

## ARTICLE OPEN



# Data-driven causal model discovery and personalized prediction in Alzheimer's disease

Haoyang Zheng<sup>1</sup>, Jeffrey R. Petrella<sup>2</sup>, P. Murali Doraiswamy<sup>3</sup>, Guang Lin<sup>1,4</sup>✉, Wenrui Hao<sup>5</sup> and for the Alzheimer's Disease Neuroimaging Initiative\*

With the explosive growth of biomarker data in Alzheimer's disease (AD) clinical trials, numerous mathematical models have been developed to characterize disease-relevant biomarker trajectories over time. While some of these models are purely empiric, others are causal, built upon various hypotheses of AD pathophysiology, a complex and incompletely understood area of research. One of the most challenging problems in computational causal modeling is using a purely data-driven approach to derive the model's parameters and the mathematical model itself, without any prior hypothesis bias. In this paper, we develop an innovative data-driven modeling approach to build and parameterize a causal model to characterize the trajectories of AD biomarkers. This approach integrates causal model learning, population parameterization, parameter sensitivity analysis, and personalized prediction. By applying this integrated approach to a large multicenter database of AD biomarkers, the Alzheimer's Disease Neuroimaging Initiative, several causal models for different AD stages are revealed. In addition, personalized models for each subject are calibrated and provide accurate predictions of future cognitive status.

*npj Digital Medicine* (2022)5:137; <https://doi.org/10.1038/s41746-022-00632-7>

## INTRODUCTION

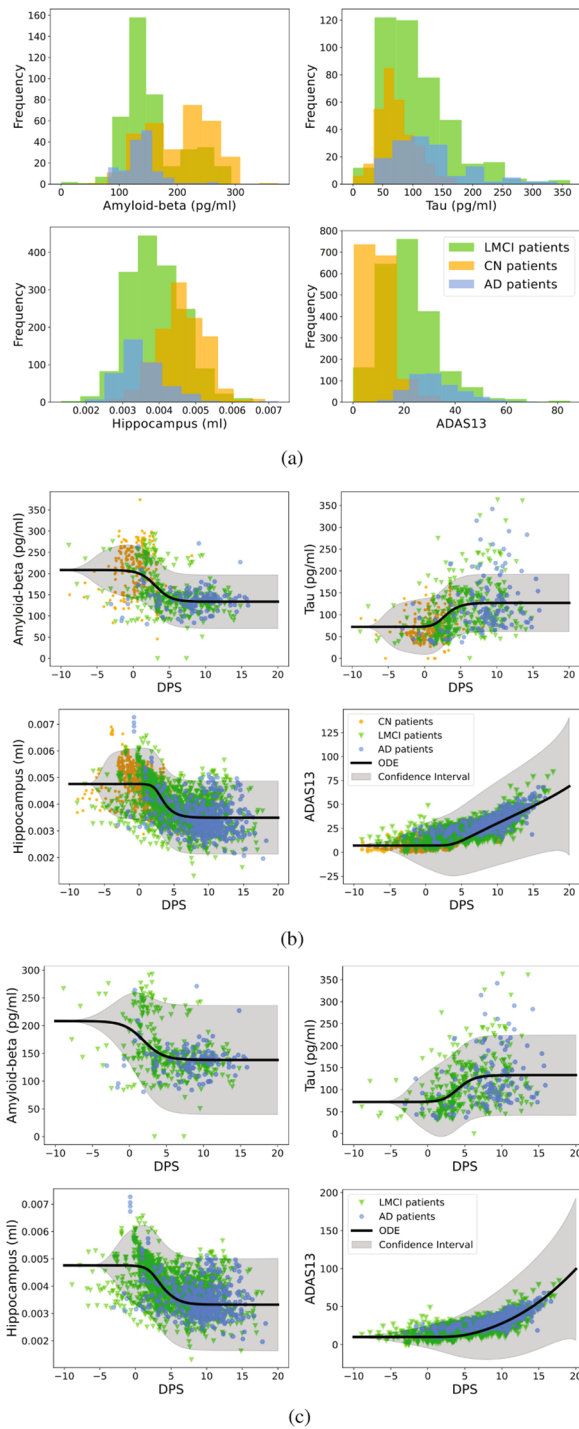
Among the top 10 causes of death in the United States, Alzheimer's disease (AD) is the only condition without a viable treatment to cure or prevent it, or even significantly slow its progression<sup>1</sup>. Failure to develop a successful disease-modifying therapy for AD, despite large investments of public and private resources, is rooted in its complexity<sup>2–4</sup>. For instance, signaling pathway analyses of AD pathophysiology has implicated over 30 metabolic pathways and over 1000 chemical species<sup>4</sup>. Our incomplete understanding of how these mechanisms vary and interact at an individual level to create a clinically and biologically heterogeneous phenotype has resulted in an attempt to treat patients with varying underlying pathophysiology in a similar fashion<sup>5,6</sup>. Thus, failure to characterize and subtype AD at an individual level has represented a major roadblock in the development of effective therapeutic strategies to slow or halt AD progression. Recent biological classification of AD, based on imaging and cerebral spinal fluid (CSF) biomarkers, represents a major step toward the future development of personalized prognoses and therapeutic strategies<sup>7,8</sup>. The increasing availability of such data in large cohorts of subjects has made possible the development and testing of rigorous quantitative models of AD pathophysiology. For example, the Alzheimer's Disease Neuroimaging Initiative (ADNI), a multicenter, prospective, naturalistic study, began in 2003, comprises four sequential studies—ADNI-1, ADNI-GO, ADNI-2, and ADNI-3—which followed subjects up to 15 years, using genetic, blood- and CSF-based, imaging, and cognitive biomarkers. The abundance of data from this and similar

multinational biomarker studies in AD will require a rigorous quantitative data-driven modeling approach to analyze, integrate and interpret data at the level of the individual, where it can have maximum clinical impact.

Several mathematical models of AD progression have been developed recently. For example, one mathematical model includes a cellular biologic system of neurons, glia, macrophages, amyloid $\beta$  aggregation, and tau to simulate and validate at a cellular level the mechanisms underlying the failure of several drugs in recent clinical trials, and suggest alternative approaches<sup>9</sup>. Moreover, a mathematical modeling approach has also been used to describe the key AD clinical biomarkers including pathologic hallmark biomarkers (beta-amyloid and tau), neuronal loss biomarkers, and cognitive impairment<sup>10</sup>. This model was parameterized and tested to successfully simulate the natural history scenarios of three sub-types of AD presented in<sup>11</sup>: (1) early-onset autosomal dominant AD, (2) late-onset amyloid-first AD, and (3) late-onset tau-first AD.

Although these mathematical models bring new insights in understanding AD progression and enable simulation of therapeutics, the current models are built upon a priori hypotheses of the AD pathophysiological network which still is an open area of research<sup>12</sup>. In fact, there are dozens of pathophysiological pathways implicated in AD by systems biologists, and our understanding of these networks and their interactions remains incomplete<sup>13</sup>. Moreover, there has been limited work on mechanistic modeling of clinically measurable AD biomarkers. Most research to date on the key AD biomarkers has been observational or correlational. Such modeling

<sup>1</sup>School of Mechanical Engineering, Purdue University, West Lafayette 47907 IN, USA. <sup>2</sup>Department of Radiology, Duke University Health System, Durham 27710 NC, USA. <sup>3</sup>Departments of Psychiatry and Medicine, Duke University School of Medicine and Duke Institute for Brain Sciences, Durham 27710 NC, USA. <sup>4</sup>Department of Mathematics, Purdue University, West Lafayette 47907 IN, USA. <sup>5</sup>Department of Mathematics, Penn State University, University Park 16802 PA, USA. \*A list of authors and their affiliations appears at the end of the paper. ✉email: [guanglin@purdue.edu](mailto:guanglin@purdue.edu)



**Fig. 1** The ADNI dataset histogram and the calibrated causal model plots. **a** Histogram of four biomarkers in the ADNI dataset. Top left is amyloid-beta; top right is the tau; bottom left is the normalized hippocampal volume; bottom right is the cognitive subscale. X-axes are the corresponding magnitudes of each biomarker, and y-axes are their frequencies. The subjects are classified into “CN”, “LMCI”, and “AD”. **b** The calibrated causal model on three groups of patient data. X-axes are fitted DPS of biomarkers, and y-axes are the corresponding magnitudes of each biomarker. The orange circles, green triangles, and blue hexagons are data from “CN”, “LMCI”, and “AD” subjects correspondingly. The black solid lines are the solutions of the causal model. **c** The calibrated causal model on the dataset of LMCI and AD groups.

approaches do not benefit from the tools of a more integrative systems approach that address disease mechanism<sup>14,15</sup>.

