NeuroImage 226 (2021) 117508

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Identifying neural signatures mediating behavioral symptoms and psychosis onset: High-dimensional whole brain functional mediation analysis

Oliver Y. Chén[a,1,*], Hengyi Cao[b,l,m,1], Huy Phan[c], Guy Nagels[d], Jenna M. Reinen[e], Jiangtao Gou[f], Tianchen Qian[g,n], Junrui Di[h], John Prince[a], Tyrone D. Cannon[b,i], Maarten de Vos[j,k,o]

[a] Department of Engineering Science, University of Oxford, Oxford OX1 4AR, United Kingdom
[b] Department of Psychology, Yale University, New Haven 06510, CT, United States
[c] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom
[d] Department of Neurology, Universitair Ziekenhuis Brussel, 1090 Jette, Belgium
[e] IBM Watson Research Center, Yorktown Heights, NY 10598, United States
[f] Department of Mathematics and Statistics, Villanova University, PA 19085, United States
[g] Department of Statistics, Harvard University, Cambridge 02138, MA, United States
[h] Department of Biostatistics, Johns Hopkins University, Baltimore 21205, MD, United States
[i] Department of Psychiatry, Yale University, New Haven 06510, CT, United States
[j] Faculty of Engineering Science, KU Leuven, Leuven 3001, Belgium
[k] Faculty of Medicine, KU Leuven, Leuven 3001, Belgium
[l] Center for Psychiatric Neuroscience, Feinstein Institute for Medical Research, Hempstead 11030, NY, United States
[m] Division of Psychiatry Research, Zucker Hillside Hospital, Glen Oaks 11004, NY, United States
[n] Department of Statistics, University of California Irvine, Irvine 92697, CA, United States
[o] KU Leuven Institute for Artificial Intelligence, Leuven B-3000, Belgium

A B S T R A C T

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 α brain atlas and a β 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. 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 independent 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 (Cannon et al., 2008; Demjaha et al., 2012; Carrión et al., 2013). 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.

* Corresponding author.
E-mail address: yibing.chen@seh.ox.ac.uk (O.Y. Chén).
[1] These authors contributed equally to this paper.

https://doi.org/10.1016/j.neuroimage.2020.117508
Received 14 August 2020; Received in revised form 19 October 2020; Accepted 22 October 2020
Available online 4 November 2020

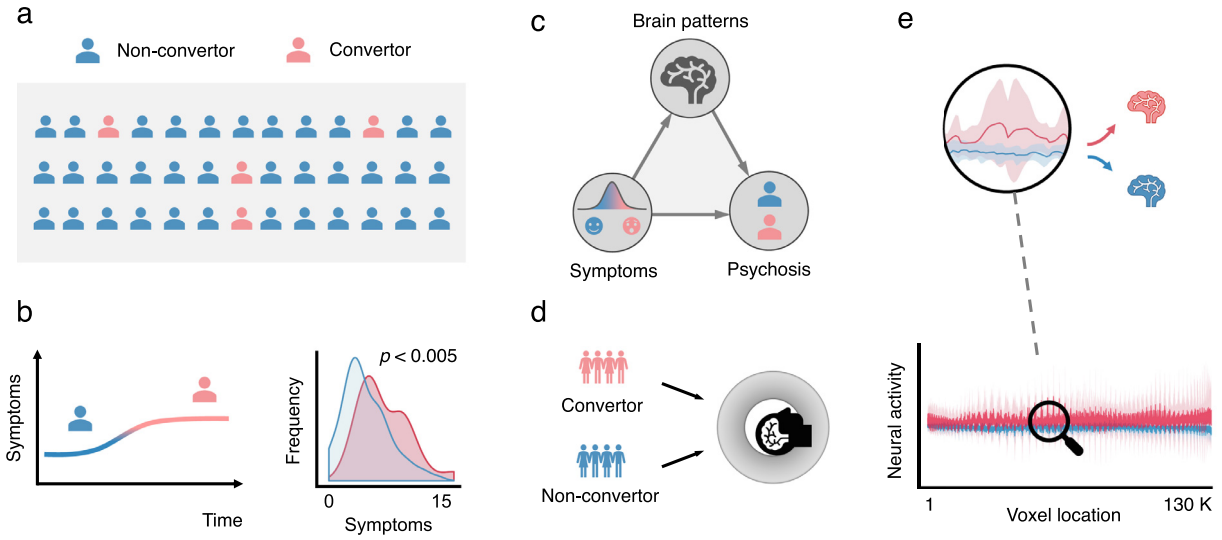

Fig. 1. *The study layout of the neural mediation analysis.* **(a)** We considered a sample of 263 subjects recruited from eight study sites across the United States and Canada who met criteria for a prodromal risk syndrome at the point of recruitment and had been clinically followed up for two years 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 along the whole-brain space and the blue shade represented those from the non-convertor group.

