soly (1641).
5. Spinoza, B. de. *Ethica, ordine geometrico demonstrata* (Ethics, demonstrated in geometrical order). Rieuwertsz, Amsterdam, The Netherlands (1677).
6. Kant, I. *Kritik der reinen Vernunft, 1st edition, translated by WS Pluhar (1996) as Critique of pure reason*. Hackett, Indianapolis, IN, USA (1781).
7. Darwin, C. *The expression of the emotions in Man and animals*. John Murray, London, UK (1872).
8. S. Zeki. *A theory of multi-stage integration in the visual cortex*. In *The principles of design and operation of the brain*, edited by Eccles, J.C. and Creutzfeldt, O. pp. 137–154. Pontifical Academy of Science, Vatican City. (1988).
9. Bartels, A. & Zeki, S. The theory of multistage integration in the visual brain. *Proc. R. Soc. B Biol. Sci.* **265**, pp. 2327–2332 (1998).
10. de Lau, L. M. & Breteler, M. M. Epidemiology of Parkinson's disease. *Lancet Neurology* **5**, pp. 525–535 (2006).
11. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, pp. 695–699 (2005).
12. Goetz, C. G. *et al.* Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord.* **23**, pp. 2129–2170 (2008).
13. Bhidayasiri, R. & Martinez-Martin, P. Clinical assessments in Parkinson's disease: Scales and monitoring. *International Review of Neurobiology*, **132**, pp. 129–182 (2017).
14. Samii, A., Nutt, J. G. & Ransom, B. R. Parkinson's disease. *Lancet* **363**, pp. 1783–1793 (2004).
15. Ives, N. J. *et al.* Monoamine oxidase type B inhibitors in early Parkinson's disease: Meta-analysis of 17 randomised trials involving 3525 patients. *Br. Med. J.* **329**, 593 (2004).
16. Connolly, B. S. & Lang, A. E. Pharmacological treatment of Parkinson disease: A review. *JAMA* **311**, pp. 1670–83(2014).
17. National Collaborating Centre for Chronic Conditions (UK). Symptomatic pharmacological therapy in Parkinson's disease. In *Parkinson's Disease: National Clinical Guideline for Diagnosis and Management in Primary and Secondary Care*. NICE Clinical Guidelines, No. 35. Royal College of Physicians (UK), London, UK (2006).
18. Schuepbach W. M. *et al.* Neurostimulation for Parkinson's disease with early motor complications. *N. Engl. J. Med.* **368**, pp. 610–622 (2013).
19. Deuschl, G. *et al.* A randomized trial of deep-brain stimulation for Parkinson's disease. *N. Engl. J. Med.* **355**, pp. 896–908 (2006).
20. Weaver, F. M. *et al.* Randomized trial of deep brain stimulation for Parkinson disease: Thirty-six-month outcomes. *Neurology* **79**, pp. 55–65 (2012).

21. Lipsmeier, F. *et al.* Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov. Disord.* **33**, pp. 1287–1297 (2018).
22. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). *American Journal of Psychiatry* (2013).
23. Moreno-Küstner, B., Martín, C. & Pastor, L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS ONE* **13**, e0195687 (2018).
24. Carpenter, W. T., Heinrichs, D. W. & Wagman, A. M. I. Deficit and nondeficit forms of schizophrenia: The concept. *Am. J. Psychiatry* **145**, pp. 578–583 (1988).
25. Green, M. F., Kern, R. S., Braff, D. L. & Mintz, J. Neurocognitive deficits and functional outcome in schizophrenia: Are we measuring the 'right stuff'? *Schizophrenia Bulletin* **26**, pp. 119–136 (2000).
26. Harvey, P. D., Reichenberg, A., Bowie, C. R., Patterson, T. L. & Heaton, R. K. The course of neuropsychological performance and functional capacity in older patients with schizophrenia: Influences of previous history of long-term institutional stay. *Biol. Psychiatry* **67**, pp. 933–939 (2010).
27. Hamilton, S. H., Edgell, E. T., Revicki, D. A. & Breier, A. Functional outcomes in schizophrenia: A comparison of olanzapine and haloperidol in a European sample. *Int. Clin. Psychopharmacol.* **15**, pp. 245–255 (2000).
28. Bond, G. R. *et al.* Response to vocational rehabilitation during treatment with first- or second-generation antipsychotics. *Psychiatr. Serv.* **55**, pp. 59–66 (2004).
29. Rosenheck, R. A. *et al.* Cost-effectiveness of second-generation antipsychotics and perphenazine in a randomized trial of treatment for chronic schizophrenia. *Am. J. Psychiatry* **163**, pp. 2080–2089 (2006).
30. Oyebode, F. The neurology of psychosis. *Medical Principles and Practice* 17, pp. 263–269 (2008).
31. Overall, J. E. & Gorham, D. R. The brief psychiatric rating scale. *Psychol. Rep.* **10**, pp. 799–812 (1962).
32. National Institute for Health and Care Excellence (NICE). *Psychosis and schizophrenia in adults: prevention and management*. (2014). Available at: https://www.nice.org.uk/guidance/cg178
33. Addington, J. *et al.* North American Prodrome Longitudinal Study (NAPLS 2): Overview and recruitment. *Schizophr. Res.* **142**, pp. 77–82 (2012).
34. McGlashan, T. H., Miller, T. J., Woods, S. W., Hoffman, R. E. & Davidson, L. Instrument for the assessment of prodromal symptoms and states. In *Early Intervention in Psychotic Disorders*, edited by T. Miller, S.A. Mednick, T.H. McGlashan, J. Libiger, and J.O. Johannessen, pp. 135–149. Springer, Dordrecht, The Netherlands (2001).
35. Bell, C. C. DSM-IV: Diagnostic and statistical manual of mental disorders. *JAMA* **272**, pp. 828–829 (1994).
36. Miller, T. J. *et al.* Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: Predictive validity, interrater reliability, and training to reliability. *Schizophrenia Bulletin* **29**, pp. 703–715 (2003).
37. Miller, T. J. *et al.* Prospective diagnosis of the initial prodrome for schizophrenia based on the structured interview for prodromal syndromes: Preliminary evidence of interrater reliability and predictive validity. *Am. J. Psychiatry* **159**, 863–865 (2002).
38. van Schependom, J. & Nagels, G. Targeting cognitive impairment in multiple sclerosis-the road toward an imaging-based biomarker. *Front. Neurosci.* **11**, 380 (2017).

39. Compston, A. & Coles, A. Multiple sclerosis. *The Lancet* **372**, pp. 1502–1517, (2008).
40. Murray, E. D., Buttner, E. A. & Price, B. H. Depression and psychosis in neurological practice. In *Neurology in clinical practice*, edited by W.G. Bradley, R.B. Daroff, G.M. Fenichel and J. Jankovic, pp. 103–116. Butterworth- Heinemann, Oxford, UK. (2012).
41. Trapp, B. D., Ransohoff, R. M., Fisher, E. & Rudick, R. A. Neurodegeneration in multiple sclerosis: Relationship to neurological disability. *Neuroscientist* **5**, pp. 48–57 (1999).
42. Benedict, R. H. B. *et al.* Frontal cortex atrophy predicts cognitive impairment in multiple sclerosis. *J. Neuropsychiatry Clin. Neurosci.* **4**, pp. 44–51 (2002).
43. Benedict, R. H. B. *et al.* Regional lobar atrophy predicts memory impairment in multiple sclerosis. *Am. J. Neuroradiol.* **26**, pp. 1824–1831 (2005).
44. Morgen, K. *et al.* Evidence for a direct association between cortical atrophy and cognitive impairment in relapsing-remitting MS. *Neuroimage* **30**, pp. 891–898 (2006).
45. Houtchens, M. K. *et al.* Thalamic atrophy and cognition in multiple sclerosis. *Neurology* **69**, pp. 1213–1223 (2007).
46. Sicotte, N. L. *et al.* Regional hippocampal atrophy in multiple sclerosis. *Brain* **131**, pp. 1134–1141 (2008).
47. Batista, S. *et al.* Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis. *J. Neurol.* **259**, pp. 139–146 (2012).
48. Damjanovic, D. *et al.* Hippocampal and deep gray matter nuclei atrophy is relevant for explaining cognitive impairment in MS: A multicenter study. *Am. J. Neuroradiol.* **38**, pp. 18–24 (2017).
49. Preziosa, P. *et al.* Structural MRI correlates of cognitive impairment in patients with multiple sclerosis: a multicenter study. *Hum. Brain Mapp.* **37**, pp. 1627–1644. (2016).
50. Rocca, M. A. *et al.* Regional hippocampal involvement and cognitive impairment in pediatric multiple sclerosis. *Mult. Scler.* **22**, pp. 628–640 (2016).
51. Calabrese, M. *et al.* Cortical lesions and atrophy associated with cognitive impairment in relapsing-remitting multiple sclerosis. *Arch. Neurol.* **66**, pp. 1144–1150 (2009).
52. Roosendaal, S. D. *et al.* Regional DTI differences in multiple sclerosis patients. *Neuroimage* **44**, pp. 1397–1403 (2009).
53. Kern, K. C. *et al.* Fornix damage limits verbal memory functional compensation in multiple sclerosis. *Neuroimage* **59**, pp. 2932–2940 (2012).
54. Mesaros, S. *et al.* Diffusion tensor MRI tractography and cognitive impairment in multiple sclerosis. *Neurology* **78**, pp. 969–975 (2012).
55. Fink, F. *et al.* The association between California Verbal Learning Test performance and fibre impairment in multiple sclerosis: Evidence from diffusion tensor imaging. *Mult. Scler.* **16**, pp. 332–341 (2010).
56. Tsang, B. K. T. & Macdonell, R. Multiple sclerosis: Diagnosis, management and prognosis. *Aust. Fam. Physician* **40**, pp. 948–955 (2011).
57. Riazi, A., Hobart, J. C., Lamping, D. L., Fitzpatrick, R. & Thompson, A. J. Multiple Sclerosis Impact Scale (MSIS-29): Reliability and validity in hospital based samples. *J. Neurol. Neurosurg. Psychiatry* **73**, pp. 701–704 (2002).
58. Hobart, J. The Multiple Sclerosis Impact Scale (MSIS-29): A new patient-based outcome measure. *Brain* **124**, pp. 962–973 (2001).
59. McGuigan, C. & Hutchinson, M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. *J. Neurol. Neurosurg. Psychiatry* **75**, pp. 266–269 (2004).
60. Rae-Grant A. *et al.* Practice guideline recommendations summary: Disease-modifying therapies for adults with multiple sclerosis: Report of the guideline development,

dissemination, and implementation subcommittee of the American Academy of Neurology. *Neurology* **90**, pp. 777–788 (2018).
61. Burton, J. M., O'Connor, P. W., Hohol, M. & Beyene, J. Oral versus intravenous steroids for treatment of relapses in multiple sclerosis. *Cochrane Database Syst. Rev.* **12**, CD006921 (2009).
62. Citterio, A. *et al.* Corticosteroids or ACTH for acute exacerbations in multiple sclerosis. *Cochrane Database Syst. Rev.* **4**, CD001331 (2000).
63. National Collaborating Centre for Chronic Conditions (UK). *Multiple Sclerosis: National Clinical Guideline for Diagnosis and Management in Primary and Secondary Care*. Royal College of Physicians (UK), London, UK (2004).
64. Rice, G. P. *et al.* Interferon in relapsing-remitting multiple sclerosis. *Cochrane Database Syst. Rev.* **4**, CD002002 (2001).
65. Pucci, E. *et al.* Natalizumab for relapsing remitting multiple sclerosis. *Cochrane Database Syst. Rev.* **10**, CD007621 (2011).
66. La Mantia, L. *et al.* Fingolimod for relapsing-remitting multiple sclerosis. *Cochrane Database Syst. Rev.* **4**, CD009371 (2016).
67. He, D. *et al.* Teriflunomide for multiple sclerosis. *Cochrane Database Syst. Rev*. **3**, CD009882 (2016).
68. Xu, Z. *et al.* Dimethyl fumarate for multiple sclerosis. *Cochrane Database Syst. Rev.* **4**, CD011076 (2015).
69. Riera, R., Porfírio, G. J. M. & Torloni, M. R. Alemtuzumab for multiple sclerosis. *Cochrane Database Syst. Rev.* **4**, CD011203 (2016).
70. Faissner, S. & Gold, R. Progressive multiple sclerosis: latest therapeutic developments and future directions. *Therapeutic Advances in Neurological Disorders* **12**, pp. 1–11(2019).
71. Derfuss, T. *et al.* Advances in oral immunomodulating therapies in relapsing multiple sclerosis. *The Lancet Neurology* **19**, pp. 336–347 (2020).
72. Rasche, L. & Paul, F. Ozanimod for the treatment of relapsing remitting multiple sclerosis. *Expert Opin. Pharmacother.* **19**, pp. 2073–2086 (2018).
73. Winslow, R. After 40-year odyssey, first drug for aggressive MS wins FDA approval. *STAT* (2017). Available at: https://www.statnews.com/2017/03/28/multiple-sclerosis-ms-drug-ocrelizumab/#:~:text=After%2040%2Dyear%20odyssey%2C%20first,aggressive%20MS%20wins%20FDA%20approval&text=Forty%20years%20ago%2C%20one,brush%20fire%20through%20her%20brain.
74. Baker, M., Beek, J. van & Gossens, C. Digital health: Smartphone-based monitoring of multiple sclerosis using Floodlight. *Scientific American* (2020). Available at: https://www.nature.com/articles/d42473-019-00412-0
75. Montalban, X. *et al.* Floodlight: Remote self-monitoring is accepted by patients and provides meaningful, continuous sensor-based outcomes consistent with and augmenting conventional in-clinic measures. *Neurology* **90**, P4.382 (2018).
76. Kieseier, B. C. & Pozzilli, C. Assessing walking disability in multiple sclerosis. *Multiple Sclerosis Journal* **18**, pp. 914–924 (2012).
77. Kierkegaard, M., Einarsson, U., Gottberg, K., Von Koch, L. & Holmqvist, L. W. The relationship between walking, manual dexterity, cognition and activity/participation in persons with multiple sclerosis. *Mult. Scler. J.* **18**, pp. 639–646 (2012).
78. Scalfari, A. *et al.* The natural history of multiple sclerosis, a geographically based study 10: Relapses and long-term disability. *Brain* **133**, pp. 1914–1929 (2010).
79. Feys, P. *et al.* The Nine-Hole Peg Test as a manual dexterity performance measure for

multiple sclerosis. *Multiple Sclerosis* **23**, pp. 711–720 (2017).
80. Midaglia, L. *et al.* Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study. *J. Med. Internet Res.* **21**, e14863 (2019).
81. Hulka, B. S. & Wilcosky, T. Biological markers in epidemiologic research. *Arch. Environ. Health* **43**, pp. 83–89 (1988).
82. Naylor, S. Biomarkers: Current perspectives and future prospects. *Expert Review of Molecular Diagnostics* **3**, pp. 525–529 (2003).
83. Mayeux, R. Biomarkers: Potential uses and limitations. *NeuroRx* **1**, pp. 182–188 (2004).
84. Friston, K. J. Functional and effective connectivity in neuroimaging: A synthesis. *Hum. Brain Mapp.* **2**, pp. 56–78 (1994).
85. Zeki, S. *A Vision of the Brain*. Blackwell Scientific, Oxford, UK (1993).
86. Zeki, S. The visual image in mind and brain. *Sci. Am.* **267**, pp. 68–76 (2010).
87. Aertsen, A. & Preissl, H. Dynamics of activity and connectivity in physiological neuronal networks. In *Nonlinear Dynamics and Neuronal Networks*, edited by Schuster H., pp. 281–301. VHC Verlag, Weinheim, Germany (1991).
88. Friston, K. J. Functional and effective connectivity: A review. *Brain Connect.* **1**, pp. 13–36 (2011).
89. Gerstein, G. L. & Perkel, D. H. Simultaneously recorded trains of action potentials: Analysis and functional interpretation. *Science* **164**, pp. 828–830 (1969).
90. Gerstein, G. L., Bedenbaugh, P. & Aertsen, M. H. J. Neuronal assemblies. *IEEE Trans. Biomed. Eng.* **36**, pp. 4–14 (1989).
91. Gochin, P. M., Miller, E. K., Gross, C. G. & Gerstein, G. L. Functional interactions among neurons in inferior temporal cortex of the awake macaque. *Exp. Brain Res.* **84**, pp. 505–516 (1991).
92. Chén, O. Y. *et al.* High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19**, pp. 121–136 (2018).
93. Geuter, S. *et al.* Multiple brain networks mediating stimulus-pain relationships in humans. *Cereb. Cortex,* **7**, pp. 4204–4219 (2020).
94. VanderWeele, T. J. & Vansteelandt, S. Mediation analysis with multiple mediators. *Epidemiol. Method.* **2**, pp. 95–115 (2014).
95. Huang, Y. T. & Pan, W. C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72**, pp. 402–413 (2016).
96. Miller, D., Barkhof, F., Montalban, X., Thompson, A. & Filippi, M. Clinically isolated syndromes suggestive of multiple sclerosis, part I: Natural history, pathogenesis, diagnosis, and prognosis. *Lancet Neurology* **4**, pp. 281–288 (2005).
97. Lublin, F. D. & Reingold, S. C. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology* **46**, pp. 907–911 (1996).
98. Hobart, J. C., Riazi, A., Lamping, D. L., Fitzpatrick, R. & Thompson, A. J. Measuring the impact of MS on walking ability: The 12-item MS Walking Scale (MSWS-12). *Neurology* **60**, pp. 31–36. (2003).
99. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, pp. 289–310 (2010).
100. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45**, pp. 427–437 (2009).

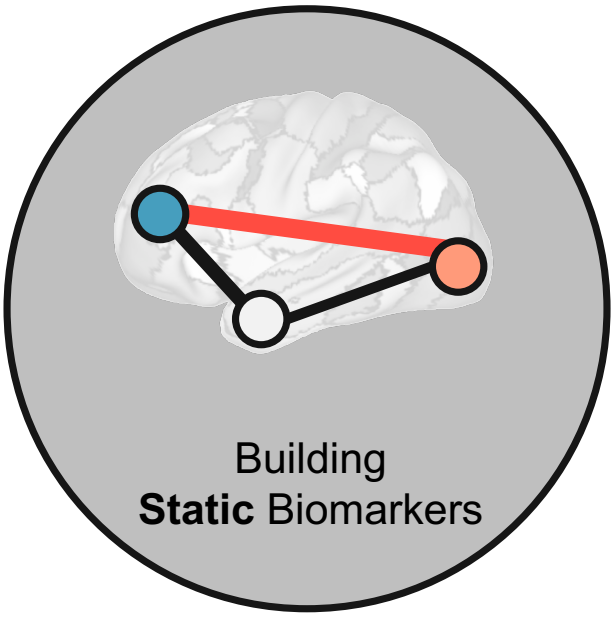
101. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*. Springer, New York, NY, USA (2001).
102. Hastie, T., Tibshirani, R. & Andreas, B. Flexible discriminant and mixture models. In *Statistics and neural networks: advances at the interface,* edited by J.W. Kay and D.M Titterington. Oxford University Press, Oxford, UK. (1995).
103. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, pp. 386–408 (1958).
104. Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag, New York, NY, USA (2000).
105. Gill, P. E., Murray, W. & Wright, M. H. *Practical optimization*. Academic Press, Cambridge, MA, USA (1981).
106. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, PP. 436–444 (2015).
107. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. MIT press, Cambridge, MA, USA (2017).
108. Friedman, J. H. & Tukey, J. W. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **9**, pp. 881–890 (1974).
109. Friedman, J. H. & Stuetzle, W. Projection pursuit regression. *J. Am. Stat. Assoc.* **76**, pp. 817–823 (1981).
110. Huber, P. J. Projection pursuit. *Ann. Stat.* 13, pp. 435–475 (1985).
111. Breiman, L. & Ihaka, R. Nonlinear discriminant analysis via scaling and ACE. *Tech. Rep. No.40.* Department of Statistics, University of California, Berkeley, CA, USA (1984).
112. Ripley, B. D. *Pattern recognition and neural networks*. Cambridge University, Cambridge, UK (2014).
113. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, pp. 123–140 (1996).
114. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
115. Breiman, L. Arcing classifiers. *Ann. Stat.* **26**, pp. 123–140 (1998).
116. Breiman, L. Random forests. *Mach. Learn.* **45**, pp. 5–32 (2001).
117. Bühlmann, P. & Hothorn, T. Boosting algorithms: Regularization, prediction and model fitting. *Stat. Sci.* **22**, pp. 477–505 (2007).
118. McCullagh, P. & Nelder, J. A. *Generalized linear models*. Chapman & Hall/CRC, Boca Raton, FL, USA (1989).
119. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, pp. 13–22 (1986).
120. Ramsay, J. O. & Silverman, B. W. *Functional data analysis*. Springer-Verlag New York, NY, USA (1997).
121. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modelling. *Neuroimage* **19**, pp. 1273–1302 (2003).
122. Wiener, N. The theory of prediction. In *Modern Mathematics for the Engineer,* edited by E.F. Beckenbach, pp. 165–190. Dover Publications, Mineola, New York, USA (1956).
123. Granger, C. W. J. Economic processes involving feedback. *Inf. Control* 6, pp. 28–48 (1963).
124. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, pp. 424–438 (1969).
125. Geweke, J. Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.* **77**, pp. 304–313 (1982).

126. McLntosh, A. R. & Gonzalez-Lima, F. Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.* **2**, pp. 2–22 (1994).
127. Jensen, F. V. *An introduction to Bayesian networks*. Springer, New York, USA (1996).
128. Nielsen, T. D. & Jensen, F. V. *Bayesian network and decision graph*. Springer-Verlag, New York, NY, USA (2007).
129. Charniak, E. Bayesian networks without tears. *AI Mag.* **12**, pp. 50–63 (1991).
130. Patel, R. S., Bowman, F. D. & Rilling, J. K. A Bayesian approach to determining connectivity of the human brain. *Hum. Brain Mapp.* **27**, pp. 267–276 (2006).
131. Brodersen, K. H. *et al.* Generative embedding for model-based classification of fMRI data. *PLoS Comput. Biol.* **7**, e1002079 (2011).
132. Zhang, T., Wu, J., Li, F., Caffo, B. & Boatman-Reich, D. A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *J. Am. Stat. Assoc.* **110**, pp. 93–106 (2015).
133. VanderWeele, T. J. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* **37**, pp. 17–32 (2016).
134. Mackinnon, D. P. & Dwyer, J. H. Estimating mediated effects in prevention studies. *Eval. Rev.* **17**, pp. 144–158 (1993).
135. VanderWeele, T. J. & Vansteelandt, S. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology* **172**, pp. 1339–1348 (2010).
136. VanderWeele, T. J. Explanation in causal inference: Developments in mediation and interaction. *Int. J. Epidemiol.* **45**, pp. 1904–1908 (2016).
137. Hyman, H. H. *Survey design and analysis: principles, cases and procedures*. Free Press, New York, NY, USA (1955).
138. Alwin, D. F. & Hauser, R. M. The decomposition of effects in path analysis. *Am. Sociol. Rev.* **40**, pp. 37–47 (1975).
139. Judd, C. M. & Kenny, D. A. Process analysis: Estimating mediation in treatment evaluations. *Eval. Rev.* **5**, pp. 602–619 (1981).
140. Sobel, M. E. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* **13**, 290–312 (1982).
141. Baron, R. M. & Kenny, D. a. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, pp. 1173–1182 (1986).
142. Lindquist, M. A. Functional causal mediation analysis with an application to brain connectivity. *J. Am. Stat. Assoc.* **107**, pp. 1297–1309 (2012).
143. Fan, J. & Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **70**, pp. 849–911 (2008).
144. Zhang, H. *et al.* Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32**, pp. 3150–3154 (2016).
145. Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. *Analysis of longitudinal data*. Oxford University Press, Oxford, UK (2013).
146. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**, pp. 1–22 (1977).
147. Tsay, R. S. Regression models with time series errors. *J. Am. Stat. Assoc.* **79**, pp. 118–124 (1984).
148. Cox, D. R. & Snell, E. J. *Analysis of binary data*. Chapman & Hall/CRC, Boca Raton, FL, USA (1989).
149. Zeger, S. L. & Diggle, P. J. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, pp. 689–699

(1994).
150. Rice, J. A. & Silverman, B. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc. Ser. B* **53**, pp. 233–243 (1991).
151. Silverman, B. W. Spline smoothing: The equivalent variable kernel method. *Ann. Stat.* **12**, pp. 898–916 (1984).
152. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **82**, pp. 371–386 (1986).
153. Staniswalis, J. G. & Lee, J. J. Nonparametric regression analysis of longitudinal data. *J. Am. Stat. Assoc.* **93**, pp. 1403–1418 (1998).
154. Dayan, P., Hinton, G. E. E., Neal, R. M. M. & Zemel, R. S. S. The helmholtz machine. *Neural Comput.* **7**, pp. 889–904 (1995).
155. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, pp. 79–87 (1999).
156. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.* **20**, pp. 1434–48 (2003).
157. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, pp. 271–304 (2004).
158. Knill, D. C. & Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences* **27**, pp. 712–719 (2004).
159. Yuille, A. & Kersten, D. Vision as Bayesian inference: Analysis by synthesis? *Trends Cogn. Sci.* **10**, pp. 301–308 (2006).
160. Friston, K., Mattout, J. & Kilner, J. Action understanding and active inference. *Biol. Cybern.* **104**, pp. 137–160 (2011).
161. Friston, K. The history of the future of the Bayesian brain. *Neuroimage* **62**, pp. 1230–1233 (2012).
162. Botvinick, M. & Toussaint, M. Planning as inference. *Trends in Cognitive Sciences* **16**, pp. 485–488 (2012).
163. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, pp. 181–253. (2013).
164. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: Knowns and unknowns. *Nat. Neurosci.* **16**, pp. 1170–1178 (2013).
165. Friston, K. *et al.* Active inference and epistemic value. *Cogn. Neurosci.* **6**, pp. 187–214 (2015).
166. Zeki, S. & Chén, O. Y. The Bayesian-Laplacian brain. *Eur. J. Neurosci.* **51**, pp. 1441–1462 (2019).

Chapter 2

Building static biomarkers



Building a Machine-Learning Framework to Remotely Assess Parkinson's Disease Using Smartphones[1]

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder that affects multiple neurological systems. Traditional PD assessment is conducted by a physician during infrequent clinic visits. Using smartphones, remote patient monitoring has the potential to obtain objective behavioral data semi-continuously, track disease fluctuations, and avoid rater dependency. Smartphones collect sensor data during various active tests and passive monitoring, including balance (postural instability), dexterity (skill in performing tasks using hands), gait (the pattern of walking), tremor (involuntary muscle contraction and relaxation), and voice. Some of the features extracted from smartphone data are potentially associated with specific PD symptoms identified by physicians. To leverage large-scale cross-modality smartphone features, we propose a machine-learning framework for performing automated disease assessment. The framework consists of a two-step feature selection procedure and a generic model based on the elastic-net regularization. Using this framework, we map the PD-specific architecture of behaviors using data obtained from both PD participants and healthy controls (HCs). Utilizing these atlases of features, the framework shows promises to (a) discriminate PD participants from HCs, and (b) estimate the disease severity of individuals with PD. Data analysis results from 437 behavioral features obtained from 72 subjects (37 PD and 35 HC) sampled from 17 separate days during a period of up to six months suggest that this framework is potentially useful for the analysis of remotely collected smartphone sensor data in individuals with PD.

[1]This chapter is published as: Chén, O.Y., Lipsmeier, F., Phan, H., Prince, J., Taylor, K., Gossens, C., Lindemann, M. and de Vos, M., 2020. Building a Machine-learning Framework to Remotely Assess Parkinson's Disease Using Smartphones. *IEEE Transactions on Biomedical Engineering* **67**, pp. 3491-3500.

1. Introduction

Parkinson's disease (PD) affects seven million people 40 worldwide; the prevalence increases from 1% of the population for those over 60 years of age to 4% over 80[1]. A reliable, objective, fast, and remote method to quantify the presence and severity of PD symptoms would benefit a large number of people who are affected by, or are at risk to develop, PD. Previous studies have measured common PD symptoms with object- and technology-based tests, such as sustained phonation (*i.e.*, voice)[2,3], rest tremor[4–6], postural tremor[7,8], dexterity[9,10], balance[11–13], and gait[12,14]. Advancement in digital technologies makes data collection using smartphones increasingly convenient and accurate. Smartphones are small, portable, and widely used. The data captured from various smartphone sensors can be remotely transferred via wireless networks, facilitating out-clinic data collection and assessment. Their relatively low cost and high penetration rate (compared to hospital devices and physician availability, respectively) make them affordable and available to a broad population. Because of these attractive properties, researchers have begun to explore the possibilities of studying PD using smartphone data, and have brought in new avenues to remote PD assessment[9,15–24].

Despite these promises, remote PD assessment using smartphones is still in its infancy. **Table 1** gives an overview of recent PD studies using machine-learning approaches on smartphone features. Although existing methods and analyses have used different datasets with various sample sizes, the overview shows that, in general, studies have considered few and inconsistent feature modalities, and reported performance accuracy via varying statistical approaches. Additionally, most models were developed with a relatively limited scope that was either restricted to disease classification or disease severity estimation. Moreover, some studies only considered PD samples. Here, in light of existing efforts, we propose a unified machine-learning framework that (1) extracts disease- or symptom-specific features from a rich

variety of sensor data, (2) takes into account the differences between PD participants and HCs, (3) builds the selected features into a relevant feature map, (4) differentiates PD cases from HCs, and (5) estimates disease severity.

Recent studies	Sample size (PD/HC)	Number of repetitions*	Out clinic	Modalities considered							Accuracy	Ensemble† improvement
				Voice	Gait	Balance	Dexterity	Rest tremor	Postural tremor	Others		
Current study	37/35	4,883*	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	0.973/0.971 (Sen/Spec) 0.987 (Accuracy) 0.993 (AUC)	Yes
Prince *et al.* (2018)[15]	949/866	NA	Yes	No	No	No	Yes	No	No	No	0.65 (Accuracy)	No
Zhan *et al.* (2018)[17]	129/0	6,148	Yes	Yes	Yes	Yes	Yes	No	No	Reaction time	0.81 (Pearson correlation)	Yes
Prince *et al.* (2018)[18]	312/236	48,892	Yes	No	No	No	Yes	No	No	Memory	NA	No
Bot *et al.* (2016)[19]	1087/5581	78,887	Yes	Yes	Yes	No	Yes	No	No	Memory	NA	No
Zhan *et al.* (2016)[20]	121/105	1,600	Yes	Yes	Yes	Yes	Yes	No	No	Reaction time	0.693/0.727 (Sens/Spec)	Yes
Neto *et al.* (2017)[21]	23/23	NA	Yes	Yes	Yes	No	Yes	No	No	No	0.5-0.6 (AUC)	No
Arora *et al.* (2015)[22]	10/10	18	Yes	Yes	Yes	Yes	Yes	No	No	Reaction time	0.962/0.969 (Sens/Spec)	No
Lee *et al.* (2016)[23]	57/87	432	No	No	No	No	Yes	No	No	No	0.92 (AUC)	No
Arroyo-Gallego *et al.* (2017)[24]	21/23	51	No	No	No	No	Yes	No	No	No	0.810/0.810 (Sens/Spec)	No
Kassavetis *et al.* (2016)[9]	14/0	14	No	No	No	No	Yes	No	No	No	NA	No
Printy *et al.* (2014)[16]	18/0	54	No	No	No	No	Yes	No	No	No	NA	No

Table 1. An overview of PD studies using smartphone features. We selected twelve recent and representative studies that used smartphone data to study PD. We listed key characteristics, including the sample size, the total number of repetitions, whether the study was conducted outside of clinics, the types of test used, the estimation accuracy (if any), and ensemble improvement, whether the study was conducted outside of clinics is important because collecting measurements frequently in clinics is inconvenient for large-scale examination in practice. *The repetition means the total number of data points across all features and subjects. In other words, if the j^{th} $(1 \leq j \leq N)$ feature of subject i $(1 \leq i \leq P)$ was measured over T_{ij} days, the repetition is $\sum_{i=1}^{N}\sum_{j=1}^{P} T_{ij}$. †If ensemble improvement equals to yes, it means that using cross-modality features (*i.e.*, features obtained from different behaviors) improves the estimation accuracy. Shaded orange *vs.* grey color indicates if a study covers a specific component.

The framework first employs a two-step feature selection procedure and identifies features that are potentially associated with the disease (in terms of diagnostic group or severity). The selected features then enter the elastic-net regularized regression model to construct a feature map consisting of parameter estimates. Subsequently, the model links the feature map with features from the training subjects to estimate their diagnostic group status and severity. To evaluate the reproducibility of the framework, the model tests the feature map on features from

the novel (testing) subjects to perform out-of-sample PD assessment. The proposed framework is illustrated in **Fig.** 1. We arrange the rest of the article as follows. In **Section** 2, we introduce the smartphone data used in this study. In **Section** 3.1, we define notations and describe data organization. In **Section** 3.2, we provide the main methodological framework and its building blocks. **Section** 3.3 highlights the framework's applications in PD/HC classification and PD severity estimation. In **Section** 4, we present experimental and data analysis results. We discuss future work in **Section** 5 and conclude the article in **Section** 6.

2. The home-based PD data collected by smartphones

We use data collected from two independent smartphone99 based remote monitoring studies[25]. The first study was a six-month-long phase 1b clinical drug trial of PRX002/RG7935 (now known as prasinezumab) conducted by Prothena and Roche, which consisted of 44 PD participants (NCT02157714). The second study was a six-week-long observational study of 35 age- and sex-matched healthy controls (HCs). The respective local ethics committees approved both studies. Written informed consent was obtained from all participants (patient study: IRB00010809, H-35018, WOR1-14-143; control study: EKNZ-BASEC-2016-00596). All controls scored $\geq$ 26 points on the Montreal Cognitive Assessment (MoCA)[26] and were free of a cardiovascular, neurological, or psychiatric condition, and had no first-degree relative with PD. The study also included the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). MDS-UPDRS scores measure the progression of an individual's Parkinson's disease, and serve as the gold standard for validation[27]. Throughout, we used the MDS-UPDRS total scores. The total score equals the summation of sub-scores obtained from 42 items covering four sub-scales: Part I: mentation, behavior, and mood; Part II: activities of daily living; Part III: motor examination; Part IV: complications of therapy. The total score

ranges from 0 to 199 points, where a patient with a higher score would be considered to have more severe PD[28]. In this study, the scores were administered by trained raters (Parts I and II) and physicians (Part III) to subjects during screening (study days -42 to -1) and days 8 and 64. Trained raters tested controls at baseline and day 42.

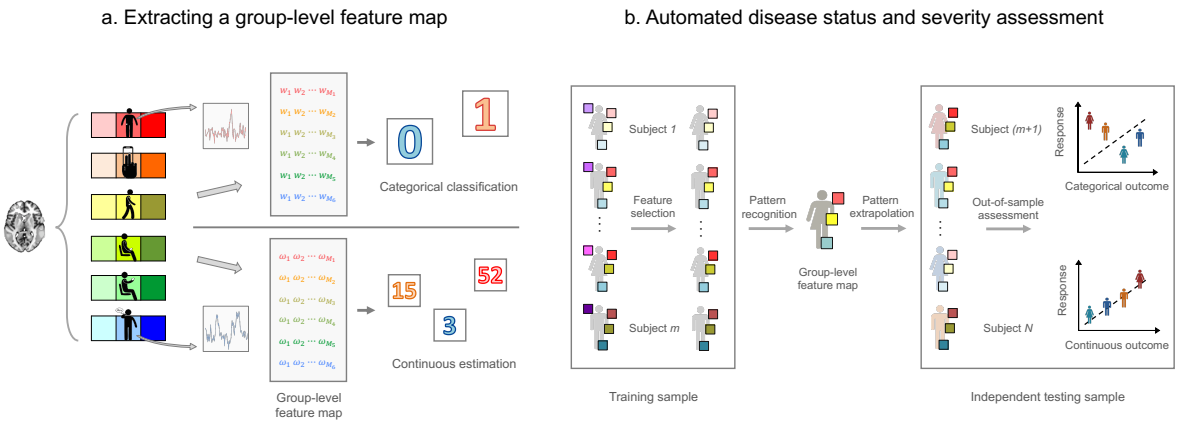


Fig. 1. (a) Extracting a group-level feature map. Each color specifies a feature modality. Boxes with the same color but with different hues indicate multiple behavioral features from the same feature modality. The red, orange, yellow, chartreuse green, and blue boxes refer to balance, dexterity, gait, rest tremor, postural tremor, and voice, respectively. The distinctive performance patterns on these tasks correspond to the functioning of specific functional-neuroanatomic circuits, which cannot be directly assessed (indicated by a gray bracket). The model couples behavioral features with a trained group-level feature map, yielding estimated outcomes. The colored Latin and Greek letters with subscripts (*e.g.*, w_1 and ω_1) represent feature weights across features. **(b) Automated disease group and severity assessment.** During the model building step, features that are relevant to the targeted outcome are selected. Subsequently, a feature map consisting of weights across selected features is developed using data from individuals in the training sample. The weights indicate how to integrate features to yield estimation for the targeted, discrete or continuous, outcome. During the prospective testing step, the efficacy of the feature map is verified by applying the map to features from previously unseen individuals without further model fitting, which yields estimations for each subject. The model produces one estimated outcome (binary disease group or continuous disease severity) per subject. The consistency of the features and the reproducibility of the model can then be evaluated by comparing the observed and estimated outcomes in the testing sample. For binary classification, we report statistics such as accuracy, kappa, sensitivity, and specificity. For continuous disease score estimation, we report RMSE and correlation between estimated and observed disease scores as measured by the MDS-UPDRS.

Both the PD and HC studies followed identical procedures. During the initial in-clinic visit, all subjects received a smartphone (Galaxy S3 mini; Samsung, Seoul, South Korea) with the Roche PD Mobile Application v1 (Roche, Basel, Switzerland) preinstalled. They also received a belt containing a pouch that carried the phone. Smartphones were “locked-down” (*i.e.*, configured so patients could only run the Roche PD Mobile Application v1 and WiFi connection software). Site staff provided the subjects training on the active tests. Subsequently, subjects

were instructed to complete the active tests at home once daily (in the morning), to carry the phone with them throughout the day, and to recharge the phone overnight. A full description of the study and data processing can be found Lipsmeier *et al.* (2018)[25].

3. Methods

3.1 Notations and data organization

We begin by defining the notations used throughout this article. To ensure that the estimation power is not influenced by the amount of data that was available to each individual, unless otherwise specified, we truncate the raw data such that every subject has data from the same number of days (17 in our study). A thorough treatment of missing data, such as imputation, is available elsewhere[29,30].

Let N denote the number of subjects in the study, where $N = 72$. The i^{th}subject, for $1 \leq i \leq N$, has T days, where $T = 17$. During each day, features from K modalities are measured, where $K = 6$ in the study. Each modality contains further features. Specifically, the k^{th} $(1 \leq k \leq K)$ modality contains M_k features, where M_k ranges from 37 to 178. The m^{th} feature of the kth modality, is measured at time points $1, 2, \ldots, T$, for the ith subject during the j^{th} day. Thus, each feature takes the form $x_{ikm}(t_j)$, for $1 \leq i \leq N$, $1 \leq k \leq K$, $1 \leq m \leq M_k$, and $1 \leq t_j \leq T$. Let $\sum_{k=1}^{K} M_k = P$, where $P = 437$ in the study. That is, there are a total of P features. Thus, the feature data $\boldsymbol{X}$ is a data cube of size $N \times P \times T$.

Similarly, we denote the outcome as $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, where y_i, for $1 \leq i \leq N$, is a categorical label in case of binary classification (*i.e.*, PD vs. HC) and a continuous value in case of PD severity estimation (*i.e.*, MDS-UPDRS total score).

To discover features useful for estimating an outcome, we first summarize each feature by its first moment (arithmetic mean). Formally, the first moment of the m^{th} feature from the k^{th} modality of the i^{th} subject is defined as $\xi_{ikm} = \frac{\sum_{t_j}^{\tilde{T}_i} x_{ikm}(t_j)}{\tilde{T}_i}$, where $\tilde{T}_i$ indicates the number of days during which features are averaged for each individual.

Throughout the article, we use the first-moment approach to summarize features for model building, because the mean conveniently provides the fundamental information of the features. In **Section** 5, we will discuss the advantages and limitations of using the mean to summarize the features.

3.2 Machine-learning framework

Our framework consists of two parts: (1) feature selection; (2) model building and automated disease assessment.

1) Feature Selection: For a $P \gg N$ problem (also known as the "short, fat data problem," where the number of features P is much larger than the number of samples N), there are commonly two difficulties. First, it suffers from the "curse of dimensionality," where the curse is twofold: (i) the number of samples (N) needed to yield a reliable statistical result grows exponentially as P grows, but a large number of samples is usually difficult to obtain; (ii) when P is large, all subjects appear to be dissimilar. This makes extracting common features (*e.g.*, features shared by PD participants or HCs) difficult. Second, there is an insufficient number of degrees of freedom to estimate the full model. To alleviate these challenges, we demonstrate that only a small number of the P features are needed to build the model (see **Section** 4.4 for

details). This assumption is supported, in part, by the strong correlation observed in the PD feature data, which we illustrate below in detail.

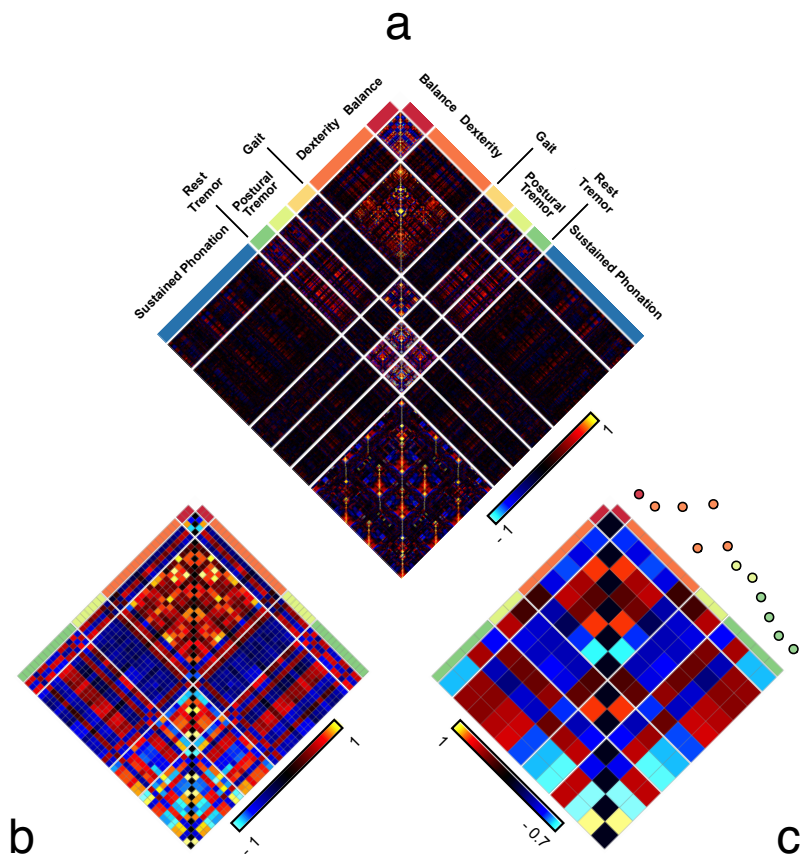


Fig. 2. An example of a two-step feature selection procedure. (a) A 437×437 matrix consisting of pair-wise Pearson correlations between 437 features obtained from six behavior modalities. The features are more correlated within their respective modality than they are with features out of that modality. (b) A 41×41 matrix consisting of pair-wise Pearson correlations between 41 features, selected from step-one feature selection. (c) A 12×12 matrix consisting of pair-wise Pearson correlations between 12 features, selected from step-two feature selection. The height of each colored circle above the heatmap indicates the weighted contribution (in the sense the higher the more important in terms of disease assessment) of each corresponding feature. Here, weighted means the features are scaled (mean 0, standard deviation 1) so that the magnitude of the weights is not biased by features with large means or variances.

Some of the PD sensor feature data are strongly correlated; the intra-modality features are more correlated than inter-modality features (see **Fig.** 2). When several highly correlated features are associated with an outcome, choosing one of them is analytically sufficient, and will give the most parsimonious model. The discarded (relevant) features, however, may uncover an underlying property that is meaningful for interpreting the biological system. For

example, consider two modalities, say voice and tremor, each with 100 features. Suppose there are 10 highly correlated voice features and 50 highly correlated tremor features that are associated with the disease outcome, and, for simplicity, suppose voice features are not highly correlated with tremor features. A model built on 1 (out of the 10 selected) voice feature and 1 (out of the 50 selected) tremor feature, therefore, is as sufficient as a model built using all selected features. The discarded features, however, may offer a better (and easier) biological explanation for the outcome. In this regard, it is important to consider a model that can account for both parsimony (*i.e.*, removing redundant features) and biological interpretation (*i.e.*, allowing a few correlated features).

To that end, we introduce a two-step feature selection procedure tailored for large-scale data. During the first step, we eliminate features that are not significantly related to the outcome (in the training data) using a mass-univariate approach.

For a continuous outcome (in our study the MDS-UPDRS total score), the first step involves a feature-wise correlation test to examine whether or not each feature is significantly correlated with the MDS-UPDRS total score, using a correlation test. Since the overall model incorporated an identity link function for continuous outcome assessment (coupled with the elastic-net, a regularized linear model), we used the Pearson correlation test to identify features that were linearly associated with the continuous outcome. Although the selected features are significantly correlated with the outcome to various degrees, they are not necessarily significantly correlated with each other (see **Fig.** 2). The heterogeneous groups of features, therefore, may each address a proportion of the variability of the outcome.

For a binary, or categorical, outcome (in our study the binary disease status), since a correlation test is inappropriate, the first step involves a feature-wise *t*-test to examine whether or not each feature varies significantly across groups.

During the second step, the selected features are further pruned via regularization. Common regularization approaches include the Lasso (least absolute shrinkage and selection operator) regularization[31] and the Ridge regularization (or the Tikhonov regularization)[32]. The Lasso picks one feature among all correlated ones, on which a single non-zero weight is imposed, whereas the Ridge imposes weights on all correlated features and then, averages their coefficients to reduce the effect of multiple correlated features on the full model.

2) Model Building and Automated Disease Assessment: Chief to automated disease assessment when the number of features (denoted as P) is much larger than the number of samples (denoted as N) is a modeling technique called regularization. A regularized model, such as Lasso and Ridge, shrinks the estimated parameters of irrelevant features (and therefore suggests either removing, or punishing the weights of, these features in the model output). The elastic-net regularization combines the Lasso and the Ridge regularizations and offers a compromise between them[33]. It chooses a small number of features (like the Lasso), some of which are correlated (like the Ridge), which may provide a useful biological interpretation of PD data. Because of its balance between interpretability and parsimony, we use the elastic-net during the second-step feature selection.

Consider a feature $\boldsymbol{\xi}$. Denote $\rho_{(\boldsymbol{\xi},\boldsymbol{y},N-2)}$ as the result from a statistical test during the step-one feature selection between the feature $\boldsymbol{\xi}$ and the outcome $\boldsymbol{y}$. The value of ρ can be a *t*-statistic

from a *t*-test for a binary outcome or a correlation for a continuous outcome; equivalently, it could be the corresponding *p*-values. Let ϵ be a pre-specified threshold for ρ. Although there is a one-to-one mapping between a statistic and its *p*-value, sometimes it may be convenient to evaluate the *p*-value, whereas other times it may be convenient to evaluate the *t*-statistic or the correlation. For example, in this study we threshold the *t*-statistic of binary *t*-tests at 5 during disease classification, and threshold the *p*-value of correlation tests at 0.01 during disease severity estimation.