Computational data-driven modeling approaches have already achieved success in analyzing multi-dimensional clinical data in diseases such as cancer<sup>16,17</sup> and cardiovascular disease<sup>18,19</sup>. Such data-driven approaches employ mathematical models for patient populations using clinical, omics, and biomarker data, as well as powerful and new means to personalize such models based on individual data, yielding personal risk profiles. These data-driven modeling approaches can simulate complex systems, helping to elucidate complex physiological interactions and optimize personalized prevention and treatment strategies. Examples of such work include statistical approaches, such as Bayesian generalized linear models<sup>20</sup>, Bayesian hierarchical models<sup>21,22</sup>, and those based on Markov chain Monte Carlo simulations<sup>23–26</sup> to analyze genome sequencing and biomarker dynamics. Recently, machine learning techniques, such as deep recurrent neural networks, have been used to predict AD progression<sup>27</sup>.

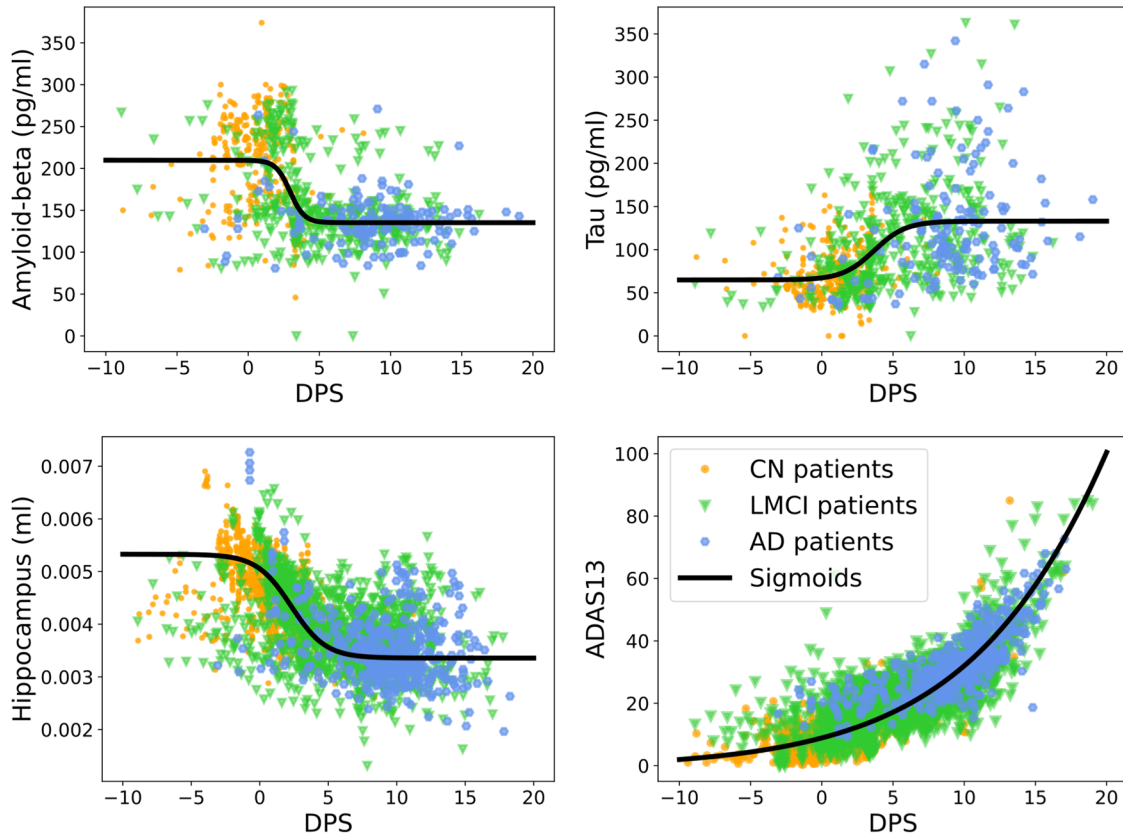
In this paper, we propose to develop a computational data-driven modeling framework to predict AD biomarker progression. We propose a methodology to construct data-driven causal models at a group and individual patient level. This method does not depend on any specific hypothesis of AD progression and extracts the causal model completely from the empirical data. More specifically, we derive the causal model based on clinical biomarkers in the ADNI dataset. In this data-driven modeling approach, the causal model is learned from four biomarkers (amyloid-beta pathology, total-tau pathology, hippocampal volume, and cognitive decline) to describe AD progression. Moreover, we incorporate a disease progression score (DPS) in the causal model<sup>28</sup> to unify AD progression for different subjects since the onset age and rates of progression may markedly vary within and across the different subject classes in ADNI.

## RESULTS

We elaborate on the effectiveness of the proposed data-driven causal model here. First, we construct a population-based causal model that describes the biomarker dynamics for all eligible subjects in ADNI-1, including normal controls. By fitting the population parameters via the ADNI dataset, the population model describes the transition of AD biomarkers between three different disease stages, cognitively normal (CN), late mild cognitive impairment (LMCI), and Alzheimer’s disease (AD). Second, we derive a population model for LMCI and AD subjects only. Third, we analyze the Sobol sensitivity<sup>29,30</sup> of the parameter space of the population model, which identifies the attribution of each model parameter. Based on the sensitivity analysis results, we finally construct a personalized model for each subject and provide personalized biomarker predictions for subjects who have more than four longitudinal biomarker data points.

### A population model

We construct a causal model by fitting biomarkers of all subjects across the ADNI dataset. Since the causal model is a dynamic system expressed as ordinary differential equations (ODEs), we require at least two longitudinal data points for each subject. More specifically, we remove patients who do not provide at least two measurements for any one of the four biomarkers. The histograms in Fig. 1(a) summarize the available biomarker data in the ADNI dataset.



**Fig. 2 Subject biomarkers in ADNI data and the fitted sigmoid function.** Top left is amyloid-beta; top right is the tau; bottom left is the normalized hippocampal volume; bottom right is the cognitive subscale. X-axes are fitted DPS of biomarkers, and y-axes are the corresponding magnitudes of each biomarker. The subjects are classified into “CN”, “LMCI”, and “AD” according to ADNI diagnostic groups, where orange is “CN”, green is “LMCI”, and blue is “AD”. The black solid lines are fitted with sigmoid functions<sup>28</sup>.

**Algorithm 1.** Population model calibration algorithm to compute the population parameters  $w^{(1)}$  and DPS parameters  $(\alpha, \beta)$ . See details in Methods section.

```

Input  $y = \{y_{ijk}\}_{ijk}, t = \{t_{ij}\}_{ij}$ .
Initialize  $\alpha^0, \beta^0$ , and  $w^0$ .
1: for  $l=1$  to  $L$  do
2:   for  $k \in \{A, T, N, C\}$  do  $\triangleright$  Population parameter calibration
3:      $w_k^l = \operatorname{argmin}_{w_k} \sum_{(i,j) \in \mathcal{I}_k} (y_{ijk} - f_k(\alpha_i^l t_{ij} + \beta_j^l; w_k^{l-1}))^2$ .
4:      $\sigma_k^l = \frac{1}{|\mathcal{I}_k - 2l - 4|} \sum_{(i,j) \in \mathcal{I}_k} (y_{ijk} - f_k(\alpha_i^l t_{ij} + \beta_j^l; w_k^{l-1}))^2$ .
5:   end for
6:
7:   for  $i=1$  to  $l$  do  $\triangleright$  Update DPS parameters
8:      $(\alpha_i^l, \beta_i^l) = \operatorname{argmin}_{\alpha_i, \beta_i} \sum_{(j,k) \in \mathcal{I}_i} \frac{1}{\sigma_k^l} (y_{ijk} - f_k(\alpha_i^l t_{ij} + \beta_j^l; w_k^{l-1}))^2$ .
9:   end for
10: end for
Output  $w^l$  as the population parameter  $w^{(1)}$ ,  $\alpha^l, \beta^l$ .
    
```

By using Algorithm 1, the initial value of  $\alpha_i$  is randomly chosen in  $(0, 4)$  and the initial value of  $\beta_j$  satisfies  $-10 \leq s_j(t) \leq 20$  on all the measurement. Then we obtain the population model in terms of the fitted DPS shown in Fig. 1(b). The population model (black solid) is learned on three different disease stages, namely, CN (orange circle), LMCI (green circle), and AD (blue hexagons). The gray area is the confidence interval of the population model. More specifically, we sample the population parameters,  $w^{(1)}$ , from the posterior distribution (given by the simulation study) and run the model with the same initial condition 1000 times. Then the 95% confidence interval at every time point is plotted. The simulation

study and diagnostic plots corresponding to the population model are shown in the Supplementary Materials. From this figure, we can separate biomarkers into three stages according to the population model. In particular, CN and AD patients correspond to  $s < 0$  and  $s > 0$ , respectively while LMCI patients locate around  $s = 0$ . Moreover, the first three biomarkers ( $A_\beta, \tau$ , and  $N$ ) start at steady-states when  $s < 0$  (CN), change gradually when  $s = 0$  (LMCI), and finally approach another steady-state (AD). Different from other biomarkers, ADAS continues to grow which means that cognitive symptoms get worse as AD progresses.