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, however, 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 outcomes. For example, existing multi-level mediation models assume that the outcomes are continuously distributed (Chén et al., 2018; Geuter et al., 2020; Huang and Pan, 2016; VanderWeele and Vansteelandt, 2014); mediation frameworks concerning binary outcomes are at present restricted to a relatively small number of mediators (VanderWeele and Vansteelandt, 2014; Nguyen, 2016); 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., 2018)). Second, although functional mediation analysis (Lindquist, 2012) has considerably advanced knowledge about the functional signal organization of the brain in relation to independent 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 non-orthogonal. 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 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 different groups in relation to behavior symptoms; (c) quantify each neural mediator's effect on disease outcome; and (d) identify and separate brain areas that are potentially positively and negatively mediating brain disorders.

In clinical practice, one assumes that an irregular change of 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 one 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 arises, 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 independent variables; and (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

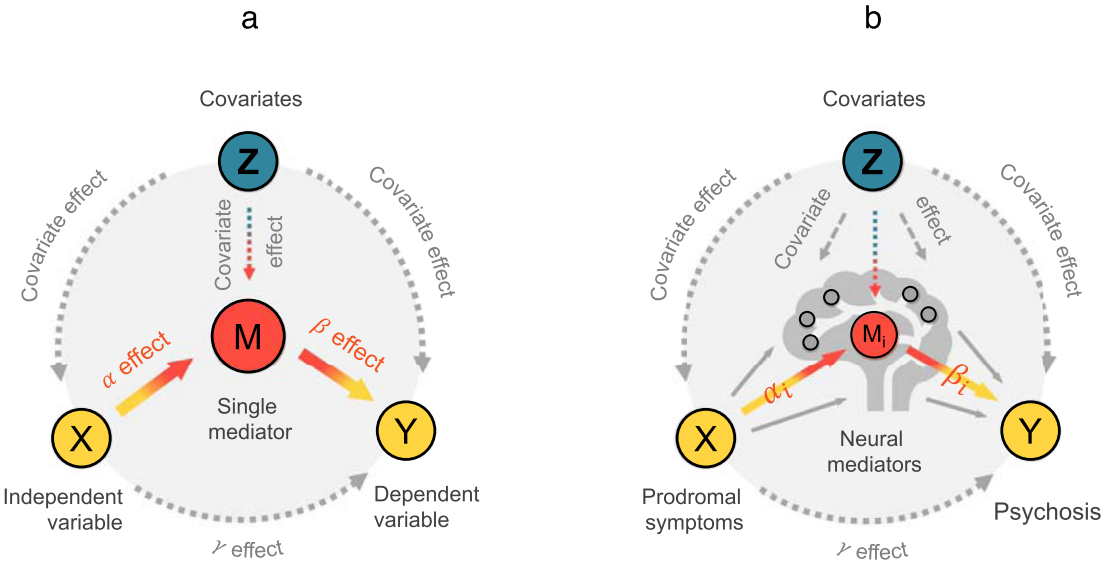

Fig. 2. *A schematic representation of mediation analysis. (a) Univariate mediation analysis.* The circles indicate an independent variable, a univariate mediator, an outcome variable, and covariates. The arrows denote pathways. The letter α denotes the effect from the independent variable to the mediator, after accounting for the covariate effect. The letter β denotes the effect of the mediator on the outcome, after controlling the independent variable and covariates. The letter γ denotes the effect from the independent 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 $(1 \le i \le V)$ denotes the effect from the independent 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 independent variable and covariates. The letter γ indicates the direct effect from the independent variable to the outcome, after accounting for the covariate effect.

(TMS)), under some identification conditions, the proposed model may unveil potential causal direct effect 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.

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 independent variable X on an outcome variable Y is at least partially carried through M (Baron and Kenny, 1986; Robins and Greenland, 1992). In the following, we use upper cases (*e.g.*, M) to indicate random variables, and lower cases (*e.g.*, m) to indicate observed values. We use nonbold letters (*e.g.*, M and m) to represent univariate variables and observations; we use bold letters (*e.g.*, $\mathbf{M}$ and $\mathbf{m}$) to represent multivariate variables and observations. Examples of univariate mediators are pain catastrophizing, which mediates the clinical treatment and disability status (Whittle et al., 2017), and intention, which mediates attitudes and behavior (Fishbein and Ajzen, 1975).

The identification of a univariate mediator consists of two steps (Alwin and Hauser, 1975; Baron and Kenny, 1986; Hyman, 1955; Judd and Kenny, 1981; Sobel, 1982) (see **Supporting Information** for a comparison between two common mediation analysis frameworks). The first step examines if the independent variable X has an effect on the mediator M after controlling for the covariates $\boldsymbol{Z}$, using the following conditional model:

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

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

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

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

where θ_0', γ, β and $\boldsymbol{\tau}$ are coefficients for the intercept, the observed independent variable x, the observed value of the mediator m, and the observed covariates $\boldsymbol{z}$. The prime in θ_0' is to differentiate it from its counterpart in (i). If β is significantly different from zero, then the mediator has an effect on 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 in a different way, if the product $\alpha\beta$ is non-zero, then M is a mediator for X and Y, and the value of $\alpha\beta$ 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.

