BRAIN
ORIGINAL ARTICLE

Machine learning reveals connections between preclinical type 2 diabetes subtypes and brain health

Fan Yi,[1,†] Jing Yuan,[2,†] Fei Han,[2] Judith Somekh,[3] Mor Peleg,[3] Fei Wu,[1] Zhilong Jia,[4] Yi-Cheng Zhu[2] and Zhengxing Huang[1]

[†]These authors contributed equally to this work.

Previous research has established type 2 diabetes mellitus as a significant risk factor for various disorders, adversely impacting human health. While evidence increasingly links type 2 diabetes to cognitive impairment and brain disorders, understanding the causal effects of its preclinical stage on brain health is yet to be fully known. This knowledge gap hinders advancements in screening and preventing neurological and psychiatric diseases. To address this gap, we employed a robust machine learning algorithm (Subtype and Stage Inference, SuStaIn) with cross-sectional clinical data from the UK Biobank (20 277 preclinical type 2 diabetes participants and 20 277 controls) to identify underlying subtypes and stages for preclinical type 2 diabetes.

Our analysis revealed one subtype distinguished by elevated circulating leptin levels and decreased leptin receptor levels, coupled with increased body mass index, diminished lipid metabolism, and heightened susceptibility to psychiatric conditions such as anxiety disorder, depression disorder, and bipolar disorder. Conversely, individuals in the second subtype manifested typical abnormalities in glucose metabolism, including rising glucose and haemoglobin A1c levels, with observed correlations with neurodegenerative disorders. A >10-year follow-up of these individuals revealed differential declines in brain health and significant clinical outcome disparities between subtypes. The first subtype exhibited faster progression and higher risk for psychiatric conditions, while the second subtype was associated with more severe progression of Alzheimer's disease and Parkinson's disease and faster progression to type 2 diabetes. Our findings highlight that monitoring and addressing the brain health needs of individuals in the preclinical stage of type 2 diabetes is imperative.

1 College of Computer Science and Technology, Zhejiang University, Hangzhou 310008, China
2 Department of Neurology, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100730, China
3 Department of Information Systems, University of Haifa, Haifa 3303219, Israel
4 Medical Innovation Research Division of Chinese PLA General Hospital, Beijing 100853, China

Correspondence to: Zhengxing Huang
College of Computer Science and Technology, Zhejiang University
No. 38, Zheda Road, Xihu District, Hangzhou 310008, Zhejiang Province, China
E-mail: Zhengxinghuang@zju.edu.cn

Correspondence may also be addressed to: Yi-Cheng Zhu
Department of Neurology, Peking Union Medical College Hospital
Peking Union Medical College, Chinese Academy of Medical Sciences, No. 1,
Shuifuyuan, Dongcheng District, Beijing 100730, China
E-mail: zhuyc@pumch.cn

Received July 22, 2024. Revised December 28, 2024. Accepted January 23, 2025. Advance access publication February 11, 2025

Zhilong Jia
Medical Innovation Research Division of Chinese PLA General Hospital
No. 28, Fuxing Road, Haidian District, Beijing 100853, China
E-mail: jiazhilong@plagh.org

Keywords: preclinical-T2DM; subtype and stage inference; machine learning; brain health

Introduction

Type 2 diabetes mellitus (T2DM) represents a significant global public health challenge, with its prevalence steadily increasing. In 2021, it affected 537 million adults worldwide.[1] Projections from the International Diabetes Federation indicate that by 2030, this number will rise to 643 million, reaching a staggering 783 million by 2045.[1,2] The phase during which clinical symptoms remain absent but biological irregularities hint at the potential development of T2DM, namely, the preclinical stage of type 2 diabetes (preclinical-T2DM), plays an important role in the development of T2DM.[3] Recent research has indicated a link between T2DM and deterioration in brain health, including accelerated rates of neurological and cognitive decline.[4,5] In terms of preclinical-T2DM, these aspects remain poorly understood.

First, there is limited knowledge regarding how preclinical-T2DM progresses over time before the onset of T2DM and how the phenotypic and genetic aetiologies of preclinical-T2DM vary. While previous studies have clustered prediabetes into subtypes and explored their connections with T2DM and multi-systemic impairments,[6] these approaches often overlook the longitudinal diversity of preclinical-T2DM. Conversely, progressive heterogeneity characterizes T2DM, potentially emerging from its preclinical stages and subtly influencing brain health, encompassing neurological, psychiatric and cognitive functions. Understanding these complexities is crucial to elucidating how preclinical-T2DM evolves and affects various aspects of health, particularly brain structures and functions.