We also compare the population model with the sigmoid function fitting (black solid in Fig. 2). First, the population model provides relatively smooth transitions from one stage to another while the sigmoid function fitting gives more abrupt changes for  $A_\beta$  at  $s \approx 4$ . Second, the population model follows the biomarker cascade theory which is that  $\tau$  rises after  $A_\beta$  starts decreasing,  $N$  increases after  $\tau$ , and  $C$  rises after  $N$ . However, the sigmoid function fitting makes  $A_\beta$  and  $\tau$  change after  $s = 0$ , while  $N$  and  $C$  change at  $s \approx -3$  and  $s \approx -10$ .

Since the CN group might not follow the same disease trajectory, we next derive a population model based on the LMCI and AD groups only. In order to better identify the biomarker dynamics among LMCI and AD group patients, we fix the parameters of DPS,  $(\alpha, \beta)$ , that we obtained before and only update the causal model parameters,  $w$ , by using Algorithm 1. Figure 1(c) shows the population model based on LMCI and AD groups.

The transitions for different biomarkers shown in Fig. 1(c) are similar to Fig. 1(b). But Fig. 1(c) advances the onset of changes since the LMCI and AD groups are prone to suffer from cognitive

**Table 1.** Population parameters  $w^{(1)}$  of the calibrated causal models based on the ADNI dataset.

Biomarkers	Parameters	Included subjects		
		CN, LMCI, AD	LMCI, AD	
$A_\beta$	$w_{A0}$	0	0	
	$w_{A1}$	0.917	0.745	
	$w_{A2}$	-0.873	-0.749	
	$\tau$	$w_{\tau0}$	0	0
		$w_{\tau1}$	0.788	0.689
$w_{\tau2}$		-0.246	-0.679	
$w_{\tau3}$		0.002	0.000	
$w_{\tau4}$		3.066	0.185	
$N$	$w_{N5}$	-3.650	-0.101	
	$w_{N0}$	0	0	
	$w_{N1}$	1.627	0.899	
	$w_{N2}$	-1.253	-0.927	
	$w_{N3}$	0.018	0.554	
$C$	$w_{N4}$	2.342	1.792	
	$w_{N5}$	-4.015	-2.127	
	$w_{C0}$	0	0	
	$w_{C1}$	0.159	0.134	
	$w_{C2}$	0.202	-0.067	
	$w_{C3}$	0.010	0.004	
	$w_{C4}$	0.019	0.007	
	$w_{C5}$	-0.176	-0.008	
Initial conditions	$y_0$	6.35e-6	1.41e-4	

Superscripts of parameters are omitted from the table. See details in Methods section.

decline earlier. We summarize the parameters of the above-mentioned causal models in Table 1, which corresponds to the results given in Fig. 1(b) & (c).

### Sensitivity analysis

The quasi-Monte Carlo method is applied to compute sensitivity indices. For more details about Sobol sensitivity analysis, please refer to<sup>29,30</sup>. By taking C(0) as the output, Fig. 3(a) list the top nine most sensitive parameters for the first-order effects and total order sensitivity index. We see that the weight with greater first-order impact ( $S_m > 0.4$ ) is associated with  $A_\beta(\cdot)$ .

The right figure in 3(a) shows the second-order interaction between two parameters. We see that the parameters associated with  $A_\beta^2$  are always positively related to other terms. While  $A_\beta$  with parameter  $w_{A1}$  is almost positively related with other terms, the  $A_\beta$  term with parameter  $w_{T3}$  are negatively related with other parameters except  $A_\beta^2$ . But compared to their first-order sensitivity contribution, the second-order ones contribute slightly.

Figure 3(b, c) shows the dynamics of sensitivities with respect to DPS. From the figures, we can see the first-order sensitivity value of  $w_{A1}$  drops down over DPS which implies that the effect of  $A_\beta$  on cognitive decline switches from linear to nonlinear in later-stage disease. At the same time, the first-order sensitivity values of some other parameters increase gradually, with a notable increase of  $w_{C3}$  and  $w_{C5}$ . The second-order sensitivities between different parameters eventually converge to zero thus the interactions among different parameters become less as the biomarkers reach equilibrium. Based on the results shown in Fig. 3(a), we select

$w_{A1}, w_{A2}, w_{T4}, w_{T5}, w_{N4}, w_{N5}, w_{C3}$ , and  $w_{C5}$  as the most sensitive parameters for personalization by setting the threshold, Tol, as 0.01 in Algorithm 2.

### Personalized model and biomarker prediction

**Algorithm 2.** Personalized model calibration algorithm. The personalized parameters are initialized by the population model. The personalized models are applied for subjects who meet the requirement denoted as  $i \in \Omega$ .

**Input** longitudinal biomarker data  $\{y_{ijk}\}$  at  $\{t_{ij}\}$  with  $i \in \Omega$ ;  
**Input** the DPS parameter values  $(a_i, \beta_i)$  for each subject  $i \in \Omega$ ;  
**Input** the population parameter values  $w^{(1)}$  ( $w$  for simplicity);  
**Input** sensitivity threshold, TOL.

```

1: for  $m=1$  to 21 do ▷ First order sensitivity.
2:    $S_m(z) = \frac{\text{Var}_{w_m}[\mathbb{E}_{w_{-m}}(z|w_m)]}{\text{Var}(z)}$ .
3:   if  $S_m(z) \geq \text{TOL}$  then
4:     set  $w_m$  as a personalized parameter and denote as  $w_m^{(2)}$  else
5:     keep  $w_m$  as a population parameter.
6:   end if
7: end for
8:
9: for  $i=1$  to  $|\Omega|$  do ▷ Personalized model calibration.
10:  for  $k \in \{A, T, N, C\}$  do
11:    Denote the personalized parameters in  $k$ -th equation as  $w_k^{(2)}$ .
12:    ▷ Select parameters to calibrate.
13:     $w_k^{(2)} = \arg \min_{w_k^{(2)}} \sum_{j=1}^{M-1} (\hat{y}_{ijk} - f_k(a_i t_{ij} + \beta_i; w_k^{(2)}))^2$ .
14:     $PA_{ik} = \frac{\hat{y}_{imk} - f_k(a_i t_{(M)} + \beta_i; w_k^{(2)})}{\hat{y}_{imk}} \times 100\%$ .
15:    ▷ Compute prediction accuracy.
16:  end for
17: end for
Output  $PA_{ik}$  for  $i \in \Omega$  and  $k \in \{A, T, N, C\}$ .

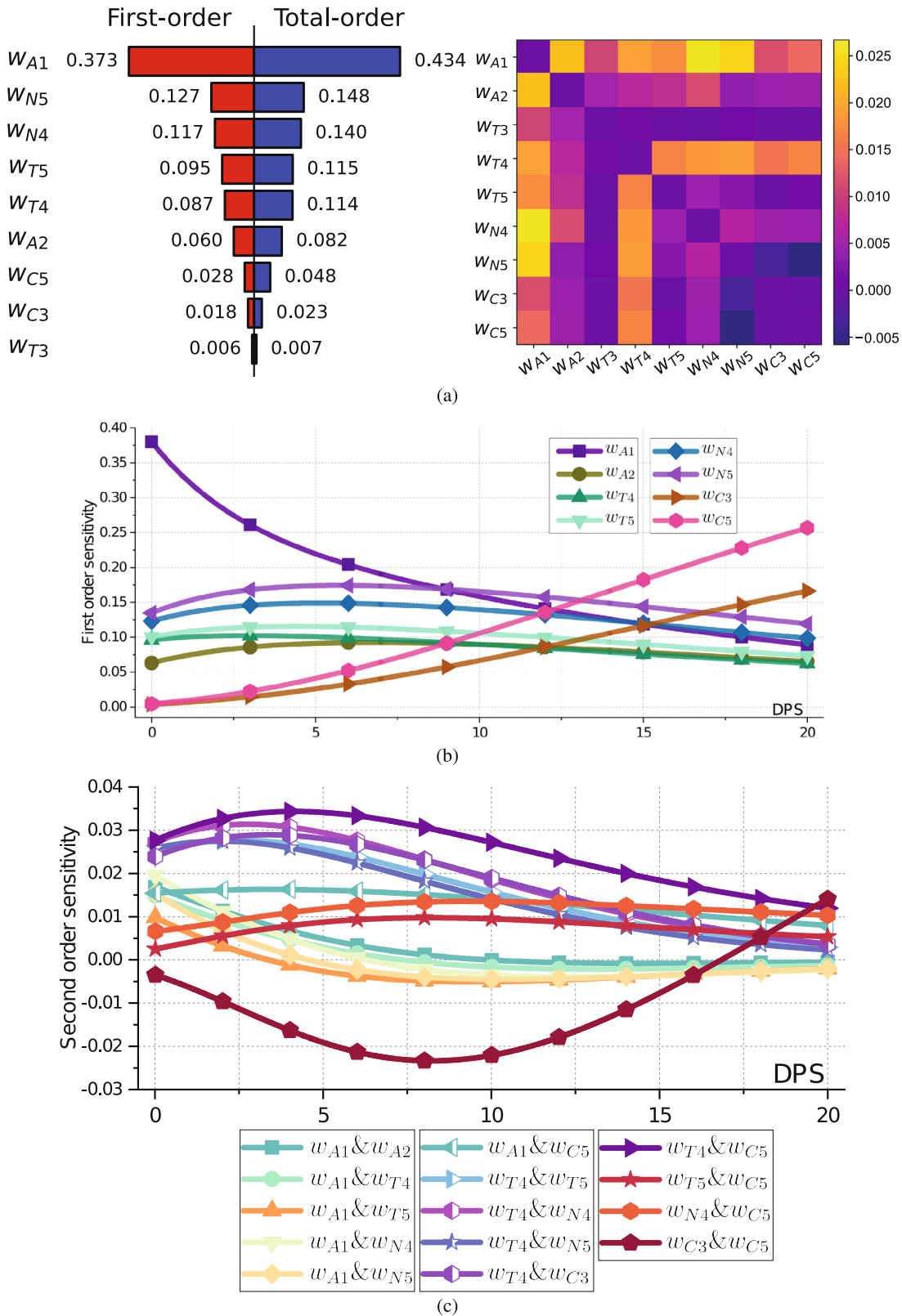
```