1.2. Multivariate mediation analysis

Mediation analysis concerning a multivariate mediator can be conducted using structural equation models (SEMs) (Lindquist, 2012; VanderWeele and Vansteelandt, 2014) (see **Fig. 2(b)**). Formally, consider V mediators ($V \ge 2$), denoted as $M(1), M(2), \ldots, M(V)$, an independent variable X, and an outcome variable Y. Multivariate mediation

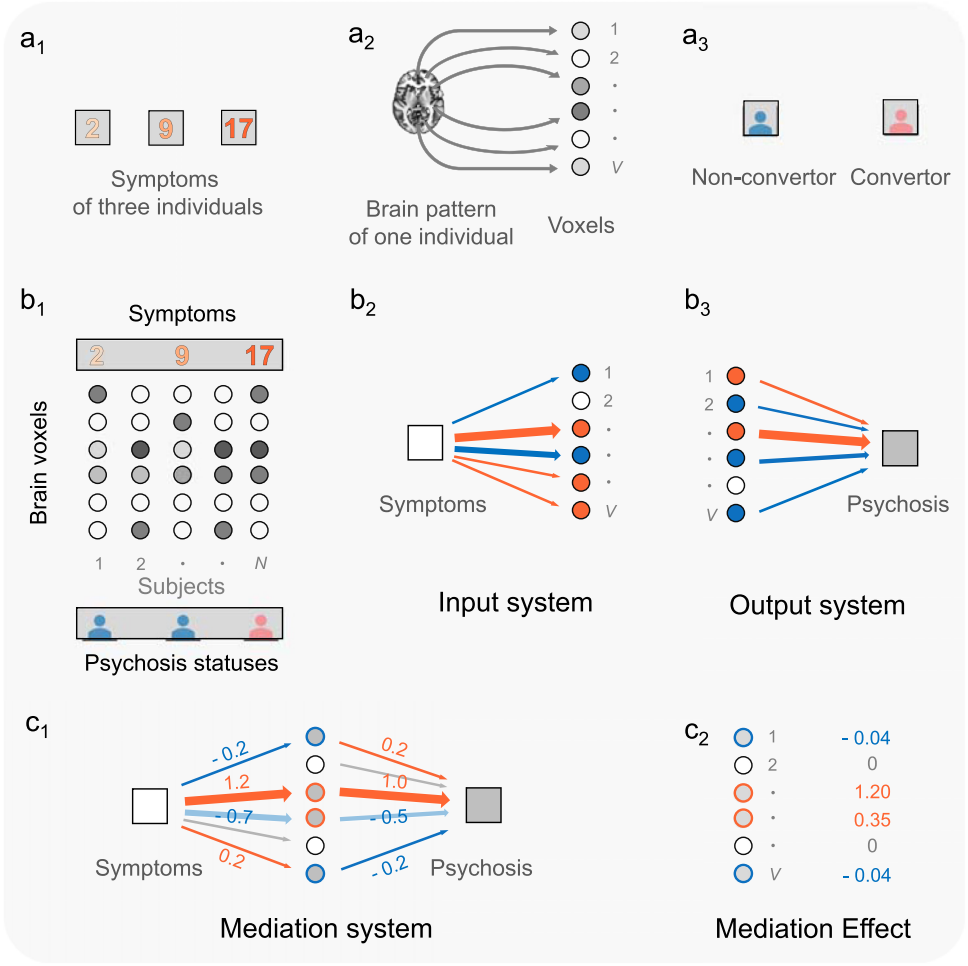

Fig. 3. *A hypothetical experiment and how multivariate mediation analysis can be used to study brain mediation in health and disease.* **(a1)** Three individuals' behavioral symptom scores are measured at baseline. **(a2)** Resting-state brain activities for one individual is collapsed into activities 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 a blue icon to refer to a non-convertor and a red icon to refer to 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 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 pathways corresponding to the voxel.

analysis considers two conditional models as follows.

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

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

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

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 independent 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 independent 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 independent 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 independent and outcome variables for subject i, respectively. Let $\boldsymbol{z}_i = (z_{1i}, z_{2i}, z_{3i}, z_{4i}$ and $z_{5i})$ 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 ($\mathbf{b_2}$)):

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

where $\theta_0(j)$, $\alpha(j)$, and $\boldsymbol{t}(j)$ are coefficients for the intercept, the independent variable, and covariates that are associated with the j^{th} mediator. It is worthwhile mentioning that the estimated parameters $\boldsymbol{\alpha} = (\alpha(1), \alpha(2), \ldots, \alpha(V))$ (see the alpha atlas in the **Results** section) are closely related to canonical correlation (CCA) between multivariate brain signals $\boldsymbol{M}$ and the independent 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 ($\mathbf{b_3}$)):