Formally, we define our model as:

$$\mathbb{E}(y_i|\,\boldsymbol{\xi}_i, \boldsymbol{\delta}_i) = g^{-1}\big(\mu + \mathbf{f}_i^T\mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^T\boldsymbol{\gamma}\big) + \lambda_2|\mathbf{P}|_2^2 + \lambda_1|\mathbf{P}|_1 \qquad (1)$$

where $g(\cdot)$ is a link function, μ is the intercept, $\mathbf{f}_i^T = [\mathrm{f}_{i1}^T, \mathrm{f}_{i2}^T, \ldots, \mathrm{f}_{iK}^T]$, $\mathrm{f}_{ik}^T = (\xi_{ik1}, \xi_{ik2}, \ldots, \xi_{ikM_k})$. Recall that there are K total modalities, with the k^{th}modality containing M_k features. Here, $\mathbf{S} = blockdiag(\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_K)$, and $\mathbf{I}_k = diag\{i_{k1}, i_{k2}, \ldots, i_{kM_k}\}$, wherein $i_{km} = 1$ if $\rho_{(\boldsymbol{\xi}_{km}, \boldsymbol{y}, N-2)} < \varepsilon$, and 0 otherwise, where $\boldsymbol{\xi}_{km} = (\xi_{1km}, \xi_{2km}, \ldots, \xi_{Nkm})$ (a particular feature across all subjects); $\boldsymbol{\delta}_i$ is a vector containing all covariates for the i^{th} subject (in this study the covariates are age and gender), and $\boldsymbol{\gamma}$ is its coefficient; $\mathbf{P} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\gamma}]^T$, where $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kM_k})$, for $k = 1, 2, \ldots, K$. Finally, λ_1 and λ_2 are penalty parameters.

Through standard linear algebraic manipulation[33], the solution for **Eq.** (1) is

$$\widehat{\boldsymbol{\beta}} = \sqrt{1+\lambda_2}\left\{argmin_{\boldsymbol{\beta}^*}|\boldsymbol{y}^* - \boldsymbol{Z}^*\boldsymbol{\beta}^*|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}|\boldsymbol{\beta}^*|_1}\right\} \qquad (2)$$

where $\boldsymbol{y}_{n+p}^* = \left\{g\big(\mathbb{E}(y_1|\boldsymbol{\xi}_1, \boldsymbol{\delta}_1)\big), g\big(\mathbb{E}(y_2|\boldsymbol{\xi}_2, \boldsymbol{\delta}_2)\big), \ldots, g\left(\mathbb{E}\big(y_p|\boldsymbol{\xi}_p, \boldsymbol{\delta}_p\big)\right), \mathbf{0}_p\right\}^T$, and $\boldsymbol{Z}_{(n+p)\times p}^* = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathbf{F}^T\mathbf{S} \\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix}$.

The choice of λ_1 and λ_2 are determined in two steps: for each fixed λ_2, we find the optimal λ_1; subsequently we find the optimal λ_2 along the selected λ_1[33]. When $\lambda_1 = 0$, **Eq.** (2) reduces to the Lasso solution; when $\lambda_2 = 0$, **Eq.** (2) reduces to the Ridge (or the Tikhonov) solution. Any other (elastic) choices of λ_1 and λ_2 form a compromise between the Ridge and the Lasso regularizations. The compromise can be illustrated by rewriting the penalty terms in **Eq.** (1) as:

$$(1-\alpha)|\mathbf{P}|_2^2 + \alpha|\mathbf{P}|_1 \quad (3)$$

where α is called a mixing parameter[33], which controls how much "Lasso-ness" and "Ridge-ness" the regularization chooses. Specifically, when $\alpha = 0$, the regularization is strictly Ridge, and when $\alpha = 1$, the regularization is strictly Lasso.

(a) Results of PD/HC classification

	Multivariate logistic regression		XGBoost	Elastic-net			
				Raw biomarkers		Mean biomarkers	
	One-step	Two-step		One-step	Two-step	One-step	Two-step
Accuracy	0.493	0.813	0.944	0.947	0.947	0.960	0.973
Kappa	0.0035	0.625	0.889	0.894	0.894	0.920	0.947
Specificity	0.629	0.800	0.943	1.000	1.000	1.000	1.000
Sensitivity	0.375	0.825	0.946	0.900	0.900	0.925	0.950
Misclassified Subjects	38	14	4	4	4	3	2
Computing Time	2.46 mins	2.76 mins	< 1 min	42.26 mins	42.23 mins	3.09 mins	3.31 mins

(b) Results of PD severity assessment

		Multivariate logistic regression		XGBoost	Elastic-net			
					Raw biomarkers		Mean biomarkers	
		One-step	Two-step		One-step	Two-step	One-step	Two-step
Both PD and HC data	RMSE	503.21	553.53	23.21	602.31	48.59	23.88	16.58
	Pearson correlation	0.12	0.28	0.67***	-0.08	0.41***	0.57***	0.72***
	Computing time	<1 min	< 1 min	< 1 min	71 mins	14 mins	< 1 min	< 1 min
Only PD data	RMSE	1927.66	2827.81	30.81	567.55	71.16	27.66	17.19
	Pearson correlation	-0.14	0.30	0.11	-0.26	0.42**	0.16	0.54***
	Computing time	< 1 min	< 1 min	< 1 min	37 mins	4 min	< 1 min	< 1 min

Table 2. Results of PD/HC classification and disease severity assessment. We compared the performance of our framework with it using the baseline approaches. All results were cross-validated; RMSE refers to the root mean square error and ** and *** indicate that The Pearson correlations were significant at $p < 0.01$ and $p < 0.001$, respectively. The computing time was calculated using a Macintosh computer with a 2.4 GHz intel core i5 processor.

3.3 PD assessment

In the following, we apply the framework outlined in **Eq.** (1) in two specific scenarios: (i) PD/HC classification, and (ii) PD 303 severity estimation.

i) PD/HC Classification: When the outcomes are binary (*e.g.*, diseased vs. healthy), the link function in **Eq.** (1) is $g(x) = \ln\left(\frac{x}{1-x}\right)$ (*i.e.*, the inverse of logistic function). Formally,

$$\mathbb{P}(y_i = 1|\boldsymbol{\xi}_i, \boldsymbol{\delta}_i) = \frac{\exp\left(\mu + \mathbf{f}_i^T \mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^T \boldsymbol{\gamma}\right)}{1 + \exp\left(\mu + \mathbf{f}_i^T \mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^T \boldsymbol{\gamma}\right)} + \lambda_2|\mathbf{P}|_2^2 + \lambda_1|\mathbf{P}|_1 \quad (4)$$

where i refers to the i^{th} subject. The estimated conditional disease propensity, or $\mathbb{P}(y_i = 1|\boldsymbol{\xi}_i, \boldsymbol{\delta}_i)$, is further classified to be 1 if it is greater than 0.5, or 0 otherwise. The results are shown in **Section** 4.4.

ii) Estimation of PD severity: When the outcomes are continuous (*e.g.*, the MDS-UPDRS total scores), the link function in **Eq.** (1) is $g(x) = x$ (*i.e.*, an identity mapping). Formally,

$$\mathbb{E}(y_i = 1|\boldsymbol{\xi}_i, \boldsymbol{\delta}_i) = \mu + \mathbf{f}_i^T \mathbf{S}\boldsymbol{\beta} + \boldsymbol{\delta}_i^T \boldsymbol{\gamma} + \lambda_2|\mathbf{P}|_2^2 + \lambda_1|\mathbf{P}|_1 \quad (5)$$

where i refers to the i^{th} subject. The results are shown in **Section** 4.3.

4. Experiment and results

4.1 Cross-Validation Setup and Model's Parameters

To evaluate the performance of the framework, we split the data from N subjects described in **Section** 2 into four folds and conducted four-fold cross-validation. We used four statistics (accuracy, kappa, specificity, and sensitivity) to evaluate binary disease classification performance (*i.e.*, PD vs. HC); we used RMSE and correlation between observed and estimated outcomes to evaluate continuous PD severity assessment performance. The observed outcomes are individual mean (over time) MDS-UPDRS total scores and estimated outcomes are estimated mean MDS-UPDRS total scores.

We summarize the experimental set-up in **Algorithm** 1. The analyses were performed using *R* software via customized codes. The second step feature selection was conducted using the elastic-net regularization provided by *R* package *glmnet*[34]. Two parameters were tuned for the two-step feature selection and predictive framework: ϵ, a threshold used during the first step of feature selection, and α $(0 \leq \alpha \leq 1)$, a mixing parameter controlling how much Ridge-ness or Lasso-ness the elastic-net was. For disease classification, ϵ was used to threshold *t*-statistics and was set at 5; namely, a feature would be selected if its t-statistic was above 5. For disease severity estimation, ϵ was used to threshold *p*-values and was set at 0.01; namely, a feature would be selected if its *p*-value was below 0.01. We also provided the computing time needed to evaluate the model efficiency on the same computer (a standard Macintosh computer with a 2.4 GHz Intel Core i5 processor).

Algorithm 1: A generalized two-step feature selection and predictive framework for automated disease assessment.

Step 0: Reshape $\boldsymbol{X}$ to be of size $N \times P \times T$, where $P = \sum_{k=1}^{K} M_k$.

Step 1: For every feature m of modality k of subject i, compute the temporal mean ξ_{ikm}. We stack all subject's mean feature as an $N \times P$ matrix $\mathbf{F}^T$.

Step 2: Conduct the step-one feature selection to obtain estimate $\hat{\mathbf{S}}$ of $\mathbf{S}$ in **Eq.** (1). The selected features are then $\mathbf{F}^T\hat{\mathbf{S}}$.

Step 3: Conduct the step-two feature selection via (the elastic-net) regularization. The remaining features are those whose estimated parameters in **Eq.** (2) are non-zero.

Step 4: Run out-of-sample disease assessment using estimates from **Eq.** (1).

To demonstrate the efficacy of the proposed framework, we compared it to multivariate logistic regression (MLR) and XGBoost models in the same cross-validation strategy. To show the advantage of using mean features, we applied the proposed framework to the raw features (where repeated measurements of one feature are considered as multiple samples). We

recorded the accuracy statistics from each of the alternative approaches in the following section, with a discussion.

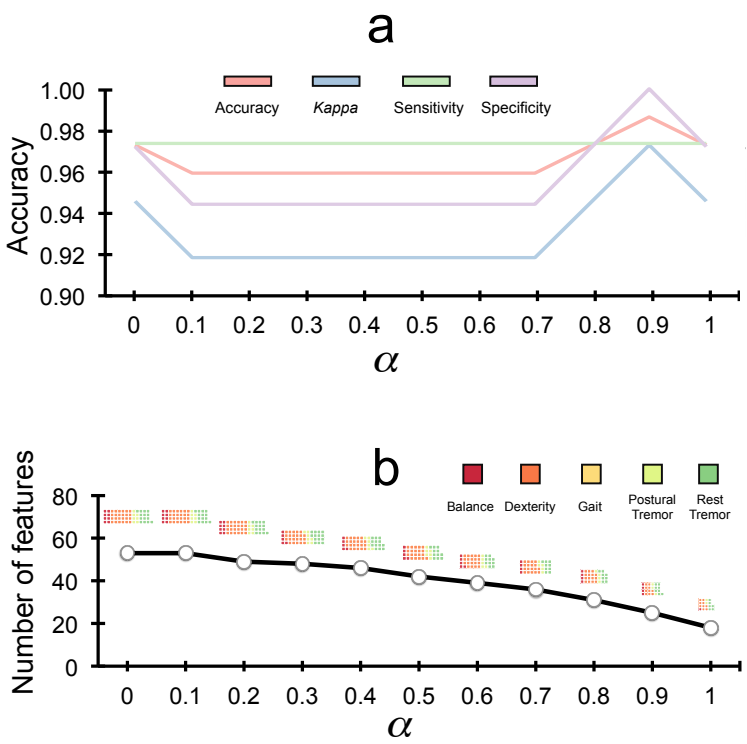


Fig. 3. Estimation accuracy and the number of features selected when the elastic-net mixing parameter α takes values from 0 to 1. (a) We examine four statistics (accuracy, kappa, sensitivity, and specificity) for evaluating model estimation accuracy. (b) When α increases, the number of selected features reduces. We use color to distinguish the disproportional representation of features selected from each modality. Of note, when $\alpha = 0$, the elastic-net regularization reduces to the Ridge regularization; when $\alpha = 1$, the elastic-net regularization reduces to the Lasso regularization.

4.2 Binary PD/HC classification results

In **Table** 2, as an initial step to understanding the machine learning framework we introduced in this article, we presented the model's performance on binary PD/HC classification using **Eq.** (4). There are three points to note. First, across multiple models, the two-step feature selection procedure yielded a higher estimation accuracy than a one-step feature selection procedure. Even with regularization, the two-step feature selection procedure still marginally improved accuracy and sensitivity. Second, using mean features significantly reduced computing time from 40 minutes (using raw features) to 3 minutes (using mean features), meanwhile improving

estimation accuracy mildly. Third, our framework outperformed the baseline MLR and XGBoost models in identifying PD participants and HCs (see **Table** 2 (a)).

The disease assessment accuracy and the number of selected features depend on the mixing parameter α in **Eq.** (3). Nevertheless, across α values, a majority of selected features belong to dexterity and rest tremor modalities. Specifically, when setting $\alpha = 0$ (*i.e.*, the Ridge), 38 out of the 53 final features are from dexterity and rest tremor modalities; when setting $\alpha = 1$ (*i.e.*, the Lasso), 18 out of the 25 final features are from them (see **Fig.** 3). The contribution each feature modality makes to the disease assessment is highlighted in **Fig.** 2(c), where dexterity features present the largest weight followed by rest tremor features. Taken together, our results suggest the importance of dexterity and rest tremor features in PD assessment.

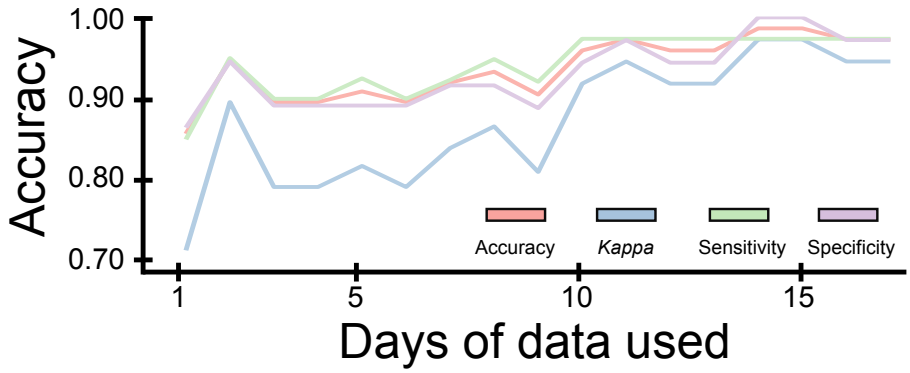


Fig. 4. Determining the minimal amount of data needed to build a reliable model. Each colored curve represents a function of how many days of data are averaged. The results suggest that accuracy improves and stabilizes once more than 10 non-contiguous days of data are used.

Although we showed that it is possible to identify PD participants from HCs using data sampled from 17 non-contiguous days with high accuracy, it remained unclear how many days of data are required to yield a stable estimate of the disease status. To check for minimal data requirement, we applied **Eq.** (4) to data obtained from an increasing number of days and

demonstrated that PD can be reliably identified using behavioral data sampled from 10 non-contiguous days (see **Fig.** 4).

In summary, our results suggest that (i) the two-step feature selection procedure generally outperforms more traditional approaches in classification accuracy; (ii) using mean features is computationally more efficient than using raw features; (iii) a model trained using behavioral data obtained from 10 non-contiguous days can reliably distinguish PD participants from HCs.

4.3 PD severity model results

We assessed continuous PD severity (*i.e.*, the MDS-UPDRS total scores) using **Eq.** (5) in two experiments. During the first experiment, we conducted a disease assessment using data from both PD participants and HCs. Note that (a) not all HCs' MDS-UPDRS total scores are 0; (b) all MDS-UPDRS total scores are non-negative, and a large MDS-UPDRS total score indicates the disease is more severe. Our analysis showed a Pearson correlation (between the observed and predicted scores) of 0.72 ($p < 0.0001$) **Fig.** 5. Assessing subjects' continuous MDS-UPDRS total score using individual features. (a) Estimation result of MDS-UPDRS total scores using data from both PD participants and HCs (N = 72). (b) Estimation result of MDS-UPDRS total scores using data from PD participants only (N = 37). Each dot represents one subject; the gray area represents a 95% confidence interval for the best-fit line. and an RMSE of 16.58 between the estimated and the observed MDS-UPDRS total scores (see **Fig.** 5(a) and **Table** 2). Although the estimated MDS-UPDRS total scores in the novel samples are significantly correlated with the observed scores with a small RMSE, it remained possible that the high estimation accuracy was merely driven by obvious differences in features between the PD participants and HCs. To check for this possibility, we sought to perform an additional

experiment. During the second experiment, we only used features from PD participants (N = 37). The analysis yielded a significant (although smaller) Pearson correlation of 0.54 ($p < 0.001$) and a similar RMSE of 17.19 (see **Fig.** 5(b) and **Table** 2).

For PD severity assessment, a two-step feature selection procedure (namely, a step-one feature selection via a pair-wise Pearson correlation test, followed by a step-two feature selection via the elastic-net regularization on selected features) yielded a higher estimation accuracy than a one-step feature selection procedure (namely, only using the elastic-net regularization) across multiple models. Additionally, using mean features significantly reduced computing time while improving estimation accuracy modestly. Finally, our framework outperformed the baseline MLR and XGBoost models in assessing continuous MDS-UPDRS total scores (see **Table** 2 (b)).

Taken together, for disease classification, the proposed framework selected features with heterogeneous profiles between PD participants and HCs. Further, for disease severity estimation, the framework reliably discovered features related strongly to continuous disease severity measures. Finally, the significant correlation between the estimated and observed MDS-UPDRS total scores shows that our framework is capable of identifying relevant features useful for disease severity assessment.

4.4 Selected features for diagnostic categorization and disease severity estimation

In the experiments, a small number of features were consistently and reliably selected across subjects and cross-validation iterations. For disease group classification, two dexterity features (one measures the variation in the number of touch events during a “right” tap, quantifying the variability of fine finger movement, and the other measures the time spent on-screen during

taps), two rest tremor features (one measures tremor frequency and the other measures tremor power in 3-8 Hz band) were consistently selected across the four different cross-validations, suggesting their reliability in distinguishing PD participants from HCs.

For MDS-UPDRS estimation, a dexterity feature (which indicates the variation in the number of touch events during a “right” tap, quantifying the variability of fine finger movement), a dexterity feature (which measures the distribution of time spent on-screen during a tap), a gait feature (which quantifies the variability between the execution of different step phases), and a rest tremor feature (which measures the ratio of power in 5-6 Hz to total, quantifying acceleration power) were consistently selected, suggesting their reliability in evaluating PD severity.

Although selected features came from balance, dexterity, postural tremor, and rest tremor modalities, our results suggested that dexterity features made the most contribution to disease estimation. In **Fig.** 2(c), 5 out of 12 selected features were dexterity-related, and their weights (which determined the contribution to outcome estimation) were higher than those of others. Rest tremor, postural tremor, and balance features had smaller, but significant, weights and accounted for 4, 2, and 1 of the 12 selected features, respectively.

In feature selection, there is a trade-off to be made between accuracy, interpretability, and model parsimony. When α increases, most of the accuracy statistics improve (see **Fig.** 3(a)). Meanwhile, as α increases, the number of selected features decreases (see **Fig.** 3(b)). A large α would yield a parsimonious model consisting of a small number of features. The elastic-net approach and the Lasso yield similar classification and estimation results. Typically, the Lasso arbitrarily picks a feature from a cluster of highly correlated features. The elastic-net, on the

other hand, retains feature clusters in the model (where features within each cluster have similar influence), thereby avoiding making random within-cluster feature selection. This property is particularly attractive when investigators are interested in extracting several clusters of features across modalities, and in examining how consistent they are across different models.

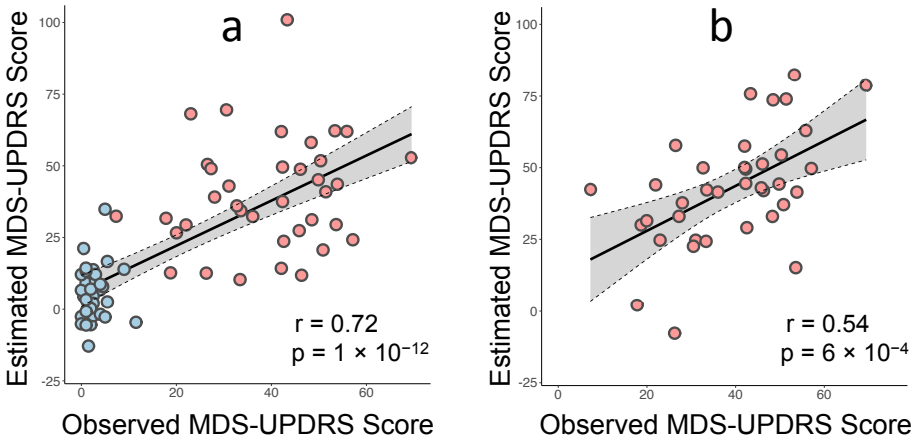


Fig. 5. Assessing subjects' continuous MDS-UPDRS total score using individual features. (a) Estimation result of MDS-UPDRS total scores using data from both PD participants and HCs ($N = 72$). (b) Estimation result of MDS-UPDRS total scores using data from PD participants only ($N = 37$). Each dot represents one subject; the gray area represents a 95% confidence interval for the best-fit line.

5. Discussion

In this article, we introduced a machine-learning framework for conducting automated analyses of PD symptoms using smartphone data. Our proposed framework yielded accurate and meaningful results. Under this framework, we identified a reliable and PD-specific feature profile among individuals, using a dataset consisting of 437 features across six modalities measured on 17 unevenly sampled non-contiguous days during a period of up to six months. For PD classification, we obtained accuracy, specificity, and sensitivity statistics of 0.972, 0.971, and 0.973, respectively; for PD severity assessment, our estimated MDS-UPDRS total scores were significantly correlated with the observed scores ($r = 0.72$, $p < 0.0001$ using both PD and HC data and $r = 0.54$, $p < 0.001$ using only PD data) and yielded small RMSEs (16.58 using

both PD and HC data and 17.19 using only PD data). The selected features were consistent with previous reports, where several distinct symptom domains, such as dexterity and rest tremor, present different patterns between PD participants and controls and are associated with disease outcomes[2–14,35]. Caution is warranted given the limited sample size, and that the *p*-values correlation tests could be affected by the sample size; our analyses, however, suggest that the introduced model and the identified features are promising to assess PD in out-of-sample individuals. Future studies with even larger sample sizes should independently verify the extent to which these features are optimal to assess PD.

We have introduced and used a two-step feature selection procedure in the proposed machine-learning framework. Compared to a one-step feature procedure (*i.e.*, using regularization alone), a two-step feature selection not only improves, but also balances, specificity and sensitivity (see **Table** 2).

For binary classification, we used a feature-wise *t*-test (which required a Gaussian assumption) during the first step feature selection (see **Section** 3.2). Naturally, one would ask, by doing so, was it likely to overlook biomarkers that were not normally distributed but may be useful for identifying patients? To evaluate this possibility, we performed an additional analysis considering a *U*-test (or the Mann–Whitney test, which did not require a Gaussian assumption)[36] during the first step feature selection, while keeping everything else the same. Our results showed that it yielded slightly worse classification performance (Accuracy= 0.96 or 72/75 and AUC=0.99) than it of the proposed method (Accuracy = 0.973 or 73/75 and AUC = 0.991). A possible explanation can be made using a bias-variance trade-off argument. We achieved better results using a *t*-test than using a *U*-test because (1) when the (Gaussian) assumption held, the *t*-test may be more powerful than (at least as powerful as) the *U*-test; (2)

in the case where the distributional assumption did not hold, we were trading variance with bias by using the *t*-test instead of the *U*-test. More precisely, by doing so we were more likely to incur wrong hypothesis testing results during the first step feature selection - although it was, in part, protected by using regularization in the second step feature selection - but the overall power was higher. To sum up, combining a *t*-test and regularization achieved higher bias (regarding training data, namely, it may not capture some regularity of the training data) and lower variance (in out-of-sample testing), and combining a *U*-test and regularization had lower bias (regarding training data, namely, it may better capture the regularities of the training data) and higher variance (in out-of-sample testing). As prediction was the main goal of automated disease assessment in mobile health, we chose to consider the combination of a *t*-test and regularization during the two-step feature selection procedure in disease classification.

We used the first moment to summarize features. The reasons for doing this were twofold. First, it is a simple, but effective, de-noising approach, evidenced by the experimental results in **Section** 4. The rationale is that the mean can capture intrinsic properties of features for disease assessment problems, such as PD/HC classification, as some mean features differ significantly between PD participants and HCs. Second, it is computationally efficient. Compared to using the raw data directly, the mean approach incurs far less computing time, and yields improved estimation accuracy (see **Table** 2).

We would like to note several limitations of our framework. First, it only captures linearly associated behavioral features. Future work using non-linear parametric approaches may be useful to uncover the non-linear feature architecture related to PD. Second, it may overestimate disease severity when responses are zero-inflated (*i.e.*, outcomes contain many entries equal or close to zero while other entries are very different from zero). Future work using the Zero-

Inflated Poisson (ZIP) model, and mixed-effect models may be useful to address this issue. Finally, our framework does not capture the features' temporal dynamics. Future work considering sliding-window analysis, generalized estimation equation (GEE) and generalized linear models (GLMs) with repeated measures[37,38], and functional principal component analysis (fPCA)[39,40] may be useful to unveil semi-continuous (windowed) and continuous temporal dynamic of the disease profile.

There are also a few challenges related to digital health in general and smartphone-based disease assessment and prediction in particular. I will list a few key ones in the following, along with their potential solutions. First, of major concern are data privacy[41] and quality[42]. In terms of data privacy, how can one ensure that smartphones are capturing, transferring, storing, and analyzing personal data securely? In terms of data quality, do collected data contain useful signal (versus noise); are they biologically and clinically meaningful; do the analyses and conclusions overfit the data and overstate the effect? For the former line of questions, mitigation is already underway through improving government policies[43], company regulations[44], and (computational and storage) technology[45]. For the latter, integrating protective measures, introducing rigorous quality controls, developing efficient, targeted algorithms, and performing comprehensive out-of-sample or out-of-population cross-validation and testing, are critical.

The next set of challenges concerns accuracy, flexibility, and robustness[46]. A human doctor has the advantage of interacting directly with a patient, including observing complex body language and emotion, to make personalized medical judgments. Smartphones have the advantage of objectivity and sensitivity, but they generally lack the ability to directly detect and integrate emotional information and personal feedback. Improved, more holistic smartphone

applications and algorithms that can incorporate dynamic, interactive user feedback, such as emotions and verbal input, are important, yet scarce, in literature or are under development in relative isolation[47–52].

Related to digital accuracy are ethics and the ability to handle subject-level, group-level, and stochastic variability (or noise) appropriately. Regarding data ethics[53,54], how can we ensure results derived from smartphone data are not biased and do not disadvantage a particular group? The solution, we think, needs to start at the foundation, including collecting representative samples, and building comprehensive modelling pipelines. By collecting representative samples, we refer to considering subjects that reflect the target population in as many dimensions as is practicable, including the distribution of sexes, ethnicities, ages, technologies, instruments, *etc*. If the goal is to study a specific disease, one needs to carefully coordinate the ratio between healthy and diseased subjects. By building comprehensive modelling pipelines, we mean to develop powerful methodological frameworks that comprise rigorous cross-validation and testing mechanisms to ensure that the boundaries of the models are well understood, and that the models are reproducible for the intended use. Regarding variability or noise, a human doctor is very good at filtering noise and extracting essential information generated by traditional medical devices in a controlled environment. Smartphones we carry around will, inevitably, be affected by environmental noise. This needs to be carefully accounted for to maintain sensitivity, specificity, and accuracy. Unfortunately, it is at present difficult for a smartphone to make accurate detections in a live environment, for example, to distinguish hand tremor from hand movement during a walk or to identify subtle changes in voice in live settings[55,56]. Thus, it is important to develop next-generation smartphones that can perform robust health monitoring, assessment, and diagnosis in a live, potentially noisy environment.

Smartphone-based disease assessment and prediction are at present still domain-specific, as those performed by most digital health devices do[46,57–60]. In other words, each device is specialized to perform one or a small number of tasks. This is acceptable, if one's goal is to gain insight into a particular illness – similar to the situation where one goes to see a specialized human doctor. Yet, for the majority who wish to perform comprehensive self-checks and -monitoring at home, with a goal to detect early signs of disease without *a priori* knowledge about which disease(s) may develop, a domain-specific device becomes less convenient and effective. For this purpose, a comprehensive battery of bio-signals would need to be analyzed. The analysis, ideally, needs to be performed on a single smartphone. For example, to give a relatively comprehensive picture of one's daily metabolism, the device needs to not only integrate multivariate cardiovascular digital biomarker data, hormonal, and metabolic data, but also consider a modelling architecture that enables interoperability and interpretability in combination[61,62]. Although the concept of combining datasets and algorithms may seem straightforward, given the fragmentation of healthcare and the lack of comprehensive global data standards, it will likely face major challenges, such as complex semantic integration, meta-analysis, and model integration[61,63–65], which need to be addressed before moving forward. Beginnings, however, are already being made, evident in the digital ward with the integration of multiple wearable sensors[66], and attention drawn to the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles[67] in relation to health data. The FAIR data principles, coupled with ensemble learning[61], potentially pave the way for interoperability between different smartphones, algorithms, datasets, and manufacturers.

While lasting rapport and trust can be established between a family doctor and a patient, it is difficult at present to form any personal relationship between smartphones and patients. Future smartphones may consider incorporating more sophisticated and generalized algorithms that

can ask and answer, with a degree of empathy, more targeted and relevant questions with more natural and human-like qualities.

Another issue is the contrast between the usefulness and the (low) acceptance of smartphones among the aging population. On the one hand, the aging population will benefit tremendously from digital health, receiving semi-continuous, home-based, and automated digital assessment and monitoring[68,69]. On the other hand, there is a stark gap between the younger and the older generations in terms of smartphone usage[70,71]. The low penetration in the aging population is, in part, due to the fact that smartphones were only introduced relatively recently, especially in emerging and developing countries: many did not grow up with such devices and are unlikely to adopt them as they age. The relative complexity of such devices and applications also makes it challenging for older generations (compared to the “digital natives”) to adapt and incorporate them into their lives. In the long run, the low acceptance of smartphones is likely to be mitigated, as the current “digital generation” grows older. Yet, in the present and near future, to enable the elderly and others with cognitive impairment to benefit from smartphone-based health solutions, developing ease-of-use smartphone devices and apps remains as important as making functionally powerful ones.

Last, but not least, is the talent challenge - the rarity of professionals and organizations with integrated interdisciplinary knowledge. At present, a doctor may need to consult with data scientists to interpret the output of automated smartphone algorithms. Equally, an engineer with state-of-the-art know-hows may rely on domain experts to build biologically and medically relevant applications. Disciplinary experts together with multi-disciplinary teams are essential in developing and improving smartphone-based disease assessment and prediction today and will continue to be an integral, indispensable part of the digital age. Nevertheless, a new

generation of professionals with interdisciplinary training in life science, medicine, public health, and the computational domains will be in high demand in the coming years. Positioned at the nexus of disciplines, they will be the integrators in complex teams, bridging multiplex interactions, multi-domain operations, and disciplinary experts.

6. Conclusion

Throughout this article, we have introduced and demonstrated an automated disease assessment framework that has the potential for remote PD classification and PD severity estimation at home using smartphones. Using this framework, we showed the presence of a disease-specific feature profile across multiple behavioral modalities.

Selected features reflected individual-specific traits that were unique and generally reliable among subjects across 17 unevenly sampled non-contiguous days during a period of up to six months. Our results showed that the proposed framework is possible, with relatively high accuracy, to identify PD participants solely using sensor feature data from their remote, smartphone-based behavioral measurements. In addition, data analyses using this framework revealed that individual variability in extracted features (in particular, dexterity, rest tremor, and gait) were informative for the continuous disease severity estimation in an independent group of participants.

In conclusion, we proposed a machine-learning framework for automated disease assessment and provided preliminary evidence for a PD-specific behavioral architecture, in this case the extracted features, that may be associated with PD. Extensive data analyses suggest that this framework has the potential for identifying behavioral signatures to advance automated and remote assessment of PD.

References

1. de Lau, L. M. & Breteler, M. M. Epidemiology of Parkinson's disease. *Lancet Neurology* **5**, pp. 525–535 (2006).
2. Benba, A., Jilbab, A. & Hammouch, A. Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines. *Int. J. Electr. Eng. Informatics* **7**, pp. 297–307 (2015).
3. Kapoor, T. & Sharma, R. K. Parkinsons disease diagnosis using Mel frequency cepstral coefficients and vector auantization. *Int. J. Comput. Appl.* **14**, pp. 43–46 (2011).
4. Timmer, J., Gantert, C., Deuschl, G. & Honerkamp, J. Characteristics of hand tremor time series. *Biol. Cybern.* 70, pp. 75–80 (1993).
5. Deuschl, G. *et al.* The pathophysiology of parkinsonian tremor: A review. *J. Neurol.* **247**, pp. V33–V48 (2000).
6. Deuschl, G., Raethjen, J., Lindemann, M. & Krack, P. The pathophysiology of tremor. *Muscle and Nerve* **24**, pp. 716–735 (2001).
7. Vaillancourt, D. E. & Newell, K. M. The dynamics of resting and postural tremor in Parkinson's disease. *Clin. Neurophysiol.* **111**, pp. 2046–2056 (2000).
8. Salarian, A. *et al.* Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system. *IEEE Trans. Biomed. Eng.* **54**, pp. 313–322 (2007).
9. Kassavetis, P. *et al.* Developing a tool for remote digital assessment of Parkinson's disease. *Mov. Disord. Clin. Pract.* **3**, pp. 59–64 (2016).
10. Tavares, A. L. T. *et al.* Quantitative measurements of alternating finger tapping in Parkinson's disease correlate with UPDRS motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation. *Mov. Disord.* **20**, pp. 1286–1298 (2005).
11. Mayagoitia, R. E., Nene, A. V. & Veltink, P. H. Accelerometer and rate gyroscope measurement of kinematics: An inexpensive alternative to optical motion analysis systems. *J. Biomech.* **35**, pp. 537–542 (2002).
12. Horak, F. B. & Mancini, M. Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors. *Movement Disorders* **28**, pp. 1544–1551 (2013).
13. Mancini, M. *et al.* Postural sway as a marker of progression in Parkinson's disease: A pilot longitudinal study. *Gait Posture* **36**, pp. 471– 476 (2012).
14. Cheng, W. Y. *et al.* Human activity recognition from sensor-nased large-scale continuous monitoring of parkinson's disease patients. In *Proc. IEEE/ACM Int. Conf. Connected Health: Appl. Syst. Eng. Technol.*, Philadelphia, PA, USA, pp. 249–250 (2017).
15. Prince, J. & de Vos, M. A deep learning framework for the remote detection of Parkinson's sisease using smartphone sensor data. In *Proc. 40th Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc.*, Honolulu, HI, USA, pp. 3144-3147 (2018).
16. Printy, B. P. *et al.* Smartphone application for classification of motor impairment severity in Parkinson's disease. In *Proc. 36th Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc.*, Chicago, IL, USA, pp. 2686–2689 (2014).
17. Zhan, A. *et al.* Using smartphones and machine learning to quantify Parkinson disease severity the mobile Parkinson disease score. *JAMA Neurol.* **75**, pp. 876–880 (2018).
18. Prince, J., Arora, S. & De Vos, M. Big data in Parkinson's disease: Using smartphones to remotely detect longitudinal disease phenotypes. *Physiol. Meas.* **39**, 044005. (2018).

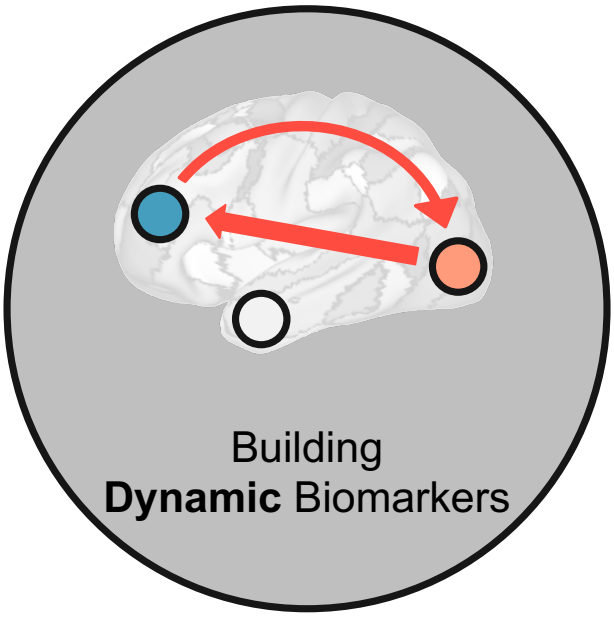
19. Bot, B. M. *et al.* The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* **3**, 160011 (2016).
20. Zhan, A. *et al.* High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection. *arXiv Prepr.* **arXiv:1601**, (2016).
21. Neto, E. C. *et al.* On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: The mpower case study. *arXiv Prepr.* **arXiv:1706** (2017).
22. Arora, S. *et al.* Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. *Park. Relat. Disord.* **21**, pp. 650–653 (2015).
23. Lee, C. Y. *et al.* A validation study of a smartphone-based finger tapping application for quantitative assessment of bradykinesia in Parkinson's disease. *PLoS One* **11**, e0158852 (2016).
24. Arroyo-Gallego, T. *et al.* Detection of motor impairment in Parkinson's disease via mobile touchscreen typing. *IEEE Trans. Biomed. Eng.* **64**, pp. 1994–2002 (2017).
25. Lipsmeier, F. *et al.* Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov. Disord.* **33**, pp. 1287–1297 (2018).
26. Nasreddine, Z. S. *et al.* The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, pp. 695–699 (2005).
27. Goetz, C. G. *et al.* Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord.* **23**, pp. 2129–2170 (2008).
28. Bhidayasiri, R. & Martinez-Martin, P. Clinical assessments in Parkinson's disease: Scales and monitoring. In *International Review of Neurobiology*. New York, NY, USA: Elsevier **132**, pp. 129–182 (2017).
29. Little, R. J. a & Rubin, D. B. *Statistical Analysis With Missing Data.* Hoboken, NJ, USA: Wiley (2002).
30. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley (2004).
31. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, pp. 267–288 (1996).
32. Tikhonov, A. N. Solution of incorrectly formulated problems and the regularization method. *Dokl. Akad. Nauk SSSR* **151**, 501–504 (1963).
33. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, pp. 301–320 (2005).
34. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, pp. 1–20 (2010).
35. Salarian, A. *et al.* Analyzing 180° turns using an inertial system reveals early signs of progression of Parkinson's disease. In *Proc. Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc.*, Minneapolis, MN, USA, pp. 224– 227 (2009).
36. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, pp. 50–60 (1947).
37. Zeger, S. L., Liang, K.-Y. & Albert, P. S. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* 44, pp. 1049–1060 (1988).
38. Moulton, L. H. & Zeger, S. L. Analyzing repeated measures on generalized linear models via the Bootstrap. *Biometrics* **45**, pp. 381–394 (1989).
39. Yao, F., Müller, H. G. & Wang, J. L. Functional linear regression analysis for longitudinal data. *Ann. Stat.* **33**, pp. 2873–2903 (2005).

40. He, G., Müller, H. G. & Wang, J. L. Extending correlation and regression from multivariate to functional data. In *Asymptotics in Statistics and Probability*, edited by M. Puri, pp. 197–210. Zeist, The Netherlands: VSP (2019).
41. Luxton, D. D., Kayl, R. A. & Mishkind, M. C. mHealth data security: The need for HIPAA-compliant standardization. *Telemed. e-Health* 18, pp. 284–288 (2012).
42. Fadahunsi, K. P. *et al.* Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth. *BMJ Open* **9**, e024722 (2019).
43. Benjamin, K. & Potts, H. W. Digital transformation in government: Lessons for digital health? *Digit Heal.* **3**, pp. 1–5 (2018).
44. Wolf, M., Semm, A. & Erfurth, C. Digital transformation in companies – challenges and success factors. In *International Conference on Innovations for Community Services*, Žilina, Slovakia, pp. 178-193 (2018).
45. Bhavnani, S. P., Narula, J. & Sengupta, P. P. Mobile technology and the digitization of healthcare. *European Heart Journal* **37**, pp. 1428–1438 (2016).
46. Agrawal, R. & Prabakaran, S. Big data in digital healthcare: Lessons learnt and recommendations for general practice. *Heredity* **124**, pp. 525–534 (2020).
47. Albanie, S., Nagrani, A., Vedaldi, A. & Zisserman, A. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, Seoul, Republic of Korea, pp. 292–301 (2018).
48. Avots, E., Sapiński, T., Bachmann, M. & Kamińska, D. Audiovisual emotion recognition in wild. In *Machine Vision and Applications* **30**, pp. 975–985 (2019).
49. Martínez-Miranda, J. & Aldea, A. Emotions in human and artificial intelligence. *Comput. Human Behav.* **21**, pp. 323–341 (2005).
50. Conati, C., Chabbal, R. & Maclaren, H. A study on using biometric sensors for monitoring user emotions in educational games. (2003). Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.3171
51. Butow, P. & Hoque, E. Using artificial intelligence to analyse and teach communication in healthcare. *Breast* **50**, pp. 49–55 (2020).
52. Laird, J. E. & Van Lent, M. Human-level AI's killer application interactive computer games. *AI Mag.* **22**, pp.15-25 (2001).
53. Richards, N. M. & King, J. H. Big data ethics. *Wake Forest Law Rev.* **49**, 393–432 (2014).
54. Fan, J., Han, F. & Liu, H. Challenges of big data analysis. *National Science Review* **1**, 293–314 (2014).
55. Musliner, D. J. *et al.* The challenges of real-time AI. *Computer* **28**, 58–66 (1995).
56. Stoica, I. *et al.* A Berkeley view of systems challenges for AI. *arXiv Prepr.* **arXiv:1712**, (2017).
57. Widmer, R. J. *et al.* Digital health interventions for the prevention of cardiovascular disease: A systematic review and meta-analysis. *Mayo Clin. Proc.* **90**, pp. 469–480 (2015).
58. Chén, O. Y. *et al.* Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans Biomed Eng* **67**, pp. 3491–3500 (2020).
59. Li, X. *et al.* Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biol.* **15**, e2001402 (2017).
60. Kim, J. A. *Creative destruction of medicine: How the digital revolution will create better health care*. Basic Books, New York, New York, USA (2013).
61. Polikar, R. Ensemble learning. In *Ensemble Machine Learning: Methods and Applications*, Springer, Berlin, Germany (2012).

62. Roberts, B. Vision of a digital health ecosystem. *LinkedIn* (2019). Available at: https://www.linkedin.com/pulse/vision-digital-health-ecosystem-bryn-roberts/.
63. LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J. & Derderian, F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N. Engl. J. Med.* **337**, pp. 536–542 (1997).
64. Slavin, R. E. Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educ. Res.* **15**, pp. 5–11 (1986).
65. Hunter, J. E., Schmidt, F. L. & Jackson, G. B. Meta-analysis: cumulating research findings across studies. *Educ. Res.* **15**, pp. 20–21 (1986).
66. Institute of Biomedical Engineering, University of Oxford. Using wearable technology to monitor COVID-19 patients. Available at: http://www.ibme.ox.ac.uk/news-events/news/using-wearable-technology-to-monitor-covid-19-patients#:~:text=A%20vHDU%20is%20a%20general,computers%20and%20smart%20alerting%20algorithms (2020).
67. Wilkinson, M. D. *et al.* Comment: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, pp. 1–9 (2016).
68. Lupton, D. The digitally engaged patient: Self-monitoring and self-care in the digital health era. *Soc. Theory Heal.* **11**, pp. 256–270 (2013).
69. Bhattarai, P. & Phillips, J. L. The role of digital health technologies in management of pain in older people: An integrative review. *Archives of Gerontology and Geriatrics* **68**, pp. 14–24 (2017).
70. Poushter, J. *Smartphone ownership and internet usage continues to climb in emerging economies*. Pew Research Center (2016). Available at: https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/
71. Berenguer, A. *et al.* Are smartphones ubiquitous?: An in-depth survey of smartphone adoption by seniors. *IEEE Consum. Electron. Mag.* 6, pp.1 04–110 (2017).

Chapter 3

Building dynamic biomarkers



Resting-state brain information flow predicts cognitive flexibility in humans[1]

Abstract

The human brain is a dynamic system, where communication between spatially distinct areas facilitates complex cognitive functions and behaviours. How information transfers between brain regions and how it gives rise to human cognition, however, are unclear. In this article, using resting-state functional magnetic resonance imaging (fMRI) data from 783 healthy adults in the Human Connectome Project (HCP) dataset, we map the brain's directed information flow architecture through a Granger-Geweke causality prism. We demonstrate that the information flow profiles in the general population primarily involve local exchanges within specialized functional systems, long-distance exchanges from the dorsal brain to the ventral brain, and top-down exchanges from the higher-order systems to the primary systems. Using an information flow map discovered from 550 subjects, the individual directed information flow profiles can significantly predict cognitive flexibility scores in 233 novel individuals. Our results provide evidence for directed information network architecture in the cerebral cortex and suggest that features of the information flow configuration during rest underpin cognitive ability in humans.

[1]This chapter is published as: Chén, O.Y., Cao, H., Reinen, J.M., Qian, T., Gou, J., Phan, H., de Vos, M. and Cannon, T.D., 2019. Resting-state brain information flow predicts cognitive flexibility in humans. *Scientific Reports* **9**, pp. 1-16.

1. Introduction

The human brain is a dynamic system whose function is built upon communications between spatially distinct areas. Although traditional approaches using undirected functional connectivity have advanced our understanding of the functional architecture of the brain and its relationship to cognition and behaviour[1,2], the directional pattern of information flow across the brain and its relationship to human behaviour are largely unknown.

Answering these questions requires another approach, so-called “effective connectivity”, which uses functional imaging data to estimate causal relationships between separated brain regions and thus quantifies information flow directions[3,4]. Previous studies using data from magnetoencephalography (MEG)[5] have shown that brain information flow during rest is not random, but follows a posterior-to-anterior flow in high-frequency bands and an anterior-to-posterior flow in low-frequency bands. Data from electroencephalography (EEG)[6,7] and functional magnetic resonance imaging (fMRI)[8,9] have also demonstrated that the information flow in the brain is associated with the embedded network topology, such that the flow is more likely to occur from low-degree regions to high-degree regions. These results suggest that information flow in the brain possibly relates to the anatomical and functional bases of the brain architecture. Nevertheless, the exact pattern of information transfer at the regional and systems level is still unclear.

The human brain is efficiently organized to facilitate information exchange in order to support higher cognitive functions[10,11]. Consequently, the underlying information flow should subserve human cognition. Prior work indicates that executive functioning — the ability to modulate behaviour to achieve a certain goal — strongly relates to the strength of information flow in the brain. For instance, effective connectivity patterns of subregions in

the frontoparietal network could distinguish distinct domains of executive functions[12,13]. Research using effective connectivity approaches has also linked cognitive control to information flow in a top-down network from the rostral and caudal prefrontal cortex to premotor regions[14]. Further, alterations in cortical effective connectivity have been widely reported in patients with psychiatric disorders that are characterized by executive functional deficits, such as attention-deficit hyperactivity disorder (ADHD)[15] and obsessive compulsive disorder (OCD)[16]. These lines of evidence point to a potential link between the strength of information flow and executive ability in humans.

Using resting-state fMRI data acquired from the Human Connectome Project (HCP) in 783 healthy adults, we investigated the pattern and cognitive correlates of whole-brain information flow under a modified Granger-Geweke causality framework that accounts for the subject- and edge-specific lags in brain effective connectivity. We first obtained the directionality, strength and variability of information flow across the entire brain using a predefined atlas of 268 regions[1,17] in a subsample of 550 healthy subjects (70% of the total sample). We then trained a regression model to identify edges correlated with measures of cognitive flexibility and employed these edges to predict cognitive flexibility scores in an independent sample of 233 healthy subjects (30% of the total sample). We hypothesized that (1) heteromodal cortical regions may show higher information flow strength and variability as well as differential information directions compared with unimodal and subcortical regions; and (2) information flow profiles associated with higher between-subject variability are predictive of executive ability in humans.