Next, we build personalized models and provide biomarker prediction for subjects whose data satisfies the following two criteria: (1) There are at least four measurements for each biomarker; (2) Each biomarker measurement changes monotonically with respect to DPS. Based on the first-order sensitivity analysis results shown in Fig. 3(a), we chose the eight most sensitive parameters as personalized parameters by choosing  $\text{TOL} = 0.01$  in Algorithm 2. For each subject, we denote the biomarker data as  $\hat{y}(s_i) = [\hat{A}_\beta(s_i) \hat{\tau}(s_i) \hat{N}(s_i) \hat{C}(s_i)]^T$  ( $i = 1, \dots, M$ ), fit the sensitive personalized parameters of the population model  $w^{(1)}$  by using the first  $M - 1$  data points, and test the prediction accuracy on the last data point by  $\frac{\hat{y}(s_M) - \mathcal{Y}(s_M)}{\hat{y}(s_M)} \times 100\%$ . A detailed procedure is outlined in Algorithm 2.

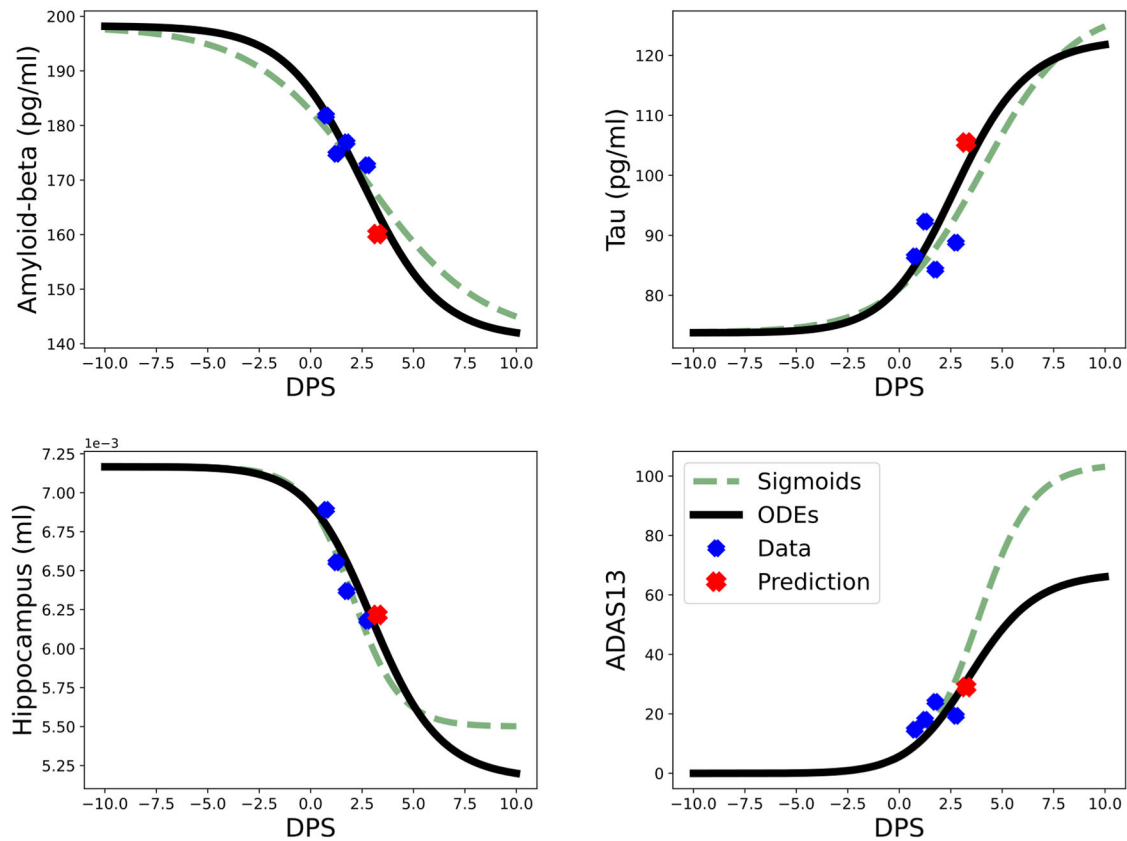
Figure 4 shows the biomarker trajectories of the personalized model by training (blue) and testing (red) data for one subject (pseudo ID = 18). We also compare the personalized model with the sigmoid function fitting, the personalized model provides a better prediction accuracy. In fact, the prediction accuracies given by the personalized model are 97.3% ( $A_\beta$ ), 95.9% ( $\tau$ ), 98.4% ( $N$ ), and 95.1% ( $C$ ), respectively while the ones given by the sigmoid function fitting are 95.5% ( $A_\beta$ ), 90.8% ( $\tau$ ), 95.7% ( $N$ ), and 63.4% ( $C$ ), respectively. Since the sigmoid function fitting predicts by using the longitudinal information of the current biomarker only, it provides a less accurate cognitive score.

Furthermore, we build personalized models for the CN and LMCI groups (there are not enough data points in the AD group) with different numbers of longitudinal data points and summarize the predictive results in Tables 2–3. The tables indicate that our personalized models can provide high predictive accuracy compared to the sigmoid function fitting. Moreover, the accuracy





**Fig. 3 The Sobol sensitivity analysis plots with respect to cognitive decline.** **a** First-order, second-order, and total-order Sobol sensitivities of  $C(0)$ . Left: the red rectangles are assigned first-order sensitivities of model parameters, and the blue rectangles are their total-order sensitivities. The length of the rectangles represents the attribution of sensitivities to outputs; Right: each square represents the second-order sensitivity correlations of two model parameters. The lighter the color, the stronger the positive correlation while the darker the color, the stronger the negative correlation. **b** The dynamics of first-order Sobol sensitivities with respect to DPS. Each curve corresponds to the first-order sensitivity values with an output  $C(s)$ . Only first-order sensitivity values greater than 0.01 are plotted. **c** The dynamics of second-order Sobol sensitivities for two parameters with respect to DPS. Only the maximum absolute second-order sensitivity values greater than 0.01 are plotted.



**Fig. 4** The personalized model for one LMCI subject with pseudo ID = 18. The green dashed lines are fitted by sigmoid functions, and black solid lines are the solutions of the personalized model. The blue markers are training data points while the red markers are for testing.

of predicting biomarker dynamics increases as the number of biomarkers data points increases.

## DISCUSSION

Different from the existing pathophysiological AD network which is based on a priori assumptions about biomarker trajectories, this work develops a data-driven causal modeling approach informed by AD clinical biomarker data and demonstrates both population and personalized models. The proposed population model traces the general biomarker dynamics for all patient data without any specific assumptions regarding the form of the model and enables personalized AD risk prediction via incorporating historical clinical data such as CSF protein and imaging biomarkers as well as cognitive scores. By introducing a DPS for each subject, we calibrate and scale AD biomarker progression across the ADNI population and derive population parameters. We also compare the proposed data-driven modeling approach to an empirical fitting approach with a sigmoid function fitting and conclude that the proposed causal model is able to better capture disease progression with a smoother transition over time. Moreover, this causal model allows us to explore the underlying cascade relationship among biomarkers, while the empirical sigmoid function approach considers each biomarker as an independent term. The population model not only provides a means to classify different stages of AD progression for each biomarker, but also lays the foundation for personalized modeling.