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

where β_0, γ, and $\boldsymbol{\tau}$ are coefficients for the intercept, the independent variable, and covariates. Here, $\boldsymbol{\Delta}_i = (1, x_i, \mathbf{M}_i^T, \mathbf{z}^T)$ denotes the data ξ_{ik} and $\varphi_k(j)$ are the Karhunen-Loève expansion (Loève, 1945; Karhunen, 1947) 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 $\xi_{ik} \sim N(0, \lambda_k)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_\infty$, 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 $g(y) = y$, $g(y) = \log(\frac{y}{1-y})$, 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 multivariate neural mediator $\boldsymbol{M}$. Although by Mercer's theorem (see Chapter 4 of Indritz, 1963), 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 independent variable on each brain area; this yields an $\boldsymbol{\alpha}$ brain atlas (see Fig. 3($\mathbf{b_2}$)). 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 independent 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 Fig. 3 ($\mathbf{b_3}$)), **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($\mathbf{c_1}$) - ($\mathbf{c_2}$) 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 Logitudinal Study (NAPLS-2) consortium (Addington et al., 2012), 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) (McGlashan et al., 2001) and were clinically followed-up for two years. During follow-up, 25 subjects developed one type of the 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) (Bell, 1994) criteria for an Axis-I psychotic disorder or had at least one fully psychotic symptom assessed by the Structured Interview for Prodromal Syndromes (SIPS) at follow-up (Miller et al., 2002; Miller et al., 2003). 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), 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-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 × 240 mm × 176 mm FOV, TR/TE 2300/2.91 ms, 9° flip angle; 2) GE scanners: spoiled gradient recalledecho (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 (Cao et al., 2018; Cao et al., 2019; Cao et al., 2019; Cao et al., 2019), 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 potential head motion effect, 8 subjects (1 converter, 7 non-converters) with mean FD > 0.35 mm were excluded from further analysis. The activities of each gray matter voxel during resting state were quantified by fractional amplitude of low-frequency fluctuation (fALFF), a well-established measure evaluating the ratio of power spectrum of low-frequency (0.01–0.1 Hz) to that of the entire frequency range (Zou et al., 2008). After computation, fALFF values were extracted from a total of 130,992 voxels covering

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

	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

the gray matter of the entire brain for each subject. Secondary data analysis was conducted using the *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 independent variable (*e.g.*, behavioral symptoms) has an effect on each brain area (see Fig. 3($\mathbf{b_2}$)). The output system examines if each brain area has an effect on the outcome (see Fig. 3($\mathbf{b_3}$)).

Formally, consider N subjects and V brain areas, where $N = 263$ and $V = 130{,}992$ in this study. Let $x_i \in \mathbb{R}$ be the independent 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) (Ramsay and Silverman, 2005; Wang et al., 2016). 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, $\mu(j) = \mathbb{E}(m(j))$ and $K(s, t) = cov(m(s),\ m(t))$. By Mercer's theorem (see Chapter 4 of Indritz, 1963), 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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_\infty$ are decreasingly ordered nonnegative eigenvalues and φ_k's are their corresponding orthogonal eigenfunctions with unit $\mathcal{L}^2$ norms.

Karhunen-Loève expansion (Loève, 1945; Karhunen, 1947) 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 inquire 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 independent 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 input system consists of the following model:

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

where $\theta_0(j)$, $\alpha(j)$, and $\boldsymbol{t}(j)$ are coefficients for the intercept, the independent variable, and covariates that are associated with the j^{th} mediator. Specifically, $\alpha(j)$ captures the effect of the independent 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 $j \in \{\frac{1}{V}, \frac{2}{V}, \ldots, \frac{V}{V}\}$.

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

$$f\left(y_i|\ \boldsymbol{\Delta}_i\right) = \exp\left\{\left(y_i\theta_i - a\left(\theta_i\right) + b\left(y_i\right)\right)\phi\right\} \tag{vii}$$

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, $\boldsymbol{m}_i^{\mathrm{T}} = (m_i(1),\ m_i(2),\ \ldots,\ m_i(V))$, and $m_i(j) = \sum_{k=1}^{\infty} \xi_{ik}\varphi_k(j) \approx \sum_{k=1}^{K} \xi_{ik}\varphi_k(j)$, where $\xi_{ik} \sim N(0,\ \lambda_k)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_\infty$, for $i \in \{1,\ 2, \ldots, N\}$ and $j \in \{\frac{1}{V}, \frac{2}{V}, \ldots, \frac{V}{V}\}$, and $\boldsymbol{\varphi} = \{\varphi_1,\ \varphi_2, \ldots,\ \varphi_K\}$ 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 independent 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, Eq. (vi) has the following form:

$$y_i \sim Bernoulli\left(p_i\right) \tag{viii}$$

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

Thanks to spectral decomposition of $\boldsymbol{m}$, parameter estimation of $\boldsymbol{\beta}$ can be performed on $\boldsymbol{\xi} = (\boldsymbol{\xi}_1,\ \boldsymbol{\xi}_2,\ \ldots,\ \boldsymbol{\xi}_N)^{\mathrm{T}}$, where $\boldsymbol{\xi}_i = \{\xi_{i1}, \xi_{i2}, \ldots, \xi_{iK}\}$. 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 Eq. (vi) 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 $\tilde{\boldsymbol{\beta}} = (\sum_{j=1}^{p} \tilde{\beta}_1(j),\ \sum_{j=1}^{p} \tilde{\beta}_2(j),\ \ldots,\ \sum_{j=1}^{p} \tilde{\beta}_K(j))$. Subsequently, the estimation of $\boldsymbol{\beta}$ can be retrieved by projecting the estimated $\hat{\tilde{\boldsymbol{\beta}}}$ back to the brain space using the estimated basis functions $\hat{\boldsymbol{\varphi}}$; in other words, $\hat{\boldsymbol{\beta}} = \hat{\tilde{\boldsymbol{\varphi}}}^{-}\hat{\tilde{\boldsymbol{\beta}}}$, where $^{-}$ 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 $j = 1,\ 2,\ \ldots,\ V$. 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 independent variable. When V is small, Sobel's Test (Sobel, 1982) 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 (Efron, 1979) (see **Supporting Information**); the results are further adjusted for multiple comparisons.

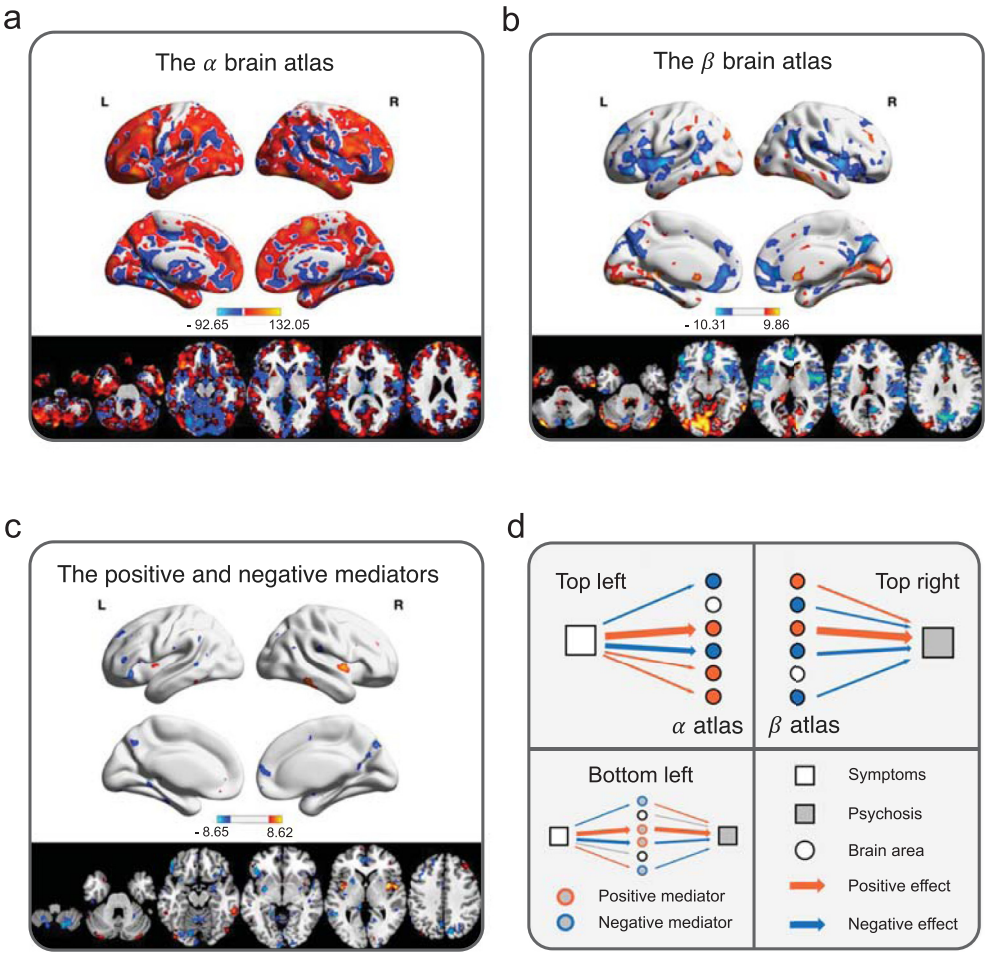

Fig. 4. *Brain areas potentially mediating behavioral symptoms and the development of psychosis.* A high-dimensional mediation analysis on 130,992 brain voxels of 263 subjects suggests that the pathway between behavioral symptoms and psychosis is potentially positively mediated by the right lateral prefrontal 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 orange 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.

2.5. Brain-wide functional mediation analysis

We conducted 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 independent 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 α brain atlas. We then estimated the effect each brain area had on the outcome, controlling for the independent variable and all covariates, using Eq. (vi) and obtained the β 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 independent 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 and disease outcomes using Eqs. (v) and (vi), respectively (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.