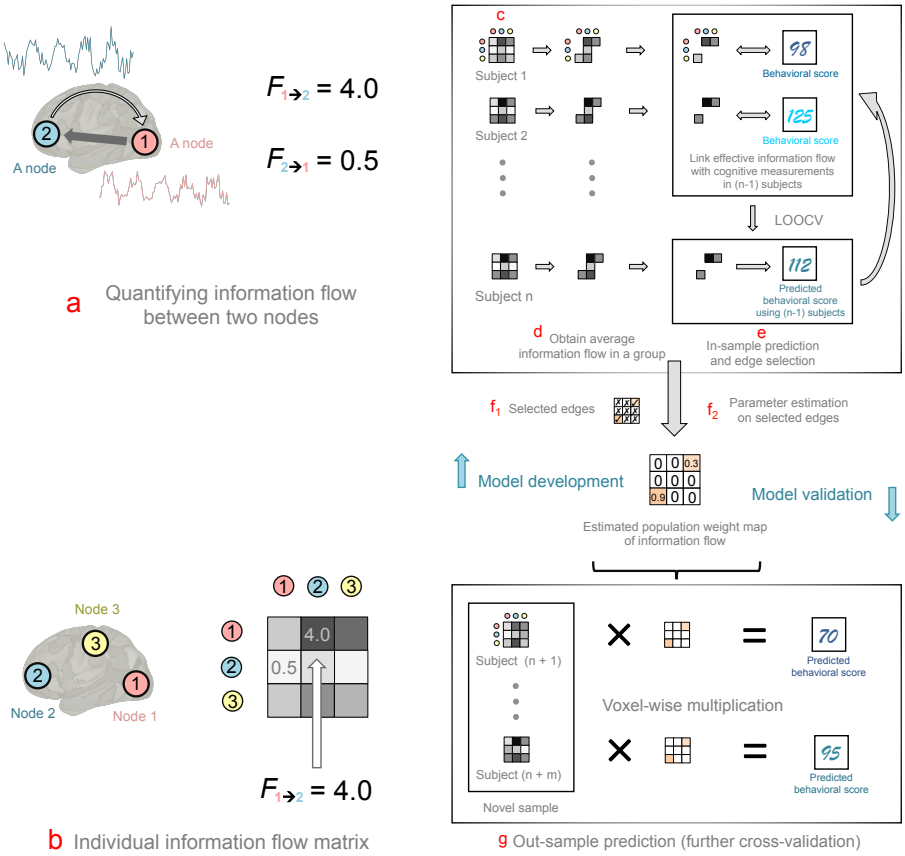


Fig. 1. A flow chart for extracting population information map and conducting out-of-sample prediction for cognitive measurement using information flow in the human brain. (a) Obtaining information flow metrics (*F* values) between every pair of regions. For time courses from brain regions 1 and 2, we obtained two directed values $F_{1\to2}$ (information flow from region 1 to 2) and $F_{2\to1}$ (information flow from region 2 to 1). (b) Arranging *F* values as an information flow matrix. Specifically, *F* values obtained from (a) were arranged corresponding to brain regions in an asymmetrical matrix. (c) Obtaining individual information flow matrix. For every subject, we computed the *F* values for every pair of time courses. This yielded a 268 × 268 subject-specific asymmetrical *F* value matrix and a 268 × 268 asymmetrical *p*-value matrix (corresponding to the *F* value matrix) for each subject. (d) Obtaining average information flow in a group. We first computed the average *p*-values for edge across subjects. We recorded edges with corresponding *F* values that had an average *p*-value less than a threshold (*e.g.,* 0.05). (e) Edge (*F* value) selection. We used a two-step feature selection procedure. During the first step, we selected an *F* value if its average *p*-values across subjects were below a threshold (see **Methods** for details). We then used a leave-one-subject-out cross-validation (LOOCV) analysis to conduct a further feature selection on *F* values and build a predictive model. The training and testing in LOOCV were performed iteratively for *n* times. (f_1) Due to the large variability of LOOCV, we took the union of the selected features from each LOOCV iteration. (f_2) Using selected *F* values, we built a feature weight map of information flow. (g) Out-of-sample prediction. We multiplied the information weight map with *F* values from previously unseen subjects, without further model fitting, to predict their cognitive scores. The efficacy and predictive power of information flow in predicting cognitive flexibility were evaluated by correlating the predicted and observed cognitive measurements.

2. Results

2.1 Individual-level whole brain information flow map

We used eyes-open resting-state scans of 783 subjects from the HCP 1200 data release (age 22-36 years, 383 males and 400 females). The HCP 1200 data release had 1096

subjects in total. We excluded 313 subjects due to loss of time points and/or high head motion during the scans (see **Methods**). Each subject was scanned for two resting sessions (REST 1 and REST 2) over two days. During each session, data were collected using both the left-right (LR) and right-left (RL) phase-encoding runs. For each subject, we concatenated the data (REST 1 LR, REST 1 RL, REST 2 LR, REST 2 RL) into a 3,456 seconds time-course (containing 4800 time points, with TR = 720ms). In this study, the information flow strength between two brain regions was quantified by the directed Geweke *F* values[18] between their time courses (see **Methods**). The directed Geweke *F* value, or $F_{i\to j}$, is a feedback measure, which quantifies the Wiener-Granger[19-21] causal effect of a time series *i* on another time series *j*. A large $F_{i\to j}$ indicates that, using past information of both *i* and *j* better predicts future values of *j*, than using only past information of *j*. This is called *i* causes *j* in the sense of Wiener-Granger causality[18]. In previous studies, the *F* values were used to measure the "strength of causality" between two time series[18,20]. Here we extended this concept by introducing $F^k_{i\to j}(l)$ as an information flow metric in the brain. Specifically, $F^k_{i\to j}(l)$ quantified subject-specific (indexed by subject *k*) directed connectivity strength between two brain regions (denoted by $i \to j$) with lag-length (indexed by l). The optimal lag l between two regions was determined by the Akaike information criterion (AIC)[22], which essentially balanced the trade-off between the goodness of fit and the parsimony of the prediction model. The *F* value varies in brain space (denoted by different *i's* and *j's*) and across subjects (specified by different *k's*). A large F value indicates potential strong effective connectivity between two brain regions (see **Fig.** 1 (a)).

Using this approach, we constructed a whole-brain information flow map, based on a functional brain atlas of 268 nodes across the whole brain[17]. For each subject, we performed Granger-causality analysis (GCA) on time courses recorded from every pair of nodes in the atlas. This yielded a 268 × 268 *asymmetrical* information flow matrix for each

subject. Each entry of the information matrix represented a directed edge between two nodes. For example, the $(i, j)^{th}$ entry of the matrix represents the strength of the information flow from node *i* to *j*; similarly, the $(j, i)^{th}$ entry denotes the strength of the information flow from node *j* to *i* (see **Fig.** 1 (b)). Additionally, GCA yielded one *p*-value for every *F* value.

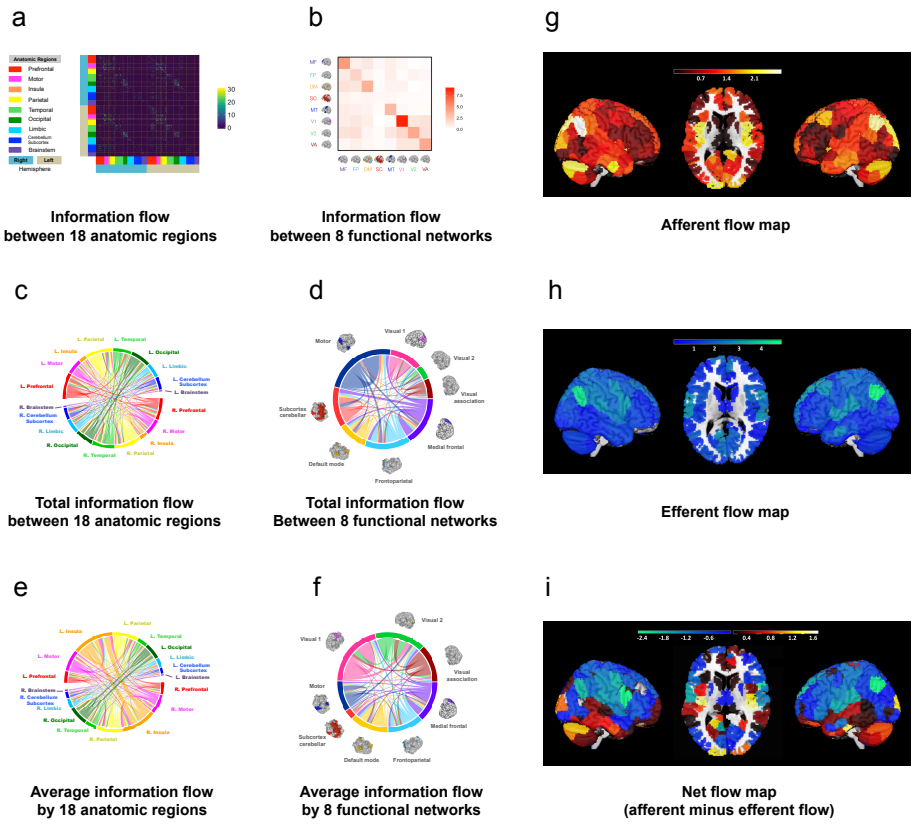


Fig. 2. Average whole-brain information flow in a group. (a) The 268 × 268 *asymmetrical* information flow matrix averaged from 550 individual information matrices. The matrix is defined on 18 anatomic regions using a 268-node functional atlas. The atlas is based upon an independent data set of healthy control subjects using a group-wise spectral clustering algorithm[17]. Every entry contains a selected *F* value from the lag-adjusted Granger-Geweke test between two time courses, or 0, if the *F* value is not selected. An edge is selected if its average *p*-value across all subjects is smaller than 0.05. (b) The 8 × 8 asymmetrical information flow matrix defined on 8 functional networks. The value for each entry is the mean *F* values associated with a functional brain network summarized from the 268 × 268 matrix in (a). (c) Total information flow map between 18 anatomic regions. The map contains 3,927 significant directed edges organized on 18 regions as in (a). The color of the edge indicates the origin of the flow. For example, the red curved line crossing the circle starting from the left (red bar) to the lower right (light orange bar) indicates information flow from the left prefrontal region to the right insula region. (d) Total information flow map between 8 functional networks. The same 3,927 edges in (c) now visualized on 8 functional brain networks. (e) Average information flow between 18 anatomic regions. The total amount of information flow from each anatomic region specified in (c) divided by the number of nodes in the region. (e) Average information flow between 8 functional networks. The total amount of information flow from each functional network specified in (d) divided by the number of nodes in the region. (g)-(i) Afferent, efferent, and net information flow maps of the whole brain. The brain images in (b), (d), and (f) are adapted by permission from RightsLink Permissions Springer Customer Service Centre GmbH: Springer *Nature Neuroscience* "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity" by Finn *et al.* Copyright 2015.

The 268 nodes are located in a total of 18 anatomic regions; they can be further allocated into eight functional networks according to previous work (*i.e.,* the medial prefrontal, frontoparietal, default-mode, subcortical-cerebellum, motor, primary visual (V1), secondary visual (V2), and visual association)[17]. Since different anatomic regions have various numbers of nodes, we examined both the total and the average *F* values with regards to each region to reduce bias caused by the parcel size difference between anatomic regions. (see **Fig.** 2).

2.2 The magnitude of the information flow

At the regional level, we observed relatively large mean *F* values in the prefrontal ($\mu_F = 2.18$ [L] and 1.77 [R]; we used L and R to denote left and right hemisphere and μ_F to denote the mean *F* value in an anatomic region), motor ($\mu_F = 3.58$ [L] and 2.31 [R]), temporal ($\mu_F = 2.92$ [L] and 1.77 [R]), parietal ($\mu_F = 3.26$ [L] and 4.04 [R]) and occipital cortex ($\mu_F = 4.67$ [L] and 6.59 [R]) (see **Fig.** 2). Compared to the global mean of *F* values at 0.78, this suggests that there exists substantial information exchange between cerebral regions. In comparison, the brainstem had the least information exchange with the cerebral cortex ($\mu_F = 0.004$ [L] and 0.03 [R]) (see **Fig.** 2 (a), (c) and (e)).

At the network level, our analyses demonstrated that (1) compared to the global mean of *F* values at 0.78, there were relatively large information exchanges within each functional network ($\mu_F > 1.24$), and (2) between functional networks, larger information exchange existed between the medial prefrontal (MF) and default mode (DM) ($\mu_F = 1.64$ [MF → DM] and $\mu_F = 1.32$ [DM → MF]), between medial prefrontal and frontoparietal (FP) ($\mu_F = 1.28$ [MF → FP] and $\mu_F = 1.12$ [FP → MF]), from medial prefrontal and frontoparietal areas to the primary visual area (V1) and secondary visual area (V2) ($\mu_F = 1.17$ [MF → V1], $\mu_F = 1.46$ [MF → V2], $\mu_F = 1.54$ [FP → V1], and $\mu_F = 1.97$ [FP → V2]), from the motor (MT) to the secondary visual area ($\mu_F = 1.53$ [MT → V2]) and between visual areas ($\mu_F = 2.71$ [V1 →

V2], $\mu_F = 2.73$ [V2 → V1], $\mu_F = 1.68$ [VA → V1], and $\mu_F = 2.53$ [VA → V2], where VA stands for the visual associate area). As for comparison, other between functional networks had mean *F* values that were less than or close to 1 (see **Fig.** 2 (b), (d) and (f)).

2.3 The directionality of the information flow

We further examined the directionality of information flow between different nodes, regions, and networks, and statistically compared the strength of the information flow from different directions. We found that the strongest afferent flow was primarily present in the posterior and ventral parts of the brain, including the occipital cortex, temporal cortex, insula, and cerebellum. In contrast, the afferent flow was relatively small in the lateral prefrontal cortex, sensorimotor area, and posterior temporal region. The large efferent flow was observed in the dorsal and lateral parts of the brain including the frontal cortex, sensorimotor cortex, and the parietal cortex, while the relatively small efferent flow was shown in the cerebellum, subcortex, and temporal cortex (see **Fig.** 2 (g)-(i)).

Since each node is associated with both afferent and efferent flows, we further subtracted the efferent flow *F* values from the afferent flow *F* values for each node to calculate the “net” information flow across the whole brain. We found that the ventral part of the brain (lateral and medial temporal cortex, insula, cerebellum) was predominantly associated with the net afferent flow, while the dorsal part of the brain (mostly frontal and parietal cortices) showed net efferent flow. These results suggest that the directionality of net information flow is differentially distributed across the whole brain, and the distinct information flow patterns between the dorsal and ventral brain may be attributed to their functionality differences (see **Fig.** 2 (g)-(i)).

Next, we examined the directionality of information flow in the 18 anatomical regions (see **Fig.** 3 (a) and (b)). The results revealed that the motor ($\mu_F = 4.42$ [L], $\mu_F = 4.46$ [R]), parietal ($\mu_F = 4.74$ [L], $\mu_F = 4.95$ [R]), and prefrontal ($\mu_F = 4.00$ [L], $\mu_F = 3.88$ [R]) regions were primarily associated with net efferent information, whereas the insula ($\mu_F = 4.63$ [L], $\mu_F = 4.26$ [R]), limbic ($\mu_F = 3.27$ [L], $\mu_F = 3.52$ [R]), brainstem ($\mu_F = 2.45$ [L], $\mu_F = 2.45$ [R]), cerebellar and subcortical ($\mu_F = 3.48$ [L], $\mu_F = 3.47$ [R]) regions were primarily associated with the net afferent flow. The temporal and visual regions were associated with a more balanced information inflow and outflow. These findings suggest that the overall directionality of resting-state information flow is from the dorsal brain to the ventral brain.

The network-level results revealed that the medial frontal ($\mu_F = 4.56$), frontoparietal ($\mu_F = 4.17$), and visual association ($\mu_F = 3.87$) subnetworks were primarily associated with net efferent flow, while the secondary visual (V2) ($\mu_F = 4.20$) and subcortico-cerebellar ($\mu_F = 3.00$) subnetworks were primarily associated with net afferent information. The other subnetworks (default-mode, motor, and primary visual) were associated with a more balanced information inflow and outflow (see **Fig.** 3 (d) and (e)). These findings suggest substantial information flow from high-order cognitive systems to primary functional systems during resting state.

Finally, to investigate whether the net information flow was statistically significant, we conducted a pair-wise ANOVA (analysis of variance) study of the net information flows (specifically, we performed ANOVA for the *F* values associated with each pair of brain regions). We tested if there was a difference between information flow from area A to area B and that from area B to area A, for each pair of networks (see **Fig.** 3 (c) and (f)). Our analyses showed that there was significant net information flow associated with medial frontal, default mode, and visual networks ($P_{FWE} < 0.001$, where P_{FWE} denotes adjusted *P*-

value controlling for family-wise error rate (FWER)). Taken together with the directionality of the net information flow from **Fig.** 3 (b), these results suggest that the medial frontal and frontoparietal subnetworks were associated with the significant net efferent flow, and the visual association subnetwork was associated with significant net afferent information flow.

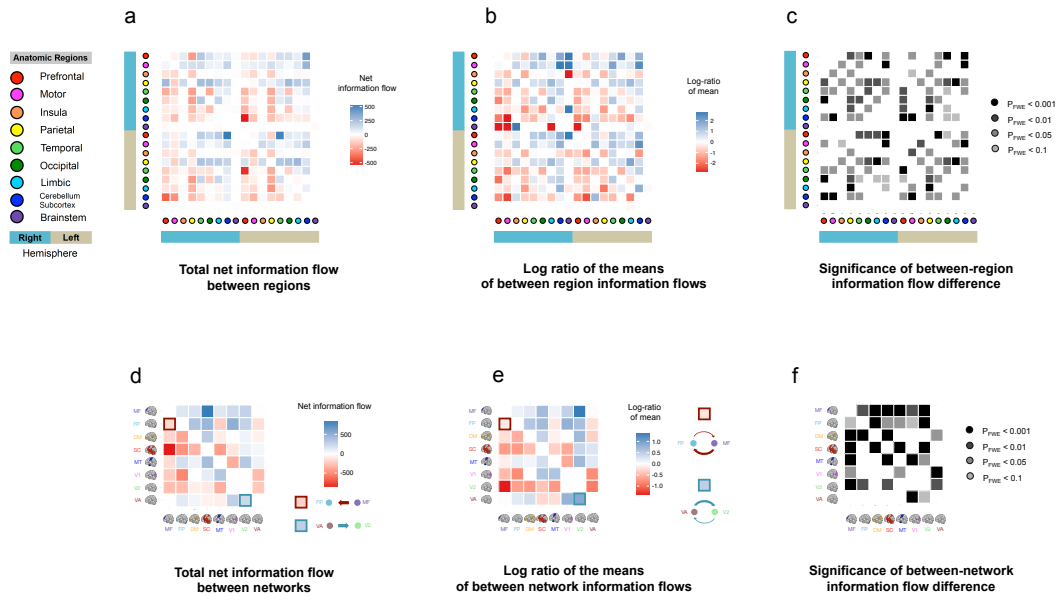


Fig. 3. Net total information flows of the whole-brain and mean information flow difference between regions and networks. (a) The net total information flows of the whole brain organized by regions. Each square represents the net total amount of information flows from one brain area to the other. The darker the blue square was, then the more (net) information was flowing out from the brain region denoted by the reference brain image in the horizontal direction into the brain area denoted by the reference brain image on the vertical direction; the darker the red was, the more (net) information was flowing into the brain region denoted by the reference brain image in the horizontal direction from the region denoted by the reference brain image on the vertical direction. (b) The mean difference between region information flows. Each square represents the difference between the mean of information flows from the brain area denoted by the reference brain image in the horizontal direction to the brain area denoted by the reference brain image on the vertical direction. Color conventions as in (a). (c) The *P* values for the corresponding mean difference in (b), corrected by FWER. (d) The net total information flows of the whole brain organized by networks. Color conventions as in (a). (e) The mean difference between network information flows. Color conventions as in (a). (f) The *P* values for the corresponding mean difference in (e), corrected by FWER. The brain images in (d), (e), and (f) are adapted by permission from RightsLink Permissions Springer Customer Service Centre GmbH: Springer *Nature Neuroscience* "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity" by Finn *et al.* Copyright 2015.

2.4 Analysis of variability of the whole-brain information flows

In humans, the brain is the most variable and fastest evolving organ[23]. Within the brain, different regions may be associated with different variabilities in information flow. The characteristics of the variability associated with information communication across different

brain networks, however, are not well-charted. To inquire into this property, we performed an analysis to investigate the variability of the information flow.

To quantify between-subject variability, we first calculated the variance of *F* values (directed edges) across all subjects in this sample. We found that edges with relatively high afferent flow variance generally involved those from the whole brain to the parietal ($Ratio = 1.34$ [L] and 1.65 [R], both P_{FWE}'s < 0.001, where $Ratio$ refers to the ratio between the average variability of the afferent information flow of a particular region and that of the whole brain, and [L] and [R] refer to the left and right hemispheres, respectively), from the whole brain to insula ($Ratio = 1.43$ [L] and 1.22 [R], both P_{FWE}'s < 0.01), and from the whole brain to occipital ($Ratio = 1.35$ [L] and 1.47 [R], both P_{FWE}'s < 0.001) regions, suggesting high between-subject variability of information flow. In contrast, edges within the cerebellar-subcortical ($Ratio = 0.86$ [L] and 0.86 [R], both P_{FWE}'s < 0.001), brain stem ($Ratio = 0.39$ [L] and 0.37 [R], both P_{FWE}'s < 0.001), and limbic ($Ratio = 0.75$ [L] and 0.90 [R], both P_{FWE}'s < 0.001) regions showed relatively low variance, suggesting that afferent flows in these regions are relatively stable across subjects (see **Fig.** 3).

Edges with relatively high efferent flow variance generally involved those from the PFC region to the whole brain ($Ratio = 1.25$ [L] and 1.16 [R], both P_{FWE}'s < 0.001), from the motor region to the whole brain ($Ratio = 1.35$ [L] and 1.37 [R], both P_{FWE}'s < 0.001), from the parietal region to the whole brain ($Ratio = 1.62$ [L] and 1.88 [R], both P_{FWE}'s < 0.001), and from the occipital to the whole brain ($Ratio = 1.35$ [L] and 1.24 [R], both P_{FWE}'s < 0.001), suggesting high between-subject variability of information flow. In contrast, edges within the cerebellar-subcortical ($Ratio = 0.69$ [L] and 0.74 [R], both P_{FWE}'s < 0.001), brain stem ($Ratio = 0.28$ [L] and 0.25 [R], both P_{FWE}'s < 0.001), and limbic ($Ratio = 0.56$ [L] and 0.60

[R], both P_{FWE}'s < 0.001) regions showed relatively low variance, suggesting that efferent flows in these regions are relatively stable across subjects (see **Fig.** 3).

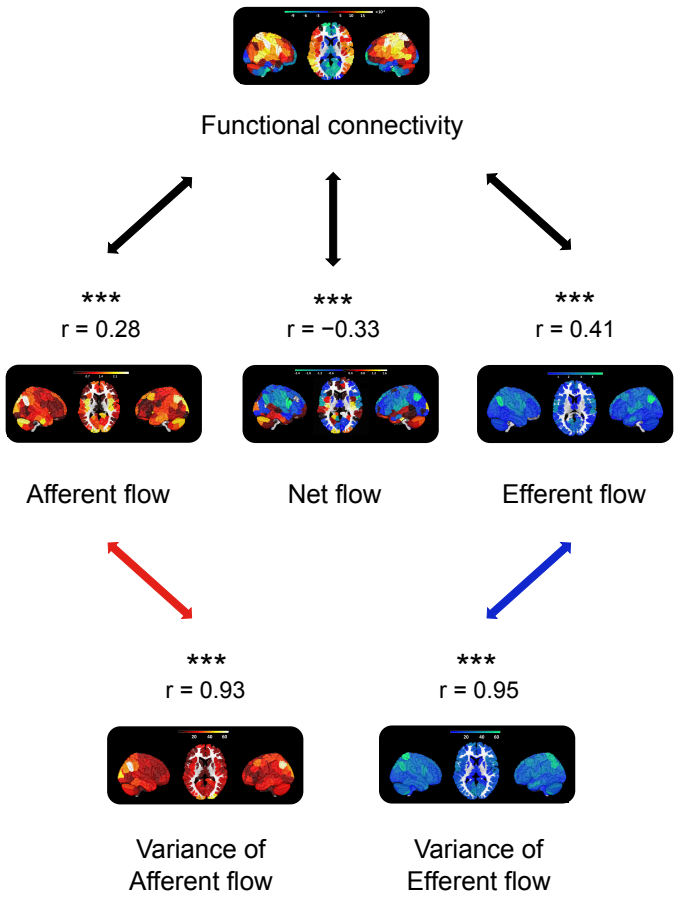


Fig. 4. A comparison between functional and effective connectivity. The functional connectivity map was obtained by first averaging 550 268×268 Pearson correlation matrices obtained from 550 subjects. Subsequently, the resulting 268×268 matrix was further summarized according to 268 brain areas. This yielded a vector containing 268 values, each of which corresponds to the "connectedness" of a brain area to the rest of brain areas. The effective connectivity map was obtained by first averaging 550 268×268 F matrices obtained from 550 subjects. Subsequently, the resulting 268×268 matrix was further summarized according to 268 brain areas, into three vectors. The first vector contained 268 values, each of which corresponds to how much information flow is flowing into a particular brain area from other brain areas (afferent flow). The second vector also contained 268 values, each of which corresponds to how much information flow was flowing out from a particular brain area to other brain areas (efferent flow). The third vector was obtained by subtracting the efferent flow from the afferent flow, indicating the net information flow entering each brain area. Finally, the association between functional and effective connectivities can be quantified by correlating the functional connectivity vector with each of the three effective connectivity vectors.

Interestingly, we observed high correlations between the variability maps and the information flow strength (F value) maps (**Fig.** 4), suggesting that regions with high information flow strength also had high information flow variability, and *vice versa*.

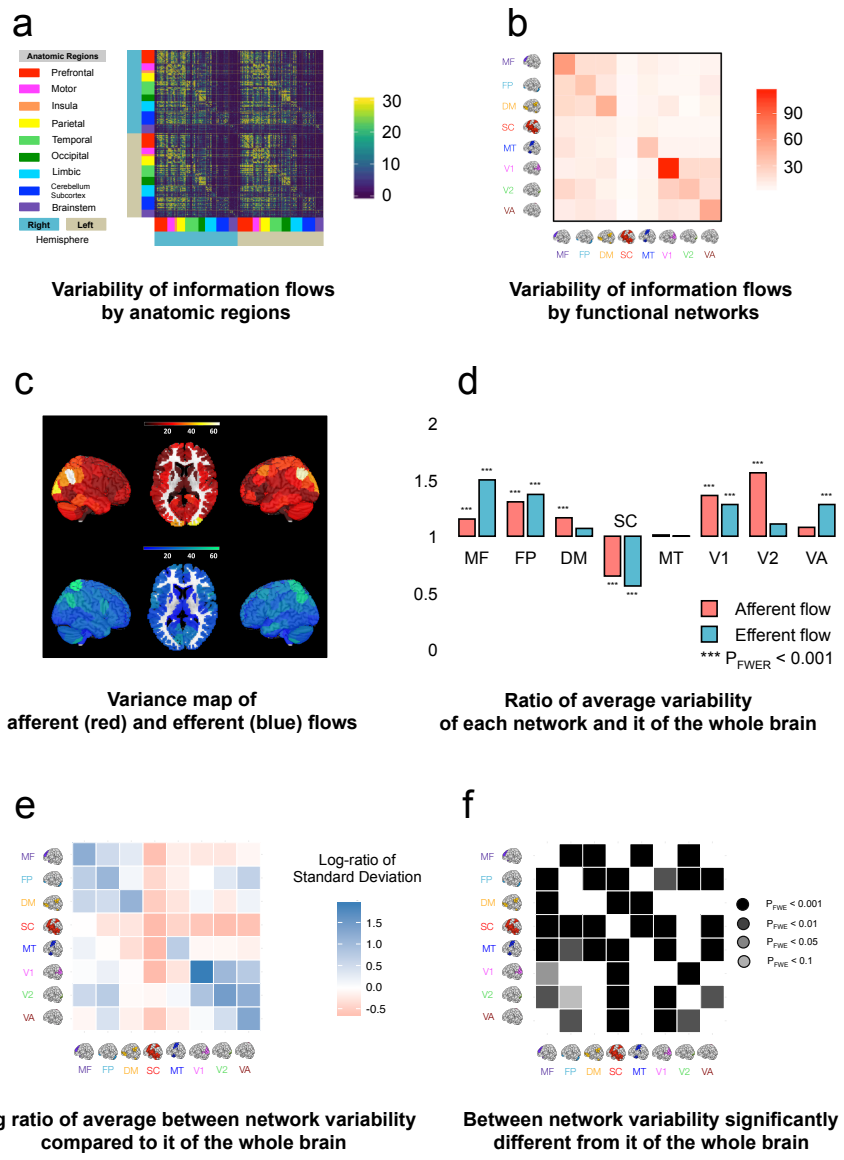


Fig. 5. Analysis of variance of the whole-brain information flows. (a) The variance of 268 × 268 asymmetrical information flow map across all subjects. Every entry contains the variance of the *F* value between two time courses across subjects. In the figure, for color contrast convenience, if an *F* value is greater than 100, we fix it at 100. (b) The variance of 8 × 8 asymmetrical information flow map across all subjects defined on 8 functional networks. Every entry is the mean variance of *F* value associated with a functional brain network summarized from the 268 × 268 matrix in (a). (c) The variability map of afferent (in red color) and efferent flows (in blue color). (d) The variability of afferent (red) and efferent flows (blue) associated with each network compared to it of the whole brain. The height of each histogram quantifies the average variance of F values corresponds to each network, which was visualized in panel (c), compared to it of the whole brain. A pair-wise ANOVA test determines the significance level. All P-values were adjusted for FWER. (e) The log-ratio of the average variation of information flows between each pair of networks. Each log-ratio was calculated in two steps. First, we found the average variations of the efferent flow from nodes linking regions A and B - it measures the average variability of the efferent flow from region A to region B. We also found the average variations of the afferent flow from regions B to A - it measures the average variability of the afferent flow from region A to region B. Second, we calculated the natural log of these two average variations. The darker the blue square was, the more variability of information flowing out from the brain region denoted by the reference brain image on the horizontal direction than the variability of the opposite information flows; the darker the red was, the more variability of information flowing into the brain region denoted by the reference brain image on the horizontal direction than the variability of the opposite information flows. (f) The significance of the between-network variability. The darker the green square, there was more significant a difference between the variability of the information flows from opposite directions (namely, the variability of information flows from regions A to B compared to it from region B to A). The size of the difference

(namely, whether there is more variability of the information flows from regions A to B than it from B to A) can be determined from the map in (e). The comparison was done using a pair-wise ANOVA test. All P-values were adjusted for FWER. The brain images in (e) and (f) are adapted by permission from RightsLink Permissions Springer Customer Service Centre GmbH: Springer *Nature Neuroscience* "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity" by Finn *et al.* Copyright 2015.

Next, we compared the variability of the information flow obtained from the eight functional networks. We began by taking the ratio of the variability within each network and that of the whole brain. A positive ratio indicated that the within network variability was greater than that of the whole brain, and a negative ratio indicated that the within network variability was smaller than that of the whole brain. To quantitatively demonstrate whether the above differences were significant, we performed an ANOVA test between the variability within each network and that of the whole brain. As **Fig 4** (d) shows, most networks had significantly higher variability ($P_{FWE} < 0.001$) than the whole brain, except the subcortical cortex.

Finally, we analyzed the variability of between-network information flows. To that end, we computed the log ratios of the variability between the afferent and efferent flows in each pair of networks. Specifically, for each two networks A and B, we calculated the natural logs of the average variance from A to B and that from B to A (see **Fig.** 5 (e)). The derived log values were further divided by each other to generate a ratio. Next, we conducted a pair-wise ANOVA test to compare the log ratios of every pair of networks with that of the whole brain to examine whether the differences between variability of information inflow and outflow were statistically significant. Our results showed that, compared to the variability of information flow in the whole brain, there was significantly higher information flow variability among the medial frontal cortex, frontal-parietal regions, the default mode regions (all P_{FWE}'s between every two networks < 0.001), suggesting that information flow between association cortices is highly variable. Additionally, there was significantly higher

information flow variability among different visual networks (all P_{FWE}'s between every two networks < 0.001, see **Fig.** 5 (f)).

As a whole, our analyses both qualitatively and quantitatively suggest that there is substantial variability of information flow within each within brain networks compared to the whole brain. Between-network information flow is associated with higher variability between higher-order cognitive systems (*i.e.,* medial prefrontal, frontoparietal, and the default-mode) and between different visual areas.

2.5 Prediction of cognitive flexibility

Inspired by prior work, which showed that information flow in the brain was related to executive functioning[13,14], we further investigated whether directed information flow among those edges selected during the first step edge selection could predict individual executive ability in the healthy population. The executive ability was evaluated by the Dimensional Change Card Sort (DCCS) Test[24] in the HCP data. The DCCS test is a widely used neurocognitive task that assesses individuals' cognitive flexibility by rapidly switching card-sorting rules between different dimensions. A higher performance score reflects greater flexibility in monitoring and switching thoughts and behaviour to facilitate the attainment of target goals.

The behaviour prediction procedure can be, broadly, summarized in three steps. First, we conducted the first step edge selection to remove spurious connections possibly caused by noise. During the first step, we selected an edge, if its average *p-value* (obtained from the GCA) across all subjects was smaller than a threshold (*e.g., p* < 0.1) (see **Remarks** in the supplementary materials). Next, we performed a second step edge selection and model building. We followed prior work[1] and used a machine-learning-based framework with

leave-one-subject-out cross-validation (LOOCV) on a randomly selected subsample of 550 healthy subjects (70% of the total samples). The selected edges were employed to fit a linear regression model, where edge strengths (*F* values) and covariates (*i.e.,* age and gender) were entered as regressors. Edges that were significantly correlated with the DCCS scores at $P < 0.005$[25,26] were identified as "effective information flows" and were selected as neural biomarkers. The identified effective flows were subsequently isolated into positive and negative groups, based on the sign of their correlations with DCCS scores. During each LOOCV, we applied the estimated regression weights using data from 549 training subjects to data from the holdout and obtained a predicted DCCS test score. We iterated the LOOCV analysis 550 times, where each subject's DCCS score in the sample was predicted once using other subjects' data. We built the predictive edges into an *F* value weight map (as the weights from a regression analysis), where each entry of the map indicated the weight of a predictive *F* value (see **Fig.** 1 f).

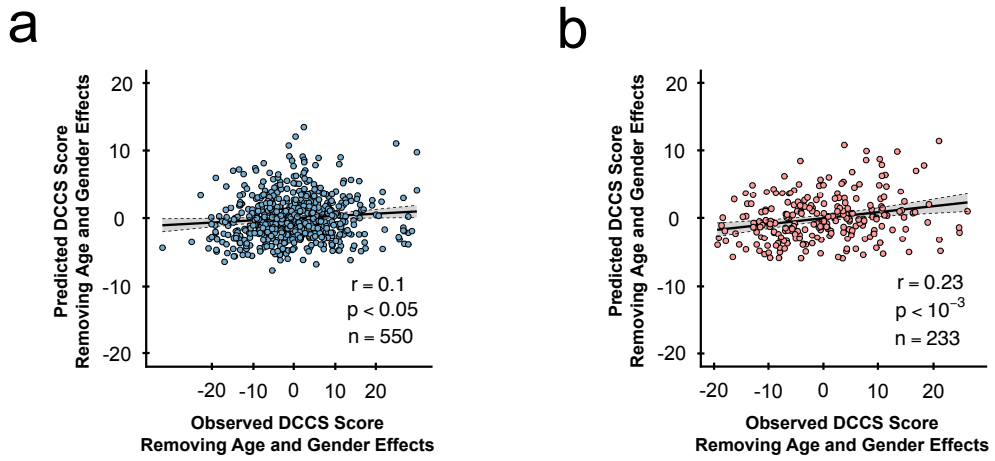


Fig. 6. Individual effective information flow predicts cognitive flexibility. (a) LOOCV prediction result comparing predicted and observed DCCS scores (*n* = 550 subjects). Confounds, such as age and gender, are regressed out before prediction. Scatter plot shows prediction based upon the whole brain positive effective information flows (threshold at *p* <0.005). Each dot represents one subject; the grey area indicates the 95% confidence band for the best-fit line. (b) Out-of-sample prediction result comparing predicted and observed adjusted DCCS scores (*n* = 233 subjects). Confounds, such as age and gender, are regressed out before prediction. Scatter plot shows prediction based upon the whole brain positive effective information flows (threshold at *p* <0.005). Each dot represents one subject; the grey area indicates the 95% confidence band for the best-fit line. (c) The information map of positive total effective information flow edges based upon feature selection. The flows are arranged according to 18 brain regions. The brain images in (d) are adapted by permission from RightsLink Permissions Springer Customer Service Centre GmbH: Springer *Nature Neuroscience* "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity" by Finn *et al.* Copyright 2015.

To investigate the robustness and reproducibility of the results, we extended the prediction analysis to the held-out sample of 233 subjects (testing set, or 30% of the total sample). *F* values for each novel subject in the testing set were extracted and then multiplied by the weight map obtained from the training set, yielding a predicted DCCS test score for each subject in the testing set (see **Fig.** 1). To evaluate prediction power, we correlated all predicted and observed DCCS scores of subjects from the training dataset using Pearson correlation. One thousand nonparametric permutation tests were performed to evaluate model fitting and out-of-sample prediction. During each test, we randomly permuted the DCCS scores, and then conducted model fitting using the same procedure as described above. Prediction power for each permutation test was recalculated.

Our analysis revealed that effective information flow (extracted *F* values) positively correlated with the DCCS scores in the training sample significantly predicted cognitive flexibility in the test sample. In particular, at the threshold of $p < 0.005$, 178 edges were found to be positively predictive of DCCS score. The LOOCV in the training sample showed a prediction accuracy of $r = 0.1$ ($n = 550$, $p < 0.05$) (see **Fig.** 6 a). The correlation between predicted and true values in the out-of-sample prediction was $r = 0.23$ ($n = 233$, $p < 0.001$) (see **Fig.** 6 b). Notably, the observed Pearson correlation *r* was larger than any permuted *r* values in the permutation distribution ($p < 0.001$), suggesting highly significant predictability than one would expect by chance. In contrast, we did not observe a significant predictive effect on edges with negative correlations with the DCCS scores at any of the above thresholds ($p > 0.05$).

Because the choices of statistical thresholds during the two-step edge selection are somewhat arbitrary, a range of thresholds was tested to ensure that results were consistent. In an additional analysis, we maintained the second step edge selection threshold at 0.005

and changed the first step threshold from 0.1 to 0.05. In the training sample, the prediction is $r = 0.11$ ($p = 0.01$, $n = 550$) and out-sample prediction is $r = 0.20$ ($p = 0.002$, $n = 233$). The more conservative threshold selected fewer edges (68, as compared to 178 using the more moderate P level). In another analysis, we maintained the first step edge selection threshold at 0.1 and considered the second step edge selection threshold at 0.01. In the training sample, the prediction is $r = 0.07$ ($p = 0.1$, $n = 550$) and out-sample prediction is $r = 0.21$ ($p = 0.001$, $n = 233$). However, the increasing threshold selected significantly more edges (303, as compared to 178 using the more conservative threshold).

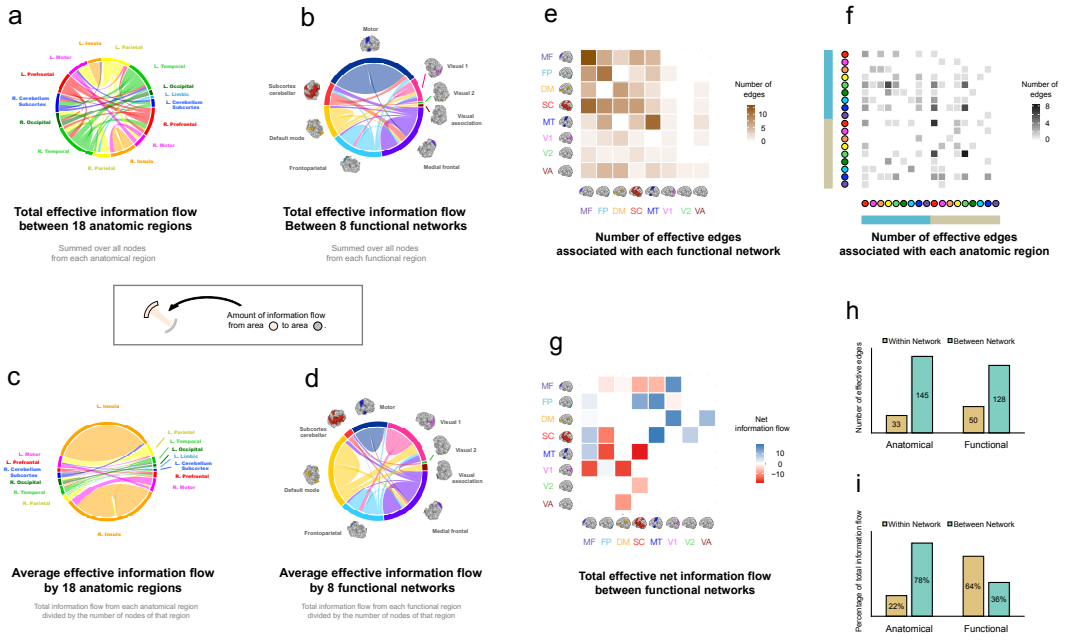


Fig. 7. Individual effective information flows. (a) The information map of positive total effective information flow edges based upon feature selection. The flows are arranged according to 18 brain regions. (b) The same edges in (a) now organized according to 8 functional brain networks. (c) Average effective information flow (total effective information flow in figure (a) divided by the number of nodes in each anatomic region). (d) Average effective information flow (total effective information flow in figure (b) divided by the number of nodes in each functional network). (e) The number of effective edges arranged according to 8 functional brain regions. (f) The number of effective edges arranged according to 18 anatomic brain regions. Color convention as in Fig. 3 (a). (g) Net total effective information flow between 8 functional brain regions. The darker the black square was, then the more information is flowing out from the brain region denoted by the reference brain image on the horizontal direction to the brain region denoted by the reference brain image on the vertical direction; the darker the red was, then the more information is flowing into the brain region denoted by the reference brain image on the horizontal direction from the brain region denoted by the reference brain image on the vertical direction. (h) The number of effective edges arranged by within and between anatomic regions, as well as it arranged by within and between functional networks. (i) The total information flow (summed F values) arranged by within and between anatomic regions, as well as it arranged by within and between functional networks. The brain images are adapted by permission from RightsLink Permissions Springer Customer Service Centre GmbH: Springer *Nature Neuroscience* "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity" by Finn *et al.* Copyright 2015.

Next, we aimed to unveil the details of the extracted directed edges predictive of DCCS scores. Our results showed that they were primarily present within the higher-order cognitive systems (*i.e.*, medial prefrontal ($\mu_F = 13.28$), frontoparietal ($\mu_F = 18.98$) and the default-mode ($\mu_F = 12.41$)), and from higher-order systems to the primary systems (*e.g.*, frontoparietal to the motor ($\mu_F = 6.09$), frontoparietal to V1 ($\mu_F = 9.14$), medial prefrontal to V1 ($\mu_F = 8.60$), and frontoparietal to the visual association ($\mu_F = 11.85$)) (**Fig.** 7). In addition, information flow was also shown from the motor network to the medial prefrontal ($\mu_F = 8.41$) and frontoparietal networks ($\mu_F = 7.51$), and from the subcortical network to the medial prefrontal network ($\mu_F = 9.82$). Taken together, these results suggest that information flow associated with the prefrontal cortex is critical to, and predictive of, human cognitive flexibility.

2.6 Comparison between functional connectivity and effective connectivity

An interesting question would be whether the effective connectivity patterns examined here can be largely explained by undirected functional connectivity. If so, the effective connectivity measures may have limited practical value. For this purpose, we conducted a further analysis comparing the patterns of the functional connectivity with that of the afferent flow (*i.e.*, information flow towards each brain area), the efferent flow (*i.e.,* information flow outwards each brain area), and the net flow (*i.e.,* afferent flow minus efferent flow). Our analysis showed that the effective connectivity maps were only moderately correlated with the functional connectivity map (r = 0.28, 0.41, and -0.33 for afferent, efferent and net flow, respectively). As a result, functional connectivity can only explain between 0.08 and 0.17 (r^2) of the total variance in the effective connectivity matrices. These findings suggest the uniqueness of the results derived from GCA analysis, which cannot be fully accounted for by functional connectivity analysis.

Furthermore, we compared the behaviour prediction using effective information flow (*F* values) and the traditional functional connectivity[1]. Using functional connectivity (*i.e.,* Pearson correlation between times course from different brain areas), we obtained a prediction of r = 0.1 in the training set ($p = 0.001$, $n = 550$) and r = 0.22 in the test set ($p < 0.001$, $n = 233$). This was similar to our "effective" edge-based prediction performance. These prediction results, however, were based on 342% more edges (787 functional edges as compared to 178 "effective" edges identified using information flow), suggesting that many of the correlation-based edges did not further improve behaviour prediction.

3. Discussion

In this study, we employed a Granger-Geweke causality framework on fMRI data to study resting-state idiosyncratic whole-brain information flows in healthy individuals. The analyses revealed that information flows were mainly distributed within each specialized functional system, from the dorsal brain to the ventral brain, and from the higher-order cognitive systems to the primary functional systems. The regions with high information flow strength also showed high between-subject information flow variability. In addition, we showed that the strength of information flow between the higher-order cognitive systems and from these systems to the primary systems (visual, motor, and subcortical) predict cognitive flexibility scores in humans.

Substantial evidence has demonstrated that within-system functional connectivity is particularly strong in the brain, as compared to between-system connectivity[27,28]. This finding is consistent across different brain states[29-31], data processing methods[32,33] and populations[34,35], suggesting a robust network organizational feature of the brain. In this article, we extended previous findings by showing strong effective connectivity within each functional system. Our result suggests that system-level information flow is strong between

regions that are involved in the same function. In contrast, between-system effective connectivity was much weaker than within-system connectivity. Strong between-system connectivity, however, was mainly observed in the higher-order cognitive systems and from these systems to more primary functional systems. This agrees with prior findings that brain hubs are primarily distributed in the frontoparietal and the default-mode networks[36,37], which have the most connections with other parts of the brain. Since the top-down projections from the prefrontal and higher-order systems to the primary functional systems are critical to control and coordinate human behaviour[38-41], the abundant information flow at such direction during resting state may suggest a cognitive effort to focus the attention and to predict and adapt one's mind and behaviour during the scan.

Interestingly, directed edges with the highest *F* values also showed the largest information flow variability across subjects. The variability was particularly prominent in the frontal cortex, parietal cortex, and visual cortex. Much evidence has shown that connectivity in the frontoparietal system is associated with low within-subject variability but high between-subject variability, which may act as "fingerprints" to distinguish individuals[1,42,43]. Further, connectivity in the cognitive control systems shows the highest vulnerability to subject-specific psychological and physiological factors, such as mood, stress, fatigue, personal experience, among others[44-47]. These findings support our results and suggest that the high variability of information flows in the frontal and parietal cortices across subjects may serve as a neural signature to discern and predict individual behaviours. Apart from these regions, high variability was also observed in the visual systems, which may relate to the continuous sensation of environmental change during the eyes-open resting-state scan. This finding, however, is different from previous findings that the primary functional systems are associated with relatively small between-subject variability[48]. Such discrepancy may to

some degree suggest a higher sensitivity of effective connectivity in assessing individual differences in the primary functional systems.