Before constructing the personalized model, we performed a sensitivity analysis for the population parameters. From a clinical standpoint, the sensitivity analysis provides insights on AD progression in terms of which parameters play the greatest role

in disease progression, and when during the disease course they are most relevant. From a computational standpoint, the sensitivity analysis aids the subsequent personalized parameter selection. Based on the sensitivity analysis, we see that change in cognition is driven primarily by first-order effects and is time-dependent. Initially, the greatest effects are by amyloid, represented by  $w_{A1}$ , and to a lesser extent tau and neuronal vulnerability to tau, represented by  $w_{N4}$  and  $w_{N5}$ , respectively. The amyloid parameter  $w_{A1}$  is most sensitive when the disease starts ( $DPS = 0$ ) and the sensitivity diminishes as DPS increases. On the other hand, the sensitivity of parameters related to  $N$  and  $C$ , namely  $w_{C3}$  and  $w_{C5}$ , increase significantly as the disease progresses. Thus, the sensitivity analysis suggests that at the early stage of AD cognitive decline is driven by  $A\beta$  levels and sensitivity decreases linearly as the disease progresses. Whereas at the later stages, cognitive decline is driven mainly by downstream effects including the level of neuronal degeneration, represented by  $w_{C3}$ , and the interaction of cognition and neuronal degeneration, represented by  $w_{C5}$ . These results are consistent with prior observational studies based on ADNI and other longitudinal cohorts, which suggest that cognitive decline is driven primarily by high amyloid levels at earlier disease stages and by neurodegeneration at later stages<sup>31</sup>.

Sensitivity analysis also provides key insights in terms of personalized parameter selection. The paucity of longitudinal biomarker data and the relatively larger number of model parameters can easily lead to overfitting for personalized models. Based on the sensitivity analysis results, we chose the eight most sensitive population parameters as personalized parameters and set the rest of the parameters at the mean population parameter values. In this case, calibration of personalized parameters based

**Table 2.** The prediction accuracy summary for CN subjects using different numbers of longitudinal biomarker datapoints (*n*).

PseudolDs ( <i>n</i> )	DPS Diff	Model	Accuracy			
			CSF Abeta42	CSF tTau	HIPPv	ADAS13
1 (4)	0.13	ODE	98.3%	93.6%	99.4%	92.6%
		Sigmoid	74.0%	79.8%	70.5%	84.8%
2 (4)	3.00	ODE	99.8%	93.2%	98.7%	93.0%
		Sigmoid	93.9%	61.5%	90.7%	80.4%
3 (5)	0.52	ODE	86.6%	98.8%	95.9%	85.3%
		Sigmoid	90.3%	82.6%	71.1%	56.9%
4 (5)	0.59	ODE	98.8%	96.1%	88.3%	96.6%
		Sigmoid	76.8%	76.9%	86.1%	66.7%
5 (5)	0.39	ODE	97.8%	90.0%	99.7%	94.8%
		Sigmoid	84.3%	79.5%	79.9%	81.9%
6 (4)	0.46	ODE	96.3%	93.6%	90.9%	92.7%
		Sigmoid	75.4%	91.1%	91.2%	84.0%
7 (4)	0.55	ODE	99.8%	88.2%	98.7%	90.3%
		Sigmoid	96.5%	86.0%	92.0%	72.3%
8 (4)	0.63	ODE	95.9%	98.9%	92.0%	92.6%
		Sigmoid	85.8%	86.8%	91.7%	96.6%
9 (4)	0.71	ODE	99.6%	96.1%	97.1%	87.5%
		Sigmoid	89.4%	80.3%	79.2%	69.5%
10 (5)	1.04	ODE	83.4%	81.2%	98.7%	85.5%
		Sigmoid	88.3%	78.4%	74.4%	80.1%
11 (6)	1.04	ODE	98.2%	99.8%	86.5%	85.1%
		Sigmoid	75.7%	76.3%	67.6%	72.6 %
12 (4)	0.40	ODE	94.6%	91.3%	96.5%	91.7%
		Sigmoid	89.7%	81.5%	88.9%	75.1%
13 (6)	0.88	ODE	97.0%	92.8%	96.1%	98.8%
		Sigmoid	97.4%	85.4%	85.3%	84.3%
14 (4)	0.75	ODE	98.4%	99.1%	99.1%	87.1%
		Sigmoid	90.9%	79.7%	88.6%	79.8%
15 (4)	0.55	ODE	99.6%	96.8%	90.9%	81.5%
		Sigmoid	93.8%	95.1%	81.5%	59.2%
Average	0.78 ± 0.64	ODE	96.3% ± 4.9%	94.0% ± 5.0%	95.2% ± 4.4%	90.3% ± 4.8%
		Sigmoid	86.8% ± 7.6%	81.4% ± 7.4%	82.6% ± 8.2%	76.3% ± 10.1%

The last four columns list the prediction accuracy of both the personalized model and the sigmoid function fitting for each biomarker. The last four rows summarize the mean and the standard deviation of prediction accuracy.

on sparse longitudinal biomarker data for each patient avoids the overfitting issue and provides a high-precision personalized prediction for each subject, as outlined in Results section.

Limitations of this work include sampling bias. Because the ADNI dataset is a research cohort from academic clinics, only one-third of ADNI subjects agreed to provide CSF biomarkers. Thus we need to replicate these findings using data from more general practice settings in the future. Despite these limitations, this model advances our understanding of the complexity of AD biomarker pathophysiology over that of current biomarker models which have primarily been independent and ad hoc in nature, with inherent assumptions regarding the shape of individual biomarker trajectories. Our current approach is integrative and based on the cascade mechanism, yet without assumptions regarding the exact mathematical form of the individual biomarker models or the resulting shape of the biomarker trajectories. In the future, we intend to extend the current approach to the spatiotemporal domain by utilizing longitudinal

imaging data to determine mechanisms driving the spread of pathology in time and space.

## METHODS

We propose a pathophysiology and data-driven modeling approach to construct a causal model of AD clinical biomarkers. We construct a causal model from the serial clinical biomarker measures across 819 subjects from the ADNI-1 datasets with mild AD ( $N = 192$ ), late mild cognitive impairment (LMCI,  $N = 398$ ), and normal cognition ( $N = 229$ ) (more details are shown in Table 4). We use PseudolDs instead of RIDs to link across all clinical biomarker data belonging to a patient. The CSF proteins measured in ADNI are the following A-Beta 42 and Phosphorylated tau 181 (p-tau 181)<sup>32,33</sup>. These measures were obtained through serial spinal taps on subjects over approximately two-year intervals. Of note, A-Beta in the CSF goes down, and total and phosphorylated tau go up as the disease progresses. Hippocampal volume, a measure of neurodegeneration, was measured through volumetric analysis of serial MRI images obtained at approximately one-year intervals. It goes down as the disease progresses. Finally, cognitive decline was measured through a pencil-and-paper neuropsychological test,

**Table 3.** The prediction accuracy summary for LMCI subjects using different longitudinal data points (shown in the first column).

PseudoIDs (n)	DPS Diff	Model	Accuracy			
			CSF Abeta42	CSF tTau	HIPPv	ADAS13
16 (5)	0.70	ODE	99.0%	87.8%	99.1%	88.2%
		Sigmoid	91.2%	87.4%	74.5%	70.1%
17 (6)	0.51	ODE	97.8%	86.2%	92.2%	90.7%
		Sigmoid	94.1%	83.3%	81.6%	88.7%
18 (5)	0.50	ODE	97.3%	95.9%	98.4%	95.1%
		Sigmoid	95.5%	90.8%	95.7%	63.4%
19 (4)	0.55	ODE	88.9%	96.3%	82.8%	96.6%
		Sigmoid	79.1%	87.7%	94.3%	85.1%
20 (5)	0.30	ODE	96.8%	96.7%	97.0%	84.9%
		Sigmoid	80.3%	89.1%	80.1%	82.1%
21 (6)	0.75	ODE	95.3%	85.8%	98.3%	84.4%
		Sigmoid	86.3%	72.9%	90.5%	79.8%
22 (4)	0.82	ODE	97.0%	95.7%	89.8%	84.6%
		Sigmoid	72.5%	68.5%	96.6%	81.6%
23 (4)	0.43	ODE	99.2%	91.0%	94.2%	86.0%
		Sigmoid	88.8%	61.3%	80.8%	62.1%
24 (6)	0.49	ODE	98.2%	98.9%	99.9%	96.9%
		Sigmoid	83.5%	89.9%	88.8%	81.1%
25 (6)	0.70	ODE	97.0%	94.4%	93.3%	89.3%
		Sigmoid	93.2%	80.2%	74.1%	70.7%
26 (4)	0.98	ODE	97.7%	93.5%	96.7%	93.8%
		Sigmoid	85.2%	89.1%	91.8%	81.3%
27 (4)	0.38	ODE	96.4%	97.4%	94.4%	96.3%
		Sigmoid	97.9%	79.5%	92.1%	72.6%
28 (5)	0.52	ODE	97.0%	90.6%	87.7%	96.6%
		Sigmoid	85.1%	78.4%	84.6%	83.1%
29 (4)	0.50	ODE	94.2%	91.4%	91.3%	95.0%
		Sigmoid	76.3%	82.7%	75.1%	72.1%
30 (4)	0.57	ODE	94.7%	90.0%	93.6%	83.1%
		Sigmoid	73.1%	80.6%	86.3%	85.8%
Average	0.58 ± 0.17	ODE	97.1% ± 1.5%	92.3% ± 4.1%	94.8% ± 3.6%	89.8% ± 5.4%
		Sigmoid	85.5% ± 7.8%	81.4% ± 8.2%	85.8% ± 7.5%	77.3% ± 7.9%

The last four columns list the prediction accuracy of both the personalized model and the sigmoid function fitting for each biomarker. The last four rows summarize the mean and the standard deviation of prediction accuracy.

the thirteen-item Alzheimer's Disease Cognitive Assessment Scale (ADAS13). This measures function in several cognitive domains affected by AD, including memory, language, and praxis and is the de facto primary outcome measure in AD clinical trials. It goes up as the disease progresses.