Overall, the proposed framework was able to uncover (simulated) brain areas 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 α was better than it on estimating the output map β. With larger samples, however, the estimation of β improved. Both estimations of α and β 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 criteria for a prodromal risk syndrome (Miller et al., 2003) at the point of recruitment and were clinically followed up for two years as part of the NAPLS-2 project (Addington et al., 2012) (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) (Zou et al., 2008) 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) (McGlashan et al., 2001), 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 (Cannon et al., 2008; Demjaha et al., 2012; Carrión et al., 2013), 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. 1** (**b**)). We then continued to investigate which brain regions would functionally mediate this association. First, we tested if behavioral symptoms had an effect on 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 α 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 β 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 α and β brain atlases and classified them into two categories: the positive and negative neural mediators (see **Fig. 4** (**c**)). Finally, we conducted bootstrap experiments to test whether the mediation effect corresponding to 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 α and β brain atlases

We further inquired into the α brain atlas (**Fig. 4**(**a**)) and the β brain atlas (**Fig. 4**(**b**)) in order to investigate how the input and output systems contribute to overall mediation. Specifically, the α brain atlas included brain regions that were associated with behavioral symptoms; the β brain atlas consisted of brain areas that were associated with psychosis status when controlled for behavioral symptoms. On the α 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 β 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 parallel, 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.

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' 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 independent 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 independent 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 cerebellum (crus 1 and 2). With extensive connections with both sensory areas and 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 (Wylie and Tregellas, 2010; Namkung et al., 2017). 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 (Dierks et al., 1999; Shergill et al., 2000; Powers et al., 2017). Increased connectivity was found between insula and multiple perceptual areas such as sensorimotor cortex and visual cortex, together with decreased connectivity between insula and prefrontal cortex (Tian et al., 2019). 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 (Braff, 1990). The frontoparietal network, inferior temporal gyrus, and cerebellar crus 1 and crus 2 are key cognitive areas in humans (Dosenbach et al., 2007; Dosenbach et al., 2008; Buckner et al., 2011; Marek et al., 2018); increased activity and connectivity in these regions have been shown to be predictive of psychosis onset in previous studies (Cao et al., 2018; Cao et al., 2019). 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 or DMN (including medial prefrontal cortex, anterior cingulate cortex, and precuneus), together with 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 the DMN in patients is the failure to deactivate this network during cognitive tasks (Pomarol-Clotet et al., 2008; Fryer et al., 2013), which may relate to exaggerated internally-focused thoughts and lack of sufficient suppression of these thoughts during cognition (Whitfield-Gabrieli and Ford, 2012). 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 insula (as a positive mediator) on the DMN has been reported in individuals at risk for psychosis (Wotruba et al., 2014; Bolton et al., 2020). Cerebellar lobule 8 is a key region for processing sensorimotor-related errors (Buckner et al., 2011; Schmahmann, 2019). 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 (Sollychin et al., 2019), 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 (Hamilton et al., 2019); it nevertheless may not capture as much variability that occurs in spontaneous brain activity as 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 (Ingvar et al., 1976; Guich et al., 1989). 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 (Chung et al., 2019); 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 independent variable and the mediator may be useful to expand current analysis (see **Supporting Information** for an example and see (Muller et al., 2005) 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, 2014)). Fifth, the proposed model focused on neural markers by averaging the 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 effect related to brain development during childhood and adolescence, brain aging between health and disease, and brain degeneration along the trajectory of a neurodegenerative disease development. We are currently investigating how to extend the techniques used in our framework to improve our understanding about large-scale longitudinal mediation; potential directions include combing functional data analysis (FDA) and a dual mediation system with autoregressive models (Gollob and Reichardt, 1991; Cole and Maxwell, 2003), latent growth curve (LGM) (Muthén and Curran, 1997), parallel process models (Cheong et al., 2009), latent difference score (LDS) models (McArdle, 2001), and autoregressive LGM models (Bollen and Curran, 2004).

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 mediation analysis concept and neuroscientific insights to inquire into the properties of large-scale functional neural markers that mediate the relationship between an independent 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.

Author contributions

O.Y.C. designed the model and performed the mediation analysis. H.C. provided neurobiological intepretations. G.N. and J.M.R. provided neurobiological support. H.P. and J.P. provided simulation support. J.G., T.Q., and J.D. provided statistical support. T.D.C. and M.D.V. provided funding, support, and guidance. O.Y.C. and H.C. wrote the manuscript, with comments from all other authors.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgement

The research was supported by the Wellcome Trust SCNI (098461/Z/12/Z) and the Flemish Government (AI Research Program) to Dr. de Vos, the NARSAD Young Investigator Grant (No. 27068) to Dr. Cao, and the NIH grant U01 MH081902 to Dr. Cannon. T.C. Dr. Cannon would like to acknowledge gifts from the Staglin Music Festival for Mental Health and International Mental Health Research Organization. The authors wish to thank the principal investigators of the North American Prodrome Longitudinal Study for allowing use of the project's imaging data in this demonstration analysis: Jean Addington, Carrie Bearden, Kristin Cadenhead, Barbara Cornblatt, Diana Perkins, Larry Seidman, Elaine Walker, and Scott Woods. The authors thank the Editor and three anonymous reviewers for their constructive suggestions which improved the quality and readability of the paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117508.