The information flow that predicts cognitive flexibility involved connectivity within the higher-order systems (frontoparietal, medial prefrontal, and the default-mode) and from these systems to the primary sensory-motor systems. This reflects the key role of higher-order systems in human cognition. The frontoparietal system is a crucial system for cognitive control in humans – a set of cognitive processes that coordinates and monitors goal-directed behaviours[49,50], and the medial frontal system participates in cognitive control processing mainly by evaluating the value of input information[51,52], error detection[51,52], conflict monitoring[53], and decision making[54]. Previous research has revealed that functional connectivity in the frontoparietal system has the highest predictability for fluid intelligence in humans[1]. In line with prior results, we further show the highest predictability of information flow in higher-order cognitive systems for human cognitive flexibility – a critical feature of cognitive control ability. These findings together suggest that both functional and effective connectivity of higher-order systems are critical to cognitive functioning in humans.

We would like to note several limitations of our study. First, the sampling rate for fMRI data is slower than the timescale of the underlying neuronal responses. This is a common limitation for GCA studies using fMRI data[55,56]. In this work, we sought to attenuate the effect of this limitation by using HCP data with relatively high temporal resolution (TR = 720 ms). This potentially reduces the biases caused by the mismatch of timescales. We acknowledge, however, that the findings might still be affected by this factor to a certain degree. Future work replicating these results using other data acquisition techniques, such as EEG and MEG, is important. Second, the information flow profiles and the prediction power were calculated based on the integration of large time series acquired from two days

lasting for a total of one hour. The findings reported in the current study, therefore, are likely to reflect brain information configuration over a long period. Given the dynamic nature of connectivity measures[57,58], information flows may dynamically change over time during the whole scan. Third, our study using GCA only captures linear information flows in the brain. Future work using non-linear parametric approaches may be useful to uncover the non-linear information flow architecture of the brain. Fourth, our results are derived from a sample of young healthy adults. Whether these findings would generalize to other populations is an open question and needs to be investigated in future studies. Fifth, although BOLD signals possess a strong neural basis as they are highly correlated with local field potential, which is a direct measure of synaptic activity[59], we acknowledge that our results may to a certain degree be influenced by the vasculature. Finally, although we used the Pearson correlation test to verify if the predicted values were significantly associated with the observed values, the assumptions for the test, such as normality and homoscedasticity, however, may be violated. Additionally, the *p*-values may be affected by sample sizes. Even though these issues were partly addressed by (1) showing the results can be reproduced in novel subjects and (2) the unveiled meaningful neurobiological areas, future studies need to verify the generality of brain information flows in new samples and in predicting other human behaviours, and to apply and develop other metrics that may be more suitable for testing the predictive performance.

In conclusion, our study provides evidence for the directionality and patterns of information flow in the human brain and highlights the importance of information flow in the high-order systems in relation to cognitive flexibility. Since these information flow patterns cannot be fully explained by traditional functional connectivity measures, the data presented here extend current knowledge of human brain functional organization and open a new avenue towards investigating the neurobiological basis for individual executive functional ability.

These findings may also show the potential to help advance our understanding of brain disorders characterized by deficits in executive functioning, such as schizophrenia, ADHD, and OCD.

4. Methods

4.1 Subject information

The data set used in this article was from the Human Connectome Project (HCP) 1200 data release. A total of 783 subjects from the HCP 1200 data release were used for this analysis (383 males and 400 females). 180 subjects are 22- to 25-year-old, 343 subjects are 26- to 30-year-old, 252 subjects are 31- to 35-year-old, and 8 subjects are above 36-year-old. The HCP 1200 data release had 1096 subjects in total. The Pearson correlation between head motion and the cognitive scores is r = -0.17. In other words, some subjects with low cognitive scores have high head motion. While establishing the association between information flow and cognitive scores, we sought to remove the confounding effects from head motion. We excluded 211 subjects with significantly high head motion and low cognitive scores during the scans. Specifically, we excluded subject i, if $\frac{HM_i}{DCCS_i} \times 10^3 > 1$, where HM_i is the frame-to-frame head motion estimate (averaged across both day 1 rest runs; HCP: Movement_RelativeRMS_mean) and $DCCS_i$ is the DCCS score for subject i, respectively. The correlation between head motion and the cognitive scores is r = -0.005 after exclusion. The other excluded subjects had missing time points. The data were acquired in two separate sessions (REST 1 and REST 2) on two different days. Each session contained data from both the left-right (LR) and right-left (RL) phase-encoding runs. The data (REST 1 LR, REST 1 RL, REST 2 LR, REST 2 RL) were concatenated to calculate the information flow metrics. All participants provided written informed consent. Subject recruitment procedures and informed consent forms, including consent to share de-identified data, were approved by the Washington University in St. Louis Institutional

Review Board (IRB). All experimental procedures were performed under the guidelines of the HCP, which adhered to the relevant IRB processes related to that project; full details on the HCP have been published previously[60]. The datasets analyzed during the current study are available on the HCP page (https://www.humanconnectome.org/study/hcp-young-adult).

4.2 Data acquisition and preprocessing

The functional imaging data (the WU-Minn HCP Phase II data) were acquired at 2 mm isotropic on a 3T Magnetom Skyra Connectom scanner using a 32-channel head coil. Parameters for the functional scans were: TR = 720 ms, TE = 33 ms, echo spacing (spin echo field) = 0.58 ms, FOV = 208 mm × 180 mm, Matrix =104 × 90 with 72 slices covering the whole brain, multiband factor of 8, and FA = 52°. There were two resting-state functional runs for each participant, each lasted 14.39 mins (1200 time points). The structural data included a pair of T1- weighted image, all acquired at 0.7 mm isotropic voxel resolution, plus ancillary scans, for a session duration of ~40 min: TR = 2400 ms, TE = 2.14 ms, TI = 1000ms, voxel size 0.7*0.7*0.7 mm^3, FOV = 224 mm, FA = 8°.

Data were preprocessed using the standard pipeline implemented in the Statistical Parametric Mapping (SPM12, http://www.fil.ion.ucl.ac.uk/spm/), following the previously published work[61-64]. The procedure included slice-timing correction, realignment, individual structural-functional image coregistration, Montreal Neurological Institute (MNI) template normalization and spatial smoothing. Additional noise corrections were applied to be consistent with previous work[3], including removal of linear components related to the 12 motion parameters (six motion parameters plus their first derivatives), regression of mean time courses of the white matter, cerebrospinal fluid and global signal, removal of the linear trend, and low-pass filtering (< 0.12 Hz). Secondary data analysis, including calculating

information flow metrics, feature selection, and model building, were conducted using the *R* software. For each subject, we obtained a 268 × 268 asymmetrical information flow matrix.

4.3 A three-step procedure for identifying information flow

Preamble: Granger-Geweke causality analysis. A random variable X *Granger causes* another variable Y, if the prediction of Y is improved using the information of its past and the past of X, compared with when using only the past information of Y. Formally, let $\{X_t\}_{t \in Z}$ be a process, where Z denotes the set of integers each of which corresponds to a time point, and $X_t = (X_{t1}, \cdots, X_{td})^T$. For $j \neq k \in \{1, \cdots d\}$, $\{X_{tk}\}_{t \in Z_+}$ *Granger causes* $\{X_{tj}\}_{t \in Z_+}$ if and only if there exists a measurable set A such that:

$$P\left(X_{t+1,j} \in A \mid \{X_s\}_{s \leq t}\right) \neq P\left(X_{t+1,j} \in A \mid \{X_{s,\backslash k}\}_{s \leq t}\right)$$

for all $t \in Z_+$, where Z_+ denotes positive integers and $\{X_{s,\backslash k}\}$ is the subvector obtained by removing X_{sk} from X_s.

Given that the processes are stationary, Granger causality can be implemented via autoregressive (AR) modelling. Formally, consider the following AR processes:

$$X_t = \sum_{j=1}^{\infty} \alpha_j X_{t-j} + \varepsilon_j, \quad Var(\varepsilon_j) = \Sigma_1.$$

$$Y_t = \sum_{j=1}^{\infty} \beta_j Y_{t-j} + \eta_j, \quad Var(\eta_j) = T_1.$$

To inquire into the potential predictability of X_t on Y_t, and *vice versa*, we further define:

$$X_t = \sum_{j=1}^{\infty} a_j X_{t-j} + \sum_{j=1}^{\infty} b_j Y_{t-j} + e_j, \quad Var(e_j) = \Sigma_2.$$

$$Y_t = \sum_{j=1}^{\infty} c_j X_{t-j} + \sum_{j=1}^{\infty} d_j Y_{t-j} + z_j, \quad Var(z_j) = T_2.$$

where the covariance between the noises is $Cov(e_j, z_j) := \gamma_2$.

Under this framework, the Geweke test[18] provides a linear-feedback measure. Denote $F_{X\to Y}$, $F_{Y\to X}$, and $F_{X\cdot Y}$ as the feedback measure from X to Y, from Y to X, and the instantaneous feedback measure between X and Y, respectively. Formally, they can be written as:

$$F_{Y\to X} = \ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)$$

$$F_{X\to Y} = \ln\left(\frac{|T_1|}{|T_2|}\right)$$

where |·| denoted matrix determinant.

Step 1: Granger-Geweke test and Individual-level information flow map. The information flow varied between different subjects and across brain regions. To quantify the subject-specific information flow, we introduced the node-wise optimal lag Granger-Geweke test to obtain the whole brain information flow metrics. For two corresponding time courses corresponding to a node pair (*i, j*) we conducted the Granger-Geweke test to uncover the information feedback between them: $F_{i\to j}$ and $F_{j\to i}$. The choice of lag *l*, where x_{t-l} was the lag *l* value of x_t, was critical to detect the Granger causality between two time courses. Here, we define an optimal choice of lag *l* as one that not only explains a reasonable amount of variance of the data, but also is easy to estimate for large-scale data, using the node-wise Akaike information criterion (AIC)[22]. Formally, we define the optimal lag length between two time courses from nodes *i* to *j*, for a specific subject *k*, as:

$$l^{opt}_{i\to j,k} = \min_l\{\alpha^l_{i\to j,k}\}$$

where $\alpha^l_{i\to j}$ is the AIC score of two times courses for nodes *i* to *j* with lag *l*.

Step 2: Edge selection. The subject-level information flow maps share a common pattern (**Fig.** 1 a), indicating that there are consistent directed edges (*F* values) in a population. To uncover these common information flow paths, we further investigated information flow in a group. We conducted a node-wise feature selection as follows.

We select an edge from *i* to *j* (namely we select $F_{\mathrm{i}\to\mathrm{j}}$), if $\xi_{\mathrm{i}\to\mathrm{j}} \leq \alpha$, where $\xi_{\mathrm{i}\to\mathrm{j}}$ is a feature calculated as the average of the P-values across all subjects ($1 \leq k \leq N$) resulting from the Granger-Geweke tests regarding two time courses obtained from nodes *i* and *j*, namely, $\xi_{\mathrm{i}\to\mathrm{j}} = \frac{\sum_{\mathrm{k}=1}^{\mathrm{N}} \mathrm{p}_{\mathrm{i}\to\mathrm{j}}^{\mathrm{k}}}{\mathrm{N}}$, and α is the feature selection rejection threshold (*e.g.,* 0.05). Particularly, the feature selection was conducted based on $p_{\mathrm{i}\to\mathrm{j}}^{\mathrm{k}}$, for every subject *k*, and for every pair of nodes (*i, j*). Specifically, $p_{\mathrm{i}\to\mathrm{j}}^{\mathrm{k}}$ is the P-value from a Granger-Geweke analysis between two time courses from region *i* to *j*, for subject *k*, under the optimal lag length $\mathrm{l}_{\mathrm{i}\to\mathrm{j},\mathrm{k}}^{\mathrm{opt}}$ for the same subject. The resulting directed (significant) edges were information flow paths that were likely to be present across different subjects. They were candidate features used to predict cognition.

Step 3: Prediction of executive functioning. Next, we examined if their variability was associated with variability in subject-specific cognition. In the HCP protocol, executive functioning was evaluated by the Dimensional Change Card Sort (DCCS) test. Subjects were asked to sort a series of bivalent test cards, first according to one dimension (*e.g.,* color), and then according to the other (*e.g.,* shape).

During the training state, we used leave-one-subject-out cross-validation (LOOCV) in a sample of 550 subjects (70% of the total samples) to further prune edges extracted from Step 2 above. During each iteration, we first conducted a feature selection on edges extracted from 549 subjects and built a regression model. Afterwards, we validated the model using data from the holdout subject. This was iterated 550 times, where each subject was predicted once. Specifically, for the feature selection during LOOCV, we correlated the *F* values from each edge with the DCCS scores in the training set. We chose the edges whose *F* values had a significant Pearson correlation with the DCCS scores as neurological signatures. Selected edges were subsequently separated into two groups: ones with a

positive correlation (positive signatures, or F^+ edges), and ones with a negative correlation (negative signatures, or F^- edges). As a result of the model development, we built two regression models using the F^+ and F^- edges in the training data, as follows:

$$\hat{y}^+ = \hat{\beta}_0^+ + \hat{\beta}_1^+ \sum_{i=1}^{q} \sum_{j=1}^{q} F_{i\to j}^+ \hat{I}_{i\to j}^+$$

$$\hat{y}^- = \hat{\beta}_0^- + \hat{\beta}_1^- \sum_{i=1}^{q} \sum_{j=1}^{q} F_{i\to j}^- \hat{I}_{i\to j}^+$$

where $F_{i\to j}^+$ and $F_{i\to j}^-$ are observed information flow metric from node *i* to node *j*. $\hat{I}_{i\to j}^+$ is an indicator function, where $\hat{I}_{i\to j}^+ = 1$ if $F_{i\to j}^+$ is associated with a significant information flow, and 0 otherwise. $\hat{I}_{i\to j}^-$ is similarly defined. $\hat{\beta}_0^+$, $\hat{\beta}_0^-$, $\hat{\beta}_1^+$, and $\hat{\beta}_1^-$ are estimated regression weights. *q* is the number of nodes. Finally, $\hat{y}^+$ and $\hat{y}^-$ are predicted DCCS scores using each model.

To conduct out-of-sample prediction, we first computed the *F* values ($F_{i\to j}^+$ and $F_{i\to j}^-$) for each individual in the testing data consisting of 233 novel subjects (30 % of the total samples). We then integrate the parameter ($\hat{\beta}_0^+$, $\hat{\beta}_0^-$, $\hat{\beta}_1^+$, $\hat{\beta}_1^-$, $\hat{I}_{i\to j}^+$, and $\hat{I}_{i\to j}^-$) estimated from the training data with the *F* values from the testing data to predict DCCS scores ($\hat{y}^+$ and $\hat{y}^-$) for the novel subjects. The prediction was done without any further modeling of the testing data. Finally, we accessed the efficacy and power of the model by correlating the predicated and observed DCCS scores in the training data.

References

1 Finn, E. S. *et al.* Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci* **18**, pp. 1664–1671 (2015).

2 Rosenberg, M. D. *et al.* A neuromarker of sustained attention from whole-brain functional connectivity. *Nat Neurosci* **19**, pp. 165–171 (2016).

3 Friston, K. J. Functional and effective connectivity in neuroimaging: A synthesis. *Human brain mapping* **2**, pp. 56–78 (1994).

4 Friston, K. J. Functional and effective connectivity: A review. *Brain connectivity* **1**, pp. 13-36 (2011).

5 Hillebrand, A. *et al.* Direction of information flow in large-scale resting-state networks is frequency-dependent. *Proc Natl Acad Sci* **113**, pp. 3867–3872 (2016).

6 Stam, C. J. & van Straaten, E. C. W. Go with the flow: Use of a directed phase lag index (dPLI) to characterize patterns of phase relations in a large-scale model of brain dynamics. *NeuroImage* **62**, pp. 1415–1428 (2012).

7 Moon, J.-Y., Lee, U., Blain-Moraes, S. & Mashour, G. A. General relationship of global topology, local dynamics, and directionality in large-scale brain networks. *PLOS Computational Biology* **11**, e1004225 (2015).

8 Yan, C. & He, Y. Driving and driven architectures of directed small-world human brain functional networks. *PLoS One* **6**, e23460 (2011).

9 Liao, W. *et al.* Evaluating the effective connectivity of resting state networks using conditional Granger causality. *Biological cybernetics* **102**, pp. 57–69 (2010).

10 Bullmore, E. & Sporns, O. The economy of brain network organization. *Nat Rev Neurosci* **13**, pp. 336–349 (2012).

11 Bertolero, M. A., Yeo, B. T. & D'Esposito, M. The modular and integrative functional architecture of the human brain. *Proc Natl Acad Sci* **112**, pp. E6798–6807 (2015).

12 Shokri-Kojori, E., Motes, M. A., Rypma, B. & Krawczyk, D. C. The network architecture of cortical processing in visuo-spatial reasoning. *Sci Rep* **2**, 411 (2012).

13 Harding, I. H., Yucel, M., Harrison, B. J., Pantelis, C. & Breakspear, M. Effective connectivity within the frontoparietal control network differentiates cognitive control and working memory. *Neuroimage* **106**, pp. 144–153 (2015).

14 Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, pp. 1181–1185 (2003).

15 Posner, J. *et al.* Abnormal amygdalar activation and connectivity in adolescents with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* **50**, pp. 828–837 (2011).

16 Schlosser, R. G. *et al.* Fronto-cingulate effective connectivity in obsessive compulsive disorder: a study with fMRI and dynamic causal modeling. *Hum Brain Mapp* **31**, pp. 1834–1850 (2010).

17 Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* **82**, pp. 403–415 (2013).

18 Geweke, J. Measurement of linear-dependence and feedback between multiple time-series. *J Am Stat Assoc* **77**, pp. 304–313 (1982).

19 Wiener, N. The theory of prediction. *Modern Mathematics for the Engineer*, edited by E.F. Beckenbach. McGraw-Hill, New York (1956).

20 Granger, C. W. J. Economic processes involving feedback. *Inform Control* **6**, pp. 28–48 (1963).

21 Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, pp. 424–438 (1969).

22 Akaike, H. A new look at the statistical-model identification. *IEEE Transactions on Automatic Control*, **19**, pp. 716–723 (1974).

23 Zeki, S. Artistic creativity and the brain. *Science* **293**, pp. 51–52 (2001).

24 Zelazo, P. D. The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nat. Protocols* **1**, pp. 297–301 (2006).

25 Benjamin, D. J. *et al.* Redefine statistical significance. *Nature Human Behaviour* **2**, pp. 6–10 (2018).

26 Ioannidis, J. P. A. The proposal to lower P value thresholds to .005. *JAMA* **319**, pp. 1429–1430 (2018).

27 Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med* **34**, pp. 537–541 (1995).

28 Greicius, M. D., Krasnow, B., Reiss, A. L. & Menon, V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc Natl Acad Sci* **100**, pp. 253–25 (2003).

29 Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S. & Petersen, S. E. Intrinsic and task-evoked network architectures of the human brain. *Neuron* **83**, pp. 238-251 (2014).

30 Fox, M. D., Snyder, A. Z., Zacks, J. M. & Raichle, M. E. Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nat Neurosci* **9**, pp. 23–25 (2006).

31 Fox, M. D. *et al.* The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci* **102**, pp. 9673–9678 (2005).

32 Fox, M. D., Zhang, D., Snyder, A. Z. & Raichle, M. E. The global signal and observed anticorrelated resting state brain networks. *Journal of Neurophysiology* **101**, pp. 3270–3283 (2009).

33 Wu, C. W. *et al.* Frequency specificity of functional connectivity in brain networks. *NeuroImage* **42**, pp. 1047–1055 (2008).

34 Whitfield-Gabrieli, S. & Ford, J. M. Default mode network activity and connectivity in psychopathology. *Annual Review of Clinical Psychology* **8**, pp. 49–76 (2012).

35 Greicius, M. D., Srivastava, G., Reiss, A. L. & Menon, V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proc Natl Acad Sci* **101**, pp. 4637–4642 (2004).

36 Power, Jonathan D., Schlaggar, Bradley L., Lessov-Schlaggar, Christina N. & Petersen, Steven E. Evidence for hubs in human functional brain networks. *Neuron* **79**, pp. 798–813 (2013).

37 van den Heuvel, M. P. & Sporns, O. Network hubs in the human brain. *Trends in Cognitive Sciences* **17**, pp. 683–696 (2013).

38 Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* **24**, pp. 167–202 (2001).

39 Groenewegen, H. J. & Uylings, The prefrontal cortex and the integration of sensory, limbic and autonomic information. *Progress in Brain Research* **126**, pp. 3–28 (2000).

40 Romanski, L. M. Convergence of auditory, visual, and somatosensory information in ventral prefrontal cortex. in *The neural bases of multisensory processes,* edited by M. M. Murray and M. T. Wallace. CRC Press/Taylor & Francis, Boca Raton, Floriada, USA (2012).

41 Mesulam, M. M. From sensation to cognition. *Brain* **121**, pp. 1013–1052 (1998).

42 Laumann, Timothy O. *et al.* Functional system and areal organization of a highly sampled individual human brain. *Neuron* **87**, pp. 657–670 (2015).

43 Mueller, S. *et al.* Individual variability in functional connectivity architecture of the human brain. *Neuron* **77**, pp. 586–595 (2013).

44 Gómez-Robles, A., Hopkins, W. D., Schapiro, S. J. & Sherwood, C. C. Relaxed genetic control of cortical organization in human brains compared with chimpanzees. *Proc Natl Acad Sci* **112**, pp. 14799–14804 (2015).

45 Arnsten, A. F. T. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat Rev Neurosci* **10**, pp. 410–422 (2009).

46 Aoki, R. *et al.* Relationship of negative mood with prefrontal cortex activity during working memory tasks: An optical topography study. *Neuroscience Research* **70**, pp. 189–196 (2011).

47 Soares, J. M., Marques, P., Magalhães, R., Santos, N. C. & Sousa, N. The association between stress and mood across the adult lifespan on default mode network. *Brain Structure and Function* **222**, pp. 101–112 (2017).

48 Laumann, T. O. *et al.* Functional system and areal organization of a highly sampled individual human brain. *Neuron* **87**, pp. 657–670 (2015).

49 Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* **3**, pp. 201–215 (2002).

50 Cole, M. W., Repovš, G. & Anticevic, A. The frontoparietal control system. *The Neuroscientist* **20**, pp. 652–664 (2014).

51 Ridderinkhof, K. R., Ullsperger, M., Crone, E. A. & Nieuwenhuis, S. The role of the medial frontal cortex in cognitive control. *Science* **306**, pp. 443–447 (2004).

52 Ridderinkhof, K. R., van den Wildenberg, W. P. M., Segalowitz, S. J. & Carter, C. S. Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition* **56**, pp. 129–140 (2004).

53 Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. Conflict monitoring and cognitive control. *Psychological review* **108**, pp. 624–652 (2001).

54 Euston, David R., Gruber, Aaron J. & McNaughton, Bruce L. The role of medial prefrontal cortex in memory and decision making. *Neuron* **76**, pp. 1057–1070 (2012).

55 Seth, A. K., Barrett, A. B. & Barnett, L. Granger causality analysis in neuroscience and neuroimaging. *J Neurosci* **35**, pp. 3293–3297 (2015).

56 Stokes, P. A. & Purdon, P. L. A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proc Natl Acad Sci* **114**, pp. E7063–E7072 (2017).

57 Hutchison, R. M. *et al.* Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage* **80** pp. 360–378 (2013).

58 Bassett, D. S. *et al.* Dynamic reconfiguration of human brain networks during learning. *Proc Natl Acad Sci* **108**, pp. 7641–7646 (2011).

59 Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, pp. 150–157 (2001).

60 Glasser, M. F. *et al.* The human connectome project's neuroimaging approach. *Nature Neuroscience* **19**, pp. 1175–1187 (2016).

61 Cao, H. *et al.* Altered functional subnetwork during emotional face processing: A potential intermediate phenotype for schizophrenia. *JAMA psychiatry* **73**, pp. 598–605 (2016).

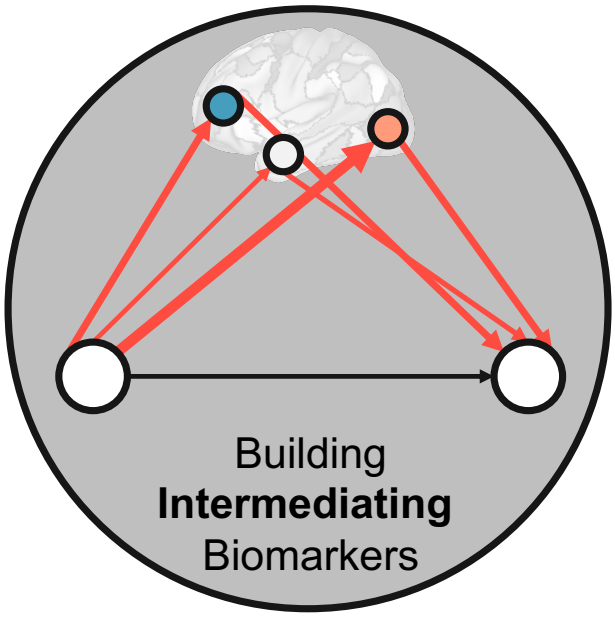
62 Cao, H. *et al.* The 5-HTTLPR polymorphism affects network-based functional connectivity in the visual-limbic system in healthy adults. *Neuropsychopharmacology* **43**, pp. 406–414 (2017).

63 Cao, H. *et al.* Toward leveraging human connectomic data in large consortia: Generalizability of fMRI-based brain graphs across sites, sessions, and paradigms. *Cereb Cortex* 29, pp. 1263–1279(2018).

64 Cao, H. *et al.* Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. *Nat Commun* **9**, 3836 (2018).

Chapter 4

Building intermediate biomarkers



Identifying Neural Signatures Mediating Behavioral Symptoms and Psychosis Onset: High-Dimensional Whole-Brain Functional Mediation Analysis[1]

Abstract

Along the pathway from behavioral symptoms to the development of psychotic disorders sits the multivariate mediating brain. The functional organization and structural topography of large-scale multivariate neural mediators among patients with brain disorders, however, are not well understood. Here, we design a high-dimensional brain-wide functional mediation framework to investigate brain regions that intermediate between baseline behavioral symptoms and future conversion to full psychosis among individuals at clinical high risk (CHR). Using resting-state functional magnetic resonance imaging (fMRI) data from 263 CHR subjects, we extract an $\boldsymbol{\alpha}$ brain atlas and a $\boldsymbol{\beta}$ brain atlas: the former underlines brain areas associated with prodromal symptoms, and the latter highlights brain areas associated with disease onset. In parallel, we identify and separate mediators that potentially positively and negatively mediate symptoms and psychosis, respectively, and quantify the effect of each neural mediator on disease development. Taken together, these results paint a brain-wide picture of neural markers that are potentially mediating behavioral symptoms and the development of psychotic disorders; additionally, they underscore a statistical framework that is useful to uncover large-scale intermediating variables in a regulatory biological system.

[1]This chapter is published as: Chén, O.Y., Cao, H., Phan, H., Nagels, G., Reinen, J.M., Gou, J., Qian, T., Di, J., Prince, J., Cannon, T.D. and de Vos, M., 2021. Identifying neural signatures mediating behavioral symptoms and psychosis onset: High-dimensional whole-brain functional mediation analysis. *NeuroImage* **226**, 117508.

1. Introduction

How does the human brain intermediate between behavioral symptoms and the development of brain diseases? Which brain areas are involved in this process? Can we chart these areas' functional characteristics and structural organization?

Researchers studying brain diseases often observe that brain signals are on the one hand associated with behavioral symptoms, and on the other hand linked to disease status. Conventionally, the former is called an explanatory variable, the latter is called a dependent variable (or an outcome), and the brain areas interposed in-between are called mediators. A central problem in neural mediation analysis is to identify which brain regions are positioned along the pathway between behavioral symptoms and disease status. Equally important is to quantify the effect of each identified brain area on developing the disease and to determine its relative prominence in the mediation system.

Disorganization symptoms, such as bizarre thoughts and behaviors, are considered to be associated with conversion to psychosis among individuals at clinical high risk (CHR); empirical studies have shown a significantly higher hazard ratio for psychosis onset in CHR subjects with higher disorganization symptoms at baseline[1–3]. Yet, as properties associated with a mental disorder, the disorganization symptoms and disease development are reflected by the measured brain signals. Probing into the neural basis of human behavior and disease development, mediation analysis can help us to understand the functional attributes and structural topography of the brain areas that potentially mediate behavioral symptoms and disease development. But it can only do so by first charting the neural pathways that make brain mediation possible.

A beginning in this direction can be made by identifying and isolating neural mediators that are interposed between behavioral symptoms and disease development (see **Fig.** 1). Central to this enquiry is a high-dimensional brain mediation analysis: examining hundreds of thousands of brain areas to find a subset of potential mediators. To uncover high-dimensional functional neural mediators with binary outcomes (*e.g.,* whether one has a full-blown psychotic disorder or not), one must confront several challenges. First, although existing mediation models have made the search for mediators fruitful, they are not suitable for studying high-dimensional mediation analysis with binary disease outcomes. For example, existing multi-level mediation models assume that the outcomes are continuously distributed[4–7]; mediation frameworks concerning binary outcomes are at present restricted to a relatively small number of mediators[6,8]; high-dimensional mediation models whose outcomes are not normally distributed do not have a closed-form solution (therefore it is difficult to estimate parameters analytically, as, for example, in Chén *et al.*[4]). Second, although functional mediation analysis[9] has considerably advanced knowledge about the functional signal organization of the brain in relation to explanatory and outcome variables, it remains unclear whether it is suitable for analyzing high-dimensional brain data, and if so, how the underlying data configuration, such as the sample size and noise level, would affect parameter estimation. In parallel, its efficacy needs to be evaluated for brain disease studies. Third, signals from brain mediators could be orthogonal or nonorthogonal. Whether and how their orthogonality would affect mediation analysis is an as-of-yet less-well-charted area. If not properly treated, this set of circumstances could generate inconsistent methodological results and confusing interpretations. Finally, the search for functional neural mediators among subjects with severe behavioral symptoms raises the question of which mediators are positively, and which are negatively, mediating the development of brain disorders.

To address these questions, here we propose a high-dimensional functional mediation model. Through simulation studies and empirical data analysis, we demonstrate that the model may be useful to (a) analyze large-scale intermediating brain signals (*e.g.,* resting-state brain activities from hundreds of thousands of voxels); (b) distinguish distinctive functional brain regions between disease groups in relation to behavior symptoms; (c) quantify each neural mediator's effect on disease outcome; (d) identify and separate brain areas that are potentially positively and negatively mediating brain disorders.

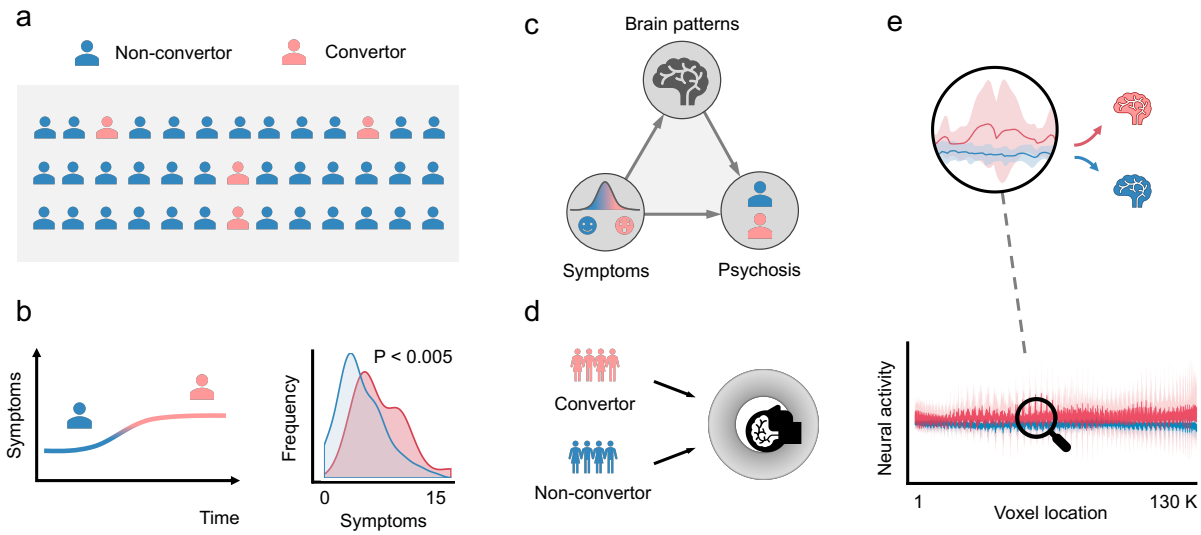


Fig. 1. The study layout of the neural mediation analysis. (a) We considered a sample of 263 subjects recruited from eight study sites across the United States and Canada who met criteria for a prodromal risk syndrome at the point of recruitment and had been clinically followed up for two years as part of the NAPLS-2 project. During the follow-up period, 25 subjects developed a full-blown psychotic disorder (CHR convertors); 238 did not (CHR non-convertors). **(b)** The behavioral symptoms of convertors were significantly more severe than those of non-convertors. **(c)** The neural mediation analysis investigated which brain regions were intermediating between psychosis symptoms and disease status. Once neural mediators were identified, one could further quantify the mediation effect of each mediator to determine its relative prominence in the mediation system. **(d)** Both convertors and non-convertors received an eyes-open resting-state functional magnetic resonance imaging (fMRI) scan at the point of recruitment. **(e)** Resting-state brain activities from both convertor and non-convertor samples were plotted along 130,992 brain areas. The red shade represented brain signals stacked across the convertor group across the whole-brain space and the blue shade represented those from the non-convertor group.

In clinical practice, one assumes that changing brain signals can first cause prodromal signs and symptoms, followed in some cases by later conversion to psychosis. In this paper, we aimed at studying the influence of the underlying brain signals on the link between two directly

and clinically observable sets of variables: prodromal signs and symptoms on the one hand, and conversion state on the other hand. The framework we designed to map the pathways contained directed arrows. The arrows clarified that the statistical model was a mediation model; they did not suggest definitive causal flows from prodromal signs via brain areas towards conversion status (see **Figs.** 1 and 2). When confusion about the causal direction existed, one can interpret the identified neural mediators as brain areas that are jointly associated with behavioral symptoms and psychosis conversion. In other words, the neural mediators exclude brain areas that are associated with conversion, but that are not associated with prodromal symptoms, and *vice versa*. One should also note that in cases where the brain signals first have an effect on the symptoms and then on the disease status (namely when the symptoms are the mediator), the identified brain regions are identical to the alpha atlas estimated by the proposed model (see **Results**). Although there are overlaps between the two models, the interpretations are different. The key differences between brain areas identified by these two models are (1) the identified brain areas in the current study are potential neural mediators whereas those identified using the other model (where the symptom is treated as the mediator) are multivariate explanatory variables; (2) the neural mediators from the present study are a subset of brain areas identified using the other model. In the **Supporting Information**, we extend the proposed model to causal mediation setting using counterfactuals; additionally, we discuss how to interpret the model when causal inference is concerned; in cases where brain signals can be manually controlled (*e.g.*, via transcranial magnetic stimulation (TMS)), under some identification conditions, the proposed model may unveil potential causal direct and indirect effect on the odds ratio scale.

In the following, we begin with a brief overview of the mediation analytical frameworks concerning univariate and multivariate mediators. After discussing these basic concepts, we

introduce the high-dimensional functional mediation framework. To demonstrate its utility, we perform both simulation and case studies. During the simulation study, we consider various experimental settings, including different levels of noise, sample sizes, and both orthogonal and non-orthogonal basis functions, to ensure that the proposed framework is suitable for studying high-dimensional functional mediation. During the case study, we uncover brain areas that potentially mediate psychosis symptoms and disease status from whole-brain resting-state functional magnetic resonance imaging (rs-fMRI) data obtained from 263 subjects at clinical high risk (CHR) for psychosis.

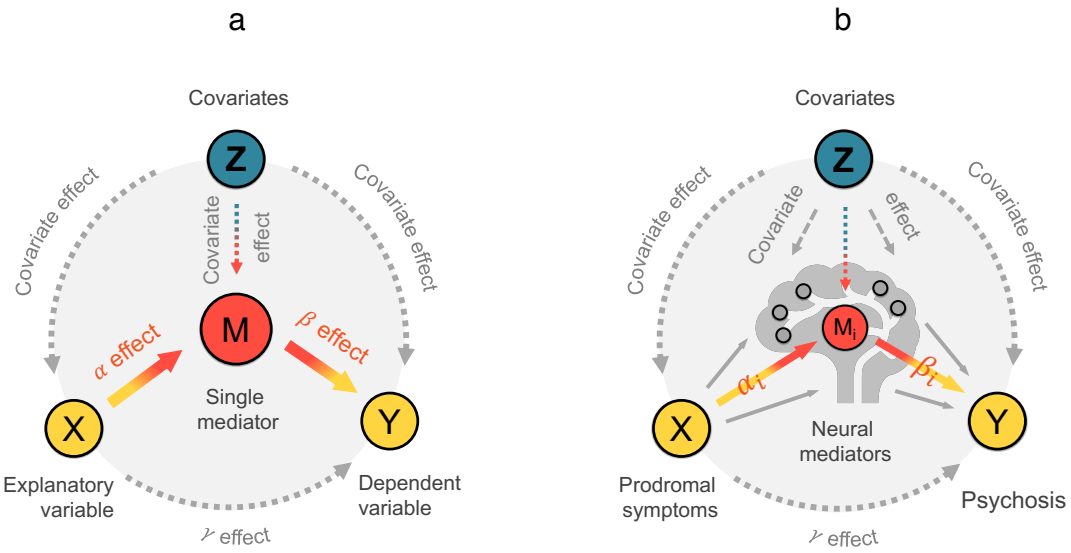


Fig. 2. A schematic representation of mediation analysis. (a) Univariate mediation analysis. The circles indicate an explanatory variable, a univariate mediator, an outcome variable, and covariates. The arrows denote pathways. The letter α denotes the effect from the explanatory variable to the mediator, after accounting for the covariate effect. The letter β denotes the effect of the mediator on the outcome, after controlling the explanatory variable and covariates. The letter γ denotes the effect from the explanatory variable to the outcome, after accounting for the covariate effect. **(b) Multivariate neural mediation analysis.** Each circle within the brain represents a potential neural mediator. The arrows denote pathways. The letter α_i $(1 \leq i \leq V)$ denotes the effect from the explanatory variable to the i^{th} neural mediator (represented by a red circle). The letter β_i denotes the effect of the neural mediator on the outcome, after controlling the explanatory variable and covariates. The letter γ indicates the direct effect from the explanatory variable to the outcome, after accounting for the covariate effect.

1.1 Univariate mediation analysis

Univariate mediation analysis considers a single mediator (M) (see **Fig.** 2 (a)). In other words, a variable M is a mediator if, after accounting for covariates $\boldsymbol{Z}$, the effect of an explanatory variable X on an outcome variable Y is at least partially carried through M[10,11]. Examples of

univariate mediators are pain catastrophizing, which mediates the clinical treatment and disability status[12], and intention, which mediates attitudes and behavior[13].

The identification of a univariate mediator consists of two steps[10,14–17] (see **Supporting Information** for a comparison between two common mediation analysis frameworks). The first step examines if the explanatory variable X affects the mediator M after controlling for the covariates $\boldsymbol{Z}$, using the following conditional model:

$$\mathbb{E}(M|x,\boldsymbol{z}) = \theta_0 + \alpha x + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{t} \qquad \text{(i)}$$

where $\mathbb{E}$ refers to the expectation operation; θ_0, α, $\boldsymbol{t}$ are coefficients for the intercept, the observed explanatory variable x, and the observed covariates $\boldsymbol{z}$. If α is significantly different from zero, then the explanatory variable affects the mediator.

The second step evaluates if the mediator M affects the outcome Y, after controlling for the explanatory variable X and covariates $\boldsymbol{Z}$, using the following conditional model:

$$\mathbb{E}(Y|x,m,\boldsymbol{z}) = \theta_0' + \gamma x + \beta m + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\tau} \qquad \text{(ii)}$$

where θ_0', γ, β and $\boldsymbol{\tau}$ are coefficients for the intercept, the observed explanatory variable x, the mediator m, and the observed covariates $\boldsymbol{z}$. The prime in θ_0' is to differentiate it from their counterparts in (i). If β is significantly different from zero, then the mediator affects the outcome.

The univariate variable M is said to significantly mediate the relationship between X and Y, if both α and β are significantly different from zero. Said differently, if the product $\alpha\beta$ is significantly different from zero, then M is a mediator for X and Y, and the product value $\alpha\beta$ quantifies the mediation effect. In the language of a graphical model, this means that both α and β edges in **Fig.** 2 (a) exist, connecting nodes X and Y via a pathway passing through node M.

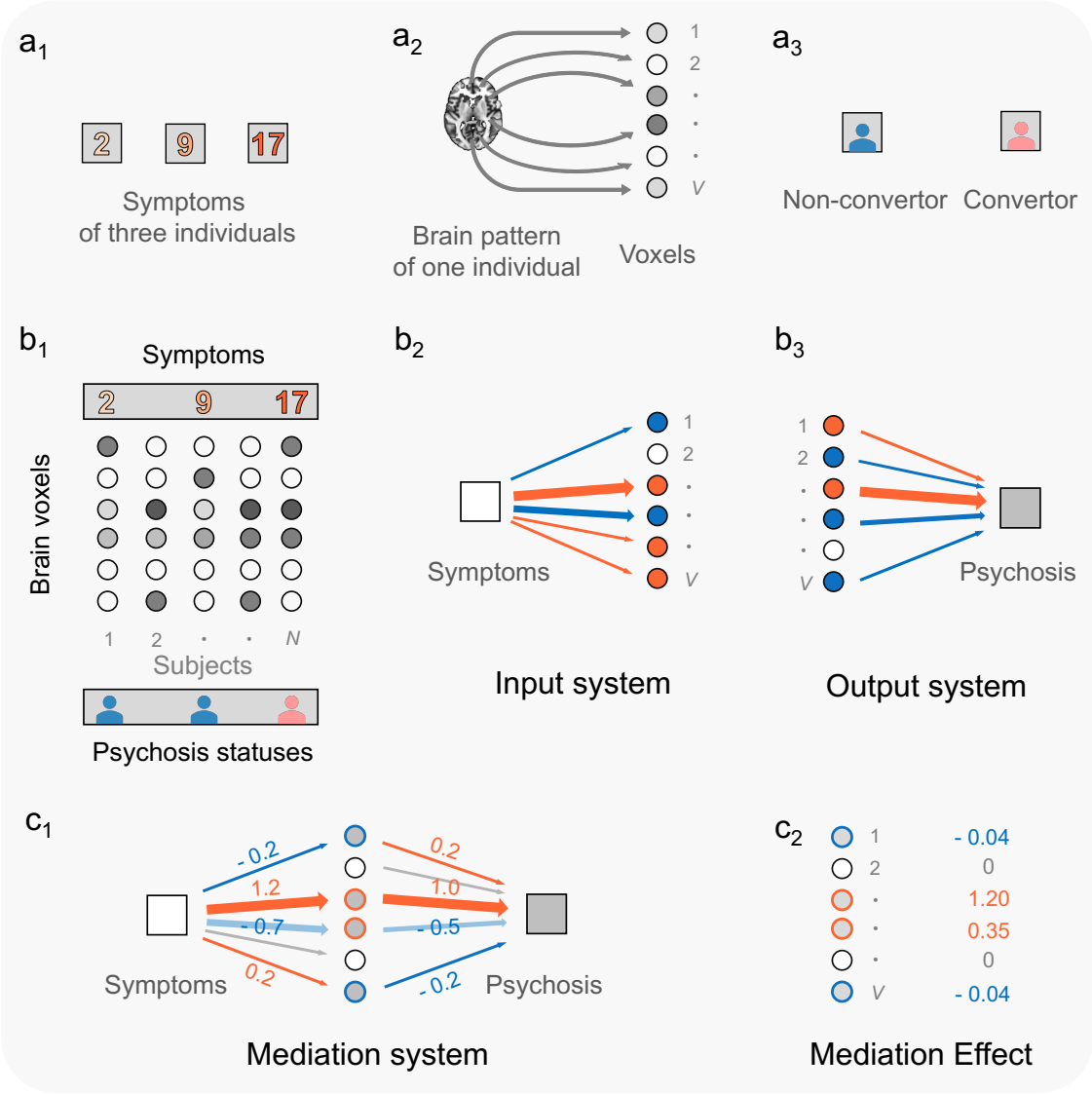


Fig. 3. A hypothetical experiment and how multivariate mediation analysis can be used to study brain mediation in health and disease. **(a1)** Three individuals' behavioral symptom scores are measured at baseline. **(a2)** Resting-state brain activities for one individual are collapsed into activities averaged over time across voxels in the whole brain. Each circle corresponds to one voxel. **(a3)** Individuals' two-year clinical outcome for psychosis. We use 0 to refer to a non-convertor and 1 for a convertor. **(b_1)** Measured individual brain signals are arranged corresponding to their behavioral symptom scores and psychosis statuses. Each column contains data from a particular subject. **(b_2)** The input system of the brain mediation framework studies the association between the behavioral symptom score (the box) and measured brain signals (the circles). The colored arrows indicate significant pathways from the behavioral symptom score to brain areas. The width of the arrows indicates effect size, and the color indicates a sign (where orange means positive and blue means negative). **(b_3)** The output system of the brain mediation framework studies the association between the measured brain signals (the circles) and disease status (the box). **(c1)** The mediation analysis framework combines the input and output systems, and studies how the effect of behavioral symptom score (the left box) on the psychosis status (the right box) is intermediated by neural mediators (the circles). A voxel significantly mediates the relationship if its signal is associated with both the behavioral symptom score and the psychosis status. **(c_2)** The mediation effect of a particular voxel is calculated by multiplying the effect sizes from the input and the output pathway corresponding to the voxel.

1.2 Multivariate mediation analysis

Mediation analysis concerning a multivariate mediator can be conducted using structural equation models (SEMs)[6,9] (see **Fig.** 2 (b)). Formally, consider V mediators ($V \geq 2$), denoted as $M(1), M(2), \ldots, M(V)$, an explanatory variable X, and an outcome variable Y. Multivariate mediation analysis considers two conditional models as follows.

$$\mathbb{E}(M(j)|x, \boldsymbol{z}) = \theta_0(j) + \alpha(j)x + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{t}(j), \quad j = 1, 2, \ldots, V \qquad \text{(iii)}$$

where $M(j)$ is the j^{th} mediator; $\theta_0(j)$, $\alpha(j)$, and $\boldsymbol{t}(j)$ are coefficients for the intercept, the observed explanatory variable x, and the observed covariates $\boldsymbol{z}$ that are associated with the j^{th} mediator.

$$\mathbb{E}(Y|x, \boldsymbol{m}, \boldsymbol{z}) = \theta_0' + \gamma x + \sum_{j=1}^{V} \beta(j)\, m(j) + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\tau} \qquad \text{(iv)}$$

where $\boldsymbol{m} = (m(1), m(2), \ldots, m(V))$ is a vector representing V observed mediators; θ_0', γ, and $\boldsymbol{\tau}$ are coefficients for the intercept, the observed explanatory variable x, and the observed covariates $\boldsymbol{z}$; $\beta(j)$ is the coefficient associated with the j^{th} mediator.

The j^{th} mediator $M(j)$, for $j = 1, 2, \ldots, V$, is said to significantly mediate the relationship between X and Y, if both $\alpha(j)$ and $\beta(j)$ are significantly different from zero after accounting for the covariates. The product, $\alpha(j)\beta(j)$, quantifies the mediation effect for the j^{th} mediator (see **Fig.** 3).

1.3 High-dimensional brain-wide functional mediation

High-dimensional mediation analysis aims at identifying mediators from a high-dimensional multivariate variable. For example, a neurobiologist is interested in searching through the entire brain to look for neural mediators using signals recorded from hundreds of thousands of brain regions. The framework introduced in the paper consists of a dual system: the input system

investigates how an explanatory variable (*e.g.*, behavioral symptoms) may be associated with signals of brain mediators, after controlling for covariates; the output system examines how signals of brain mediators may give rise to the outcome variable (*e.g.*, psychosis disease status), after controlling for the explanatory variable and covariates (see **Fig.** 3). For brevity, from now on we will refrain from mentioning covariate control in our writing; readers should note that this has been included during the mediation analysis.