After constructing and calibrating the population model with data across all ADNI subjects, we then personalize the parameters of the model using each patient's longitudinal data to provide a personalized prediction of biomarker trajectories. The overall modeling approach is outlined in Fig. 5, and each step is elaborated in the following subsections.

#### The data-driven causal model learning via ADNI dataset

Four AD biomarkers are key factors in AD diagnosis and monitoring of AD progression, and include amyloid-beta  $A_\beta$ , tau  $\tau$ , neuronal degeneration  $N$ , and cognitive decline  $C$ . Amyloid-beta is the main component of amyloid plaques and is considered to be an early event of the pathological cascade of AD. Amyloid production leads to downstream Tau phosphorylation causing the formation of neurofibrillary tangles and neuropil threads. Tau is a microtubule-associated protein, which is very common in

neurons of the central nervous system. Both amyloid-beta and tau phosphorylation contribute to neuronal degeneration and cognitive decline.

To describe the cascade relationship among the above-mentioned four biomarkers of AD progression, we consider a canonical system of ODEs to describe their relations. The amyloid-dependent cascade is initiated by amyloid-beta pathology  $A_\beta$ , and mediated via tau  $\tau$ . Neuron degeneration  $N$  starts with the rise of tau  $\tau$ , and in turn, leads to the initiation of cognitive decline  $C$ . According to the above description, we consider the causal model as the system of ODEs:

$$\begin{cases} \frac{dA_\beta}{dt} = \sum_{\ell=0}^m w_{1,\ell} \phi_\ell(A_\beta); \frac{d\tau}{dt} = \sum_{\|\leq m} w_2 \psi(A_\beta, \tau); \\ \frac{dN}{dt} = \sum_{\|\leq m} w_3 \psi(\tau, N); \frac{dC}{dt} = \sum_{\|\leq m} w_4 \psi(N, C), \end{cases} \quad (1)$$

where  $\mathcal{L} = (\ell_1, \ell_2)$ ,  $|\mathcal{L}| = |\ell_1| + |\ell_2|$ , and  $m$  is the degree of the model. We choose the polynomial basis function in the initialized ODE model, namely,

$$\phi_\ell(x) = x^\ell, \quad \psi(x, y) = x^{\ell_1} y^{\ell_2}. \quad (2)$$



**Table 4.** The characteristics of the ADNI-1 dataset in data-driven modeling.

	CN	LMCI	AD
N	229	398	192
Age	75.72 ± 4.86	74.52 ± 7.22	75.29 ± 7.41
M/F	119/110	257/141	101/91
MMSE	28.82 ± 1.78	25.54 ± 4.16	21.52 ± 4.59
ADAS-13	10.29 ± 6.44	21.71 ± 10.99	32.49 ± 10.42
CSF Ab42 (pg/ml)	201.74 ± 55.15	159.37 ± 51.53	139.79 ± 35.87
CSF Total Tau (pg/ml)	72.69 ± 31.69	104.65 ± 58.28	122.01 ± 58.30
CSF Ptau (pg/ml)	29.57 ± 16.10	38.96 ± 21.09	43.91 ± 20.97
Hipp Volume (ml)	7045.38 ± 971.27	6163.12 ± 1179.50	5488.95 ± 1132.57

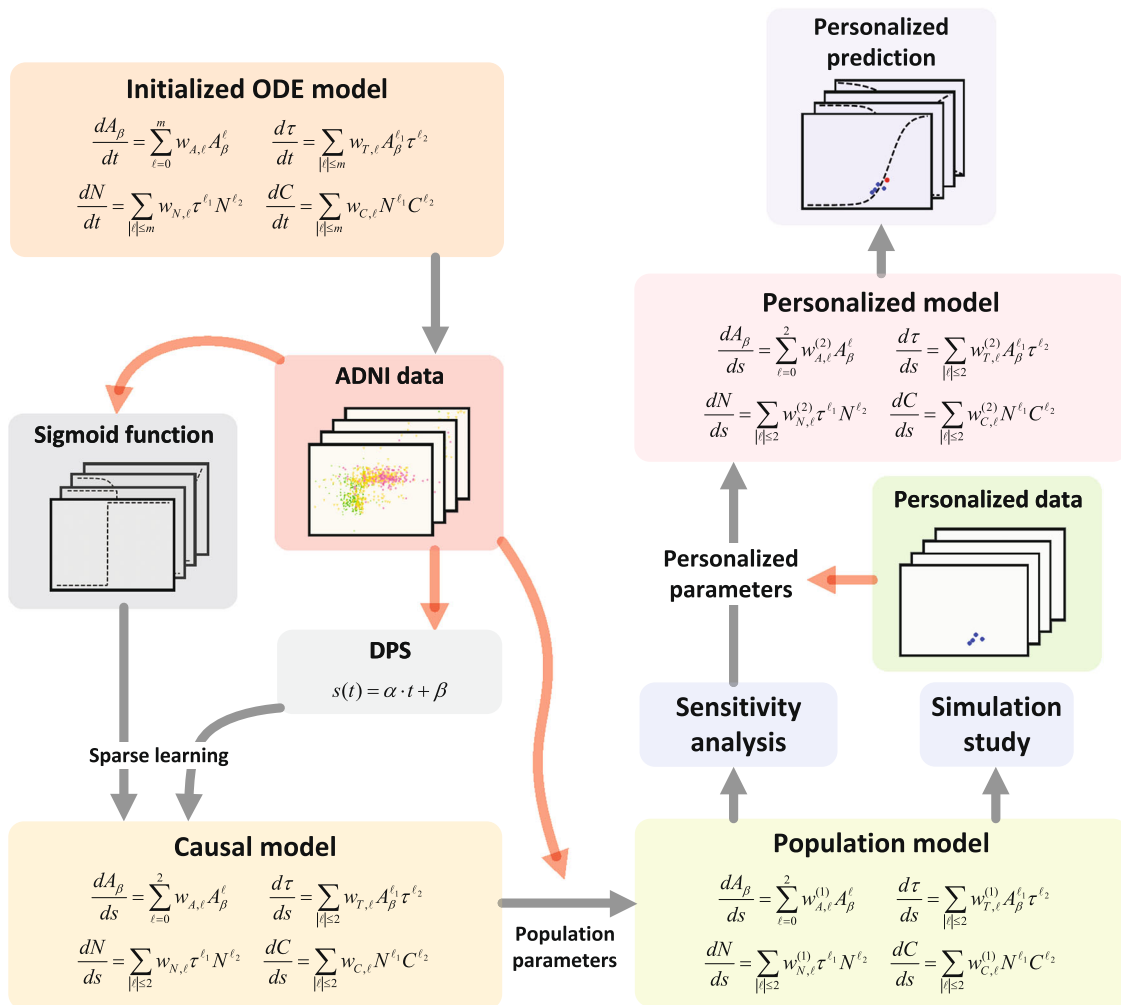
We then learn the causal model parameters in (1) by using ADNI data. More specifically, we use CSF amyloid-beta 1-42 ( $A_\beta$ ), CSF total tau ( $\tau$ ), the ratio of hippocampal volume to whole-brain volume on MRI ( $N$ ), and the Alzheimer's Disease Assessment Scale-cognitive ( $C$ ) to calibrate  $A_\beta$ ,  $\tau$ ,  $N$ , and  $C$ , respectively in the causal model. In order to denoise longitudinal data for different subjects, we applied a sigmoid interpolation for each biomarker. Moreover, because AD has a different time of onset and rate of progression for different subjects, we employ DPS<sup>28</sup> to unify the time scale across subjects in the causal model.

**Disease progression scores**

For different subjects in ADNI, the onset of disease and rate of progression are different within and among subject classes of CN, LMCI and AD. To fit the causal model for all subjects in the ADNI-1 study, we standardize the longitudinal measurement among patients by employing the DPS<sup>28</sup>. In particular, we define DPS  $s_i(t)$  as a linear function of the patient's age  $t$  for each patient:

$$s_i(t) = \alpha_i \cdot t + \beta_i, \tag{3}$$

where  $i = 1, 2, \dots, l$  is the patient index,  $\alpha_i$  is the rate of AD progression, and  $\beta_i$  is the age of AD onset.