References

Addington, J, et al., 2012. North American Prodrome Longitudinal Study (NAPLS 2): overview and recruitment. Schizophr. Res. 142, 77–82.
Alwin, D.F., Hauser, R.M., 1975. The decomposition of effects in path analysis. Am. Sociol. Rev. 40, 37.
Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. 51, 1173–1182.
Bell, C.C., 1994. DSM-IV: diagnostic and Statistical Manual of Mental Disorders. J. Am. Med. Assoc. 272, 828–829.
Bollen, K.A., Curran, P.J., 2004. Autoregressive Latent Trajectory (ALT) models: a synthesis of two traditions. Sociol. Methods Res. 32, 336–383.
Bolton, T.A.W, et al., 2020. Triple network model dynamically revisited: lower salience network state switching in pre-psychosis. Front. Physiol. 11, 66.
Braff, D.L., Geyer, M.A., 1990. Sensorimotor gating and Schizophrenia. Arch. Gen. Psychiatry 47, 181–188.
Buckner, R.L., et al., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 2322–2345.
Cannon, T.D, et al., 2008. Prediction of psychosis in youth at high clinical risk. Arch. Gen. Psychiatry 65, 28–37.
Cao, H, et al., 2018. Cerebello-thalamo-cortical hyperconnectivity as a state-independent functional neural signature for psychosis prediction and characterization. Nat. Commun. 9, 3836.
Cao, H, et al., 2019a. Altered brain activation during memory retrieval precedes and predicts conversion to psychosis in individuals at clinical high risk. Schizophr. Bull. 45, 924–933.
Cao, H, et al., 2019b. Progressive reconfiguration of resting-state brain networks as psychosis develops: preliminary results from the North American Prodrome Longitudinal Study (NAPLS) consortium. Schizophr. Res.. https://doi.org/10.1016/j.schres.2019.01.017.
Cao, H., et al., 2019c. Evidence for cerebello-thalamo-cortical hyperconnectivity as a heritable trait for schizophrenia. Transl. Psychiatry 9, 1–8.
Carrión, R.E, et al., 2013. Prediction of functional outcome in individuals at clinical high risk for psychosis. JAMA Psychiatry 70, 1133–1142.
Chén, O.Y, et al., 2018. High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics 19, 121–136.
Cheong, J., Mackinnon, D.P., Toon, S, 2009. Investigation of mediational processes using parallel process latent growth curve modeling. Struct. Equ. Model. 10, 238–262.
Chung, Y, et al., 2019. Adding a neuroanatomical biomarker to an individualized risk calculator for psychosis: a proof-of-concept study. Schizophr. Res. 208, 41–43.
Cole, D.A., Maxwell, S.E., 2003. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. J. Abnorm. Psychol. 112, 558–577.
Demjaha, A., et al., 2012. Disorganization/cognitive and negative symptom dimensions in the at-risk mental state predict subsequent transition to psychosis. Schizophr. Bull. 38, 351–359.
Dierks, T, et al., 1999. Activation of Heschl's gyrus during auditory hallucinations. Neuron 22, 615–621.
Dosenbach, N.U.F, et al., 2007. Distinct brain networks for adaptive and stable task control in humans. Proc. Natl. Acad. Sci. U. S. A. 104, 11073–11078.
Dosenbach, N.U., et al., 2008. A dual-networks architecture of top-down control. Trends Cogn. Sci. 12, 99–105.
Efron, B., 1979. Bootstrap methods: another look at the Jackknife. Ann. Stat. 7, 1–26.
Fan, J., Liao, Y., 2014. Endogeneity in high dimensions. Ann. Stat. 42, 872–917.
Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA, USA.
Fryer, S.L, et al., 2013. 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.
Geuter, S, et al., 2020. Multiple brain networks mediating stimulus-pain relationships in humans. Cereb. Cortex 30, 4204–4219.
Gollob, H.F., Reichardt, C.S, 1991. Interpreting and estimating indirect effects assuming time lags really matter. In: Collins, L.M., Horn, J.L. (Eds.), Best Methods For the Analysis of change: Recent advances, Unanswered Questions. American Psychological Association, pp. 243–259 *future directions*, edited by.
Guich, S.M, et al., 1989. Effect of attention on frontal distribution of delta activity and cerebral metabolic rate in schizophrenia. Schizophr. Res. 2, 439–448.
Hamilton, H.K, et al., 2019. Association between P300 responses to auditory oddball stimuli and clinical outcomes in the psychosis risk syndrome. JAMA Psychiatry 76, 1187–1197.
Huang, Y.T., Pan, W.C., 2016. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics 72, 402–413.
Hyman, H.H., 1955. Survey Design and Analysis: Principles, Cases and Procedures. The Free Press, Glencoe, Illinois, USA.
Indritz, J., 1963. Methods in Analysis. Macmillan, New York, USA.
Ingvar, D.H., Sjölund, B., Ardö, A., 1976. Correlation between dominant EEG frequency, cerebral oxygen uptake and blood flow. Electroencephalogr. Clin. Neurophysiol. 41, 268–276.
Judd, C.M., Kenny, D.A., 1981. Process analysis: estimating mediation in treatment evaluations. Eval. Rev. 5, 602–619.
Karhunen, K., 1947. Über lineare Methoden in der Wahrscheinlichkeitsrechnung (On linear methods in probability and statistics). Ann. Acad. Sci. Fenn. Ser. A. I. Math.-Phys. 31, 1–79.
Lindquist, M.A., 2012. Functional causal mediation analysis with an application to brain connectivity. J. Am. Stat. Assoc. 107, 1297–1309.
Loève, M, 1945. 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, 295–296.
Marek, S, et al., 2018. Spatial and temporal organization of the individual human cerebellum. Neuron 100, 977–993.
McArdle, J.J, 2001. A latent difference score approach to longitudinal dynamic structural analysis. In: Cudeck, R., du Toit, S., Sörbom, D (Eds.), Structural Equation Modeling: Present and Future. Scientific Software International, Lincolnwood, IL, USA, pp. 342–380.
McGlashan, T.H., et al., 2001. Instrument for the assessment of prodromal symptoms and states. In: Miller, T., Mednick, S.A., McGlashan, T.H., Libiger, J., Johannessen, J.O. (Eds.), Early Intervention in Psychotic Disorders. Springer, Dordrecht, the Netherlands, pp. 135–149 byNATO Science Series (Series D: Behavioural and Social Sciences).
Miller, T.J, et al., 2002. 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.
Miller, T.J, et al., 2003. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. Schizophr. Bull. 29, 703–715.
Muller, D., Judd, C.M., Yzerbyt, V.Y., 2005. When moderation is mediation and mediation is moderated. J. Pers. Soc. Psychol. 89, 852–863.
Muthén, B.O., Curran, P.J., 1997. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. Psychol. Methods 2, 371–402.
Namkung, H., Kim, S.H., Sawa, A., 2017. The insula: an underestimated brain area in clinical neuroscience, psychiatry, and neurology. Trends Neurosci. 40, 200–207.
Nguyen, T.Q., et al., 2016. Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention. Structural equation modeling 23, 368–383.
Pomarol-Clotet, E, et al., 2008. Failure to deactivate in the prefrontal cortex in schizophrenia: dysfunction of the default mode network. Psychol. Med. 38, 1185–1193.
Powers, A.R., Mathys, C., Corlett, P.R., 2017. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science 357, 596–600.
Ramsay, J., Silverman, B.W., 2005. Functional Data Analysis. Springer-Verlag, New York, USA.
Robins, J.M., Greenland, S., 1992. Identifiability and exchangeability for direct and indirect effects. Epidemiology 3, 143–155.
Schmahmann, J.D., 2019. The cerebellum and cognition. Neurosci. Lett. 688, 62–75.