In the following, we introduce the key concepts of the framework and leave derivations and discussions to the **Material and Methods** and **Supporting Information** sections.

Consider N subjects and V brain areas, where V is high-dimensional (in our study $N = 263$ and $V = 130{,}992$). Let x_i and y_i be the explanatory and outcome variables for subject i, respectively. Let $\boldsymbol{z}_i = (z_{1i}, z_{2i}, z_{3i}, z_{4i}, z_{5i})$ denote the covariates of subject i, for example, the site (from which data are collected), age, gender, frame-wise displacement (FD), and whole-brain gray matter volume, respectively. Finally, let $m_i(j)$ be the neural activity from the j^{th} brain area of subject i. The high-dimensional functional brain-wide mediation framework consists of an input system and an output system.

The following conditional model describes the input system (see **Fig.** 3 b_2):

$$\mathbb{E}(m_i(j)|x_i, \boldsymbol{z}_i) = \theta_0(j) + x_i\alpha(j) + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{t}(j) \qquad \text{(v)}$$

where $\theta_0(j)$, $\alpha(j)$, and $\boldsymbol{t}(j)$ are coefficients for the intercept, the explanatory variable, and covariates that are associated with the j^{th} mediator. It is worthwhile mentioning that the estimated parameters $\boldsymbol{\alpha} = \big(\alpha(1), \alpha(2), \dots, \alpha(V)\big)$ (see the alpha atlas in the **Results** section) are closely related to canonical correlation (CCA) between brain signals $\boldsymbol{M}$ and the explanatory

variable X and are remotely related to the Partial Least Squares (PLS) estimates (see **Supporting Information** for details).

The following generalized functional linear model (a generalized linear model containing functional PCs in its regressors) represents the output system (see **Fig.** 3 b_3):

$$\mathbb{E}(y_i|\boldsymbol{\Delta}_i) = g^{-1}\left(\beta_0 + x_i\gamma + \sum_{j=1}^{V}\sum_{k=1}^{\infty}\xi_{ik}\varphi_k(j)\beta(j) + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\tau}\right) \qquad \text{(vi)}$$

where β_0, γ, and $\boldsymbol{\tau}$ are coefficients for the intercept, the explanatory variable, and covariates. Here, ξ_{ik} and $\varphi_k(j)$ are the Karhunen-Loève expansion[18,19] of $m_i(j)$, the j^{th} mediator of subject i; in other words $m_i(j) = \sum_{k=1}^{\infty}\xi_{ik}\varphi_k(j) \approx \sum_{k=1}^{K}\xi_{ik}\varphi_k(j)$ (without loss of generality, assume $m_i(j)$ is zero centered), where $\xi_{ik}\sim N(0,\lambda_k)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\infty$, for $i \in \{1,2,\ldots,N\}$, $j \in \{1,2,\ldots,V\}$, and K is a finite integer such that $\sum_{k=1}^{K}\xi_{ik}\varphi_k(j)$ captures the importance of modes of variations of $m_i(j)$ (see **Material and Methods** for more detailed explanation). Additionally, $\boldsymbol{\varphi} = \{\varphi_1,\varphi_2,\ldots,\varphi_\infty\}$ denotes the basis functions and $\beta(j)$ is the coefficient associated with the j^{th} mediator. The link function $g(\cdot)$ takes various forms based on the outcome distribution. For example, when y_i is Gaussian, binary, or Poisson, the link function is $g(y) = y$, $g(y) = \log\left(\frac{y}{1-y}\right)$, or $g(y) = \log(y)$, respectively.

The first K basis functions or $\{\varphi_1,\varphi_2,\ldots,\varphi_K\}$ in **Eq.** (vi) are functional representations of K dominant population-specific (*i.e.*, shared by all N subjects) information of brain data, ranked decreasingly (according to $\{\lambda_1,\lambda_2,\ldots,\lambda_K\}$) based on the amount of information each basis function explains about the neural mediator $\boldsymbol{M}$. Although by Mercer's theorem (see Chapter 4 of Indritz[20]), the basis functions are orthogonal, and researchers indeed are oftentimes interested in uncovering orthogonal brain signals in relation to mediation, it remains possible that the brain signals consist of non-orthogonal information. Despite the central focus herein

being on mediation studies, it is important to understand how the underlying orthogonality of brain data's basis functions may affect the identification of neural mediators and the estimation of the mediation effect. To that end, we performed simulation studies using both orthogonal and non-orthogonal basis functions with different noise levels and sample sizes, and the results showed that our framework was successful to uncover neural mediators regardless of the orthogonality. Naturally, the mediation analysis performance was better when the underlying basis functions were orthogonal, and the estimation results improved as the noise decreased or the sample size increased (see **Supporting Information**).

We summarize the key steps of the framework as follows. First, it estimates the effect of the explanatory variable on each brain area; this yields an $\boldsymbol{\alpha}$ brain atlas (see **Fig.** 3 (b_2)). Next, it extracts subject-specific principal component (PC) scores $\boldsymbol{\xi}_i = \{\xi_{i1}, \xi_{i2}, \ldots, \xi_{iK}\}$ from each individual i's brain signals ($\boldsymbol{m}_i$), and estimates the effect of the transformed lower-dimensional mediators (*i.e.,* $\boldsymbol{\xi}_i$) on the disease outcome y_i after controlling for the explanatory variable and the covariates. Subsequently, the low-dimensional mediator-on-outcome effect is translated to the high-dimensional brain space using the estimated brain-wide basis functions (*i.e.*, $\boldsymbol{\varphi}$); this produces a $\boldsymbol{\beta}$ brain atlas (see **Material and Methods** and **Supporting Information**). Finally, it obtains the brain-wide mediation effect using the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ brain atlases and bootstrap tests (see **Fig.** 3 (c_1) - (c_2) and **Methods**).

2. Material and methods

2.1 The NAPLS-2 Data

The fMRI data were drawn from the second phase of the North American Prodrome Longitudinal Study (NAPLS-2) consortium[21], which included 263 subjects recruited from eight

study sites across the United States and Canada. All subjects met the criteria for the prodromal syndromes at the point of recruitment according to the Structured Interview for Prodromal Syndromes (SIPS)[22] and were clinically followed-up for two years. During follow-up, 25 subjects developed one type of Axis-I psychotic disorders (CHR convertors, age 18.52 ± 4.08 years, 17 males) and 238 did not (CHR non-convertors, 19.07 ± 4.16 years, 136 males). The conversion was defined as the individual either met the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV)[23] criteria for an Axis-I psychotic disorder or had at least one fully psychotic symptom assessed by the SIPS at follow-up[24,25]. See **Table** 1 for details on the studied sample. All subjects received an eyes-open resting-state functional magnetic resonance imaging (fMRI) scan at the point of recruitment.

2.2 Data Acquisition

During the 5-min eyes-open resting-state scan (154 whole-brain volumes), participants were asked to lay still in the scanner, relax, gaze at a fixation cross, and not engage in any particular mental activity. After the scan, investigators confirmed with the participants that they had not fallen asleep in the scanner. Data were acquired from 3T MR scanners located at eight study sites with identical fMRI protocols. Siemens scanners were used at Emory, Harvard, University of California Los Angeles (UCLA), the University of North Carolina at Chapel Hill (UNC), and Yale; GE scanners were used at Calgary, University of California San Diego (UCSD), and Zucker Hillside Hospital (ZHH). Functional images were collected using gradient-recalled-echo echo-planar imaging (GRE-EPI) sequences: TR/TE 2000/30 ms, 77° flip angle, 30 4-mm slices, 1-mm gap, 220-mm FOV. In addition, we also acquired high-resolution T1-weighted images for each participant with the following sequence: 1) Siemens scanners: magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence with 256 mm x 240 mm x 176 mm FOV,

TR/TE 2300/2.91 ms, 9° flip angle; 2) GE scanners: spoiled gradient recalled echo (SPGR) sequence with 260 mm FOV, TR/TE 7.0/minimum full ms, 8° flip angle.

2.3 Preprocessing of rs-fMRI Data

Data were preprocessed using the standard pipeline implemented in the Statistical Parametric Mapping (SPM12, http://www.fil.ion.ucl.ac.uk/spm/) software following previously published work[26–29], including slice-timing correction, realignment, individual structural-functional image coregistration, normalization to the Montreal Neurological Institute (MNI) template, and spatial smoothing with 8-mm full width at half maximum (FWHM). Preprocessed images were further corrected for white matter, cerebrospinal, and global signals, 24 head motion parameters (6 translation and rotation parameters, their first derivatives, and the square of these 12 parameters), and frame-wise displacement (FD). To mitigate the potential head motion effect, 8 subjects (1 converter, 7 non-converters) with mean FD > 0.35mm were excluded from further analysis. The activities of each gray matter voxel during resting state were quantified by the fractional amplitude of low-frequency fluctuation (fALFF), a well-established measure evaluating the ratio of the power spectrum of low-frequency (0.01–0.1 Hz) to that of the entire frequency range[30] . After computation, fALFF values were extracted from a total of 130,992 voxels covering the gray matter of the entire brain for each subject. Secondary data analysis was conducted using *R* software.

2.4 The Model

Here we introduce the high-dimensional functional mediation analysis framework. The input system of the framework tests if the explanatory variable (*e.g.,* behavioral symptoms) affects each mediator (see **Fig.** 3 (b_2)). The output system examines if each mediator affects the outcome (see **Fig.** 3 (b_3)).

Formally, consider N subjects and V brain areas, where $N = 263$ and $V = 130{,}992$. Let $x_i \in \mathbb{R}$ be the explanatory variable for subject i, $\boldsymbol{z}_i \in \mathbb{R}^5$ be the covariates consisting of the site (from which data are collected), age, gender, frame-wise displacement (FD), and whole-brain gray matter volume, respectively, of subject i, $\boldsymbol{m}_i \in \mathbb{R}^V$ be the measured brain signals of subject i spanning V brain areas, and y_i be the outcome for subject i (in this study $y_i \in \{0,1\}$).

We first review the basics of functional principal component analysis (fPCA)[31,32]. Let $m(j), j \in [0,1]$, be a squared integrable random function with mean $\mu(j)$ and covariance function $K(s,t)$. In other words, $\mu(j) = E(m(j))$ and $K(s,t) = cov(m(s), m(t))$. By Mercer's theorem (see Chapter 4 of Indritz[20]), one can obtain the spectral decomposition of $K(s,t)$ as:

$$K(s,t) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(s) \varphi_k(t)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\infty$ are decreasingly ordered nonnegative eigenvalues and φ_k's are their corresponding orthogonal eigenfunctions with unit $\mathcal{L}^2$ norms.

Karhunen-Loève expansion[18, 19] of the random function $m(j)$ yields:

$$m(j) = \mu(j) + \sum_{k=1}^{\infty} \xi_k \varphi_k(j)$$

where $\xi_k = \int_0^1 \{m(j) - \mu(j)\} \varphi_k(j) dj$ are uncorrelated random variables with zero mean and variance λ_k. For a given functional sample, the mean function $\mu(j)$ and covariance function $K(s,t)$ can be consistently estimated using the method of moments. The eigen-values and -functions are estimated from the empirical covariance function, and the principal component scores (ξ_k's) can be estimated by numeric integration.

Next, we enquire further into **Eqs.** (v) and (vi) from the **Introduction** section. Resume the notations used in **Eqs.** (v) and (vi). Consider N subjects and V brain areas, where V is high-dimensional. Let x_i and y_i be the explanatory and outcome variables for subject i, respectively. Let $\boldsymbol{z}_i = (z_{1i}, z_{2i}, z_{3i}, z_{4i}, z_{5i})$ denote the covariates of subject i, for example, the site (from which data are collected), age, gender, frame-wise displacement (FD), and whole-brain gray matter volume, respectively. Finally, let m_{ij} be the neural activity from the j^{th} brain area of subject i.

The input system consists of the following model:

$$E(m_{ij}|x_i, \boldsymbol{z}_i) = \theta_0(j) + x_i\alpha(j) + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{t}(j)$$

where $\theta_0(j)$, $\alpha(j)$, and $\boldsymbol{t}(j)$ are coefficients for the intercept, the explanatory variable, and covariates that are associated with the j^{th} mediator. Specifically, $\alpha(j)$ captures the effect of the explanatory variable on the j^{th} mediator, $\theta_0(j)$ indicates an intercept, and $\boldsymbol{t}(j)$ denotes the coefficients for the covariates with respect to the j^{th} mediator. Without loss of generality, consider $j \in \{\frac{1}{V}, \frac{2}{V}, \dots, \frac{V}{V}\}$.

The output system consists of the following generalized functional linear model:

$$f(y_i|\,\boldsymbol{\Delta}_i) = \exp\{(y_i\theta_i - a(\theta_i) + b(y_i))\phi\} \qquad \text{(vii)}$$

where $\theta_i = h(\eta_i)$, η_i is called a linear predictor with $\eta_i = \boldsymbol{\Delta}_i\boldsymbol{\omega}$, $\boldsymbol{\Delta}_i = (1, x_i, \boldsymbol{m}_i^{\mathrm{T}}, \boldsymbol{z}_i^{\mathrm{T}})$ denotes the data, $\boldsymbol{m}_i^{\mathrm{T}} = (m_i(1), m_i(2), \dots, m_i(V))$, and $m_i(j) = \sum_{k=1}^{\infty}\xi_{ik}\varphi_k(j) \approx \sum_{k=1}^{K}\xi_{ik}\varphi_k(j)$, where $\xi_{ik} \sim N(0, \lambda_k)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\infty$, for $i \in \{1, 2, \dots, N\}$ and $j \in \{\frac{1}{V}, \frac{2}{V}, \dots, \frac{V}{V}\}$, and $\boldsymbol{\varphi} = \{\varphi_1, \varphi_2, \dots, \varphi_K\}$ is a set of basis functions. $\boldsymbol{\omega} = (\beta_0, \gamma, \boldsymbol{\beta}, \boldsymbol{\tau})^{\mathrm{T}}$ denotes the corresponding parameters for $\boldsymbol{\Delta}_i$. $h(\cdot)$, $a(\cdot)$ and $b(\cdot)$ are proper functions. $\boldsymbol{\beta}$ is a $V \times 1$ vector, whose j^{th} entry $\beta(j)$ estimates the

effect of the j^{th} mediator on the outcome, controlling the explanatory variable and covariates. ϕ is a nuisance parameter.

By taking the expected value of y_i conditioning on $\boldsymbol{\Delta}_i$, **Eq.** (vii) yields **Eq.** (vi) in the **Introduction** section. Particularly, when y_i is binary, (vi) has the following form:

$$y_i = Bernoulli(p_i) \qquad \text{(vii)}$$

where $p_i = \frac{1}{1+\exp\left[-(\beta_0+\gamma x_i+\sum_{j=1}^{V}\sum_{k=1}^{\infty}\xi_{ik}\varphi_k(j)\beta(j)+\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\tau})\right]}$.

Thanks to the spectral decomposition of $\boldsymbol{M}$, parameter estimation of $\boldsymbol{\beta}$ can be performed on $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_N)^{\mathrm{T}}$, where $\boldsymbol{\xi}_i = \{\xi_{i1}, \xi_{i2}, \dots, \xi_{iK}\}$. To see this, rewrite $\beta_0 + \gamma x_i + \sum_{j=1}^{p}\sum_{k=1}^{K}\xi_{ik}\varphi_k(j)\,\beta(j) + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\tau}$ in (vii) as $\beta_0 + \gamma x_i + \sum_{j=1}^{p}\sum_{k=1}^{K}\xi_{ik}\tilde{\beta}_k(j) + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\tau}$, where $\tilde{\beta}_k(j) = \varphi_k(j)\beta(j)$. The estimation problem now translates to estimating the low-dimensional mediator-on-outcome effect, or $\tilde{\boldsymbol{\beta}} = \left(\sum_{j=1}^{p}\tilde{\beta}_1(j), \sum_{j=1}^{p}\tilde{\beta}_2(j), \dots, \sum_{j=1}^{p}\tilde{\beta}_K(j)\right)$. Subsequently, the estimation of $\boldsymbol{\beta}$ can be retrieved by projecting the estimated $\widehat{\tilde{\boldsymbol{\beta}}}$ back to the brain space using the estimated basis functions $\widehat{\boldsymbol{\varphi}}$; in other words, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\varphi}}^{-}\widehat{\tilde{\boldsymbol{\beta}}}$, where the minus sign $\{-\}$ represents a *generalized inverse*.

Although in real-world brain data, the (unknown) basis functions $\boldsymbol{\varphi}$ can be either orthogonal or non-orthogonal, simulation studies showed that, regardless of the orthogonality of basis functions, our framework was successful to uncover the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ brain atlases under different noise levels and sample sizes (see **Supporting Information**). Since $NIE(j) = \alpha(j)\beta(j)$ has a one-to-one relationship to the j^{th} voxel's mediation effect (see **Supporting Information**), for simplicity we estimate $NIE(j)$, for $j = 1, 2, \dots, V$. In words, $NIE(j)$ is the j^{th} voxel's natural indirect mediation effect on the *log* odds-ratio scale *per* unit increase of the explanatory variable.

When V is small, Sobel's Test[17] evaluates the statistical significance of $NIE(j)$; when V is high-dimensional, the statistical significance of $NIE(j)$ can be evaluated using a bootstrap approach[33] (see **Supporting Information**); the results are further adjusted for multiple comparisons.

2.5 Brain-wide functional mediation analysis.

We conducted a brain-wide functional mediation analysis on the rs-fMRI data from the NAPLS-2 sample using the proposed framework. We first assessed if there was any effect from the explanatory variable (SOPS disorganization symptom scores) on each voxel, controlling for age, sex, study site, FD, and whole-brain gray matter volume, using **Eq.** (v) and obtained the $\boldsymbol{\alpha}$ brain atlas. We then estimated the effect each brain area had on the outcome, controlling for the explanatory variable and all covariates, using **Eq.** (vi) and obtained the $\boldsymbol{\beta}$ brain atlas. Finally, we identified positive and negative neural mediators and estimated their mediation effect via bootstrap experiments. We presented the empirical results in **Fig.** 4.

3. Results

3.1 Simulation studies

We conducted simulation studies to ensure that the proposed framework was able to identify brain areas that intermediate the explanatory and the outcome variables under different settings. We first simulated covariates such that their distributions were similar to those from the empirical data. We then simulated symptom data using the covariates. Next, we simulated brain signals using **Eqs.** (v) and (vi) (see more details in the **Supporting Information**). When simulating brain signals, we examined different combinations of eigenfunctions, idiosyncratic noises, and sample sizes. For eigenfunctions, we considered both orthogonal and non-

orthogonal cases (to cover the two general scenarios where the brain signals consist of orthogonal and non-orthogonal information). For noises, we considered a range of magnitudes, from small, moderate, to very large scales; we considered very large noise because we wanted to see to what degree the estimates would be able to (and unable to) uncover the signals, and whether, even under very large noise level, signals could be recovered using more samples. We considered dimensionality of $130{,}992$ (the same as the empirical study) and sample sizes of 100 and 500. Together, we examined 12 simulation conditions; for each condition, we conducted 100 bootstrap simulations. From the estimation performance, one can relatively easily peer into how the model would perform under other combinations of basis functions, noise levels, and sample sizes.

	CHR converters (***N*** = 25)	CHR non-converters (***N*** = 238)	*p*-value
Age (years)	18.52±4.08	19.07±4.16	0.52
Sex (M/F)	17/8	136/102	0.30
Site	7/4/1/0/8/1/3/1	36/41/28/26/44/21/36/6	0.23
SOPS disorganized symptoms	7.52±3.38	5.04±3.15	<0.001
Mean frame-wise displacement (mm)	0.14±0.10	0.11±0.08	0.17

Table 1. Demographic and clinical information for the studied sample. The two groups showed significant differences in disorganization symptoms.

Overall, the proposed framework was able to uncover (simulated) areas of the brain involved in mediation. Particularly, (1) the framework successfully uncovered neural mediators across different combinations of eigenfunctions, idiosyncratic noises, and sample sizes. (2) The performance on estimating the input map $\boldsymbol{\alpha}$ was better than it on estimating the output map $\boldsymbol{\beta}$. With larger samples, however, the estimation of $\boldsymbol{\beta}$ improved. Both estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

deteriorated when more noises were added; but for each noise level, the estimation performance improved with more samples. At perhaps the extreme end when the noise was very large, the algorithm was still able to uncover some signals using large samples. The estimation using signals simulated from orthogonal basis functions outperformed those from non-orthogonal basis functions. Both cases saw improvement when less noise or larger a sample size was considered. The detailed simulation procedures and results can be found in the **Supporting Information**.

3.2 Empirical results

We applied the framework to an empirical study to identify and isolate functional brain regions that mediate prodromal symptoms at baseline and two-year clinical outcome in subjects at clinical high risk (CHR) for psychosis. The sample included 263 subjects recruited from eight study sites across the United States and Canada who met the criteria for a prodromal risk syndrome[24] at the point of recruitment and were clinically followed up for two years as part of the NAPLS-2 project[21] (see **Table** 1). During the follow-up period, 25 subjects developed a full-blown psychotic disorder (CHR convertors); 238 did not (CHR non-convertors). All participants received an eyes-open resting-state functional magnetic resonance imaging (fMRI) scan at the point of recruitment. After data preprocessing and noise correction, the fractional amplitude of low-frequency fluctuation (fALFF)[30] for each voxel within a binary gray-matter mask (130,992 voxels in total) was extracted. The prodromal symptoms were quantified using the Scale of Prodromal Symptoms (SOPS)[22], and the clinical outcome was labeled as convertor or non-convertor.

Since disorganization symptoms have been shown to be a potential clinical predictor for psychosis[1–3], we first investigated whether the SOPS scores of disorganization symptoms were

significantly different between convertors and non-convertors at baseline. Using the Welch two-sample *t*-test and Pearson's (product-moment) correlation coefficient test, the data confirmed that the convertors and non-convertors in the study indeed had significantly different behavioral symptom scores ($t = 3.49$, $p < 0.005$; Pearson correlation $r = -0.22$, $p < 0.001$) (see **Fig.** 2 b). We then continued to investigate which brain regions would functionally mediate this association. First, we tested if behavioral symptoms affected any of the 130,992 voxels, controlling for age, sex, study site, head motion parameter, and total gray-matter volume (see **Fig.** 3 (b_2)). This analysis yielded the $\boldsymbol{\alpha}$ brain atlas (see **Fig.** 4 (a)): each of its 130,992 elements indicated the effect of behavioral symptoms on a brain voxel. Second, we tested if activity from a brain voxel would increase (or decrease) the likelihood of developing psychosis, while controlling for behavioral symptoms and the aforementioned covariates (see **Fig.** 3 (b_3)). This was conducted using a generalized functional linear model (see **Methods**). This analysis yielded the $\boldsymbol{\beta}$ brain atlas (see **Fig.** 4 (b)): each of its 130,992 elements represented the effect from a brain voxel to the likelihood of developing psychosis, controlling behavioral symptoms and covariates. Third, we obtained the brain-wide functional mediators using the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ brain atlases and classified them into two categories: the positive and negative neural mediators (see **Fig.** 3 (c_1) - (c_2)). Finally, we conducted bootstrap experiments to test whether the mediation effect of each voxel was statistically significant (see **Methods**). Results were reported after Bonferroni correction across all voxels in the brain (see **Fig.** 4).

3.3 The $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ brain atlases

We further inquired into the $\boldsymbol{\alpha}$ brain atlas (**Fig.** 4 (a)) and the $\boldsymbol{\beta}$ brain atlas (**Fig.** 4 (b)) in order to investigate how the input and output systems contribute to overall mediation. Specifically, the $\boldsymbol{\alpha}$ brain atlas included brain regions that were associated with behavioral symptoms; the $\boldsymbol{\beta}$ brain atlas consisted of brain areas that were associated with psychosis status when controlled

for behavioral symptoms. On the $\boldsymbol{\alpha}$ brain atlas, activities of the majority of brain regions were positively associated with behavioral symptoms, while activities of the striatum, thalamus, insular and opercular areas, middle cingulate cortex, sensorimotor area, lingual gyrus, together with cerebellar lobules 4, 5, 6, crus 1, and crus 2 were negatively correlated with behavioral symptoms. On the $\boldsymbol{\beta}$ brain atlas, positive associations were present mainly in the middle cingulate cortex, inferior parietal lobule, striatum, thalamus, lingual gyrus, calcarine sulcus, inferior and middle temporal gyri, and cerebellar lobule 6, crus 1, and crus 2. In contrast, negative associations were shown in the lateral and medial prefrontal cortices, anterior cingulate cortex, posterior cingulate cortex, precuneus, insular and opercular areas, angular gyrus, fusiform gyrus, and cerebellar lobule 8.

3.4 The positive and negative mediators

Discovering brain areas that are positively and negatively mediating behavioral symptoms and disease development is a central problem in neuropathology. To promote discussion, here we defined the positive mediators as brain areas whose activities were positively mediating higher behavioral symptoms and increased chance of conversion to psychosis (see regions with positive weights in **Fig.** 4 (c)). Our results suggested that the positive mediators were mainly present at the bilateral insular and opercular areas, left inferior parietal cortex, right middle frontal gyrus, bilateral inferior temporal gyrus, and cerebellar crus 1 and 2. In contrast, we defined the negative mediators as brain areas whose activities were negatively mediating higher behavioral symptoms and increased chance of conversion to psychosis (see regions with negative weights in **Fig.** 4 (c)). Our results suggested that the negative mediators were located chiefly at the bilateral medial frontal cortex, anterior cingulate cortex, orbitofrontal cortex, precuneus, cuneus, calcarine sulcus, striatum, thalamus, and cerebellar vermis and

lobule 8. The direct effect after accounting for the mediators remained significant ($\gamma > 0$, bootstrap $p < 0.05$), suggesting that the identified mediation effect was partial mediation.

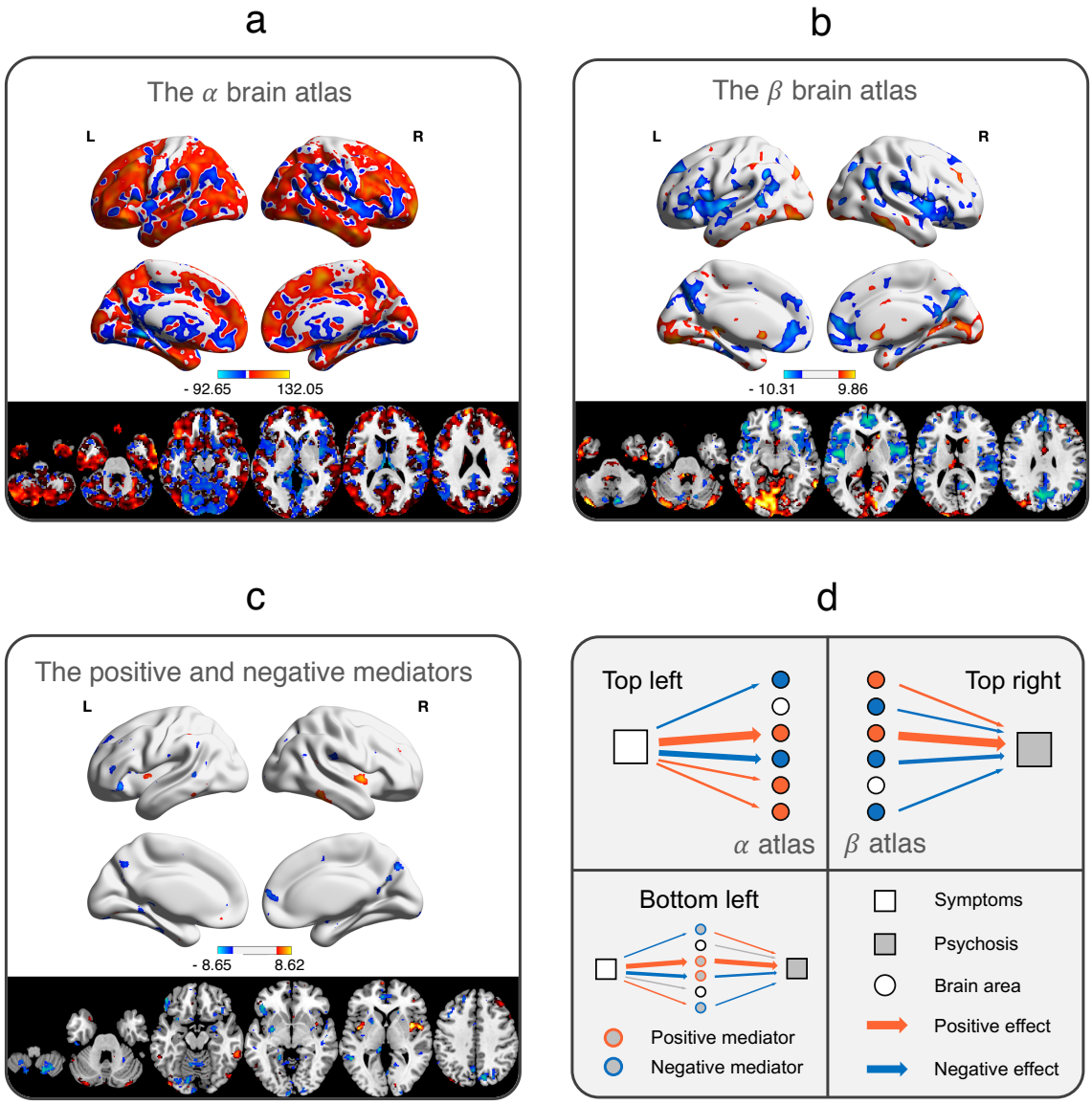


Fig. 4. Brain areas potentially mediating behavioral symptoms and the development of psychosis. A high-dimensional mediation analysis on 130,992 brain voxels of 263 subjects suggests that the pathway between behavioral symptoms and psychosis is potentially positively mediated by the right lateral prefrontal cortices, bilateral insular and opercular areas, bilateral sensorimotor areas, striatum, and cerebellar lobules 4, 5, and 6, and negatively mediated by the bilateral medial frontal and orbitofrontal cortices, left lateral prefrontal cortices, posterior cingulate, precuneus, visual cortex, and cerebellar Crus 1 and lobule 9. **(a) The α brain atlas.** The α brain atlas shows surface and subcortex areas associated with behavioral symptoms when controlled for covariates. **(b) The β brain atlas.** The β brain atlas shows surface and subcortex areas associated with psychosis status when controlled for behavioral symptoms and covariates. **(c) The positive and negative neural mediators.** The neural mediators include surface and subcortex areas that are jointly associated with behavioral symptoms and brain disease status, after all covariates are controlled. The areas highlighted in red are positive mediators, and those highlighted in blue represent negative mediators. **(d) Explanations regarding the pathways and color codes of figures (a) - (c).** The color bars in (a)-(c) indicate effect sizes from bootstrap experiments.

4. Discussion

In this study, we designed a high-dimensional brain-wide functional mediation analysis framework. Through its lenses, we identified and isolated positive and negative neural mediators that potentially mediate psychosis prodromal behavioral symptoms and disease status among individuals at CHR and quantified each mediator's effects on developing psychosis. The positive mediators consisted of brain areas associated with positive mediation, which were primarily distributed in the brain's sensorimotor system, insular and opercular areas, and striatum. The negative mediators consisted of areas associated with negative mediation, which were chiefly located in the brain's default-mode system and visual system. The identification and isolation of the positive and negative mediators provide insights regarding the neurobiological pathways from early psychotic signs to full-blown psychosis and demonstrate the potential utility of the proposed methodological framework in clinical neuroscience studies.

The proposed framework showed promise to study how brain signals may intermediate between an explanatory variable (such as prodromal behavioral symptoms) and an outcome variable (such as psychosis disease status). The framework extends multivariate mediation analysis to high-dimensional functional mediation analysis with non-Gaussian outcomes by incorporating functional data analysis and a generalized linear model. To inquire into the functional organization of the mediator, the model extracts subject-specific principal component (PC) scores and functional representations (population-specific brain-wide basis functions) of measured brain signals. The estimated effect from PC scores to psychosis status is then translated to the whole brain space via the brain-wide basis functions. A *logit* link function was employed to couple measured brain signals and the explanatory variable with the disease

outcome. Since the model allows for a variety of link functions, it may assist other mediation problems with different outcome distributions from the exponential family.

There are a few additional properties of the proposed framework that may be useful in other studies. First, the framework integrates a generalized functional linear model into a dual regulatory system connecting an input system and an outcome system. This technique may provide some insights regarding designing biological models consisting of sub-systems. Second, when the outcomes are binary, one can use the framework to evaluate controlled direct effects, and natural direct and indirect effects on the odds-ratio scale (see **Supporting Information**). Third, high-dimensional brain signals may contain multilevel information. The proposed framework allows us to extract both group-level (*i.e.*, the group-level basis functions) and subject-specific features (*i.e.*, the subject-specific PC scores) of the brain data. The subject-specific features may be used as low-dimensional neural features to assess individual differences in the future (see **Supporting Information**).

We applied the proposed framework in a psychosis neuroimaging study, where we investigated how functional brain activities may mediate prodromal behavioral symptoms and clinical outcome. The findings provided some insights into psychotic disorders. First, our results suggest that the positive mediators mainly involved the insular-opercular areas, temporal areas, frontoparietal areas, and part of the cerebellum (crus 1 and 2). With extensive connections with both sensory areas and the limbic system, the insula is a critical structure in the brain for the integrating and processing of visual and auditory emotional information and supporting subjective feeling states[34,35]. A large number of studies have demonstrated that increased activity in the insular-opercular area is strongly associated with auditory hallucinations in patients with schizophrenia[36–38]. Increased connectivity was found between the insula and

multiple perceptual areas such as the sensorimotor cortex and visual cortex, together with decreased connectivity between the insula and prefrontal cortex[39]. The exact mechanism underlying the hyperactivity state of this area is unclear; it, however, may relate to sensory gating deficits disrupted by excessive mesolimbic dopamine input[40]. The frontoparietal network, inferior temporal, and cerebellar crus 1 and crus 2 are key cognitive areas in humans[41–44]; increased activity and connectivity in these regions have been shown to be predictive of psychosis onset in previous studies[26,29]. In line with previous findings, the current study further shows that these increased activities are potential mediators positively mediating psychosis conversion in individuals with prodromal symptoms. Second, the negative mediators are primarily distributed in the brain's default mode network (including the medial prefrontal cortex, anterior cingulate cortex, and precuneus), together with the thalamus, visual cortex, and cerebellar lobule 8. The DMN is one of the most frequently reported systems whose function is strongly associated with psychosis. Perhaps the most prominent finding regarding DMN in patients is the failure to deactivate this network during cognitive tasks[45,46], which may relate to exaggerated internally-focused thoughts and lack of sufficient suppression of these thoughts during cognition[47]. Here, the finding of negative mediation effect in DMN activities during resting state is parallel to such interpretation, suggesting lower activity (indicating insufficient activation) during rest may potentially mediate prodromal symptoms and psychosis onset. In addition, this finding also corresponds well with the "triple network" model of psychosis, where the dysregulation of the insula (as a positive mediator) on DMN has been reported in individuals at risk for psychosis[48,49]. Cerebellar lobule 8 is a key region for processing sensorimotor-related errors[43,50]. Higher activity in this region and the visual cortex may therefore imply a compensatory mechanism or an amplification of neural representations of perceptual information, potentially related to resolving or attenuating the existing perceptual deficits.

A few reasons have made the blood-oxygen-level-dependent functional magnetic resonance imaging (BOLD fMRI) data suitable for studying brain-wide mediation. First, although studies have used resting-state electroencephalography (EEG) data and discovered brain areas, such as the frontal regions, that are associated with psychosis[51], imaging modalities with greater spatial resolution, such as fMRI, may both confirm and extend neural signatures beyond those identified using EEG. Second, reduced auditory P300 event-related potential (ERP) amplitude (from a functional neurophysiological test) is a primary candidate electrophysiologic biomarker of psychosis[52]; it nevertheless may not capture as much variability that occurs in spontaneous brain activity as the fMRI data to work well as a biomarker for conversion. Third, slow-wave power has been shown to correlate with reduced blood flow and glucose utilization in schizophrenia patients, and is therefore thought to reflect reduced functioning in the frontal area[53,54]. This supports the utility of fMRI BOLD data in mediation studies. Finally, structural MRI studies are beginning to discover associations between structural brain information and conversion to psychosis[55]; here we showed that functional MRI data may add new insights into studies of neural markers associated with psychosis.

There are several limitations in this study. First, the nature of the imaging and clinical data suggests that the identified mediating pathways are associative; our results do not conclude definitive causal flows from prodromal signs via brain areas towards conversion status (see **Introduction**). Second, we were mainly interested in identifying brain regions that were simultaneously mediating behavioral symptoms and psychosis status. This naturally left out the cases where some mediators were interposed before or after other mediators. Future analysis may incorporate dynamic mediating systems and information feedback components. Third, throughout, we assumed that behavioral symptoms did not interact with the mediators. Future work that includes interaction between the explanatory variable and the mediator may

be useful to expand current analysis (see **Supporting Information** for an example and see[56] for a special case). Fourth, although dimension reduction could reduce biases caused by spurious correlation, our framework cannot remove the coincidental association between some (voxel) features and the residual term (*i.e.*, incidental endogeneity). This is an active research area in high-dimensional data analysis (see, for example, Fan and Liao[57]). Fifth, the proposed model focused on neural markers by averaging brain time courses over time. This omitted the territory of mediation analysis where the mediation effect changes over time. Future work could extend the framework to longitudinal settings: such extensions are particularly useful for studying mediation effects related to brain development during childhood and adolescence, brain aging between health and disease, and brain degeneration along the trajectory of neurodegenerative disease development. We are currently investigating how to extend the techniques used in our framework to improve our understanding of large-scale longitudinal mediation; potential directions include combing functional data analysis (FDA) and a dual mediation system with autoregressive models[58,59], latent growth curve (LGM)[60], parallel process models[61], latent difference score (LDS) models[62], and autoregressive LGM models[63].

Although we demonstrated high-dimensional functional neural mediation analysis in the domain of brain studies, the framework may also be useful to study other high-dimensional mediation problems, such as how genome-wide genotypes mediate the effect of environmental factors on phenotypes. With the recent convergence in neuroimaging, genomics, health informatics, wearable, and digital sensors, the model may be useful to study a broad range of intermediating variables. For example, the model may be used to understand how gene expression, brain physiology, and circadian patterns jointly mediate environment and biological phenotypes, uncover brain regions that mediate sensory input and behavior outcome collected

by wearable devices, and study how computers may act as a mediating artificial intelligence (AI), transferring human input into computer-generated intelligent responses.

To summarize, in the present study, we propose a framework that leverages the mediation analysis concept and neuroscientific insights to inquire into the properties of large-scale functional neural markers that mediate the relationship between an explanatory variable and a binary disease outcome. The framework's capacity of treating high-dimensional data and flexibility in handling non-normally distributed outcomes make it potentially useful in a variety of scenarios to uncover intermediating pathways in complex regulative systems.

References

1. Cannon, T. D. *et al.* Prediction of psychosis in youth at high clinical risk. *Arch. Gen. Psychiatry* **65**, pp. 28–37 (2008).
2. Demjaha, A., Valmaggia, L., Stahl, D., Byrne, M. & McGuire, P. Disorganization/cognitive and negative symptom dimensions in the at-risk mental state predict subsequent transition to psychosis. *Schizophr. Bull.* **38**, pp. 351–359 (2012).
3. Carrión, R. E. *et al.* Prediction of functional outcome in individuals at clinical high risk for psychosis. *JAMA Psychiatry* **70**, pp. 1133–1142 (2013).
4. Chén, O. Y. *et al.* High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19**, pp. 121–136 (2018).
5. Geuter, S. *et al.* Multiple brain networks mediating stimulus-pain relationships in humans. *Cereb. Cortex* **30**, pp. 4204–4219 (2020).
6. VanderWeele, T. J. & Vansteelandt, S. Mediation analysis with multiple mediators. *Epidemiol. Method.* **2**, pp. 95–115 (2014).
7. Huang, Y. T. & Pan, W. C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72**, pp. 402–413 (2016).
8. Nguyen, T. Q., Webb-Vargas, Y., Koning, I. M. & Stuart, E. A. Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: simulations and application to an alcohol intervention. *Struct. Equ. Model.* **23**, pp. 368–383 (2016).
9. Lindquist, M. A. Functional causal mediation analysis with an application to brain connectivity. *J. Am. Stat. Assoc.* **107**, pp. 1297–1309 (2012).
10. Baron, R. M. & Kenny, D. a. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, pp. 1173–1182 (1986).
11. Robins, J. M. & Greenland, S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, pp. 143–155 (1992).
12. Whittle, R., Mansell, G., Jellema, P. & van der Windt, D. Applying causal mediation methods to clinical trial data: What can we learn about why our interventions (don't) work? *Eur. J. Pain* **21**, pp. 614–622 (2017).
13. Fishbein, M. & Ajzen., I. *Belief, attitude,iIntention, and behavior: An introduction to theory and researched.* Addison-Wesley, Reading, MA, USA (1975).
14. Hyman, H. H. *Survey design and analysis: Principles, cases and procedures*. Free Press, Glencoe, Illinois, USA (1955).
15. Alwin, D. F. & Hauser, R. M. The decomposition of effects in path analysis. *Am. Sociol. Rev.* **40**, pp. 37–47 (1975).
16. Judd, C. M. & Kenny, D. A. Process analysis: Estimating mediation in treatment evaluations. *Eval. Rev.* **5**, pp. 602–619 (1981).
17. Sobel, M. E. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* **13**, pp. 290–312 (1982).
18. Karhunen, K. Über lineare methoden in der wahrscheinlichkeitsrechnung (On lin- ear methods in probability and statistics). *Ann. Acad. Sci. Fenn. Ser. A. I. Math.-Phys.* **31**, pp. 1–79 (1947).
19. Loève, M. Calcul des probabilités – sur la covariance d'une fonction aléatoire (Calculating probabilities – on the covariance of a random function). Note by M. Michel Loéve, present by M. Henri Villat. In *Comptes rendus de l'Académie des Sciences* (Comptes rendus, or Proceedings of the Academy of sciences) **220**, pp. 295–296

(1945).
20. Indritz, J. *Methods in analysis*. Macmillan, New York, USA (1963).
21. Addington, J. *et al.* North American Prodrome Longitudinal Study (NAPLS 2): Overview and recruitment. *Schizophr. Res.* **142**, pp. 77–82 (2012).
22. McGlashan, T. H., Miller, T. J., Woods, S. W., Hoffman, R. E. & Davidson, L. Instrument for the assessment of prodromal symptoms and states. In *Early Intervention in Psychotic* Disorders, edited by Eds. Miller, T., Mednick, S.A., McGlashan, T.H., Libiger, J., and Johannessen, J.O., pp. 135–149. Springer, Dordrecht, the Netherlands, by NATO Science Series (Series D: Behavioural and Social Sciences) (2001).
23. Bell, C. C. DSM-IV: Diagnostic and statistical manual of mental disorders. *J. Am. Med. Assoc.* **272**, pp. 828–829 (1994).
24. Miller, T. J. *et al.* Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: Predictive validity, interrater reliability, and training to reliability. *Schizophrenia Bulletin* **29**, 703–715 (2003).
25. Miller, T. J. *et al.* Prospective diagnosis of the initial prodrome for schizophrenia based on the structured interview for prodromal syndromes: Preliminary evidence of interrater reliability and predictive validity. *Am. J. Psychiatry* **159**, pp. 863–865 (2002).
26. Cao, H. *et al.* Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. *Nat. Commun.* **9**, 3836 (2018).
27. Cao, H. *et al.* Progressive reconfiguration of resting-state brain networks as psychosis develops: Preliminary results from the North American Prodrome Longitudinal Study (NAPLS) consortium. *Schizophr. Res.* (2019). doi:10.1016/j.schres.2019.01.017
28. Cao, H., Ingvar, M., Hultman, C. & Cannon, T. Evidence for cerebello-thalamo-cortical hyperconnectivity as a her- itable trait for schizophrenia. *Transl. Psychiatry* **9**, pp. 1–8 (2019).
29. Cao, H. *et al.* Altered brain activation during memory retrieval precedes and predicts conversion to psychosis in individuals at clinical high risk. *Schizophr. Bull.* **45**, pp. 924–933 (2019).
30. Zou, Q. H. *et al.* An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *J. Neurosci. Methods* **172**, pp. 137–141 (2008).
31. Wang, J.-L., Chiou, J.-M. & Müller1, H.-G. Functional data analysis. *Annu. Rev. Stat. Its Appl.* **3**, 257–295 (2016).
32. Ramsay, J. & Silverman, B. W. *Functional data analysis*. Springer-Verlag New York, USA (2005).
33. Efron, B. Bootstrap methods: Another look at the Jackknife. *Ann. Stat.* **7**, pp. 1–26 (1979).
34. Wylie, K. P. & Tregellas, J. R. The role of the insula in schizophrenia. *Schizophrenia Research* **123**, pp. 93–104 (2010).
35. Namkung, H., Kim, S. H. & Sawa, A. The insula: An underestimated brain area in clinical neuroscience, psychiatry, and neurology. *Trends in Neurosciences* **40**, pp. 200–207 (2017).
36. Dierks, T. *et al.* Activation of Heschl's gyrus during auditory hallucinations. *Neuron* **22**, pp. 615–621 (1999).
37. Shergill, S. S., Brammer, M. J., Williams, S. C. R., Murray, R. M. & McGuire, P. K. Mapping auditory hallucinations in schizophrenia using functional magnetic resonance imaging. *Arch. Gen. Psychiatry* **57**, pp. 1033–1038 (2000).

38. Powers, A. R., Mathys, C. & Corlett, P. R. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science* **357**, pp. 596–600 (2017).
39. Tian, Y., Zalesky, A., Bousman, C., Everall, I. & Pantelis, C. Insula functional connectivity in schizophrenia: Subregions, gradients, and symptoms. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, pp. 399–408 (2019).
40. Braff, D. L. Sensorimotor gating and schizophrenia. *Arch. Gen. Psychiatry* **47**, pp. 181–188 (1990).
41. Dosenbach, N. U. F. *et al.* Distinct brain networks for adaptive and stable task control in humans. *Proc. Natl. Acad. Sci.* **104**, pp. 11073–11078 (2007).
42. Dosenbach, N. U., Fair, D. A., Cohen, A. L., Schlaggar, B. L. & Petersen, S. E. A dual-networks architecture of top-down control. *Trends Cogn. Sci.* **12**, pp. 99–105 (2008).
43. Buckner, R. L., , Krienen, F. M., Castellanos, A., Diaz, J. C. & Yeo, B. T. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, pp. 2322–2345 (2011).
44. Marek, S. *et al.* Spatial and temporal organization of the individual human cerebellum. *Neuron* **100**, pp. 977–993 (2018).
45. Pomarol-Clotet, E. *et al.* Failure to deactivate in the prefrontal cortex in schizophrenia: Dysfunction of the default mode network? *Psychol. Med.* **38**, pp. 1185–1193 (2008).
46. Fryer, S. L. *et al.* Deficient suppression of default mode regions during working memory in individuals with early psychosis and at clinical high-risk for psychosis. *Front. Psychiatry* **4**, 92 (2013).
47. Whitfield-Gabrieli, S. & Ford, J. M. Default mode network activity and connectivity in psychopathology. *Annu. Rev. Clin. Psychol.* **8**, pp. 49–76 (2012).
48. Wotruba, D. *et al.* Aberrant coupling within and across the default mode, task-positive, and salience network in subjects at risk for psychosis. *Schizophr. Bull.* **40**, pp. 1095–1104 (2014).
49. Bolton, T. A. W. *et al.* Triple network model dynamically revisited: Lower salience network state switching in pre-psychosis. *Front. Physiol.* **11**, 66 (2020).
50. Jeremy D.Schmahmann. The cerebellum and cognition. *Neurosci. Lett.* **688**, pp. 62–75 (2019).
51. Sollychin, M. *et al.* Frontal slow wave resting EEG power is higher in individuals at Ultra High Risk for psychosis than in healthy controls but is not associated with negative symptoms or functioning. *Schizophr. Res.* **208**, pp. 293–299 (2019).
52. Hamilton, H. K. *et al.* Association between P300 responses to auditory oddball stimuli and clinical outcomes in the psychosis risk syndrome. *JAMA Psychiatry* **76**, pp. 1187–1197 (2019).
53. Guich, S. M. *et al.* Effect of attention on frontal distribution of delta activity and cerebral metabolic rate in schizophrenia. *Schizophr. Res.* **2**, pp. 439–448 (1989).
54. Ingvar, D. H., Sjölund, B. & Ardö, A. Correlation between dominant EEG frequency, cerebral oxygen uptake and blood flow. *Electroencephalogr. Clin. Neurophysiol.* **41**, 268–276 (1976).
55. Chung, Y. *et al.* Adding a neuroanatomical biomarker to an individualized risk calculator for psychosis: A proof-of-concept study. *Schizophr. Res.* **208**, pp. 41–43 (2019).
56. Muller, D., Judd, C. M. & Yzerbyt, V. Y. When moderation is mediation and mediation is moderated. *J. Pers. Soc. Psychol.* **89**, pp. 852–863 (2005).
57. Fan, J. & Liao, Y. Endogeneity in high dimensions. *Ann. Stat.* **42**, pp. 872–917 (2014).
58. Gollob, H. F. & Reichardt, C. S. Interpreting and estimating indirect effects assuming

time lags really matter. In *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* Edited by Collins, L.M. and Horn, J.L., pp. 243–259. American Psychological Association (1991).