**Fig. 5** A flowchart of the pathophysiology and data-driven modeling approach. Given the initialized ODE model, a causal model is obtained by fitting the ADNI dataset and DPS model through sparse learning; secondly, the ADNI dataset is used to calibrate the population parameters in the causal model and obtain the population parameters; thirdly, a sensitivity analysis is applied to analyze the sensitivity of each population parameters and determine the sensitive personalized parameters, and a simulate study is conducted to validate the population model. Then, the personalized model is obtained by calibrating the sensitive personalized parameters with the use of personalized data. A prediction is made by the personalized model in the end.

### The sigmoid function fitting

We fit each biomarker data in ADNI to a sigmoid function. Specifically, each biomarker is parameterized by four parameters  $\theta_k = [a_k, b_k, c_k, d_k]^T$ :

$$g_k(s; \theta_k) = a_k(1 + e^{-b_k(s-c_k)})^{-1} + d_k, \quad (4)$$

where  $a_k$  is a magnitude scale of the function,  $b_k$  is a slope coefficient, and  $c_k$  and  $d_k$  determine function positions. Here we take  $g_1(s) = A_\beta(s)$ ,  $g_2(s) = \tau(s)$ ,  $g_3(s) = N(s)$ ,  $g_4(s) = C(s)$  and denote  $\mathbf{g} = (g_1, g_2, g_3, g_4)^T$ .

Next, we apply the sparse learning to reveal the causal model in (1) which is re-written as

$$\frac{d\mathbf{x}}{ds} = \sum_{|\ell| \leq m} \phi_\ell(\mathbf{x}) w_\ell, \text{ where } \mathbf{x} = (A_\beta, \tau, N, C)^T \in \mathbb{R}^4.$$

By taking uniform grid points  $\{s_i\}_{i=1}^M$  on  $s \in [-10, 20]$ , we denote

$$D_i = [\phi_{\ell_1}(\mathbf{g}(s_i)), \dots, \phi_{\ell_n}(\mathbf{g}(s_i))] \text{ and } b_i = \frac{d(\mathbf{g}(s_i))}{ds},$$

where  $\ell_1, \dots, \ell_n$  are in the set of  $|\ell| \leq m$ . By expanding

$$D = \begin{pmatrix} D_1 \\ \vdots \\ D_M \end{pmatrix} \text{ and } b = \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix},$$

we learn the causal model via the following Lasso regression, namely,

$$\min_w \|Dw - b\|_2^2 + \lambda \|w\|_1, \quad (5)$$

where  $\|w\|_1$  enforces the sparsity.

Here we keep the polynomial degrees among all the variables in the causal model be consistent and choose  $m=4$  with  $\lambda=10^{-7}$  in (5). By performing Lasso, we find the result is consistent with the causal model when  $m=2$  but different from the one with  $m=1$ , which indicates the optimal choice of the causal model is  $m=2$ . Then the general causal model of ODEs describing the progression of AD biomarkers is summarized below (All rights to the in-silico model belong to the authors and it cannot be used for any commercial purpose without permission):

$$\begin{cases} \frac{dA_\beta}{ds} = w_{A0} + w_{A1}A_\beta + w_{A2}A_\beta^2; \\ \frac{d\tau}{ds} = w_{T0} + w_{T1}\tau + w_{T2}\tau^2 + w_{T3}A_\beta + w_{T4}A_\beta^2 + w_{T5}A_\beta\tau; \\ \frac{dN}{ds} = w_{N0} + w_{N1}N + w_{N2}N^2 + w_{N3}\tau + w_{N4}\tau^2 + w_{N5}\tau N; \\ \frac{dC}{ds} = w_{C0} + w_{C1}C + w_{C2}C^2 + w_{C3}N + w_{C4}N^2 + w_{C5}NC, \end{cases} \quad (6)$$

with an initial condition  $A_\beta(-10) = y_0$  and  $\tau(-10) = N(-10) = C(-10) = 0$ , where  $y_0$  is also a parameter that we consider as a small positive value to initiate the cascade.

### Population model calibration

First, we calibrate the learned causal model by using the ADNI dataset and rewrite (6) as the following population model

$$\begin{cases} \frac{dA_\beta}{ds} = \sum_{\ell=0}^2 w_{A,\ell}^{(1)} A_\beta^\ell, \frac{d\tau}{ds} = \sum_{|\ell| \leq 2} w_{T,\ell}^{(1)} A_\beta^\ell \tau^{\ell_2}, \\ \frac{dN}{ds} = \sum_{|\ell| \leq 2} w_{N,\ell}^{(1)} \tau^{\ell_1} N^{\ell_2}, \frac{dC}{ds} = \sum_{|\ell| \leq 2} w_{C,\ell}^{(1)} N^{\ell_1} C^{\ell_2}, \end{cases} \quad (7)$$

where  $\mathbf{w} = \{w_{A,\ell}^{(1)}, w_{T,\ell}^{(1)}, w_{N,\ell}^{(1)}, w_{C,\ell}^{(1)}\}$  denote the population parameters. We also denote  $f_1(s) = A_\beta(s)$ ,  $f_2(s) = \tau(s)$ ,  $f_3(s) = N(s)$ , and  $f_4(s) = C(s)$  with the initial conditions  $f_1(-10) = y_0$ ,  $f_2(-10) = f_3(-10) = f_4(-10) = 0$ . Then the population parameters are calibrated based on the ADNI dataset by minimizing the sum of squared differences between the data and the solution of the causal model, namely

$$\min_{\mathbf{w}_k} \sum_{(i,j) \in \mathcal{I}_k} (y_{ijk} - f_k(a_i t_{ij} + \beta_j; \mathbf{w}_k))^2, \quad (i, j, k) \in \mathcal{I} \quad (8)$$

where  $y_{ijk}$  is the  $k$ -th biomarker data for  $i$ -th patient at  $j$ -th visit and  $\mathcal{I}_k$  is the set of  $(i, j)$  for  $k$ -th biomarker.

Since the biomarkers for each patient will generally increase or decrease monotonically, we consider fitting DPS as a least square linear regression problem, namely,

$$\min_{\alpha_i, \beta_i} \sum_{(j,k) \in \mathcal{I}_i} \frac{1}{\sigma_k} (y_{ijk} - f_k(a_i t_{ij} + \beta_j; \mathbf{w}_k))^2, \quad (9)$$

where  $\mathcal{I}_i$  is set of  $(j, k)$  for  $i$ -th patient and  $\sigma_k$  is the sum of squared error with respect to biomarker  $k$ , namely,

$$\sigma_k = \frac{1}{|\mathcal{I}_k - 2I - 4|} \sum_{(i,j) \in \mathcal{I}_k} (y_{ijk} - f_k(a_i t_{ij} + \beta_j; \mathbf{w}_k))^2. \quad (10)$$

The detailed procedure to fit the parameters is shown in Algorithm 1. The optimization solver employs the Levenberg-Marquardt method<sup>34</sup>, which can avoid getting stuck in a local minimum.

### Sensitivity analysis

We assume that the parameters in the population model,  $\mathbf{w}^{(1)} = [w_{A0}^{(1)}, w_{A1}^{(1)}, \dots, w_m^{(1)}, \dots, w_{C4}^{(1)}, w_{C5}^{(1)}] \in \mathbb{R}^{21}$ , are independent and identically distributed inputs, where  $m$  is the index of inputs. For sensitivity analysis, we omit the superscript of the parameters later for simplicity. The range of each input is 90–110% of their values shown in Table 1.

Then we perform Sobol sensitivity analysis, which is also called variance-based sensitivity analysis and is developed from the analysis of variance. As a global sensitivity analysis method, it analyzes the effects of each input by decomposing the variance of the output of the population model into fractions attributed to the inputs. In this paper, we perform both the first-order and second-order sensitivity analyses to the parameters. In particular, the first-order sensitivity index measures the attribution to the variance of the output considering only one input, which is calculated by:

$$S_m(y) = \frac{\text{Var}_{w_m}[\mathbb{E}_{w_{\sim m}}(y|w_m)]}{\text{Var}(y)}, \quad (11)$$

where  $w_{\sim m} = [w_{A1}, \dots, w_{m-1}, w_{m+1}, \dots, w_{C5}]$  includes all inputs except  $w_m$ . Next, the second order sensitivity with respect to  $m$  and  $n$  is measured by sum of attributing the variance of the output considering their first order effects and the second-order interaction between inputs  $m$  and  $n$ :

$$S_{(m,n)}(y) = S_m(y) + S_n(y) + \frac{\text{Var}_{(w_m, w_n)}[\mathbb{E}_{w_{\sim m,n}}(y|w_m, w_n)]}{\text{Var}(y)}. \quad (12)$$

Then we measure the total-order sensitivity index, which is calculated by attributing the variance of the output considering both the first-order effect, second-order effect, and other higher-order ones.

$$S_{T,m}(y) = 1 - \frac{\text{Var}_{w_{\sim m}}[\mathbb{E}_{w_m}(y|w_{\sim m})]}{\text{Var}(y)}. \quad (13)$$

When the sensitivity value is positive, the corresponding parameter is positively correlated with the model output. If the value is negative, they are negatively correlated. The absolute value of parameter sensitivities represents the degree of influence on the model output. If the sensitivity value is closer to 0, changing this parameter will have less influence on the model output. Based on the sensitivity values and the number of biomarker measurements, we determine the personalized parameters to fit the longitudinal data points for each patient and keep the remaining parameters the same as the population parameter values. This can avoid overfitting when providing the personalized prediction for each subject.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### DATA AVAILABILITY

Access to the ADNI dataset is publicly available via <http://adni.loni.usc.edu><sup>35</sup>.