Shergill, S.S., et al., 2000. Mapping auditory hallucinations in schizophrenia using functional magnetic resonance imaging. Arch. Gen. Psychiatry 57, 1033–1038.
Sobel, M.E., 1982. Asymptotic confidence intervals for indirect effects in structural equation models. Sociol. Methodol. 13, 290–312.
Sollychin, M, et al., 2019. 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, 293–299.
Tian, Y., et al., 2019. Insula functional connectivity in schizophrenia: subregions, gradients, and symptoms. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 4, 399–408.
VanderWeele, T.J., Vansteelandt, S., 2014. Mediation analysis with multiple mediators. Epidemiol. Method. 2, 95–115.
Wang, J.-.L., Chiou, J.-.M., Müller1, H.-.G., 2016. Functional data analysis. Ann. Rev. Stat. Appl. 3, 257–295.
Whitfield-Gabrieli, S., Ford, J.M., 2012. Default mode network activity and connectivity in psychopathology. Ann. Rev. Clin. Psychol. 8, 49–76.
Whittle, R., Mansell, G., Jellema, P., van der Windt, D, 2017. Applying causal mediation methods to clinical trial data: what can we learn about why our interventions (don't) work. Eur. J. Pain 21, 614–622.
Wotruba, D, et al., 2014. Aberrant coupling within and across the default mode, task–positive, and salience network in subjects at risk for psychosis. Schizophr. Bull. 40, 1095–1104.
Wylie, K.P., Tregellas, J.R., 2010. The role of the insula in schizophrenia. Schizophr. Res. 123, 93–104.
Zou, Q.H, et al., 2008. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. J. Neurosci. Methods 172, 137–141.