59. Cole, D. A. & Maxwell, S. E. Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology* **112**, pp. 558–577 (2003).
60. Muthén, B. O. & Curran, P. J. General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychol. Methods* **2**, pp. 371–402 (1997).
61. Cheong, J., Mackinnon, D. P. & Toon, S. Structural equation modeling : A investigation of mediational processes using parallel process latent growth curve modeling. *Struct. Equ. Model. A Multidiscip. J.* **10**, pp. 238–262 (2009).
62. McArdle, J. J. A latent difference score approach to longitudinal dynamic structural analysis. In *Structural equation modeling: Present and future.* Edited by Cudeck, R., du Toit, S., and Sörbom, D, pp. 342–380. Scientific Software International, Lincolnwood, IL, USA (2001).
63. Bollen, K. A. & Curran, P. J. Autoregressive Latent Trajectory (ALT) models: A synthesis of two traditions. *Sociol. Methods Res.* **32**, pp. 336–383 (2004).

Supporting Information for:

Identifying Neural Signatures Mediating Behavioral Symptoms and Psychosis Onset: High-Dimensional Whole-Brain Functional Mediation Analysis

1. Simulation studies

We simulate data using the following procedures.

Consider N subjects and V brain areas, where $V = 130{,}992$. Let x_i be the explanatory variable for subject i, z_{i1}, z_{i2}, and z_{i3} be the site (from which data are collected), age, and gender, respectively, of subject i, m_{ij} be the neural activity from the j^{th} brain area of subject i, and y_i be the outcome for subject i.

$$(1)\ x_i = \rho_0 + \rho_1 z_{i1} + \rho_2 z_{i2} + \rho_3 z_{i3} + e_i$$

where z_{i1} is a random sample from a set of eight sites (mimicking the distribution of eight sites considered in the study), $z_{i2} \sim \dot{N}(\mu = 19, \sigma = 4.15, l = 12, u = 32)$, where $\dot{N}()$ denotes a truncated normal distribution with mean μ, standard deviation σ, lower bound l and upper bound u, $z_{i3} \sim B(0.42)$. These choices are made to approximate the distribution of covariates used in the real dataset but can be chosen in other forms without loss of generality. We take parameters $\rho_0 = 6.765$, $\rho_1 = -0.174$, $\rho_2 = -0.072$, and $\rho_3 = -0.456$ to be the estimates from the real data using Equation (1) above. How these parameters are set only affects the simulated values of x_i; they will not affect parameter estimation of $\alpha(j)$ and $\beta(j)$ below. Consider noise $e_i \sim N(0, \sigma_e^2)$.

$$(2)\ m_{ij} = \theta_0(j) + \sum_{k=1}^{K} \xi_{ik}\varphi_k(j) + x_i\alpha(j) + z_{i1}t_1(j) + z_{i2}t_2(j) + z_{i3}t_3(j) + \varepsilon_i$$

where $\xi_{ik} \sim N(0, \lambda_k)$, $\lambda_k = 0.5^{k-1}$, for $i \in \{1,2,\dots,N\}$, $j \in \{\frac{1}{V}, \frac{2}{V}, \dots, \frac{V}{V}\}$, and $k \in \{1,2,\dots,K\}$. $\{\varphi_1, \varphi_2, \dots, \varphi_K\}$ is a set of basis functions (see **Sections** 1.1 and 1.2 below). Let $\alpha(j) = \cos(2\pi j)$, $\theta_0(j) = \sin(8\pi j)$, $t_1(j) = \cos(8\pi j)$, $t_2(j) = \sin(12\pi j)$, and $t_3(j) = \cos(12\pi j)$. Consider noise $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

$$(3)\ y_i \sim Bernoulli\ (p_i)$$

where $p_i = \frac{1}{1+\exp\left[-(\theta_0' + \sum_{j=1}^{V} m_i(j)\beta(j) + x_i\gamma + \tau_1 z_{i1} + \tau_2 z_{i2} + \tau_3 z_{i3} + \epsilon_i)\right]}$, $\beta(j) = \sin(2\pi j)$, $\gamma = -0.01$, $\tau_1 = 0.1$, $\tau_2 = 0.1$, and $\tau_3 = 0.1$. Consider noise $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. For each simulation below, θ_0' is chosen at the value so that the outcomes contain both 0's and 1's (to simulate the scenario where the dataset contains both convertor and non-convertor subjects).

We study different combinations of eigenfunctions ($\{\varphi_1, \varphi_2, \dots, \varphi_K\}$), idiosyncratic noises (σ_e^2, σ_ε^2, σ_ϵ^2), and sample sizes (N). For eigenfunctions, we consider $K = 4$ basis functions and study the cases when they are orthogonal and when they are non-orthogonal (to cover the two general scenarios where the brain patterns consist of orthogonal and non-orthogonal neural signals). For noises, we consider a range of magnitudes, at $(\sigma_e^2, \sigma_\varepsilon^2, \sigma_\epsilon^2) = (0.01, 0.01, 0.01)$ (small noise), $(0.1, 0.1, 0.1)$ (moderate noise), and $(5, 5, 5)$ (very large noise). We consider very large noise because we want to see to what degree the estimates would be able to (and unable to) uncover the signals, and whether, even under a very large noise level, signals can be recovered when there are more samples. To assess the performance of our method under different sample sizes, we consider N to be 100 and 500. Overall, we examine 12 model conditions; for each condition, we conduct 100 bootstrap simulations.

A summary of simulation studies

Overall, the framework proposed in this study was able to uncover (simulated) areas of the brain involved in mediation. Particularly,

1. The framework successfully uncovered neural mediators across different combinations of eigenfunctions, idiosyncratic noises, and sample sizes.
2. The performance on estimating the input map $\boldsymbol{\alpha}$ was better than it on estimating the output map $\boldsymbol{\beta}$. With larger samples, however, the estimation of $\boldsymbol{\beta}$ improved.
3. Both estimations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ deteriorated when more noises were added; but for each noise level, the estimation performance improved with more samples. At perhaps the extreme end when the noise was very large, the algorithm was still able to uncover some signals using large samples.
4. The estimation using signals simulated from orthogonal basis functions outperformed those from non-orthogonal basis functions. Both cases saw improvement when less noise or larger a sample size was considered.

1.1 Orthogonal basis functions

Consider $\{\varphi_k(j)\}_{k=1}^{4} = \{\sqrt{2}\sin(2\pi j), \sqrt{2}\cos(2\pi j), \sqrt{2}\sin(4\pi j), \sqrt{2}\cos(4\pi j)\}$.

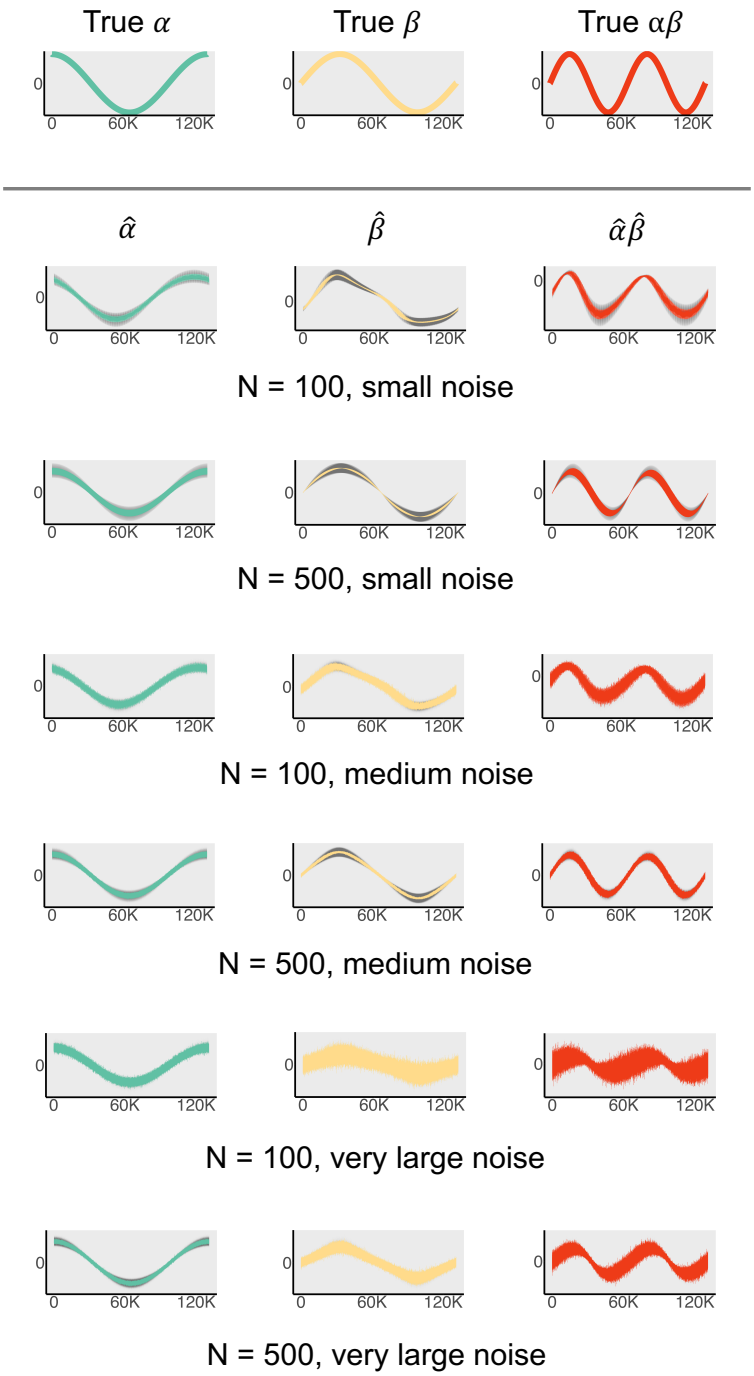


Fig. S1. Simulation studies using orthogonal basis functions at different combinations of sample size and noise level.

1.2 Non-orthogonal basis functions

Consider $\{\varphi_k(j)\}_{k=1}^4 = \{1, \sqrt{3}(2j-1), \sqrt{5}(6j^2-6j+1), \sqrt{7}(20j^3-30j^2+12j-1)\}$.

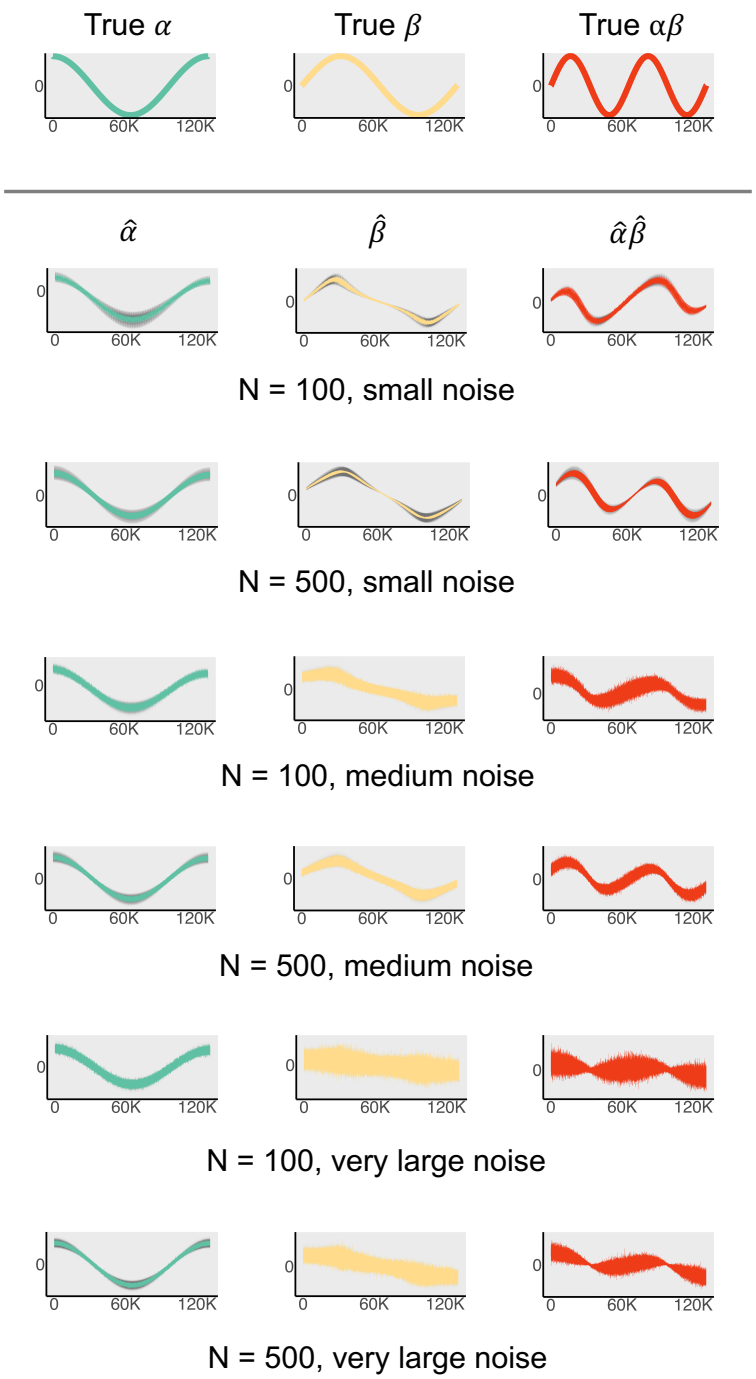


Fig. S2. Simulation studies using nonorthogonal basis functions at different combinations of sample size and noise level.

2. Theoretical properties

In the following, we demonstrate that the direct and indirect effects of a multivariate causal mediation model with binary outcomes can be captured on the odds ratio scale.

Definitions. Let X denote an explanatory variable for a given subject (*e.g.*, prodromal symptoms), and Y an outcome variable (*e.g.*, having psychosis or healthy). Suppose there are multiple mediators $\boldsymbol{M} = (M(1), M(2), \ldots, M(V))$. In the rs-fMRI context, the mediators are *V* dependent activations over the *V* voxels. Here we assume for simplicity that each subject is scanned under one condition. We use upper and lower cases to distinguish a random variable and its realization (real data), respectively. We use a bold letter to specify a vector.

Using potential outcomes notation[1–3], let $Y(x) = Y(X = x)$, $\boldsymbol{M}(x) = \boldsymbol{M}(X = x)$, and $Y(x, \boldsymbol{m}) = Y(X = x, \boldsymbol{M} = \boldsymbol{m})$. In the following, we extend the direct and indirect effects defined on the odds-ratio for a single mediator[4] to incorporate multiple (and high-dimensional) mediators.

The total effect (TEs) conditional on a covariate $\boldsymbol{Z} = \boldsymbol{z}$ on the odds ratio (OR) is defined as:

$$TE^{OR}_{x,x^*}|\boldsymbol{z} = \frac{P(Y(x)=1 \mid \boldsymbol{z})/\{1-P(Y(x)=1 \mid \boldsymbol{z})\}}{P(Y(x^*)=1 \mid \boldsymbol{z})/\{1-P(Y(x^*)=1 \mid \boldsymbol{z})\}}. \quad \text{(S.1)}$$

In words, it compares the odds of Y being 1 if X takes value x and the odds of Y being 1 had X took value x^*, on the OR scale, in a given stratum where $\boldsymbol{Z} = \boldsymbol{z}$. In a psychosis study, this will capture the effect of symptom score changing from x to x^* on developing psychosis at the odds ratio scale, for a specific age, gender, and site group.

Next, the conditional controlled direct effect (CDE) on the OR is defined as:

$$CDE^{OR}_{x,x^*}|\boldsymbol{z} = \frac{P(Y(x,\boldsymbol{m})=1 \mid \boldsymbol{z})/\{1-P(Y(x,\boldsymbol{m})=1 \mid \boldsymbol{z})\}}{P(Y(x^*,\boldsymbol{m})=1 \mid \boldsymbol{z})/\{1-P(Y(x^*,\boldsymbol{m})=1 \mid \boldsymbol{z})\}}. \quad \text{(S.2)}$$

In words, it compares the odds of Y being 1 if X takes value x and the high-dimensional mediator $\boldsymbol{M}$ is fixed (intervened to be) at $\boldsymbol{m}$, and the odds of Y being 1 had X took value x^* while $\boldsymbol{M}$ was fixed at $\boldsymbol{m}$, on the OR scale, in a given stratum where $\boldsymbol{Z} = \boldsymbol{z}$. In a psychosis study, this will capture the effect of symptom score changing from x to x^* on developing psychosis at the OR scale, with a brain region of interest fixed, via transcranial magnetic stimulation (TMS), at $\boldsymbol{M} = \boldsymbol{m}$ for a specific age, gender, and site group.

Similarly, the conditional natural direct effect (NDE) on the odds ratio is defined as:

$$NDE^{OR}_{x,\underline{x}^*}|\boldsymbol{z} = \frac{P(Y(x,\boldsymbol{M}(x^*))=1 \mid \boldsymbol{z})/\{1-P(Y(x,\boldsymbol{M}(x^*))=1 \mid \boldsymbol{z})\}}{P(Y(x^*,\boldsymbol{M}(x^*))=1 \mid \boldsymbol{z})/\{1-P(Y(x^*,\boldsymbol{M}(x^*))=1 \mid \boldsymbol{z})\}} \quad \text{(S.3)}$$

where the underlined $\underline{x}^*$ indicates that the mediator $\boldsymbol{M}$ is set at the level it would have been under x^* despite the explanatory variable being x. Hence, the natural direct effect quantifies the effect of the explanatory variable (changing from x to x^*) on the outcome on the OR scale, with mediator fixed at $\boldsymbol{M}(x^*)$ in a given stratum where $\boldsymbol{Z} = \boldsymbol{z}$.. In a psychosis study, this will capture the effect of symptom score changing from x to x^* on developing psychosis on the OR scale, with a brain region of interest fixed, via TMS, at the level had symptom score been x^* (*i.e.,* at $\boldsymbol{M}(x^*)$), for a specific age, gender, and site group.

Finally, the conditional natural indirect effect (NIE) on the OR is defined as:

$$NIE^{OR}_{\underline{x},x^*}|\boldsymbol{z} = \frac{P(Y(x,\boldsymbol{M}(x))=1 \mid \boldsymbol{z})/\{1-P(Y(x,\boldsymbol{M}(x))=1 \mid \boldsymbol{z})\}}{P(Y(x,\boldsymbol{M}(x^*))=1 \mid \boldsymbol{z})/\{1-P(Y(x,\boldsymbol{M}(x^*))=1 \mid \boldsymbol{z})\}} \quad \text{(S.4)}$$

where the underlined $\underline{x}$ indicates that the explanatory variable X is set at the level x despite the mediator is set at level $\boldsymbol{M}(x^*)$ (namely, to what the mediator would have been had the explanatory variable been set to x^*). Hence, the natural indirect effect quantifies the effect of the

mediator (changing from $\boldsymbol{M}(x)$ to $\boldsymbol{M}(x^*)$) on the outcome on the OR scale, in a given stratum where $\boldsymbol{Z} = \boldsymbol{z}$. In a psychosis study, this will capture the effect of brain measurement of a region of interest (changing from $\boldsymbol{M}(x)$ to $\boldsymbol{M}(x^*)$ while fixing symptom score at x) on developing psychosis on the OR scale, for a specific age, gender, and site group.

It follows, from Equations (S.1), (S.3), and (S.4), that the above total effect can be decomposed into a natural effect component and an indirect effect component on the OR scale:

$$TE^{OR}_{x,x^*}|\boldsymbol{z} = \{NIE^{OR}_{\underline{x},x^*} \times NDE^{OR}_{x,\underline{x}^*}\}|\boldsymbol{z}.$$

The direct effect could also be defined as $Y\big(x, \boldsymbol{M}(x)\big) - Y\big(x^*, \boldsymbol{M}(x)\big)$. In general, this would lead to a different decomposition of the total effect; this, however, is not of further concern as the modification is straightforward.

Identification. To identify the direct and indirect effects using observed data, we require a few assumptions[5,6]. In the following, we assume $\boldsymbol{Z}$ contains all relevant confounding variables.

Assumption 1: $Y(x) \perp X \mid \boldsymbol{Z}$.

Assumption 2: $Y(x, \boldsymbol{M}(x)) \perp X \mid \boldsymbol{Z}$.

Assumption 3: $Y(x, \boldsymbol{M}(x)) \perp \boldsymbol{M} \mid \{X, \boldsymbol{Z}\}$.

Assumption 4: $\boldsymbol{M}(x) \perp X \mid \boldsymbol{Z}$.

Assumption 5: $Y(x, \boldsymbol{m}) \perp \boldsymbol{M}(x^*) \mid \boldsymbol{Z}$.

Conditioning on covariate $\boldsymbol{Z}$, these assumptions imply that, there is no confounding for the relationship between: (Assumptions 1-2) explanatory variable X and outcome Y; (Assumption 3)

mediator $\boldsymbol{M}$ and outcome Y; (Assumption 4) explanatory variable X and mediator $\boldsymbol{M}$; and (Assumption 5) no confounding for the relationship between mediator and outcome that is affected by the explanatory variable. Together, they are often referred to as sequential ignorability assumptions. See Robins and Richardson[7] and Pearl[8] for a detailed discussion of these assumptions, and see Huang and Pan[9] for a critical evaluation of these assumptions in the high-dimensional setting.

Under these assumptions, it has been shown[4,10] that the average direct and indirect effects are identified from the regression function using the observed data for a single mediator and multiple mediators. Here, we show that their argument also applies to cases with high-dimensional mediators. In particular, under Assumption 1, the total effect conditional on a covariate $\boldsymbol{Z} = \boldsymbol{z}$ on the OR scale is identifiable:

$$TE^{OR}_{x,x^*}|\boldsymbol{z} = \frac{P(Y = 1\,|x, \boldsymbol{z})/\{1 - P(Y = 1\,|x, \boldsymbol{z})\}}{P(Y = 1\,|x^*, \boldsymbol{z})/\{1 - P(Y = 1\,|x^*, \boldsymbol{z})\}}.$$

Under Assumptions 2-3, the conditional direct effect on the OR scale is identifiable:

$$CDE^{OR}_{x,x^*}|\boldsymbol{z} = \frac{P(Y = 1 \mid x, \boldsymbol{m}, \boldsymbol{z})/\{1 - (Y = 1 \mid x, \boldsymbol{m}, \boldsymbol{z})\}}{P(Y = 1 \mid \ x^*, \boldsymbol{m}, \boldsymbol{z})/\{1 - (Y = 1 \mid \ x^*, \boldsymbol{m}, \boldsymbol{z})\}}.$$

Similarly, under Assumptions 2-5, the natural direct and indirect effects on the OR scales are identifiable.

Suppose Assumptions 2-5 hold, and the model in Equations (iv) in the main text is under a logit link function:

$$E(M(v)|X = x, \boldsymbol{Z} = \boldsymbol{z}) = \alpha_0 + \alpha(v)x + \boldsymbol{t}(v)\boldsymbol{z}, \quad v = 1, 2, \dots, V$$

$$logit(Y = 1|X = x, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{Z} = \boldsymbol{z}) = \beta_0 + \gamma x + \sum_{v=1}^{V} m(v)\beta(v) + \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{z}.$$

The average controlled direct effect and natural indirect effect on the OR scale can be shown to be:

$$CDE_{x,x^*}^{OR}|\boldsymbol{z} = e^{\gamma(x-x^*)} \quad \text{(S.5)}$$

$$NIE_{\underline{x},x^*}^{OR}|\boldsymbol{z} = e^{[\sum_{v=1}^{V}\alpha(v)\beta(v)](x-x^*)}. \quad \text{(S.6)}$$

Throughout, we assume that the explanatory variable does not interact with one or more of the mediators. When the explanatory variable interacts with one or more of the mediators, however, the bimodal framework considered in this paper is not appropriate for mediation analysis[11]. Instead, one may consider:

$$E(M(v)|X = x, \boldsymbol{Z} = \boldsymbol{z}) = \alpha_0 + \alpha(v)x + \tilde{\boldsymbol{t}}(v)\boldsymbol{z}, \quad v = 1, 2, \dots, V$$

$$\begin{aligned} &logit(Y = 1|X = x, \boldsymbol{M} = \boldsymbol{m}, \boldsymbol{Z} = \boldsymbol{z}) \\ &= \beta_0 + \gamma x + \sum_{v=1}^{V} m(v)\beta(v) + \sum_{v=1}^{V} x\, m(v)\, b(v) + \tilde{\boldsymbol{\tau}}^{\mathrm{T}}\boldsymbol{z}. \end{aligned}$$

Assume Assumptions 2-3 hold, we have:

$$CDE_{x,x^*}^{OR}|\boldsymbol{z} = e^{[\gamma+\sum_{v=1}^{V} m(v)\, b(v)](x-x^*)}$$

$$NIE_{\underline{x},x^*}^{OR}|\boldsymbol{z} = e^{[\sum_{v=1}^{V}\alpha(v)\beta(v) + \sum_{v=1}^{V}\alpha(v)\, b(v)x](x-x^*)}.$$

Remarks. When the counterfactuals are well defined and the Assumptions 1-5 hold, the right-hand side of Equations S.5 and S.6 identify causal mediation effects. When one or more of the assumptions in Assumptions 1-5 fail to hold, or if the counterfactuals are not well defined, the right-hand side of Equations S.5 and S.6 may still be used in an exploratory analysis to help

identify potential mediators. For example, they could identify linear combinations of voxels that correspond to specific brain functions, suggesting mediation through correlates of those brain functions. In the article, for simplicity, we use “direct effect” and “indirect effect” to refer to the right-hand sides of Equations S.5 and S.6, respectively; we are agnostic throughout as to whether these expressions can be interpreted causally or should be taken as exploratory.

Similarly, we use “mediator” agnostically to refer to variables that temporally follow the explanatory variable and precede outcome and potentially may lie on a causal pathway between them. We also note that the model only considers the case where the entire vector $\boldsymbol{M}$ is subject to a single level of intervention (x). More generally, it would be possible to define potential mediators when some elements of $\boldsymbol{M}$ are subject to x and other elements of $\boldsymbol{M}$ are subject to x^*, which could support various decompositions of direct/indirect effects due to pathways acting through different $\boldsymbol{M}$. In the fMRI setting described in the article, we anticipate that the same elements of $\boldsymbol{M}$ (brain regions) are subject to both x and x^*. For example, minor prodromal symptoms and severe prodromal symptoms are associated with activation in the same brain regions, but with different intensities.

3. A comparison between different mediation analysis approaches

To complement the presentations made in this study, in the following we provide a brief comparison between common mediation analysis. To avoid further confusion, we suppose that all covariates have been controlled.

There are, broadly, two approaches to conduct mediation analyses: the *difference* approach and the *product* approach[12,13]. Conceptually, both approaches employ a series of regression models to dissect the pathways between an explanatory variable, a mediator, and an outcome.

The difference approach determines that a variable M is a mediator as follows. Consider two regression models, one consisting of regressing an explanatory variable X on the outcome Y (call the estimated effect from the explanatory variable to the outcome as γ), and the other consisting of regressing both the explanatory variable X and the potential mediator M on the outcome Y (call the estimated effect from the explanatory variable to the outcome under this setting as γ'). If the effect of the explanatory variable X on the outcome Y when the mediator M is not included in the regression model is considerably greater than the effect of the explanatory variable X on the outcome Y when the potential mediator M is included, it suggests that there is a significant mediation effect (that alleviates the direct effect of the explanatory variable on the outcome), and thus the mediation effect can be estimated by taking the difference between γ and γ'.

The product approach also considers two regression models, one consisting of regressing the explanatory variable X on the potential mediator M (call the effect of the explanatory variable on the potential mediator as α), and the other consists of regressing both the explanatory variable X and the potential mediator M on the outcome Y (call the effect of the potential mediator on the outcome while controlling the explanatory variable as β). The product approach determines that M is a mediator if both α and β effects are significantly different from zero. Then, it suggests that there exists a pathway from the explanatory variable X to the outcome Y, connected (mediated) by M, and the mediation effect can be estimated by taking the product of α and β.

When the outcome is continuous, the estimated mediation effects from the difference and the product approaches are the same; when the outcomes are binary (such as being healthy or afflicted with a certain type of disease), and when the (disease) rate is rare ($\leq 10\,\%$), the product approach and difference approach approximately coincide[4]. When the outcomes are

binary, but the (disease) rate is common, the product approach and the difference approach diverge. The difference approach becomes particularly challenging because the coefficient of the explanatory variable using this method may not change when there is mediation (see VanderWeele[12] and see Chapter 2 of VanderWeele[14] for a discussion).

Since the product approach is suitable for both rare and common cases (the dataset used in this study contains rare outcomes), we adopt this approach in this article. For the sake of completeness, we provide a brief technical summary of both approaches below.

The product approach. The product mediation analysis framework involves two steps[15–19]. Consider the setting in **Fig.** 2 (a) of the main text. The first step examines if the explanatory variable X is associated with the mediator M. This is done using a conditional regression model:

$$E(M|x) = \theta_0 + \alpha x \qquad \text{(S.7)}$$

where E refers to the expectation operation, and θ_0 and α are coefficients for the intercept, and the explanatory variable x. In simpler terms, if the coefficient α in (S.7) is significantly different from zero, it suggests that there is potentially an effect from the explanatory variable to the mediator. In other words, the red arrow below *alpha* in **Fig.** 2 (a) exists.

The second step checks if the mediator M is associated with the outcome Y, after controlling for the explanatory variable X. This is done using another conditional regression model:

$$E(Y|x, m) = \theta_0' + \gamma x + \beta m \qquad \text{(S.8)}$$

where θ_0', γ, and β are coefficients for the intercept, the explanatory variable x, and mediator m. In simpler terms, if the coefficient β in (S.8) is significantly different from zero, it suggests that there is potentially an effect from the mediator to the outcome. In other words, the red arrow below *beta* in **Fig.** 2 (a) exists.

Taken together, when both α and β are significantly different from zero, it means that (I) a change of X affects M, and (II) a change in M further affects Y; therefore M is significantly mediating the relationship between X and Y. In simpler terms, the mediation effect exists when the product $\alpha\beta$, which quantifies the mediation effect in this setting, is non-zero. In other words, both red arrows below *alpha* and *beta* in **Fig.** 2 (a) exist, thereby connecting X and Y via a pathway that passes through (mediated by) M. The remaining coefficient γ' in (S.8) quantifies the direct effect of the explanatory variable X on outcome Y (that is not mediated by M) (see **Fig.** 2 (a)).

The difference approach. The difference approach has two modeling steps. The first step is a conditional regression model, which investigates if variability in the explanatory variable X is responsible for variability in outcome Y:

$$E(Y|x, \boldsymbol{z}) = \theta_0 + \gamma x \qquad \text{(S.9)}$$

where E stands for expectation, θ_0 and γ are coefficients for the intercept and the explanatory variable x.

The second step is a conditional regression model, which checks if variability in the explanatory variable X and mediator M are both responsible for variability in outcome Y:

$$E(Y|x, m) = \theta_0' + \gamma' x + \beta m \qquad \text{(S.10)}$$

where θ_0', γ', and β are coefficients for the intercept, the explanatory variable x, and mediator m. Note the coefficients are denoted with a prime (*e.g.,* γ' in (S.10)) to highlight that they are not necessarily the same as their counterparts (*i.e.*, γ in (S.9)).

If the coefficient of the explanatory variable on the outcome when a mediator is included, or γ', is considerably smaller than the coefficient of the explanatory variable on the outcome when the

mediator is not included, or γ, then it suggests that the mediator is at least partially responsible for explaining the variance (information) of the outcome Y. To put more concretely, this indicates that including the mediator m, the explanatory variable x explains a less amount of variance (information) of Y. The difference, or $\gamma - \gamma'$, thus quantifies the mediation effect in this setting.

4. The relationship between the alpha atlas in mediation analysis and Canonical Correlation (CC) and Partial Least Squares (PLS) coefficients

Let's begin by introducing some notations that may facilitate our discussion below. Without loss of generality, let $\boldsymbol{M}_{\boldsymbol{M}|\boldsymbol{Z}}$ and $X_{X|\boldsymbol{Z}}$ denote an $n \times p$ brain imaging data from n subjects and p brain areas and a univariate exposure variable of dimension p; both of them have regressed out the covariate effects (so we do not have to consider covariates in the following). For simplicity, let's write $\boldsymbol{M}_{\boldsymbol{M}|\boldsymbol{Z}}$ and $X_{X|\boldsymbol{Z}}$ as $\tilde{\boldsymbol{M}}$ and $\tilde{X}$. Specifically, $\tilde{\boldsymbol{M}} = (\tilde{\boldsymbol{m}}_1, \tilde{\boldsymbol{m}}_2, \ldots, \tilde{\boldsymbol{m}}_p)$, where $\tilde{\boldsymbol{m}}_j = (\tilde{m}_{1j}, \tilde{m}_{2j}, \ldots, \tilde{m}_{nj})^T$is the image from brain region j (for $1 \le j \le p$) across all subjects, and $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^T$ is the symptom across all subjects.

The alpha atlas is then the estimates of $\boldsymbol{\alpha}$ in $\tilde{\boldsymbol{M}} = \tilde{X}\boldsymbol{\alpha}^T + \boldsymbol{\varepsilon}$.

First, **the alpha atlas is proportional to the correlation coefficients** between $\tilde{\boldsymbol{M}}$ and $\tilde{X}$. This is relatively easy to see, as:

$$\hat{\alpha}_j^{atlas} = r_j \frac{s_{\tilde{m}_j}}{s_{\tilde{X}}}, \text{ for } j = 1,2,\ldots,p \qquad \text{(S.11)}$$

where r_j is the sample correlation between exposure $\tilde{X}$ and signals from area brain area j, and $s_{\tilde{m}_j}$ and $s_{\tilde{X}}$ are sample standard deviation of signals from area brain area j and that of exposure $\tilde{X}$, respectively. We use superscript "atlas" to distinguish other alphas introduced below. More specifically:

$$\widehat{\boldsymbol{\alpha}}^{atlas} = \frac{1}{s_{\tilde{X}}} diag\left\{s_{\tilde{\boldsymbol{m}}_j}\right\}_{j=1,2,..,p} \boldsymbol{r}$$

where $\boldsymbol{r} := \Sigma_{\widetilde{M},\tilde{X}} = (r_1, r_2, \ldots, r_p)^T$ is the correlation between $\tilde{X}$ and each column of $\widetilde{\boldsymbol{M}}$.

Second, **alpha atlas is related to the CCA**. To see this, first realize there is only $\min\left(\text{rank}(\tilde{X}), \text{rank}(\widetilde{\boldsymbol{M}})\right) = 1$ canonical correlation (CC). The CC between a vector $\tilde{X}$ and a matrix $\widetilde{\boldsymbol{M}}$ can be found by solving $(\widehat{\boldsymbol{\alpha}}, \hat{b}) = \underset{\boldsymbol{\alpha},b}{\text{argmax}}\{corr(\widetilde{\boldsymbol{M}}\boldsymbol{\alpha}, b\tilde{X})\}$. Keep writing, we have $\widehat{\boldsymbol{\alpha}}^{CCA} \propto \Sigma_{\widetilde{\boldsymbol{M}}}^{-1}\Sigma_{\widetilde{M},\tilde{X}}\hat{b}$, where $\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}$ is the precision for $\widetilde{\boldsymbol{M}}$, $\Sigma_{\widetilde{M},\tilde{X}} = (r_1, r_2, \ldots, r_p)^T$, and $\hat{b} \propto \Sigma_{\tilde{X},\widetilde{M}}\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}\Sigma_{\widetilde{M},\tilde{X}}$ is a scalar. Thus, we have $\widehat{\boldsymbol{\alpha}}^{CCA} \propto \left\{\left(\boldsymbol{r}\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}\right)^T\left(\boldsymbol{r}\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}\right)\right\}\boldsymbol{r}$. Together with Equation (S.11), we have:

$$\widehat{\boldsymbol{\alpha}}^{CCA} \propto \left\{\left(\boldsymbol{r}\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}\right)^T\left(\boldsymbol{r}\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}\right) diag\left\{\frac{s_{\tilde{X}}}{s_{\tilde{\boldsymbol{m}}_j}}\right\}_{j=1,2,..,p}\right\} \times \widehat{\boldsymbol{\alpha}}^{atlas}. \qquad \text{(S.12)}$$

When all brain regions are pair-wise independent (although not realistic), $\hat{\alpha}_j^{CCA} \sim (\boldsymbol{r}^T\boldsymbol{r})\frac{\sqrt{Var(\tilde{X})}}{\sqrt{Var(\tilde{\boldsymbol{m}}_j)}} \times \hat{\alpha}_j^{atlas}$, which is a weighted version of the alpha atlas. In reality when brain regions are not pair-wise independent, the CCA map is proportional to $\widehat{\boldsymbol{\alpha}}^{atlas}$ as indicated by Equation (S.12), with each entry $\hat{\alpha}_j^{CCA}$ seen as some "contaminated" version of $(\boldsymbol{r}^T\boldsymbol{r})\frac{\sqrt{Var(\tilde{X})}}{\sqrt{Var(\tilde{\boldsymbol{m}}_j)}} \times \hat{\alpha}_j^{atlas}$, with "contamination" caused by correlations between brain areas ($\Sigma_{\widetilde{\boldsymbol{M}}}^{-1}$).

Third, **the alpha atlas is remotely related to the PLS coefficients.** To see this, let's begin by defining the PLS between exposure $\tilde{X}$ and brain imaging $\widetilde{\boldsymbol{M}}$ as:

$$\tilde{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{\varepsilon}$$

$$\widetilde{\boldsymbol{M}} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{e}$$

where $\boldsymbol{T}$ is $n \times l$, $\boldsymbol{P}$ is $1 \times l$, $\boldsymbol{U}$ is $n \times l$, and $\boldsymbol{Q}$ is $p \times l$, for some l, are decompositions of $\tilde{X}$ and $\widetilde{\boldsymbol{M}}$

such that information of $\tilde{X}$ and $\widetilde{\boldsymbol{M}}$ can be accounted for by each other.

The estimates of $\boldsymbol{T}$, $\boldsymbol{P}$, $\boldsymbol{U}$, and $\boldsymbol{Q}$ can be found by the following updating rule (we need this for our last argument):

$u := \widetilde{\boldsymbol{m}}_j$, for some j
Loop:
$\boldsymbol{p} \leftarrow \tilde{X}^T\boldsymbol{u}/\left\|\tilde{X}^T\boldsymbol{u}\right\|$
$\boldsymbol{t} \leftarrow \tilde{X}\boldsymbol{p}$
$\boldsymbol{q} \leftarrow \widetilde{\boldsymbol{M}}^T\boldsymbol{t}/\left\|\widetilde{\boldsymbol{M}}^T\boldsymbol{t}\right\|$
$\boldsymbol{u} \leftarrow \widetilde{\boldsymbol{M}}\boldsymbol{q}$
Until $\boldsymbol{t}$ stops changing.

Next, let $\boldsymbol{\beta}$ be the estimate between $\boldsymbol{U}$ and $\boldsymbol{T}$, or $\boldsymbol{U} = \boldsymbol{T}\boldsymbol{\beta}$. We then have:

$$\widetilde{\boldsymbol{M}} = \boldsymbol{U}\boldsymbol{Q}^T = \boldsymbol{T}\boldsymbol{\beta}\boldsymbol{Q}^T = \tilde{X}\boldsymbol{P}^T\boldsymbol{\beta}\boldsymbol{Q}^T.$$

Let $\widehat{\boldsymbol{\alpha}}^{PLS}$ be the coefficients obtained between exposure $\tilde{X}$ and brain imaging $\widetilde{\boldsymbol{M}}$ using PLS, then:

$$\widehat{\boldsymbol{\alpha}}^{PLS} = \boldsymbol{P}^T\boldsymbol{\beta}\boldsymbol{Q}^T.$$

From the updating rule above, we have:

$$\boldsymbol{p} = \frac{\tilde{X}^T\boldsymbol{u}}{\left\|\tilde{X}^T\boldsymbol{u}\right\|} = \frac{\tilde{X}^T\widetilde{\boldsymbol{M}}\boldsymbol{q}}{\left\|\tilde{X}^T\widetilde{\boldsymbol{M}}\boldsymbol{q}\right\|} = \frac{\tilde{X}^T\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{M}}^T\boldsymbol{t}}{\left\|\tilde{X}^T\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{M}}^T\boldsymbol{t}\right\|} = \frac{\tilde{X}^T\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{M}}^T\tilde{X}\boldsymbol{p}}{\left\|\tilde{X}^T\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{M}}^T\tilde{X}\boldsymbol{p}\right\|}$$

$$= \frac{1}{\lambda}(\widetilde{\boldsymbol{M}}^T\tilde{X})^T\widetilde{\boldsymbol{M}}^T\tilde{X}\boldsymbol{p}$$

for some $\lambda > 0$.

Thus, $\boldsymbol{p}$ is an eigen-vector of the covariance matrix of $\widetilde{\boldsymbol{M}}^T\tilde{X}$. Without loss of generality, assuming that both $\widetilde{\boldsymbol{M}}$ and $\tilde{X}$ have been zero-centered, then $\widetilde{\boldsymbol{M}}^T\tilde{X} = (r_1, r_2, \ldots, r_p)^T$. Recall that $\widehat{\boldsymbol{\alpha}}^{atlas} = \frac{1}{s_{\tilde{X}}} diag\left\{s_{\tilde{\boldsymbol{m}}_j}\right\}_{j=1,2,..,p} \boldsymbol{r}$. Hence, $\widehat{\boldsymbol{\alpha}}^{PLS}$ is related to $\widehat{\boldsymbol{\alpha}}^{atlas}$ through $\boldsymbol{P}$ and more specifically through the correlation coefficients $\boldsymbol{r}$, although very remotely.

5. A brief procedure for bootstrap experiments used in this study

Denote $(\boldsymbol{x}_1^*, \boldsymbol{y}_1^*, \boldsymbol{M}_1^*)$, $(\boldsymbol{x}_2^*, \boldsymbol{y}_2^*, \boldsymbol{M}_2^*), \ldots, (\boldsymbol{x}_B^*, \boldsymbol{y}_B^*, \boldsymbol{M}_B^*)$ as B sets of bootstrap data, where the explanatory variable, the outcome variable, and the mediator are jointly sampled from ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{M}$). The j^{th} bootstrap sample has the same dimension as ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{M}$); namely $\boldsymbol{x}_j^*$ and $\boldsymbol{y}_j^*$ each contains N bootstrapped observations and $\boldsymbol{M}_j^*$ is $N \times V$. Let $NIE_1^*(j), NIE_2^*(j), \ldots, NIE_B^*(j)$ be the estimated mediation effect from each bootstrap experiment for the j^{th} voxel. Specifically, $NIE_b^*(j)$ is the estimated indirect mediation effect corresponding to the j^{th} voxel using data $(\boldsymbol{x}_b^*, \boldsymbol{y}_b^*, \boldsymbol{M}_b^*)$.

The bootstrap *t*-statistic for the mediation effect corresponding to the j^{th} voxel is defined as:

$$t_j^* = \frac{\overline{NIE}^*(j)}{SE_{NIE^*(j)}}$$

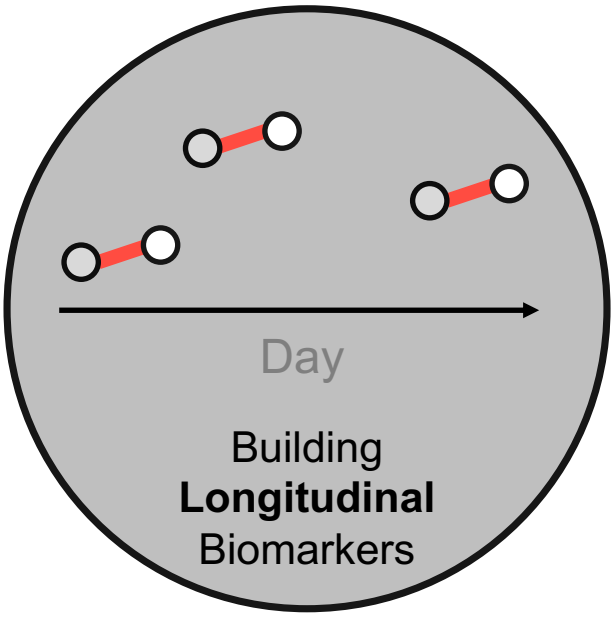
where $\overline{NIE}^*(j) = \frac{\sum_{i=1}^{B} NIE_i(j)}{B-1}$ is the mean indirect mediation effect for the j^{th} voxel, averaged across B bootstrap experiments, and $SE_{NIE^*(j)} = \frac{\sum_{i=1}^{B}(NIE_i^*(j) - \overline{NIE}^*(j))^2}{\sqrt{B}}$ is the standard error.

References

1. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, pp. 688–701 (1974).
2. Albert, J. M. Mediation analysis via potential outcomes models. *Stat. Med.* **27**, pp. 1282–1304 (2008).
3. Imai, K., Jo, B. & Stuart, E. A. Commentary: Using potential outcomes to understand causal mediation analysis. *Multivariate Behav. Res.* **46**, pp. 861–873 (2011).
4. VanderWeele, T. J. & Vansteelandt, S. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology* **172**, pp. 1339–1348 (2010).
5. Pearl, J. Direct and indirect effects. In *Proc. seventeenth Conf. Uncertain. Artif. Intell.*, edited by J. Breese and D. Koller, pp. 411–420. Morgan Kaufmann Publishers, San Francisco, CA, USA (2001).
6. Vanderweele, T. J. & Vansteelandt, S. Conceptual issues concerning mediation, interventions and composition. *Stat. Interface* **2**, pp. 457–468 (2009).
7. Robins, J. M. & Richardson, T. S. Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, edited by P.E. Shrout, K.M. Keyes and K. Ornstein. Oxford University Press, Oxford, UK (2010).
8. Pearl, J. Interpretation and identification of causal mediation. *Psychol. Methods* **19**, pp. 459–481 (2014).
9. Huang, Y. T. & Pan, W. C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72**, pp. 402–413 (2016).
10. VanderWeele, T. J. & Vansteelandt, S. Mediation analysis with multiple mediators. *Epidemiol. Method.* **2**, pp. 95–115 (2014).
11. Ogburn, E. L. Commentary of 'Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables' by Dylan Small. *J. Stat. Res.* **46**, pp. 105–111 (2012).
12. VanderWeele, T. J. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* **37**, pp. 17–32 (2016).
13. Mackinnon, D. P. & Dwyer, J. H. Estimating mediated effects in prevention studies. *Eval. Rev.* **17**, pp. 144–158 (1993).
14. VanderWeele, T. J. Explanation in causal inference: Developments in mediation and interaction. *Int. J. Epidemiol.* **45**, 1904–1908 (2016).
15. Hyman, H. H. *Survey design and analysis: Principles, cases and procedures*. Free Press, New York, NY, USA (1955).
16. Alwin, D. F. & Hauser, R. M. The decomposition of effects in path analysis. *Am. Sociol. Rev.* **40**, pp. 37–47 (1975).
17. Judd, C. M. & Kenny, D. A. Process analysis: Estimating mediation in treatment evaluations. *Eval. Rev.* **5**, pp. 602–619 (1981).
18. Sobel, M. E. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* **13**, pp. 290–312 (1982).
19. Baron, R. M. & Kenny, D. a. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51**, pp. 1173–1182 (1986).