### CODE AVAILABILITY

The sensitivity analysis code is available at <http://salib.readthedocs.io/en/latest/>. The simulation study code is available at <https://www.pymc.io/welcome>. The non-linear optimizer can be found in <https://github.com/jjhartmann/Levenberg-Marquardt>. Codes for Algorithms 1 and 2 are included in the Supplementary Information.

Received: 18 December 2021; Accepted: 16 June 2022;

Published online: 08 September 2022

## REFERENCES

- Cortes-Canteli, M. & Iadecola, C. Alzheimer's disease and vascular aging: Jacc focus seminar. *J. Am. College Cardiol.* **75**, 942–951 (2020).
- Batool, A., Kamal, M. A., Rizvi, S. & Rashid, S. Topical discoveries on multi-target approach to manage alzheimer's disease. *Curr Drug Metab.* **19**, 704–713 (2018).
- Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23 (2007).
- Lane, C. A., Hardy, J. & Schott, J. M. Alzheimer's disease. *Eur. J. Neurol.* **25**, 59–70 (2018).
- Aliev, G. et al. Alzheimer's disease—future therapy based on dendrimers. *Curr. Neuropharmacol.* **17**, 288–294 (2019).
- Milne, R. et al. At, with and beyond risk: expectations of living with the possibility of future dementia. *Soc. Health Illness* **40**, 969–987 (2018).
- Sperling, R. A. et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280–292 (2011).
- Jack, C. R. et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement* **14**, 535–562 (2018).
- Hao, W. & Friedman, A. Mathematical model on Alzheimer's disease. *BMC Syst Biol* **10**, 108 (2016).
- Petrella, J. R., Hao, W., Rao, A. & Doraiswamy, P. M. Computational causal modeling of the dynamic biomarker cascade in Alzheimer's disease. *Comput. Math. Methods Med.* **2019**, <https://doi.org/10.1155/2019/6216530> (2019).
- Jack, C. R. & Holtzman, D. M. Biomarker modeling of Alzheimer's disease. *Neuron* **80**, 1347–1358 (2013).
- Abeyasinghe, A. A. D. T., Deshapriya, R. D. U. S. & Udawatte, C. Alzheimer's disease; a review of the pathophysiological basis and therapeutic interventions. *Life Sci.* **256**, 117996 (2020).
- Guo, T., Korman, D., Baker, S. L., Landau, S. M. & Jagust, W. J. Longitudinal cognitive and biomarker measurements support a unidirectional pathway in Alzheimer's disease pathophysiology. *Biol. Psychiatry* **89**, 786–794 (2021).
- Myszczyńska, M. A. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat. Rev. Neurol.* **16**, 440–456 (2020).
- Iturria-Medina, Y., Carbonell, F. M., Sotero, R. C., Chouinard-Decorte, F. & Evans, A. C. Multifactorial causal model of brain (dis)organization and therapeutic intervention: Application to Alzheimer's disease. *Neuroimage* **152**, 60–77 (2017).
- Friedman, A. & Hao, W. The role of exosomes in pancreatic cancer micro-environment. *Bull. Math. Biol.* **80**, 1111–1133 (2018).
- Budithi, A., Su, S., Kirshstein, A. & Shahriyari, L. Data driven mathematical model of FOLFIRI treatment for colon cancer. *Cancers*. **13**, <https://doi.org/10.3390/cancers13112632> (2021).
- Hao, W. et al. A mathematical model of aortic aneurysm formation. *PLoS One* **12**, e0170807 (2017).
- Friedman, A. & Hao, W. A mathematical model of atherosclerosis with reverse cholesterol transport and associated risk factors. *Bull. Math. Biol.* **77**, 758–781 (2015).
- Wang, X. et al. A bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset alzheimer's disease. *Genetics* **209**, 51–64 (2018).
- Sun, N. et al. Multi-modal latent factor exploration of atrophy, cognitive and tau heterogeneity in alzheimer's disease. *Neuroimage* **201**, 116043 (2019).
- Schäfer, A. et al. Bayesian physics-based modeling of tau propagation in alzheimer's disease. *Front. Physiol.* 1081, <https://doi.org/10.3389/fphys.2021.702975> (2021).
- Iddi, S. et al. Estimating the evolution of disease in the parkinson's progression markers initiative. *Neurodegenerative Dis.* **18**, 173–190 (2018).
- Iddi, S. et al. Predicting the course of alzheimer's progression. *Brain Informatics* **6**, 1–18 (2019).
- Li, D. et al. The relative efficiency of time-to-progression and continuous measures of cognition in presymptomatic alzheimer's disease. *Alzheimer's & Dement.* **5**, 308–318 (2019).
- Li, D., Iddi, S., Thompson, W. K., Donohue, M. C. & Initiative, A. D. N. Bayesian latent time joint mixed effect models for multicohort longitudinal data. *Stat. Methods Med. Res.* **28**, 835–845 (2019).
- Marinescu, R. V. et al. Predicting alzheimer's disease progression: Results from the tadpole challenge: Neuroimaging: Neuroimaging predictors of cognitive decline. *Alzheimer's Dement.* **16**, e039538 (2020).
- Jedynak, B. M. et al. A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort. *Neuroimage* **63**, 1478–1486 (2012).
- Sobol, I. M. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Math. Comput. Simul.* **55**, 271–280 (2001).

- Zhang, S., Ponce, J., Zhang, Z., Lin, G. & Karniadakis, G. An integrated framework for building trustworthy data-driven epidemiological models: Application to the covid-19 outbreak in new york city. *PLoS Comput. Biol.* **17**, 1–29 (2021).
- Jack, C. R. et al. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. *Brain* **132**, 1355–1365 (2009).
- Shaw, L. M. et al. Qualification of the analytical and clinical performance of CSF biomarker analyses in ADNI. *Acta Neuropathol* **121**, 597–609 (2011).
- Shaw, L. M. PENN biomarker core of the Alzheimer's disease Neuroimaging Initiative. *Neurosignals* **16**, 19–23 (2008).
- Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.* **2**, 164–168 (1944).
- Weiner, M. W. et al. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's Dement.* **9**, e111–e194 (2013).

## ACKNOWLEDGEMENTS

G.L. and H.Z. were supported in part by NSF (DMS-1555072, DMS-1736364, DMS-2053746, and DMS-2134209) and DOE DE-SC0021142. J.R.P. was supported in part by NSF DMS-2052676. W.H. was supported in part by NSF DMS-2052685. P.M.D.'s work on this project is supported by the NIA, Karen L Wrenn Trust and Steve Aoki Fund.

## AUTHOR CONTRIBUTIONS

J.R.P. conceived the idea for in-silico modeling of AD biomarkers. P.M.D. initiated the collaboration between J.R.P., P.M.D., and W.H. to further develop this idea. W.H. developed the data-driven modeling idea. J.P. completed the idea by including DPS in consultation with P.M.D. and W.H.. P.M.D. provided the idea to apply this model separately in C.N., M.C.I. and A.D.. J.R.P. and P.M.D. provided data access and clinical constructs. G.L. and W.H. supervised the work. H.Z. implemented the code. All authors contributed to the discussions leading to the perspective presented. All authors contributed to the editing and shaping of the manuscript at various stages of preparation. All authors read and approved the final version.

## FUNDING

Funding for data collection was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904, Michael Weiner, PI) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<http://www.fnih.org>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. ADNI investigators contributed to the design and implementation of the ADNI database and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at [http://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgment\\_List.pdf](http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgment_List.pdf).

## COMPETING INTERESTS

The authors declare no Competing Non-Financial Interests but the following Competing Financial Interests: P.M.D. is a co-inventor on patents for the diagnosis or treatment of Alzheimer disease. P.M.D. owns shares in several biotechnology companies whose products are not discussed here. P.M.D. has received grants from NIH, DARPA, DOD, ONR, Bausch, Avanir, Avid, Cure Alzheimer's Fund, Karen L. Wrenn Trust, Steve Aoki Foundation, and advisory fees from Apollo, Brain Forum, Clearview, Lumos, Neuroglee, Otsuka, Verily, Vitakey, Sermo, Lilly, Vivly, and Transposon.

**ADDITIONAL INFORMATION**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-022-00632-7>.

**Correspondence** and requests for materials should be addressed to Guang Lin.

**Reprints and permission information** is available at <http://www.nature.com/reprints>


**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

**FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE**

Jeffrey R. Petrella <sup>2</sup> and P. Murali Doraiswamy<sup>3</sup>