Chapter 5

Building longitudinal biomarkers



Personalized Longitudinal Assessment and Monitoring of Multiple Sclerosis Using Smartphones[1]

Abstract

Personalized longitudinal disease assessment and monitoring are central to early diagnosis, appropriate management, and timely treatment of multiple sclerosis (MS). They are also important for identifying idiosyncratic disease profiles of each patient. Here, we design a novel longitudinal model to map individual disease trajectories in an automated way using smartphone data that may contain missing values. First, we collect digital measurements related to hand and motor functions using a smartphone application. Next, we treat missing data via imputation. We then discover potential markers of MS by employing a generalized estimation equation. Subsequently, parameters learned from multiple training datasets are ensembled to form a simple, unified longitudinal predictive model to forecast MS over time in previously unseen subjects. To mitigate potential underestimation for individuals with severe disease scores, the final model incorporates an additional subject-specific fine-tuning using data from the first day. We demonstrate the method using a smartphone-based MS study. The results show that the proposed model is promising to achieve personalized longitudinal MS assessment and monitoring; they also suggest that hand and gait features remotely collected from smartphones may be useful digital markers for predicting MS over time.

[1]This chapter is organized as a paper in preparation: Chén, O.Y., Lipsmeier, F., Phan, H., Dondelinger, F., Scotland, A., Creagh, A., Gossens, C., Lindemann, M., and de Vos, M. Personalized Longitudinal Assessment and Monitoring of Multiple Sclerosis Using Smartphones.

1. Introduction

Multiple sclerosis (MS) is the most common autoimmune disorder of the central nervous system[1]. Worldwide, approximately 2.3 million people are affected by the disease[2,3], with new cases developing at a rate ranging from 0.04 (from low-income countries) to 4.7 (from high-income countries) *per* 100,000 individuals every year[3].

The immune system of MS patients attacks the oligodendrocytes, the cells responsible for creating and maintaining the myelin sheath[4], resulting in demyelination (thinning or completely losing myelin). Such abnormal immune response is tremendously health-threatening as the myelin plays a critical role in transmitting action potentials (electrical impulses) between brain cells, critical for cell-to-cell communication[5]. As demyelination develops, communication within and between nervous systems begins to deteriorate, causing physical, mental, and sometimes psychiatric problems[4,6]. The majority of patients exhibit paresthesia, transitory abnormal sensory feelings such as numbness, prickling, or "pins and needles" sensations; most patients during later stages experience muscle weakness and have difficulties with coordination and balance[7]. Additionally, suicide is more common among MS patients; infections and other complications are especially dangerous in patients with advanced MS [4]. Eventually, most patients lose the ability to walk[8], and two-thirds of the deaths in MS patients are direct consequences of the disease[4]. Overall, the average life expectancy for MS patients is 30 years from the onset of the disease, 5 to 10 years shorter than that of unaffected individuals[4].

Traditionally, individuals at risk of developing MS are examined based on a combination of in-clinic assessments, including checking for MS-related signs and symptoms, neuroimaging studies, and laboratory testing[9]. The disease severity can be rated by clinicians using the Expanded Disability Status Scale (EDSS), which quantifies disability in functional systems and

ambulation[10]. Alternatively, the severity can be self-assessed using the Multiple Sclerosis Impact Scale (MSIS-29), which measures the physical (20 items) and psychological (nine items) impact of the disease[11–13]. The EDSS score ranges from 0 to 10 (representing normal to death due to MS). The MSIS-29 total score (physical score plus psychological score) ranges from 25 to 145 (representing best to worse physical and psychological functions).

To early diagnose, appropriately manage, and timely treat MS, it is critical to building suitable longitudinal models. Nevertheless, we must confront a few challenges when developing longitudinal models for MS studies. First, MS disease scores are at present assessed sparsely in clinics. How, then, could one paint a longitudinal picture of MS based upon infrequently measured data? Second, for some patients, the disease may monotonically progress; for some others, it may relapse-remit (in other words, a patient may enter a period of remission after presenting disease activity). For patients with primary progressive MS, how could one make an accurate forecast and assign proper treatment when the assessments made earlier may inadequately represent the disease condition at a later time? How could one distinguish a relapsing-remitting patient from a progressive patient with minor symptoms? Third, although frequent in-clinic evaluation may potentially alleviate the aforementioned two challenges, it is unfortunately rather inconvenient, costly, and at present impractical. Fourth, even if data are frequently collected over time, they oftentimes contain missing values. If not treated, one cannot make full use of the data and, particularly, cannot make disease assessments on days with incomplete data. Finally, although MS patients exhibit common patterns at the population-level, personal disease prediction on novel subjects using results discovered from others may not capture the subject-specific information of the new subjects. Can we develop a method that balances extrapolation and personalization?

The recent development in smartphone technology brings in promises to assess and monitor MS frequently over time. First, smartphone sensors collect a wealth of bio-signals that are potentially related to several critical symptom domains (such as hand function and walking ability) of the MS. Second, smartphones are portable, and allow for measuring health information semi-continuously (*i.e.*, at frequent intervals). The resulting repeated measurements make it possible to evaluate the disease longitudinally. Additionally, their relatively low cost and high penetration rate (as compared to the availability of hospital devices and physicians) make health services affordable to a large population[14,15]. Finally, smartphones can transfer data remotely to terminals and (data) cloud via wireless, ultra-fast networks, thereby facilitating out-of-clinic data collection, analysis, and remote, home-based disease assessment and monitoring[16].

Embracing these advantages, here we propose a personalized longitudinal framework to assess MS over time in an automated way using smartphones. Through data analyses and out-of-sample testing, we demonstrate that the framework is potentially useful to (1) extract MS-specific digital clinical markers, and (2) assess and monitor individual MS trajectories longitudinally.

2. MS data collected remotely at home using smartphones

We used data from a smartphone-based remote monitoring study (NIH Clinical Trial Identifier: NCT02952911). The study consisted of 76 MS patients (18–55 years; EDSS score 0–5.5). Smartphone data were collected using the Floodlight application, developed by Roche and Genentech[17]. The application contained a range of remote monitoring tests that measure a person's ability to perform simple tasks on smartphones. Each participant received a preconfigured smartphone that prompts the user to perform the Floodlight test battery,

comprising ‘active tests’ and ‘passive monitoring’, for 24 weeks[18]. The local ethics committees approved the study. Written informed consent was obtained from all participants. The exclusion criteria for MS patients were: (1) severely ill and unstable participants as per investigator's discretion; (2) change in dosing regimen or switch of Disease Modifying Therapy (DMT) in the last 12 weeks before enrollment; (3) pregnant, lactating, or intending to become pregnant during the study.

The data considered in this study are objective measurements obtained from a Two-Minute Walk (TMW) test, which contains gait features, and a Draw a Shape (DaS) test, which contains hand function features. Specifically, the gait features assess difficulties, fatigability, or unusual patterns during a two-minute walk test. The DaS test consists of drawing five shapes with circular and linear features to assess fine finger control and stroke sequencing. The reason to consider these data is twofold. First, walking ability and hand function are indispensable to support independence in activities of daily living. Their functions, unfortunately, are disrupted in MS patients[19,20]. Thus, it is important to isolate the subset(s) of gait and hand function features that are disrupted in MS patients and are potentially predictive of MS disease severity. Their identification may help improve one's understanding of the pathological aspects of MS related to motor and hand functions. Additionally, they may provide better means of prevention and interventions to mitigate gait and dexterity symptoms or to sustain/improve patients’ ability to walk and perform manual tasks. Second, the magnitude of disruption in these features may be associated with the severity of MS. For example, degenerating walking ability is associated with MS disease progression and/or one or more relapses[21]. Deteriorating hand function is frequently reported in people with MS and is increasingly present as the disease progresses[22]. Together, these lines of evidence suggest that gait and hand function features may be potentially useful markers to assess and monitor the severity and progression of the disease.

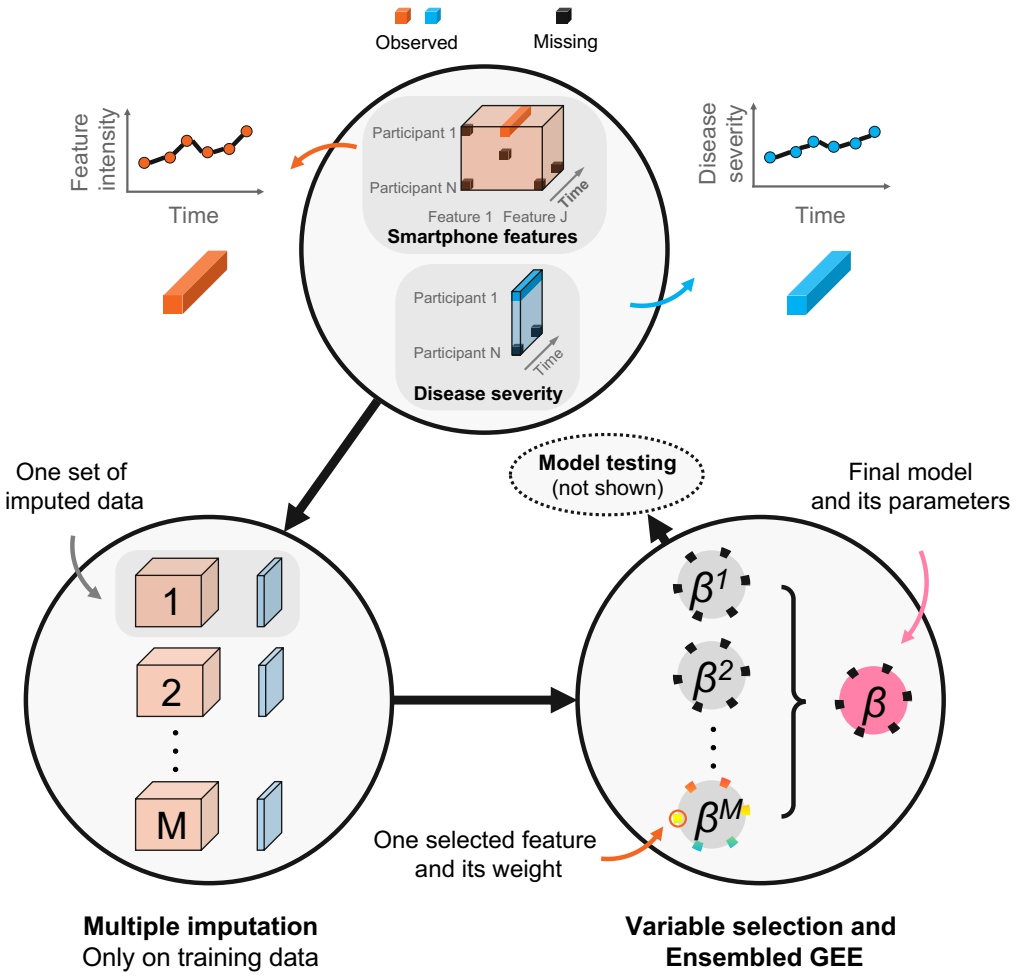


Fig. 1. A schematic representation of the longitudinal model. Top: Data organization. The smartphone data are organized in an orange cubic, where multi-dimensional feature data are collected from each participant over time. Specifically, the orange bar indicates one feature measured over time for one subject. Disease outcomes are similarly prepared, where the blue bar shows the disease scores measured over time. The black boxes refer to missing data. **Lower left: Data imputation.** Missing data are imputed via multiple imputation (MI). To accommodate the uncertainties in the imputed data, we compute multiple (M) sets of complete data. **Lower right: Model development.** A generalized estimation equation is fit on each of the fully imputed dataset to obtain parameter estimation. The M sets of estimated parameters are then pooled to generate a unified model, which will be used for further test. The (pooled) parameters are applied to features from the test sample to estimate outcomes. To prevent data imputation from influencing the test results, only complete observations will be used during the test (namely, MI will not be used during the test). The estimated results are then compared with the observed outcomes to verify the performance as well as the reproducibility of the model.

The study also included the clinician-rated Expanded Disability Status Scale (EDSS) scores, and the self-reported MS Impact Scale (29-item scale) (MSIS-29) scores. They were used to quantify disease severity. For the present study, we mainly focused on predicting MSIS-29 scores because they were collected from the same smartphone application every two weeks, and the data were available more frequently; consequently, they were more suitable for developing and evaluating a longitudinal model. Nevertheless, for completion purposes, we also evaluated and briefly discussed the model performance for the prediction of EDSS scores in this paper. One should, however, note that there were only a very small number of EDSS scores per subject for training and test. A full description of the study and data can be found in[17,18,23].

3. Methods

3.1 Treating missing data in longitudinal studies

Missing data are common to data collection and pose particular challenges for longitudinal disease prediction. When there are missing inputs (in the design matrix) or missing outputs for the training data, model development becomes difficult. Therefore, a practical longitudinal disease predictive model needs to first treat the problem of missing data.

In this paper, we used multiple imputation (MI)[24] to handle missing data. We adopted it primarily because it was relatively easy to control for model efficiency (see **Eq.** (4)) by adjusting the number of imputations performed[25,26] (see below). Additionally, it provided a principled way to estimate the uncertainty associated with the imputation[27,28]. Finally, for practitioners, MI may be easier to implement as it is readily available in standard software. One can refer to[29–32] for additional treatments of missing data.

Here, we briefly outline the steps of MI. First, a column that contains missing values is regarded as a target column (and treated as a response variable), and the remaining columns of the data set are used as predictors. The missing values in the targeted column are then filled using predicted mean matching (PMM)[32,33]. For predictors that are incomplete in themselves, the most recently generated values are used so that the predictors are complete before making imputations for the target column. Second, the first step is repeated for every column that contains missing data. Finally, the first two steps are repeated multiple (M) times to obtain M sets of complete data. These steps can be done using the publicly available *R* package *mice*[34].

3.2 Model development using the GEE

The longitudinal predictive model ensembles multiple generalized estimation equations (GEEs) followed by a subjective-specific fine-tuning. The choice of choosing the GEE is twofold. First, it can take both correlated and uncorrelated repeated longitudinal measurements (both within and between subjects). Second, even if one misspecified the correlation structure, the parameter estimates would still be consistent[35]. The subject-specific fine-tuning is added to treat potential overfitting during training and underestimating for novel subjects with severe scores during the test.

Before introducing the subject-specific aspect of the model (see **Section** 3.5), let's first outline the general longitudinal predictive model. Consider N subjects, where the i^{th} subject has p_w gait features and p_d hand function features, measured at time $t = (1, 2, \dots, T_i)$. The feature matrix for the i^{th} subject is thus $\boldsymbol{x}_i^T = \left(\boldsymbol{x}_{iw}^T, \boldsymbol{x}_{id}^T\right)$, $\boldsymbol{x}_{ij}^T = \left(\boldsymbol{x}_{ij1}, \dots, \boldsymbol{x}_{ijp_j}\right)$, for $j \in \{w, d\}$, and $\boldsymbol{x}_{ijk}$contains longitudinal measurements $\boldsymbol{x}_{ijk} = (x_{ijk}(1), \dots, X_{ijk}(T_i))^T$, for $1 \leq k \leq p_j$, where T

denotes the transpose operation. Let $\boldsymbol{y_i} = (y_{i1}, \dots, y_{iT_i})^T$ be the longitudinal outcome for the i^{th} subject and y_{it} be the outcome for the i^{th} subject at time *t*.

The study consists of $p_w = 2$ gait features from a Two-Minute Walk (TMW) test, and $p_d = 51$ hand function features from a Draw a Shape (DaS) test. For each individual i, the number of repeated measures, or T_i, ranges from 1 (two subjects) to 23 with a mean of 8.5. The model also includes covariates such as gender, age, race, and site (from where data were collected). Unless explicitly stated, the features for model development discussed in this section were from 57 randomly selected training subjects. Features from the remaining 23 subjects were held out for the test. To avoid a chance split during training and test, we will perform bootstrap experiments in **Section** 4.

We considered feature selection in two steps[36]. The first step selected features in the training set that were significantly associated with the outcome. The second step selected features using the GEE.

Consider a feature $\boldsymbol{x}$. Denote $\rho_{(x,\boldsymbol{y})}$ as the result from a statistical test during the step-one feature selection between the feature $\boldsymbol{x}$ and the outcome $\boldsymbol{y}$. For disease severity outcomes, the value of ρ is the correlation between the feature and the outcome; equivalently, it could be the corresponding *p*-values. Let ε be a pre-specified threshold for ρ. Although there is a one-to-one mapping between a statistic and its *p*-value, sometimes it may be convenient to evaluate the *p*-value, sometimes it may be convenient to evaluate the correlation, and other times, one may be interested in constraining the number of features being selected. For example, in this study we set ε to control the total number of features.

Six features (including covariates and smartphone features) were allowed in the final model; we explored other thresholds that allowed different numbers of features in the final model as a complementary analysis (see **Fig.** 4). The reason that we constrained the total number of features in the model was due to missing data in the test dataset. Although one can deal with missing data in the training set via imputation, to avoid bias brought into the test data by imputation, we chose to not perform MI on the test set and evaluated the model performance only on available observations. Naturally, if one included a larger number of features, there would be fewer numbers of subjects left with complete features. We found that a total of six features would balance the number of features for training and the number of available test data (see **Fig.** 4).

Formally, the (general) marginal model that studies the relationship between the feature and outcomes is:

$$\mathbb{E}(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{z}_i) = g^{-1}(\mu + \boldsymbol{x}_{it}^T \boldsymbol{S}\boldsymbol{\alpha} + \boldsymbol{z}_i^T \boldsymbol{\gamma}) \qquad (1)$$

where $g(\cdot)$ is a link function, μ is the intercept, $\boldsymbol{x}_{it}^T = [\boldsymbol{x}_{iwt,}^T \boldsymbol{x}_{idt}^T]$ and $\boldsymbol{x}_{ijt}^T = (x_{ij1}(t), x_{ij2}(t), \ldots, x_{ijp_j}(t))$, for $j \in \{w, d\}$. Here, $\boldsymbol{S} = blockdiag\{\boldsymbol{I}_w, \boldsymbol{I}_d\}$, and $\boldsymbol{I}_j = diag\{i_{j1}, i_{j2}, \ldots i_{jp_j}\}$, wherein $i_{jk} = 1$ if $\rho_{(\boldsymbol{x_{jk}}, \boldsymbol{y}, N-2)} < \varepsilon$ and 0 otherwise, where $\boldsymbol{x_{jk}} = (\boldsymbol{x}_{1jk}^T, \boldsymbol{x}_{2jk}^T, \ldots, \boldsymbol{x}_{Njk}^T)^T$ denotes a particular feature across all subjects over time; $\boldsymbol{z}_i$ is a vector containing all covariates for the i^{th} subject, and $\boldsymbol{\gamma}$ is its coefficient. Finally, write $\boldsymbol{\beta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$.

Denote the selected features as $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{S}$ and the outcome disease scores as $\boldsymbol{Y}$. Via MI, we obtained M sets of complete data $\left(\boldsymbol{Y}^{(1)}, \widetilde{\boldsymbol{X}}^{(1)}, \boldsymbol{Z}\right), \ldots, \left(\boldsymbol{Y}^{(M)}, \widetilde{\boldsymbol{X}}^{(M)}, \boldsymbol{Z}\right)$, where the superscript (m) denotes the m^{th} set of imputed data.

The problem in **Eq.** (1) can be solved using the GEE. Specifically, the estimation of $\boldsymbol{\beta}$ can be done by solving the following score function:

$$\boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0$$

where $\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}$, $\boldsymbol{\mu}_i$ is the expectation of $\boldsymbol{y}_i$, and $\boldsymbol{V}_i$ is the estimate (see below) of the variance-covariance matrix of $\boldsymbol{y}_i$. Let:

$$\boldsymbol{V}_i = \Delta_i^{\frac{1}{2}} \mathcal{R}_i(\alpha) \Delta_i^{\frac{1}{2}}$$

where $\Delta_i = diag\{var(y_{i1}), var(y_{i2}), \ldots, var(y_{iT_i})\}$, $\mathcal{R}_i(\alpha)$ is a T_i by T_i "working" correlation matrix[35], and α represents a vector of parameter associated with a specific model for correlation of $\boldsymbol{y}_i$. When $\mathcal{R}_i(\alpha)$ is the true correlation matrix for $\boldsymbol{y}_i$, then $\boldsymbol{V}_i = cov(\boldsymbol{y}_i)$. When the variance-covariance structure is mis-specified, the parameter estimates would still be consistent[35].

3.3 Inference

A GEE model was fit on selected features from each imputed data set. Let $\widehat{\boldsymbol{\beta}}^m = (\hat{\beta}_1^m, \hat{\beta}_2^m, \ldots, \hat{\beta}_p^m)$ denote the estimated parameter corresponding to the m^{th} set of imputed data, where p is the number of selected features plus the number of covariates, and $1 \leq m \leq M$. Then, the $(1-\alpha) \times 100\%$ confidence interval (CI) for the estimated $\hat{\beta}_j$ corresponding to the j^{th} feature, where $1 \leq j \leq p$, is:

$$\left[\bar{\hat{\beta}}_J - t_{\left(n-1,\frac{\alpha}{2}\right)} \frac{s_j}{\sqrt{M}}, \quad \bar{\hat{\beta}}_J + t_{\left(n-1,\frac{\alpha}{2}\right)} \frac{s_j}{\sqrt{M}}\right]$$

where $\bar{\hat{\beta}}_J = \sum_{m=1}^{M} \frac{\hat{\beta}_J^m}{M}$ and $s_j = \sqrt{\sum_{m=1}^{M} \left(\hat{\beta}_J^m - \bar{\hat{\beta}}_J\right)^2 / M}$.

3.4 Model ensemble and out-of-sample prediction

We obtained the ensemble parameter $\bar{\boldsymbol{\beta}}$ from M GEE models as:

$$\overline{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\beta}}^{m}.$$

Afterwards, longitudinal prediction on disease outcomes for a new subject k at time t in the test set is made using:

$$\hat{y}_{kt} = g^{-1}(\hat{\mu} + \widetilde{\boldsymbol{X}}_{kt}\overline{\boldsymbol{\beta}} + \boldsymbol{z}_k\widehat{\boldsymbol{\gamma}}) \qquad (2)$$

where $\widetilde{\boldsymbol{X}}_{kt} = \boldsymbol{X}_{kt}\widehat{\boldsymbol{S}}$ represents the selected features from subject k at time t and the selection is dependent only on the training set (as $\widehat{\boldsymbol{S}}$ is estimated from the training data).

It is worthwhile mentioning that, besides averaging M sets of parameters to yield a unified predictive model, one could alternatively average multiple sets of predictions from each of the M models. These two types of averaging techniques, however, yielded very similar outcomes in this study (see **Supplementary Information**), and we proceed with the predictive model described in **Eq.** (2).

The data organization, imputation, model development, and parameter ensemble are summarized in **Fig.** 1. As a remark, we did not perform MI for the test data to ensure that (1) the evaluation on model performance was only done on the observed (test) data; and (2) the ability of the model to conduct personalized disease assessment on new subjects was independent of the imputation mechanism.

3.5 Personalized longitudinal disease assessment and its improvement

Longitudinal disease assessment and monitoring need to address two important problems: extrapolation and personalization. For extrapolation, it needs to identify behavior markers that are present in a broad patient population so that they can be extended to novel samples. For

personalization, it needs to account for idiosyncratic information that is unique to each patient and is therefore potentially not included in the population-level parameters.

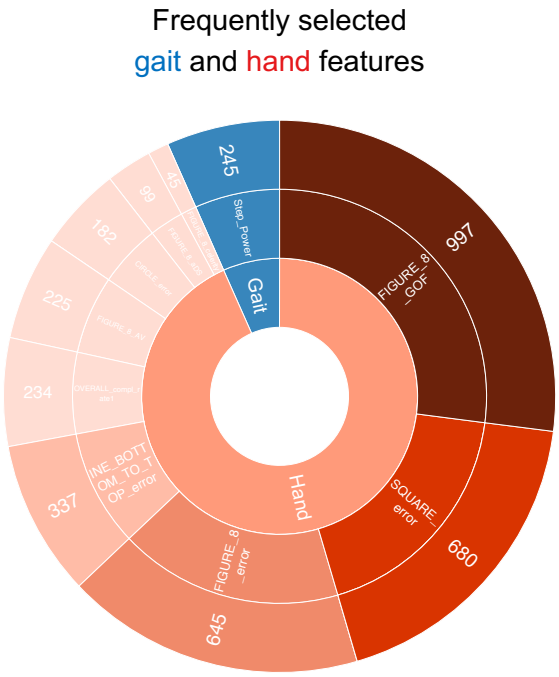


Fig. 2. Key features in smartphone-based remote longitudinal MS prediction. The distribution of selected gait and hand features. One thousand bootstrap experiments were performed to discover which gait and hand features were consistently selected for longitudinal MSIS-29 prediction. The selected gait and hand features were plotted on a pie chart: from inner circle to outer circle were **feature category** (gait *vs.* hand features), **feature name**, and **the frequency** each feature was selected from one thousand bootstrap experiments. Gait feature(s) were coded in blue and hand feature(s) were coded in red.

In methods development, these two sets of problems translate to the following tasks.

Task 1: Can we discover features whose patterns over time are associated with longitudinal outcomes and such association is generally present in all subjects in the training sample? If so, they may be useful to make forecasts for novel individuals (see **Fig.** 2); this can be further verified using out-of-sample testing.

Task 2: Can we further adjust the longitudinal predictions to reflect individual differences? Stated differently, the trained parameters during GEE model development reflect the relationship between the features and the outcomes specific to the training sample; if we were to use these parameters to make forecasts about outcomes in new subjects, they may not fully reveal the relationship between the features and outcomes in novel individuals. For example, a population-level model in practice may underestimate disease severity in novel subjects during out-of-sample testing (see **Figs.** 3 and 5). On the other hand, if one considers subject-specific parameters (such as that in the mixed effect models) or incorporates prior information about a subject (such as that in Bayesian learning) for the training model, such parameters or priors still only reflected individual information with regards to the training sample. Naturally, one would want to make an individualized adjustment to the population-level model (whose parameters were developed on the training sample) to incorporate subject-specific information that is contained in the test sample. But constructing an (independent) subject-specific model for the test sample to include idiosyncratic information may over-fit the test data.

To consider and balance extrapolation and personalization, we added the following subject-level fine-tuning into the longitudinal model:

$$\hat{y}_{kt}^{(adj)} = \begin{cases} \hat{y}_{kt}, & t = 1 \\ \hat{y}_{kt} + (y_{k1} - \hat{y}_{k1}), & t \geq 2 \end{cases} \qquad (3)$$

where $\hat{y}_{kt}$ is the estimated outcome for a new subject *k* at time *t* using **Eq.** (2), y_{k1} is the earliest ($t = 1$) observed outcome for subject k and $\hat{y}_{k1}$ is its estimation, and $\hat{y}_{kt}^{(adj)}$ is the fine-tuned estimate.

In words, the above correction means that after disease estimation has been made using the population-level parameters from the ensemble GEE on new subjects (see dashed lines in **Fig.**

3), we make a subject-level correction (by adjusting a constant $(y_{k1} - \hat{y}_{k1})$) on estimates made from day two and beyond using disease information from the first day (see solid red lines in **Fig.** 3).

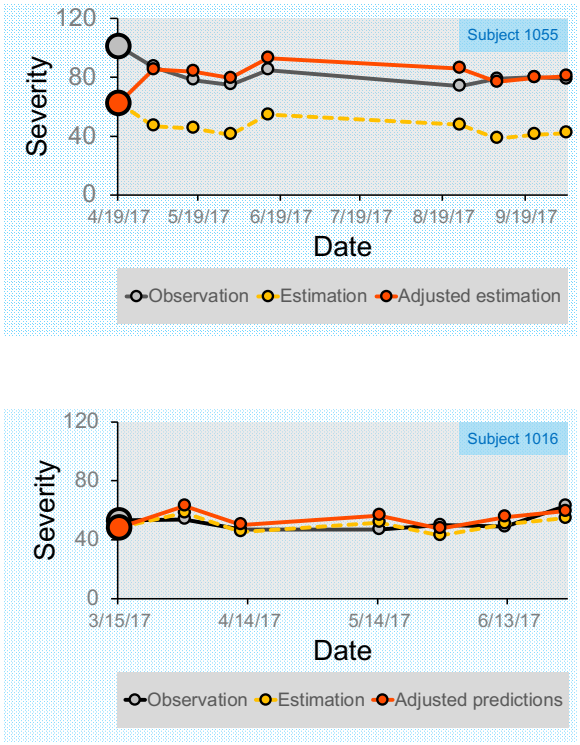


Fig. 3. Personalized longitudinal disease assessment using smartphones. Two individuals, one with severe MSIS-29 scores (top panel) and another with minor scores (bottom panel), are shown. The grey circles represent observed longitudinal scores. The yellow circles are estimated scores using the ensemble GEE parameters. The red circles are estimated scores further fine-tuned using data from the first day (see text for details).

There are a few reasons for using one day's data for fine-tuning (instead of using multiple days) in this study. First, when there are only a small number of data points recorded, using data from multiple days for tuning would result in fewer days of data left for testing (and could potentially artificially boost test results). Second, our analyses demonstrated that fine-tuning using data from one day indeed led to improved personalized disease monitoring while avoided over-fitting the test data. The reason that the fine-tuning avoided over-fitting was twofold. (a) It

did not use any information from the outcomes obtained after day one. (b) It only adjusted a constant $(y_{k1} - \hat{y}_{k1})$ on the prediction. Had the estimated longitudinal trend been inaccurate using **Eq.** (2) (see yellow line in **Fig.** 3), a constant adjustment in **Eq.** (3) would not improve the trend prediction. Finally, when longer test data become available for each individual in the future, one can potentially further improve the personalization by using a fine(r)-tuning based on more data or by extracting a subject-level prior for each novel subject using data obtained during an initial period. But our message of considering and balancing extrapolation and fine-tuning from the current proof-of-concept study remains.

3.6 Optimal number of imputed datasets

When employing MI, an interesting question arises. How many sets of imputed data are needed? Although the percentage of missing data in our MS data set is relatively small ($\gamma = 1.42\%$), missing just one entry of an important feature on a particular day would make it impossible to make a disease assessment on that day. More concretely, this question needs to be addressed in two aspects. First, how many sets of imputations would yield high prediction accuracy – this needs to be evaluated empirically. Second, theoretically, how can we determine that the so-called “high prediction accuracy” is efficiently high?

To examine the optimal number of sets of imputations empirically, we selected a variety number of choices for M and calculated their out-of-sample prediction accuracies (quantified as the correlation between true and estimated disease severity scores) (see **Fig.** 4). Our results showed that the prediction accuracy with MI was generally higher than it with missing values. Additionally, once more than 15 imputations were made, the accuracy became relatively stable.

Next, the number of imputations and efficiency has the following approximate relationship[27]:

$$\text{Efficiency} = (1 + \frac{\gamma}{M})^{-1} \qquad (4)$$

where γ refers to the percentage of missing data, and M denotes the number of imputations. When setting $M = 15$, the theoretical efficiency was 99.9% using the above formula. Since adding more imputations had not significantly improved prediction accuracy or efficiency and making too many sets of imputations would incur unnecessary computing time, we chose $M = 15$ for the remaining of the paper.

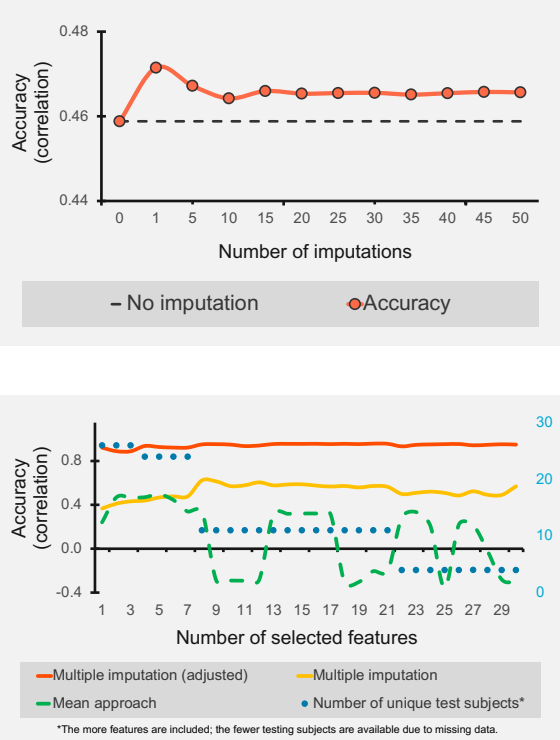


Fig. 4. Model comparison for longitudinal MSIS-29 prediction using different sets of imputations and various numbers of features. Top: Model performance using different numbers of imputations. Model performance improved with imputations compared to without, and generally stabilized when there were more than 15 imputations. **Bottom: Model performances for the mean approach, raw GEE approach, and adjusted GEE approach across different numbers of selected features.** Due to missing values in features, the size of the available independent test sample (indicated by the height of the blue dots) decreased when more features were used. This is because we only considered observed features for testing (namely we did not perform MI for the test sample to prevent it from potentially biasing prediction performance). In general, the prediction accuracy using MI and GEE (indicated by the yellow line) was relatively high and stable across various numbers of features selected. The fine-tuning further improved prediction performance (see the orange line).

4. Results

In this section, we present longitudinal MS prediction results using smartphone data. Throughout, to deal with uncertainty resulted from imputation, we generated 15 imputed values for each missing value. That is, for each experiment, 15 sets of complete data containing imputed values were used to train the model. Fifteen GEE models were subsequently fit for each of the imputed, and now full, training data set. The resulting 15 sets of estimated parameters were pooled to form a unified, ensemble GEE predictive model. The final predictive model was then used to forecast disease outcomes in independent test subjects. To ensure that the out-of-sample prediction was not affected by imputation mechanisms, we only considered imputation for the training data, and evaluated the model performance on independent subjects only on days where disease scores were available.

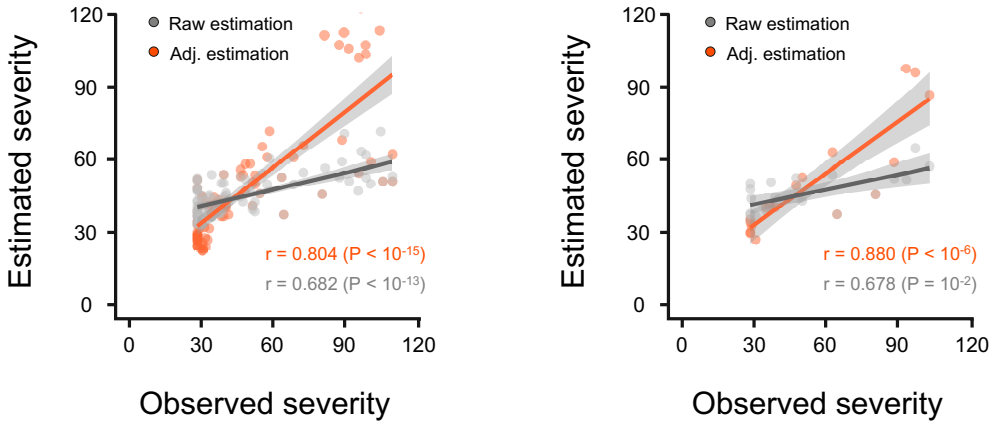


Fig. 5. Estimating multiple sclerosis (MS) severity over time in novel individuals. Left: The estimated against observed MSIS-29 scores for novel subjects. The grey dots correspond to the (raw) estimations using ensemble GEE in **Eq.** (2). The orange dots are the fine-tuned estimations using **Eq.** (3). **Right:** To examine whether the significant correlations observed in the left figure are due to repeated measurements *per* subject, the individual mean estimations are plotted against individual mean observations. The orange and grey lines represent the linear association between estimated and observed values.

To demonstrate the efficacy of the proposed framework, we compared it with mixed-effect and GLM models in the same modeling and test strategy. For each model, we applied it to predict

both the averaged outcomes and longitudinal outcomes, and for both EDSS and MSIS-29 scores. We recorded the accuracy statistics (correlation and MSE) from each model (See **Table** 1). Our results showed that for both disease types, GEE outperformed the mixed-effect model and was slightly better than GLM in mean assessment and longitudinal assessment. Since there were far fewer longitudinal measurements of EDSS score per individual compared to those of MSIS-29 score and that the main theme herein is on longitudinal disease severity prediction, in the following we will focus on examining the performance of the proposed model with regards to longitudinal MSIS-29 score prediction.

	Mean prediction						Longitudinal prediction					
	EDSS			MSIS-29			EDSS			MSIS-29		
	Mixed effect	GLM	GEE	Mixed effect	GLM	GEE	Mixed effect	GLM	GEE	Mixed effect	GLM	GEE
r	0.414	0.486	0.493	0.502	0.639	0.678	0.417	0.475	0.542	0.628	0.647	0.682
r (adjusted)	*	*	*	0.837	0.878	0.880	*	*	*	0.792	0.798	0.804
MSE	1.41	1.203	1.186	25.668	22.631	22.307	1.416	1.142	1.091	21.546	19.795	19.527
MSE (adjusted)	*	*	*	15.611	13.197	13.011	*	*	*	14.833	14.961	14.785

*Most subjects only have one to two EDSS scores. It is therefore impractical to make a further adjustment using data from day one.

Table 1. Model comparison. Left: Model comparison for mean outcome prediction. GEE model is compared with two baseline models, a mixed-effect model and a GLM, to predict both average EDSS and MSIS scores. The prediction accuracy is evaluated used two metrics: the correlation (or r) and the mean squared errors (MSE), which are respectively presented above and below the zero reference line. For longitudinal measurements, the correlation was taken assuming the repeated measurements were as if independent; we would further test the model efficacy in longitudinal assessment using two additional analyses (see text). **Right: Model comparison for longitudinal outcome prediction.** The same analysis is performed for longitudinal disease prediction, where the disease outcomes are longitudinally observed individual EDSS and MSIS-29 scores.

Overall, the estimated longitudinal MSIS-29 scores in the novel test sample were significantly correlated with their observed counterparts ($r = 0.80$, $p < 0.001$) (see left of **Fig.** 5). It, however, remained possible that the significant association was boosted by having repeated measurements in the test set. To check for this possibility, we took the means of the predicted and observed outcomes for each individual (thus there were only one estimated mean score

and one predicted mean score for each participant) and calculated the association again; the result remained significant ($r = 0.88$, $p < 0.001$).

4.1 Bootstrap experiments

Since training and test samples were obtained from a random split of the data, predictions made on some splits may perform better than those made on others. To evaluate the general performance of our model, we performed 1,000 bootstrap experiments. Specifically, each bootstrap began with a random draw of the data. The subjects in the sampled data were subsequently split into 70% (training) and 30% (test), where repeated measurements of a subject were either in the training set or the test set. Next, for each bootstrap sample, multiple imputation ($M = 15$) was made on its training set, followed by model development using the GEE on each of the imputed full set. Subsequently, parameters from the M models were pooled to form an ensemble GEE model that was specific to each bootstrap training sample, with model performance evaluated on the test data. Each bootstrap experiment ended with one correlation value between the estimated and observed outcomes with regard to the test set. Together, we obtained 1,000 correlations values from all bootstrap experiments. In parallel, we performed another 1,000 bootstrap experiments but with the personalized fine-tuning included as outlined in **Eq.** (3).

The bootstrap results showed that, first, the proposed method using GEE and MI was reproducible across all bootstrap samples with reasonable overall disease assessment results (with an average correlation of 0.34), and second, the adjusted personalized disease assessment further improved estimation accuracy (with an average correlation of $r = 0.81$) (see **Fig.** 6).

Yet, one may still question, how can one be certain that the above analyses justify the efficacy of the model for longitudinal prediction? It remained possible that, if the dataset contained many individuals without much longitudinal variability in their (within-subject) scores, then the above analyses only showed that the model was useful in estimating the average scores. To further examine the model performance in longitudinal disease prediction, we performed two additional analyses.

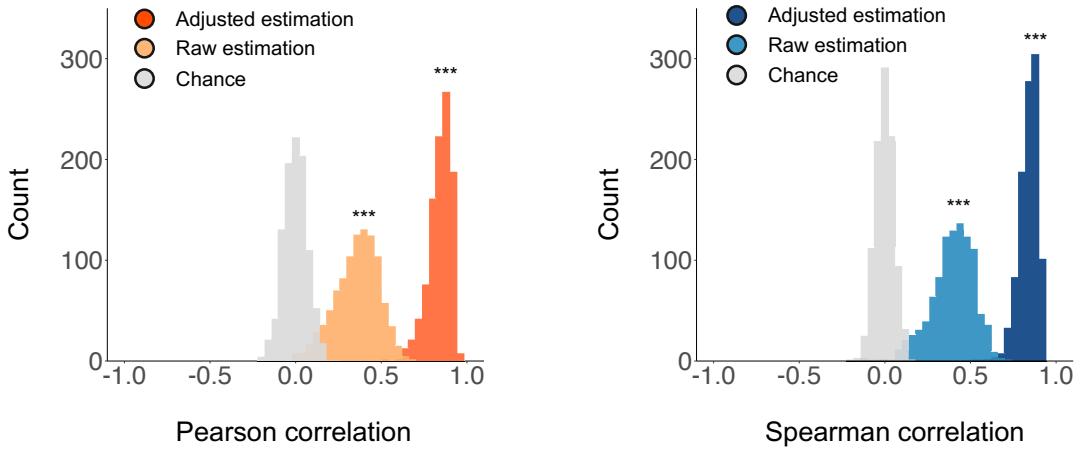


Fig. 6. Bootstrap results of out-of-sample testing performance for MSIS-29 score prediction. During each bootstrap test, data were randomly drawn from the sample. Subjects were then randomly split into a training set (containing 70% of the total samples) and a test set (containing 30% of the total samples). Next, MI was performed, and an ensemble GEE was learned from the training data. The trained model was then evaluated on the test data. Subsequently, the (longitudinal) correlation between the estimated and observed disease scores was calculated. This was repeated 1,000 times, producing 1,000 correlations. The light orange and light blue shaded histograms represent the distributions of Pearson and Spearman correlations from 1,000 bootstrap tests; the dark orange and dark blue shaded histograms represent the distributions of the fine-tuned personalized prediction outcomes. For comparison, two null distributions (the grey shaded histograms) were generated from a normal distribution with zero mean and the same standard deviations as the bootstrap Pearson and Spearman correlations.

4.2 Evaluating the longitudinal prediction performance of the model

Analysis I. Let's begin by denoting τ ($\tau > 0$) as the grouping threshold. First, we group the subjects into three categories: (1) Improved, (2) Stable, and (3) Worsened. The categories were determined based on whether the disease scores at the endpoint were at least τ units

smaller than (denoted as $\Delta < -\tau$), similar to (within plus/minus τ units, denoted as $\Delta \in [-\tau, \tau]$), or at least τ units larger than (denoted as $\Delta > +\tau$), the scores at the baseline. If the directions of estimated score changes agreed with the observed categories, then it suggests that the model was able to pick up the longitudinal disease trend. Indeed, setting $\tau = \frac{1}{2}$ and performing 1,000 bootstrap experiment, our results showed that the predicted score changes were generally in line with the observed categories (the group means between the improved and worsened groups, and those between the stable and worsened groups, were significantly different ($p < 0.001$)) (see **Fig.** 7). Note that the larger the threshold τ was set, the more observations would be classified into the stable group; but across different sets of grouping thresholds (*i.e.*, $\tau = 1$, 2, and 3), the results, in general, showed similar classification performance.

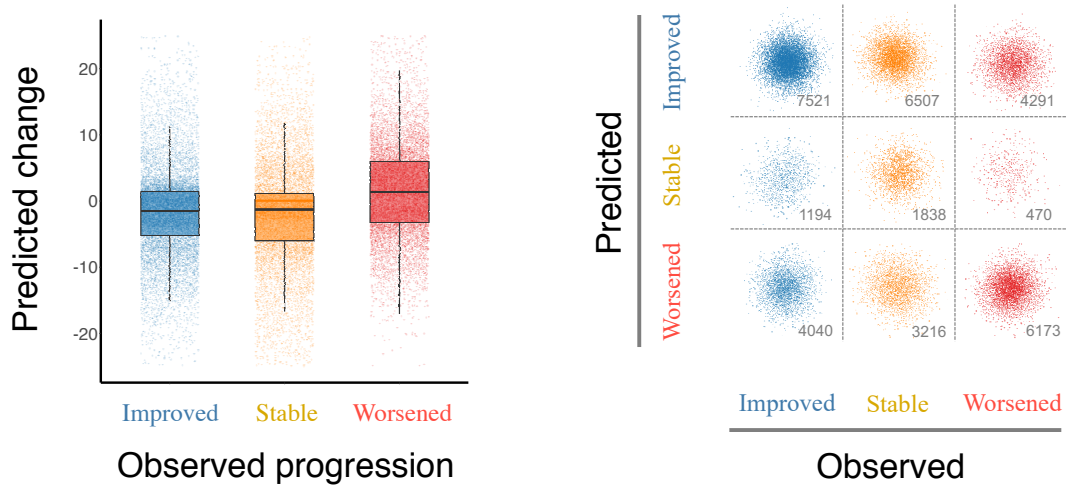


Fig. 7. Bootstrap results of (out-of-sample) predicted score changes against observed groups. Following the same bootstrap setting as in **Fig.** 6, an additional 1,000 bootstrap experiments were performed as follows. First, individuals were classified into improved, stable, or worsened groups based on whether the disease scores at the endpoint were at least a half unit smaller than, similar to (within plus/minus a half unit), or at least a half unit larger than, the score at the baseline. Longitudinal disease prediction was carried out and the difference between predicted baseline and endpoint scores for each individual was calculated. **Left:** The results of predicted score changes against the observed subject groups. **Right:** The counts of predicted categories against the counts of the observed categories. Note that some patients showed improved scores because the patients in the study were on some standard medications; additionally, in general, patients have the potential to improve after relapse events.

Analysis II. Here we calculated the repeated measures correlation (*rmcorr*)[37] to quantify the within-individual association between longitudinal observations and their estimates for multiple individuals. Specifically, we consider the *rmcorr* for estimated and observed baseline and the endpoint MSIS-29 scores; a larger positive *rmcorr* would suggest that there is more evidence that the model is able to detect the longitudinal within-individual disease score change between the baseline and endpoint. To account for variability due to splitting mechanisms (between training and test samples), we performed 1,000 bootstrap experiments and obtained 1,000 *rmcorr* values from the out-of-sample tests. The results suggest that the model is overall promising to capture the longitudinal within-individual disease trend (with mean $rmcorr = 0.42$ and $sd = 0.16$). The *rmcorr* can be calculated using the publicly available *R* package *rmcorr*[37].

5. Conclusions

Personalized longitudinal disease assessment and monitoring of MS are central to achieving an early diagnosis, appropriate management, and timely treatment of the disease. Despite advances, it is inconvenient, costly, and at present impractical to frequently obtain physician assessments of MS in clinics. Smartphones record data automatedly, remotely, and semi-continuously. They are widely available and can record meaningful signals related to MS symptoms over a broad range of behavioral domains. They are beginning to become useful digital assistants to complement traditional MS assessment and monitoring paradigms. But before implementing them into daily health service, it is important to evaluate whether, and if so, to what extent, they can distinguish between-individual differences in disease profiles and uncover within-individual disease course longitudinally.

In this study, we developed an automated longitudinal disease assessment and monitoring framework to assess and monitor MS development over time using smartphones. The

framework combines MI, the GEE, ensemble-learning, and a subject-specific fine-tuning. The MI was used to impute missing data entries; the ensemble GEE was employed for model development and longitudinal disease score prediction; the fine-tuning was introduced to adjust for idiosyncratic disease trajectory.

Utilizing smartphone and clinical data obtained from MS patients, the framework showed promise to estimate individual longitudinal MS disease profiles in previously unseen subjects. Particularly, the detected disease changes between the baseline and the endpoint using the proposed model, in general, agreed with those represented by the observed MSIS-29 scores.

Taken together, our analyses proved the concept of smartphone-based, personalized MS assessment and monitoring and demonstrated the potential of the proposed model in longitudinal MS evaluation. Future research needs to test the model using independent datasets and to verify if the framework can be extended to evaluate other MS-related clinical outcomes. Future studies may also examine the utility of the method in MS prognosis (namely, predicting the disease before its onset) and to explore whether this approach is useful to investigate and forecast other neurodegenerative diseases longitudinally.

References

1. Berer, K. & Krishnamoorthy, G. Microbial view of central nervous system autoimmunity. *FEBS Letters* **588**, pp. 4207–4213 (2014).
2. Milo, R. & Kahana, E. Multiple sclerosis: Geoepidemiology, genetics and the environment. *Autoimmunity Reviews* **9**, pp. A387–A394 (2010).
3. Browne P *et al.* Atlas of multiple sclerosis 2013: A growing global problem with widespread inequity. *Neurology* **83**, pp. 1022–1024 (2014).
4. Compston, A. & Coles, A. Seminar: Multiple sclerosis. *Lancet* **372**, pp. 1502–1517 (2008).
5. Purves, D., Augustine, G. J. & Fitzpatrick, D. (eds.) Voltage dependent membrane permeability. In *Neuroscience, Thrid Edition*. pp. 47–67 Sinauer Associates, Sunderland, MA, USA (2012).
6. Murray, E. D., Buttner, E. A. & Price, B. H. Depression and psychosis in neurological practice. In *Neurology in Clinical Practice*, edited by W.G. Bradley, R.B. Daroff, G.M. Fenichel & J. Jankovic, pp. 103–116. Butterworth- Heinemann, Oxford, UK. (2012).
7. National Institute of Neurological Disorders and Stroke. NINDS Multiple Sclerosis Information Page. (2015). Available at: https://www.ninds.nih.gov/Disorders/All-Disorders/Multiple-Sclerosis-Information-Page
8. Myhr, K. M. *et al.* Disability and prognosis in multiple sclerosis: Demographic and clinical variables important for the ability to walk and awarding of disability pension. *Mult. Scler.* **7**, pp. 59–65 (2001).
9. Tsang, B. K. T. & Macdonell, R. Multiple sclerosis: Diagnosis, management and prognosis. *Aust. Fam. Physician* **40**, pp. 948–955 (2011).
10. Kurtzke, J. F. Rating neurologic impairment in multiple sclerosis: An Expanded Disability Status Scale (EDSS). *Neurology* **33**, pp. 1444–1452 (2012).
11. Riazi, A., Hobart, J. C., Lamping, D. L., Fitzpatrick, R. & Thompson, A. J. Multiple Sclerosis Impact Scale (MSIS-29): Reliability and validity in hospital based samples. *J. Neurol. Neurosurg. Psychiatry* **73**, pp. 701– 704 (2002).
12. Hobart, J. The Multiple Sclerosis Impact Scale (MSIS-29): A new patient-based outcome measure. *Brain* **124**, pp. 962–973 (2001).
13. McGuigan, C. & Hutchinson, M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. *J. Neurol. Neurosurg. Psychiatry* **75**, pp. 266–269 (2004).
14. Poushter, J. *Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies*. *Pew Research Center* (2016). Available at: https://www.pewresearch.org/global/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/
15. Bhavnani, S. P., Narula, J. & Sengupta, P. P. Mobile technology and the digitization of healthcare. *European Heart Journal* **37**, pp. 1428–1438 (2016).
16. Chén, O. Y. & Roberts, B. Personalized healthcare and public health in the digital age. *OSF Prepr.* **10.31219**, hmves (2020).
17. Baker, M., Beek, J. van & Gossens, C. Digital health: Smartphone-based monitoring of multiple sclerosis using Floodlight. *Scientific American* (2020). Available at: https://www.nature.com/articles/d42473-019-00412-0
18. Montalban, X. *et al.* Floodlight: Remote self-monitoring is accepted by patients and provides meaningful, continuous sensor-based outcomes consistent with and augmenting conventional in-clinic measures. *Neurology* **90**, P4.382 (2018).
19. Kieseier, B. C. & Pozzilli, C. Assessing walking disability in multiple sclerosis. *Multiple*

Sclerosis Journal **18**, pp. 914–924 (2012).
20. Kierkegaard, M., Einarsson, U., Gottberg, K., Von Koch, L. & Holmqvist, L. W. The relationship between walking, manual dexterity, cognition and activity/participation in persons with multiple sclerosis. *Mult. Scler. J.* **18**, pp. 639–646 (2012).
21. Scalfari, A. *et al.* The natural history of multiple sclerosis, a geographically based study 10: Relapses and long-term disability. *Brain* **133**, pp. 1914–1929 (2010).
22. Feys, P. *et al.* The Nine-Hole Peg Test as a manual dexterity performance measure for multiple sclerosis. *Multiple Sclerosis* **23**, pp.711–720 (2017).
23. Midaglia, L. *et al.* Adherence and satisfaction of smartphone- And smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: Nonrandomized interventional feasibility study. *J. Med. Internet Res.* **21**, e14863 (2019).
24. Graham, J. W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **60**, pp. 549–576 (2009).
25. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* **338**, b2393 (2009).
26. von Hippel, P. T. New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples. *Struct. Equ. Model.* **23**, pp. 422–437 (2016).
27. Rubin, D. B. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Hoboken, NJ, United States (1989).
28. Rubin, D. B. Multiple imputation after 18+ Years. *J. Am. Stat. Assoc.* **91**, pp. 473–489 (1996).
29. Rubin, D. B. Inference and missing data. *Biometrika* **63**, pp. 581–592 (1976).
30. Allison, P. D. *Missing data*. Sage Publications, Thousand Oaks, CA, USA (2001).
31. Enders, C. K. *Applied missing data analysis*. Guilford press, New York, NY, USA (2010).
32. Little, R. J. & Rubin, D. B. *Statistical analysis with missing data*. John Wiley & Sons, Hoboken, NJ, USA (2019).
33. Andridge, R. R. & Little, R. J. A. A review of hot deck imputation for survey non-response. *International Statistical Review* **78**, pp. 40–64 (2010).
34. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, pp. 1–67 (2011).
35. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, pp. 13–22 (1986).
36. Chén, O. Y. *et al.* Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans Biomed Eng* **67**, pp. 3491–3500 (2020).
37. Bakdash, J. Z. & Marusich, L. R. Repeated measures correlation. *Front. Psychol.* **8**, 456 (2017).

Supplementary Information

A comparison between two types of averaging techniques for dealing with parameters estimated from multiple sets of date using multiple imputation (MI).

	Mean prediction				Longitudinal prediction			
	EDSS		MSIS-29		EDSS		MSIS-29	
	GEE Averaging parameters	GEE Averaging predictions	GEE Averaging parameters	GEE Averaging predictions	GEE Averaging parameters	GEE Averaging predictions	GEE Averaging parameters	GEE Averaging predictions
r	0.493	0.493	0.678	0.680	0.542	0.542	0.682	0.682
r (adjusted)	*	*	0.880	0.880	*	*	0.804	0.803
MSE	1.186	1.186	22.307	22.135	1.091	1.091	19.527	19.403
MSE (adjusted)	*	*	13.011	12.964	*	*	14.785	14.830

*Most subjects only have one to two EDSS scores. It is therefore impractical to make further adjustment using data from day one.

Table 2. Longitudinal prediction using averaged parameters versus averaged predictions. GEE model is compared with two baseline models, a mixed-effect model and a GLM, to predict both average EDSS and MSIS scores. The prediction accuracy is evaluated used two metrics: the correlation (or r) and the mean squared errors (MSE), which are respectively presented above and below the zero-reference line. For longitudinal measurements, the correlation was taken assuming the repeated measurements were as if independent; we would further test the model efficacy in longitudinal assessment using two additional analyses (see text). **Right: Model comparison for longitudinal outcome prediction.** The same analysis is performed for longitudinal disease prediction, where the disease outcomes are longitudinally observed individual EDSS and MSIS-29 scores.

Chapter 6

Conclusions and discussions

Epilogue

The biomarker-based brain study is still in its infancy. Anyone who has studied neuroscience or statistical brain science for some time is aware of how little we know about the functions and organization of the brain, how ignorant we are about the causes of neurodegenerative diseases, and how poorly equipped we are with statistical devices to uncover these myths. One should, however, be equally impressed by how much we have learned, especially in the terrain of biomarker development during recent years. Seeing through the advancement of statistical science, engineering science, and neuroscience, we are beginning to unfold the connections between brain functions, human behavior, and neurodegenerative disorders.

Building efficient, targeted, and affordable biomarkers to understand brain functions, human behavior, and brain disorders via predictive modeling has been the main theme of this thesis. It is with this goal that I begin it; it is with the same aim that I conclude the thesis.

In this section, I begin by concluding the key messages of the thesis. Extending on these messages, I make a few additional comments about biomarker developments in brain and behavior studies to complement the collection of works enclosed in the thesis. First, I discuss the projections between neural, behavior, and disease data, and how by studying one set of these data could we begin to understand another set of data. The projections between biomarkers, human behavior, and brain functions, however, do not endorse causation. Next, I examine the potential biological basis for association, explanation, prediction, and causation related to biomarker studies. Finally, I present a few topics that I briefly touch upon in previous

chapters but do not discuss in detail, including missing data, high-dimensional data analysis, and longitudinal data analysis. Standing on the shoulders of the developers, with further modifications and additions, these methodological frameworks may continue to contribute towards building better biomarkers.

1. Conclusions

In the following, I provide a summary of four types of biomarkers studied in this thesis.

i. Perhaps the simplest biomarkers are the static biomarkers. They are simple, as they summarize important aspects of biological measurements into univariate metrics such as the mean of a feature or the connectivity (*e.g.*, pair-wise correlation) between two features (*e.g.*, time series recorded from two brain areas). The mean on the one hand depicts the general magnitude of a feature with respect to it of others, and on the other hand, the variability of the means across individuals may underline a part of subject-specific disease variability. The brain connectivity paints a picture of the overall functional organization of the brain over time, and the weight of the connectivity (or edge strength) suggests the neural coupling between different brain areas. Static biomarkers, therefore, are particularly useful when one is interested in getting a “big picture” of the relationship between multivariate features and a scalar (behavior or disease) outcome. It is equally useful, by selecting static biomarkers that are associated with the outcome, to build predictive models that can classify individuals into different (*e.g.*, healthy and diseased) groups, and to predict individual behavioral or clinic scores.

In **Chapter** 2, I developed methods to build static biomarkers to identify patients with Parkinson's disease and to assess their disease severity in an automated way using smartphones.

ii. While static biomarkers solve one set of problems, the existence of directed communications between brain areas raises a host of new ones which have to be addressed by building dynamic biomarkers. Chief among those are information flows between brain areas, which recapitulate both the direction and the strength of potential neural communication between brain areas. A beginning to address the problem of building dynamic neural biomarkers can be made by first charting a diagram of directed edges between different brain areas, and then, via a predictive model, isolating a subset of directed edges that are associated with or predictive of behavior or disease outcomes. The selected directed edges are potential dynamic neural biomarkers, if they not only map the effective connectivity in the brain in a neurobiologically meaningful way, but also exist in a general population and can be used to assess outcomes in novel subjects.

In **Chapter** 3, I investigated dynamic information flows in the brain and how they subserved brain outcomes. Specifically, I mapped the information highways in the whole brain and located specific neural links that were predictive of human cognition.

iii. Sometimes, one may observe that the brain sits between an explanatory variable (such as prodromal symptom) and a dependent variable (such as conversion into psychosis). Intermediate biomarkers represent brain areas whose functional roles are on the one hand associated with the explanatory variable and on the other hand potentially give

rise to the dependent variable. They can be uncovered using a class of statistical frameworks called mediation analysis.

In **Chapter** 4, I designed a high-dimensional brain-wide functional mediation framework and identified brain regions whose functions potentially mediate baseline clinical symptoms and future conversion to full psychosis.

iv. The observed temporal variation of human behavior and disease progression suggests that, to obtain a fuller picture of the behavior trajectory and to accurately describe, monitor, and predict neurodegenerative diseases, one needs to develop suitable models to search for longitudinal biomarkers. The longitudinal biomarkers are measured repetitively, are associated with the outcome variable of interest (such as one's disease outcome) over time, and, above all, exist in a broad population and are therefore useful for assessing, monitoring, and predicting longitudinal outcomes in novel subjects.

In **Chapter** 5, I built a model that showed promise to extract longitudinal biomarkers to assess and monitor the disease trajectory over time for patients with multiple sclerosis.

There are, of course, areas within biomarker development and predictive modeling that have seen more new looks and departures. The advent of causal inference, high-dimensional statistics, and missing data analysis has led to an unprecedented explosion in studies of neural markers that underline sensation, action, and cognition in health and disease. Thus, comparing the past and the future research through the prisms of the present studies, I hope to discuss a

few additional issues that have interested statisticians, engineers, and neurobiologists in the past, and continue to do so today, and that will likely extend into the future.

2. Discussions

2.1 The projections between neural, behavior, and disease data

The central problem of developing biomarkers is to find feature variables that are associated with or predictive of other variables of interest. Thus, this is an important question related to seeking causation, namely, to discover features that potentially give rise to specific behavior and brain disorders. This is also a question, with equal importance, related to seeking reverse causation, namely, given the properties of the behavior and brain diseases, can we track the properties of the brain? There are, in general, three classes of projections related to such causal and reverse causal inquiries: the projection between neural features and behavior properties; the projection between neural features and pathological outcomes, and the projection between the behavior properties and pathological outcomes (see **Fig. C1**).

First, there are pathways between the brain (or brain patterns in terms of its measured signals) and human behavior (or behavioral measurements such as one's IQ or physical activity). The scientific questions related to these pathways can be generally summarized as to discover brain areas and their functional organization that are responsible for behavior. Once we have discovered the neural biomarkers, can we assess and predict the behavior? The reverse is equally important, namely, given observed behavior, can we trace their neural origins and pinpoint the exact cortical or subcortical areas that are responsible for it? Next, there are pathways between the brain and brain illnesses. The scientific questions are, for example, what lesions or abnormal brain patterns are accountable for each type of brain disease? Can we identify, assess, monitor, and forecast brain disorders based on the level of lesions in the brain

and abnormality of the brain patterns? Conversely, would the categorization and monitoring of brain illness in general and brain disorders whose symptoms overlapping in particular help us to understand the underlying neural mechanism that gives rise to each disease? Lastly, there are pathways between (irregular) behavioral and brain illnesses. This is scientifically meaningful in its own right, particularly in the domain of behavioral marker-based diagnosis. It is particularly useful when frequent, large-sample-based measurements of brain data are not available. These pathways, like the other two classes, are bi-directional. Can we use behavioral measurements to assess and track brain disorders? On the other direction, can we forecast (the degenerating patterns of) behavioral symptoms once an individual is diagnosed with a brain disorder, thereby allowing time to act in advance, by assigning medicine or offering physical and mental assistance, to alleviate worsening symptoms?

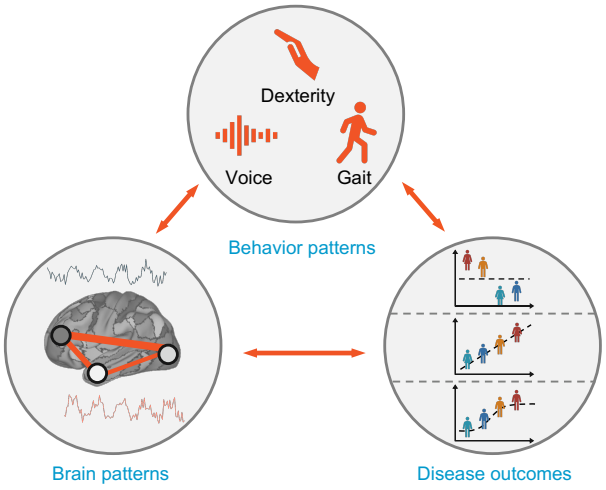


Fig. C1: Potential associations or causal pathways between the neural, behavior, and disease features. Bottom left: Neural features consist of, for example, molecule, physiological, neuroimaging measurements. Top: Behavior patterns include behavior measurements such as physical activities and behavioral phenotypes. Bottom right: Disease outcomes include disease status, severity, and longitudinal trajectory. The bidirectional arrows indicate the potential causal and reverse causation relationship between these features. In general, if one has identified the association and strength between features from each two (of the three) classes, then one can use features from one class to assess and predict the other, even for novel subjects, and *vice versa*.

All these pathways, in my view, are integral parts of the everlasting enquiry into the brain-body dualism (see **Chapter** 1): the brain generates behavior; each specific behavior linked to its neural origin reflects the functional specialization of the brain and the diverse and dynamic representations of the behavior across individuals demonstrate the variability of the brain functions; their disruptions and degenerations give rise to brain disorders, which are determined by either directly observing brain lesions or indirectly via behavior changes. Thus, because of the above-mentioned links, it seems that if one is able to uncover one set of biomarkers, one should be able to infer the previous or next sets of biomarkers along the pathways, and so on and so forth as a domino chain. Or so one would hope.

But the reality of the biomarker development is not always as straightforward, and the hypothesized links between biomarkers and the outcome variables are not as clean and clear cut as domino chains[1]. For example, observed brain and behavior measurements are contaminated with noise (for example, in functional magnetic resonance imaging (fMRI) data, multiples sources of noise can corrupt the true signals, such as scanner-related noise including thermal noise and scanner instability noise, noise due to head motion and physiology, the hemodynamic response function (HRF) model errors, and noise due to different sites). Besides noises, some biomarkers developed from a set of subjects might not offer the same implication for other subjects due to genuine biological differences between individuals. Equally troublesome are latent variables and confounds. For example, sensory inputs and cognition may be jointly associated with or mediated by “consciousness”. Thus, if irregular signals from a brain area are found to be associated with the onset of a mental disorder, one may drive into the treacherous road by concluding that signals from that area are neural markers causing the mental disorder and using them to explain mental problems. It remains possible, even after having adjusted for sampling bias, that the irregular signals are effects (but not causes) or that

the association is due to a spurious correlation (in other words, some variables in a large number of features are correlated by chance). How, then, can one say the developed biomarkers are useful and meaningful at all?

2.2 The potential biological basis for association, explanation, prediction, and causation[a]

Let us begin by defining what the problem of causal inquiry is in a biological organization in general and a neurobiological system in particular[2]. Let $\boldsymbol{\xi}$ denote a class of neurobiological random variables (where we use the bold font to indicate that $\boldsymbol{\xi}$ can be multivariate) and let ω denote an outcome variable (for simplicity consider that ω is univariate). If the properties of $\boldsymbol{\xi}$ affect the expression of ω deterministically (for simplicity without error), then we can use a link function ϕ to denote the relationship between ξ and ω, namely $\omega = \phi(\boldsymbol{\xi})$. In words, the feature variable $\boldsymbol{\xi}$ cause the outcome variable ω by means of ϕ. The link function ϕ can be a complex mathematical map, edge(s) in a graphic model, or a (set of qualitative) statements.

As $\boldsymbol{\xi}$ and ω are random (biological) variables, one can only aim at uncovering the underlying (causal) relationship by analyzing the observed values of $\boldsymbol{\xi}$ and ω. Let's denote them as $\boldsymbol{x}$ and o, respectively. Since observed data contain noises and measurement errors, the statistical model used to uncover the potential causal relationship can be written as $o = f(\boldsymbol{x}) + \varepsilon$, where ε denotes the residual term following a specific distribution (say Gaussian). Here, f specifies the model, which delineates the relationship between the observed $\boldsymbol{x}$ and O but not necessarily

[a]This section is inspired by Shmueli's paper entitled *To explain or to predict*. I shall refer interested readers to her paper for a further reading regarding studying association, explanation, prediction, and causation using statistical models. The arguments made in this section have been adopted to Section 5.2 in: Chén, O.Y, Phan, H., Nagels, G., de Vos, M., 2020. On Statistical Analysis of Brain Variability. *Preprints* **2020080428** (doi: 10.20944/preprints202008.0428.v1).

between $\boldsymbol{\xi}$ and ω; in other words, f may differ from ϕ. If $\boldsymbol{x}$ and o contain meaningful signals of $\boldsymbol{\xi}$ and ω and the noises and errors are small, then one would hope that estimated f is a reasonable proxy of the unknown causal map ϕ.

Suppose that we find a model f that depicts the relationship between one or several features, say physical activities, and an outcome, say, having Parkinson's disease. We would certainly be pleased if those features are the *only* ones (among all candidate features) that affect the outcome. We would be equally pleased if we can use the trained model f to make forecasts about future outcomes of a specific subject based on his/her past features or make forecasts about outcomes of new subjects based on subjects we have studied. Despite advances, for example, the discovery of GBA1 being present among 5% of PD patients[3], definitive or direct causal relationships are in general difficult to identify and hard to verify (sometimes one has to wait until *post mortem* for testing the causal links). Such difficulties, fortunately, can be partially addressed by performing associative or predictive analysis, which aims at unveiling features and relationships that, in a weaker sense, can explain the underlying biological system. Subsequently, one can test whether the discovered relationship can be strengthened into potential causation by performing further statistical analysis and evaluating *post hoc* empirical evidence.

Association *vs.* causation. Because the underlying causation ϕ is unknown, the estimated relationship $\hat{f}$ (or $\hat{f}(\boldsymbol{x}, o)$ as it is based upon observed features $\boldsymbol{x}$ and outcome o) can only suggest that there is association between $\boldsymbol{x}$ and o and can *at most* say that there is a potential causal relationship between feature variables $\boldsymbol{\xi}$ and outcome variable ω.

The reason is twofold. First, the observed features $\boldsymbol{x}$ and outcome o, due to noise and measurement errors, may not accurately describe the properties of the underlying $\boldsymbol{\xi}$ and ω. Second, even if features $\boldsymbol{x}$ and outcome o are accurately describing the properties of random biological entities $\boldsymbol{\xi}$ and their outcome ω, as f is an arbitrary statistical model, there may exist another model g that better depicts the relationship between observed features $\boldsymbol{x}$ and outcome o. By *better depicts*, we mean that the model g (where $o = g(\boldsymbol{x}) + e$) may produce a smaller error e than ε (as in $o = f(\boldsymbol{x}) + \varepsilon$) or yield better out-of-sample prediction than f.

But the fact that there exists an alternative model g that is better than f in terms of producing smaller an error or yield better a prediction does not conclude that g is more suitable than f to describe the causal link of ϕ. First, a model g that overfits the data may produce a smaller error e. Second, due to measurement errors and noises, a model g that yields higher prediction power may in fact be predicting signal plus noises (rather than predicting signals). Finally, sometimes an alternative model g may suggest causation when there is not. To see this, let's consider an extreme case when there is no underlying causation (say, between temporal signals obtained from two nodes ξ and ω); in other words, suppose ϕ is not causal. Due to measurement errors (*e.g.*, measurements can be temporally delayed, and the delays are node-specific), there is a spurious temporal relationship between features $\boldsymbol{x}$ and outcome o which is captured by g. Therefore, one cannot say because g exists, causation exists.

Prediction *vs.* causation. The difficulty in isolating causal relationships with regard to the brain, mind, body, behavior, and pathology as well as a lack of knowledge of ϕ should not prevent us from looking for features that are able to make sense of a neurological or behavioral phenomenon. A beginning can be made by looking for features that are predictive of the

outcome and considering them as a superset of causal features. A causal feature, such as a gene mutation, is almost surely predictive of the outcome (to a certain extent) and can therefore explain a portion of the variance of the outcome. The reverse is not always true. For example, a brain area whose signals are predictive of an outcome does not mean the former is causally related to the latter.

Suppose that observations $\boldsymbol{x}$ and outcome o have been properly preprocessed and nothing additional can be done to make them better represent the signals of the observed values of $\boldsymbol{\xi}$ and ω. The value of predictive modelling can be explained, in part, by the bias-variance trade-off. Following the above notation, the expected predictive square error is[4]:

$$\text{Expected predictive square error } = E\left(o - \hat{f}(\boldsymbol{x})\right)^2$$

$$= \underbrace{E\left(o - f(\boldsymbol{x})\right)^2}_{\text{True variance}} + \underbrace{\left(E\left(\hat{f}(\boldsymbol{x})\right) - f(\boldsymbol{x})\right)^2}_{\text{Bias}} + \underbrace{E\left(\hat{f}(\boldsymbol{x}) - E\left(\hat{f}(\boldsymbol{x})\right)\right)^2}_{\text{Estimation variance}}$$

where the first term represents the variance even if the model is correctly specified and estimated, the second term denotes the bias, and the last term is the estimation variance.

The causal inquiry, in essence, is to minimize the bias between the estimated model ($\hat{f}(\boldsymbol{x})$) and the underlying theory about the causal relationship f. Meanwhile, predictive modeling aims at minimizing both bias and the estimation variance, even if at a cost of theoretical accuracy for improved empirical precision; in other words, one would prefer a “wrong” (or less realistic) model that yields better predictions[2].

The usefulness of predictive modeling lies in that it reduces the likelihood of overfitting and *may* (my emphasis) help raise association to potential causation. This is done via out-of-sample prediction, where the estimated function $\hat{f}$ trained from a sample is used to predict outcomes in previously unseen samples. Thus, if the features and estimated function can be validated in novel subjects, it suggests that $\hat{f}$ may not be restricted to a training sample and is representing some general properties shared in a broad population. Further, if the relationship (indicated by $\hat{f}$) exists in a wide population, it hints that there exists a stronger sense of the causal relationship between the features and the outcome.

Explanation *vs.* causation. Both predictive modeling and association analysis shed useful light upon the relationship between features and the outcome. Thus, until a definitive causal map is charted, one can rely on features showing predictive or associative merits to make sense of the observed data. For example, dexterity data are known to be associated with PD. Trained in a predictive model, dexterity data can be used to predict PD in novel samples[5]. Naturally, however, irregularity in dexterity data is not the cause for PD, for PD is a neurodegenerative disease; rather, irregular dexterity is an early effect of the underlying neurodegenerative disease. Nonetheless, building an associative or predictive model between dexterity and PD is still useful for patient identification and disease severity estimation. Further, it may offer insights into the behavioral characteristics of the disease. For example, tremor, bradykinesia, rigidity, and postural instability are significantly associated with (but are not necessarily the causes for) the onset of PD, suggesting that motor functions are hindered in PD patients. Thus, observing that many PD patients have motor issues raises the question of the neural origin of the PD in the brain, for the motor disruption must be in one way or another linked to cortical abnormalities in the motor cortex of the brain.

Prediction *vs.* explanation. Predictive modeling aims to achieve accurate predictions via modeling techniques. It is useful when one is interested in extrapolating model representations learned from a sample into new observations. But predictive models are oftentimes built at the cost of their explanation capability. For example, regularization methods (*e.g.*, RIDGE, Lasso, and elastic-net regularizations) reduce estimation variance (by shrinking some feature parameters towards zero) and introduce bias into the model, thereby making the model less explainable. Similarly, ensemble methods such as bagging, boosting, and random forest (see **Chapter** 1) improve overall prediction accuracy by averaging predictions from each model meanwhile making it difficult to explain the ensembles[2]. Finally, neural networks may capture complicated associations hidden in the features space and between features and the outcome and yield accurate predictions, but the models may neither be able to accurately represent the underlying causal mechanism ϕ nor be easy to interpret the model f itself. On the other hand, including multiple highly correlated features in a model or including an insignificant feature would reduce the prediction power of the model, but the correlated features, such as different types of motor symptoms in Parkinson's studies, or insignificant (in terms of predictability) features, such as age, gender, and smoking status, may shed meaningful scientific and medical lights upon the research question; sometimes research protocols may even ask to include them mandatorily (such as the inclusion of smoking history in a cancer study).

Taken together, it is important to clarify what the research goals are during study design, model development, and data analysis. Equally, it is important to find a balance between predictability and explainability by carefully evaluating the bias-variance trade-off[6,7].

2.3 Statistical methods for dealing with high-dimensional data[b]

Thanks to the advancement in data collection, storage, and computing, biomarker studies have seen a paradigm shift during the current big data era. One has begun to see a wealth of high-dimensional data being collected and analyzed. High-dimensional data analysis is useful for two reasons. First, by analyzing high-dimensional data, one can unveil hidden structures of subpopulations of the data; such information is not easily attainable from small-size data, as underlying structures may be treated as "outliers". Second, even under circumstances where there are large individual variations, one may still be able to extract common features across subpopulations[8].

To break the barriers between old and new paradigms, high-dimensional data may spawn new challenges. Let p denote the number of features and n be the number of subjects (or experimental units) on which the features are measured. In the terrain of high-dimensional biomarker development, the chief obstacles one faces are the so-called "large p, small n" problem (where p goes to infinity faster than n) and "large p, large n" problem (where p and n go to infinity at the same rate). Under either circumstance, the classic statistical theories (developed based upon "small p, large n" paradigm) collapse.

Classic large sample theory in statistics, which is typically developed based on the Law of Large Numbers and the Central Limit Theorem, has focused on the scenario where p is smaller than n. Under such a setting, one can develop asymptotic properties of estimators of parameters, such as consistency of the estimators and their asymptotic (*e.g.*, Gaussian)

[b]The arguments made in this section have been adopted to Section 4.3 in: Chén, O.Y, Phan, H., Nagels, G., de Vos, M., 2020. On Statistical Analysis of Brain Variability. *Preprints* **2020080428** (doi: 10.20944/preprints202008.0428.v1).

distributions; with theoretical supports, one can then assign values and confidence intervals (or bands) to the estimated parameters to make the statistical inference.

The challenge brought by high-dimensional data, however, can be alleviated if the following assumption holds, namely there exists a sparse representation of the features. In other words, there are only k ($k < n$) features from all p features that are influential. Such an assumption, of course, is not purely hypothetical; rather, shreds of empirical evidence have been discovered that suggests the existence of sparse data structures in biological systems. For example, it is known that there exist a small number of network hubs (a class of highly connected nodes or dictating features) and many poorly connected nodes in cells[9], the brain[10], and the genes[11]. Under the sparse assumption, no matter how large the dimensionality is, one can apply feature selection techniques to reduce the number of features to that is smaller than n.

The two simplest classes of feature selection approaches are perhaps regularization (*e.g.*, SCAD, Dantzig selector, Lasso, elastic net, and adaptive Lasso) and stepwise feature selections[1]. In the following, I introduce a few arrays of methodological frameworks proposed during recent years to deal with high-dimensional data, although, in my view, the fundamental principles of these novel methods do not depart from performing regularization and feature selection (or dimension reduction). In the following, I shall treat feature selection or dimension reduction ambiguously; but one should note that dimension reduction includes both feature selection and feature transformation (which converts high-dimensional variables to lower-dimensional ones).

The first class of feature selection methods is suitable for linear models with ultra-high dimensional features. When dimension p grows fast with sample size n, regularized models

may not pick up the correct features. The sure independence screening (SIS)[12] addresses this issue by first reducing the dimensionality from p to relatively large number $d < n$, and subsequently applying a common regularized method for feature selection. For nonlinear parametric models in which the mean function is defined explicitly as a nonlinear function of parameters, and various types of semiparametric or nonparametric approach, one can consider sparse additive models[13] and Bayesian additive regression trees[14]. In essence, the former bears the form $y_i = \sum_{j=1}^{p} \beta_j f_j(x_{ij}) + \varepsilon_i$ and chooses (arbitrary) functions to minimize the residual sum of squares under constraints that both encourage sparsity and render the optimization problem convex. The latter treats the regression as a sum of trees $y_i = \sum_{j=1}^{p} f_j(x_i; T_j, \beta_j) + \varepsilon_i$ and uses each T_j to represent a tree structure involving a small number of the features. Finally, two roads are leading to territories where regression models cannot lead: one that explores the classic statistic devices based on eigenanalysis of sample covariance matrices, and one that looks for a nonlinear manifold of low dimension[15]. The former class includes principal components analysis, canonical correlation analysis, multivariate analysis of variance, and discriminant analysis, but newer developments such as direct thresholding[16] and sparse covariance models[17] are increasingly useful to explore the sparsity nature of the covariance matrix when the covariance matrix is particularly large. The latter include ISOMAP[18], local linear embedding[19], Laplacian eigenmaps[20], and Hessian eigenmaps[21]; the lower dimensional nonlinear manifold better represents the original data than lower dimensional linear subspace into which the projections of the original data have the largest variance (say from principal component analysis and multi-dimensional scaling).

2.4 Statistical methods for dealing with missing data

Data collection and analysis must confront the issue of missing data. The causes for missing data can, in general, be summarized into the following: nonresponse (*e.g.*, subjects do not respond to questions involving private information), measurement failure, experimental design (*e.g.*, a study considers frequent measurements of one set of variables, such as daily physical activity measurement, yet it only considers sparse measurements for another set of variables, such as disease outcomes that are only measured during in-clinic visits), drop out, and death.

Types of missing data. Statistical approaches treating missing data can be categorized according to the type of missing data for which they are designed. Specifically, the missingness (*i.e.*, whether the data are missing or not) can be categorized into three types: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR)[22].

MAR means that the missingness is not random but can be accounted for by other variables in the data set. In other words, the missingness may be dependent on the observed data, but not on unobserved data (we will see an example below). MCAR is a special case of MAR, which means the missingness is neither dependent on the observed data nor any unobserved data (or parameters of interest). When the missing data is MCAR, one can consider that the data is missing completely random (*i.e.*, by chance). With MNAR, missingness does depend on unobserved data.

It is perhaps desirable to consider an example to distinguish the three types of missing data. Suppose one is interested in analyzing a data set including self-reported multiple sclerosis (MS) severity score. If the reported MS severity is MCAR, it means that the reported MS severity is

independent of all other variables (observed or unobserved). Thus, one's missing MS severity can be considered as a random draw from all the MS severity scores. If, however, males choose to not report some of the severity scores, then the missingness is MAR. Once conditioning on the gender, the severity scores are MCAR. Finally, if people tend to not report severity scores when the scores are too high (for fear of, for example, discrimination), then the missingness is MNAR, because the missingness is dependent on the missing data itself.

Treatments for missing data. The problem of missing data is common in statistical analysis. Many attempts have been made to address this problem. So far, there has not been a universal agreement for optimal treatment, but some approaches might be more effective (*e.g.*, producing less biased estimates) than others for a given circumstance. In the following, I shall present a few useful treatments for dealing with missing data.

The simplest approach for treating missing data is perhaps the so-called *list-wise deletion*. It deletes all entries related to a sample that does not have the complete information. The advantage of list-wise deletion is that it is straightforward, and the resulting data is complete. Following list-wise deletion, one can explore the underlying scientific questions using statistical methods developed for complete data sets. Data resulting from list-wise deletion, however, may produce biased parameter estimates[23], and yield less power (because some data are thrown away).

Similar to list-wise deletion is *pair-wise deletion*. While list-wise deletion removes all subjects with a part of data that is missing, pair-wise deletion only discounts a subject when analyzing a specific feature if data of that particular feature is missing from the subject. For example,

consider ten subjects and three features X_1, X_2, and X_3, where subject 1 has missing data in X_1 and subject 2 has missing data in X_2. Then, under pair-wise deletion, when estimation the parameter for X_1, one would only consider data from subjects 2-9; similarly, when estimation the parameter for X_2, one would only consider data from subject 1 and subjects 3-9. Although pair-wise deletion uses more data than list-wise deletion, the parameter estimates for different features were based on different subsets of data, and thus the estimates may be biased.

Opposite to data deletion is data imputation, a technique that replaces missing data with artificially imputed data which leverages the underlying available data structure. The simplest imputation approach is perhaps *mean insertion*. For example, when there is a feature that contains missing data, one would fill in the missing entries with average features from observed features. This approach may yield a mean for a particular variable that is close to the true parameter value (*e.g.*, under MCAR), but other parameters using this method can be seriously biased[22].

There are a few technically relatively more advanced methods developed for treating missing data via data imputation. The first advanced imputation technique is based on expectation and maximization algorithm, *i.e.*, EM algorithm)[24,25]. In simple terms, the EM-based imputation makes guesses for the missing data during the E-step, where each missing value is estimated by using other variables as predictors in a regression fashion; subsequently, in the M-step, parameter analysis is performed using the now full data set via maximal likelihood estimation (MLE). The estimated parameters are then used to make new guesses for the missing data and replace them with the updated values. This process is repeated until convergence. The advantage of the EM-based imputation is that it achieves good parameter estimates (means, variances, and covariances), since they are derived using the maximal likelihood approach. Its

main shortcoming is that it does not provide standard errors, so confidence interval (CI) estimation and hypothesis testing are unavailable.

The problem of lacking standard error estimation can be dealt with using multiple imputation (MI)[26,27]. Suppose the data $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss})$ (where $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{miss}$ represent missing and non-missing data, respectively) follows a parametric model $\boldsymbol{Y} \sim P(\boldsymbol{Y}|\boldsymbol{\theta})$ with parameters denoted by $\boldsymbol{\theta}$. One can write the conditional probabilistic distribution of the missing data giving the observed data using the Bayesian theorem as:

$$P(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}) = \int P(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) d\boldsymbol{\theta}.$$

Subsequently, one can randomly draw a parameter $\boldsymbol{\theta}^*$ from the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$ using, for example, data augmentation[28]. Finally, one can impute a random draw of the missing value $\boldsymbol{Y}^*_{miss}$ from the (conditional) predictive model $\boldsymbol{Y}^*_{miss} \sim P(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}^*)$.

To deal with the uncertainty during imputation, multiple sets of imputed data can be generated using the method outlined above. Producing multiple sets of data sets also restores the error variance lost from regression-based single imputation. Parameter estimates can be then produced, using whichever statistical model desirable (since all missing values are filled), for each imputed full data set. Lastly, the parameters from multiple models can be pooled to generate a final, unified model (as I have demonstrated in the multiple sclerosis study in **Chapter** 5).

2.5 Longitudinal methods for studying the developmental brain

Longitudinal models are not only useful for studying neural degenerative diseases, with examples given in this thesis using multiple sclerosis as a gateway (see **Chapter** 5 and see Garcia and Marder[29] for longitudinal approaches for studying Huntington's Disease, and Bhagwat *et al.*[30] for longitudinal methods for investigating Alzheimer's disease). They are equally useful for understanding the developmental brain and the neural bases for cognitive development. Indeed, markers built from neuroimaging data have begun to paint pictures about structural and functional changes in the developing human brain and their relation to changes in cognitive processes[31]. With the advancement of longitudinal methodological frameworks, one can begin to chart a more detailed, dynamic, and specific trajectory of brain and behavior development in infants, children, and adolescents.

Biomarkers related to brain and behavior development can, in general, be summarized into two groups: those related to neural anatomical development, and those related to functional connectivity development that is associated with behavioral (and learning) development.

Biomarkers related to neural anatomical development during childhood and adolescence can be found in cortical and subcortical regions. For the former, cortical areas that underpin primary functions (such as motor and sensory areas) mature earliest; regions including associated with basic language skills and spatial attention (such as temporal and parietal association cortices) maturing next; higher-order association areas that integrate primary sensorimotor processes and modulate basic attention and language processes (such as the prefrontal and lateral temporal cortices) seem to mature last[32,33]. For the latter, there is evidence showing development in subcortical regions to which the association cortex projects. For example,

studies have shown that development in portions of the basal ganglia to which the prefrontal cortex projects[32,34].

Biomarkers related to connectivity development are primarily fine-tuned with the elimination of an excessive amount of synapses and the strengthening of relevant connections with development and experience[35]. A useful way to measure cognitive development is to quantify a child's ability to filter and suppress irrelevant information and actions (sensorimotor processes), in favor of relevant ones (*i.e.*, cognitive control)[36]; this can be done using experimental paradigms such as Stroop and Go/No-go tasks[31,37,38]. Linking brain connectivity measures with cognitive development, studies have shown that the development of working memory capacity was correlated positively with prefrontal-parietal connectivity[39]. Additionally, studies have shown that during development, children recruit distinct but often more prefrontal regions, both ventral and dorsal, when performing these tasks than do adults[35]. Taken together, areas whose brain activity correlates with task performance (reaction time and/or accuracy) become more focal or fine-tuned, whereas regions not correlated with task performance decrease in activity with age[35,40].

Before the 2000s, most developmental studies had considered relatively simple statistical methodologies. For example, cross-sectional methods comparing between childhood and adolescence were used to study gray matter loss during this period and showed how the immature brain develops into the fully-functioning mature one[33,41–43]. Cross-sectional analyses, however, may falsely suggest changes over time or fail to detect the specificity or magnitude of these changes, given the large individual variability in brain structure among individuals, especially during development[35]. Other effects were made using linear regression models. For

example, they were used to study total group and gender-specific effects of age on brain morphometry (*e.g.*, brain structure volumes); as total cerebral volume differed significantly between genders, gender differences were analyzed using ANOVA and then ANCOVA to adjust for total cerebral volume[43].

Recent years have seen a marriage between relatively more advanced statistical methods and the study of the developmental brain. For example, a mixed-model was used to study human cortical development during childhood through early adulthood region by region (one model was run for each region)[32]. Looking forward, longitudinal models that consider past and transitioning information as well as dynamic changes of covariance structures, some of which were discussed in **Chapter** 1, may further modify or add knowledge into our current understating of human brain development and bring in new insights about how such neural development may affect behavioral, learning, and cognitive development.

3. Final remarks

Naturally, this thesis gives a very general account of four classes of biomarkers and statistical frameworks suitable for studying them. It will require much more effort to divide biomarkers into further, more intricately organized groups, to go through each type of behavior paradigm within each class of biomarkers, and to explore every branch of statistical models to find the globally optimal approach. Naturally, too, there may be major flaws in my arguments and better methodological frameworks I have overlooked, which I hope my colleagues and I can remedy in future studies. I nonetheless hope that readers would agree with me that it is necessary to differentiate biomarkers based upon particular scientific interests, and that there is a need to develop and refine statistical tools to better serve the specific nature of each type of biomarkers.

The problem of identifying biomarkers is the problem of looking for a potential association, explanation, prediction, and causation: by charting the pathways between biomarkers, brain functions, and behavior landscapes, one hopes to find the neural and behavioral basis for an outcome of neurobiological interest. Thus, by developing novel methodological frameworks to uncover associative, explanatory, predictive, and potential causal pathways linking brain areas and human behavior and its normal presence and abnormal disruptions, one is contributing not only to the development of static, dynamic, intermediate, and longitudinal statistical models, but also to the equally profound problem of biological inquires, ones that may one day help us understand the functional organization of the brain, the neural basis of human behavior, the development of neurodegenerative diseases, and above all, who we are.

References

1. Chén, O. Y. The roles of statistics in human neuroscience. *Brain Sci.* **9**, 194 (2019).
2. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, pp. 289–310 (2010).
3. Stoker, T., Torsney, K. & Barker, R. Pathological mechanisms and clinical aspects of GBA1 mutation-associated Parkinson's disease. In *Parkinson's Disease: Pathogenesis and Clinical Aspects*, edited by TB, S. & Greenland JC. Codon Publications, Singapore (2018).
4. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*. Springer, New York, NY, USA(2001).
5. Chén, O. Y. *et al.* Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans Biomed Eng* **67**, pp. 3491–3500 (2020).
6. Douglas, H. E. Reintroducing prediction to explanation. *Philos. Sci.* **76**, pp. 444–463 (2009).
7. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, pp. 1100–1122 (2017).
8. Fan, J., Han, F. & Liu, H. Challenges of big data analysis. *National Science Review* **1**, pp. 293–314 (2014).
9. Barabási, A. L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* **5**, pp. 101–113 (2004).
10. van den Heuvel, M. P. & Sporns, O. Network hubs in the human brain. *Trends in Cognitive Sciences* **17**, pp. 683–696 (2013).
11. Leclerc, R. D. Survival of the sparsest: Robust gene networks are parsimonious. *Mol. Syst. Biol.* **4**, 213 (2008).
12. Fan, J. & Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B.* **70**, pp. 849–911 (2008).
13. Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. Sparse additive models. *J. R. Stat. Soc. Ser. B* **71**, 1009–1030. (2009).
14. Chipman, H. A., George, E. I. & McCulloch, R. E. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, pp. 266–298 (2012).
15. Johnstone, I. M. & Titterington, D. M. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**, pp. 4237–4253 (2009).
16. Bickel, P. J. & Levina, E. Covariance regularization by thresholding. *Ann. Stat.* **36**, pp. 2577–2604 (2008).
17. El Karoui, N. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Stat.* **36**, pp. 2717–2756 (2008).
18. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, pp. 2319–2323 (2000).
19. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, pp. 2323–2326 (2000).
20. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, pp. 1373–1396 (2003).
21. Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci.* **100**, pp. 5591–5596 (2003).
22. Graham, J. W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **60**, pp. 549–576 (2009).
23. Wothke, W. Longitudinal and multigroup modeling with missing data. In *Model.*

Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples, edited by T. D. Little, K. U. Schnabel, amd J. Baumert . pp. 219–240 & 269–281. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA (2000).
24. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm . *J. R. Stat. Soc. Ser. B* **39**, pp. 1–22. (1977).
25. Wu, C. F. J. On the convergence properties of the EM algorithm. *Ann. Stat.* **11**, pp. 95–103 (1983).
26. Rubin, D. B. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Hoboken, NJ, USA (1989).
27. Rubin, D. B. Multiple imputation after 18+ Years. *J. Am. Stat. Assoc.* **91**, pp. 473–489 (1996).
28. Tanner, M. A. & Wong, W. H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**, pp. 528–540 (1987).
29. Garcia, T. P. & Marder, K. Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Current Neurology and Neuroscience Reports* **17**, 14 (2017).
30. Bhagwat, N., Viviano, J. D., Voineskos, A. N. & Chakravarty, M. M. Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput. Biol.* **14**, e1006376 (2018).
31. Casey, B. J., Giedd, J. N. & Thomas, K. M. Structural and functional brain development and its relation to cognitive development. *Biol. Psychol.* **54**, 241–257 (2000).
32. Gogtay, N. *et al.* Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci.* **101**, pp. 8174–8179 (2004).
33. Reiss, A. L., Abrams, M. T., Singer, H. S., Ross, J. L. & Denckla, M. B. Brain development, gender and IQ in children. A volumetric imaging study. *Brain* **119**, pp. 1763–1774 (1996).
34. Sowell, E. R., Thompson, P. M., Holmes, C. J., Jernigan, T. L. & Toga, A. W. In vivo evidence for post-adolescent brain maturation in frontal and striatal regions. *Nature Neuroscience* **2**, pp. 859–861 (1999).
35. Casey, B. J., Tottenham, N., Liston, C. & Durston, S. Imaging the developing brain: What have we learned about cognitive development? *Trends in Cognitive Sciences* **9**, pp. 104–110 (2005).
36. Casey, B. J., Tottenham, N. & Fossella, J. Clinical, imaging, lesion, and genetic approaches toward a model of cognitive control. *Dev. Psychobiol.* **40**, pp. 237–254 (2002).
37. Tamm, L., Menon, V. & Reiss, A. L. Maturation of brain function associated with response inhibition. *J. Am. Acad. Child Adolesc. Psychiatry* **41**, pp. 1231–1238 (2002).
38. Adleman, N. E. *et al.* A developmental fMRI study of the Stroop color-word task. *Neuroimage* **16**, pp. 61–75 (2002).
39. Nagy, Z., Westerberg, H. & Klingberg, T. Maturation of white matter is associated with the development of cognitive functions during childhood. *J. Cogn. Neurosci.* **16**, pp. 1227–1233 (2004).
40. Brown, T. T. *et al.* Developmental changes in human cerebral functional organization for word generation. *Cereb. Cortex* **15**, pp. 275–290 (2005).
41. Pfefferbaum, A. *et al.* A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Arch. Neurol.* **51**, pp. 874–887 (1994).
42. Caviness, V. S., Kennedy, D. N., Richelme, C., Rademacher, J. & Filipek, P. A. The

human brain age 7-11 years: A volumetric analysis based on magnetic resonance images. *Cereb. Cortex* **6**, pp. 726–736 (1996).

43. Giedd, J. N. *et al.* Quantitative magnetic resonance imaging of human brain development: Ages 4–18. *Cereb. Cortex* 6**,** pp. 551–559 (1996).