Addressing this challenge can be achieved by leveraging machine learning models that are increasingly used in biomedical research.[7] One such model, the Subtype and Stage Inference (SuStaIn) model,[8] originally designed to capture disease progression patterns in chronic conditions, facilitates longitudinal inference from cross-sectional data by automatically identifying distinct spatiotemporal trajectories of cumulative pathological alterations shown by measured biomarkers.[8-13] In this study, we employed SuStaIn to decipher heterogenous progressive patterns of preclinical-T2DM, offering valuable insights into disease onset and progression. This aids in the establishment of quantitative metrics for T2DM screening and prognostication. By identifying subtypes and progressive trajectories of preclinical-T2DM and exploring systematic changes in brain health, we can enhance our ability to assess it precisely in clinical practice. This not only benefits individuals with preclinical-T2DM but also contributes to better clinical outcomes, reducing the risk of neuropathy and cognitive dysfunction etc.[14]

In this study, we embarked on a comprehensive investigation into the heterogenous progression of preclinical-T2DM and its implications for brain health using a multi-faceted research approach (Fig. 1). First, we identified 20 277 preclinical-T2DM subjects with a balanced 20 277 control group from UKB for analysis (Fig. 1A). Next, we utilized screened 18 preclinical-T2DM-associated clinical indexes and applied SuStaIn to stratify preclinical-T2DM subjects into distinct subtypes and stages, leading to two subtypes with distinct metabolic profiles (Fig. 1B). We then analysed phenotypic associations between the two subtypes and a variety of brain disorders, cognitive functions, as well as molecular phenotypes, proteins and metabolites (Fig. 1C). Additionally, using genome-wide association studies (GWAS), we identified genetic variants significantly associated with each preclinical-T2DM subtypes. Expanding on these findings, we investigated genetic relationships between the subtypes and brain disorders via genetic, genetic colocalization and Mendelian randomization (MR) analyses (Fig. 1C). By exploring the genetic and molecular landscape of preclinical-T2DM subtypes and their impact on brain health, we highlighted how underlying phenotypic and genetic variation drives the subtypes and stages of preclinical-T2DM. These insights improve our understanding of the complex interplay between preclinical-T2DM and brain health.

Materials and methods

Study cohort

In this study, we leveraged the UK Biobank (UKB) dataset, a large biomedical cohort comprising over 500 000 participants, as the primary data for our analyses. The use of UKB data was approved by UKB under application number: 85757. Approval for the UKB study was obtained from the National Research Ethics Committee (REC reference 11/NW/0382), and informed consent was obtained from all participants. The inclusion and exclusion criteria process depicted in Fig. 1A involves the identification of preclinical-T2DM cases and control participants from the UKB dataset. The preclinical-T2DM subjects included in this study were participants with T2DM diagnosis after their initial assessment at the UKB (instance = 0). The UKB dataset comprised 41 783 patients diagnosed with T2DM and 460 628 control participants without T2DM. Diagnoses of T2DM were identified through the International Classification of Diseases, 10th revision (ICD-10) codes (E11 for T2DM), as well as self-reported non-cancer T2DM diagnoses. To refine the study population, several exclusion criteria (Fig. 1A and Supplementary material, 'Methods' section) were applied, yielding a total of 20 277 individuals (aged 40–70 years, mean age 59.61 years, 42.43% female) with preclinical-T2DM.

Lastly, we employed propensity score matching (PSM) for the construction of a matched control group to ensure the computability of the SuStaIn model (Supplementary material, 'Methods' section). Thereafter, we established a balanced control group of 20 277 individuals without any diabetes (aged 40–73 years, mean age 59.96 years, 41.75% female) with the least standard mean difference (SMD) compared with the preclinical-T2DM group for subsequent analyses (Supplementary Table 1).

Feature selection process

Initially, we collected 62 biomarkers from the UKB dataset, including blood biochemistry markers, urine assay results, blood count, physical measurements and blood pressure, for SuStaIn modelling (Supplementary Table 2). These biomarkers were accessed for more than 420 000 individuals from their initial assessment at the UKB.

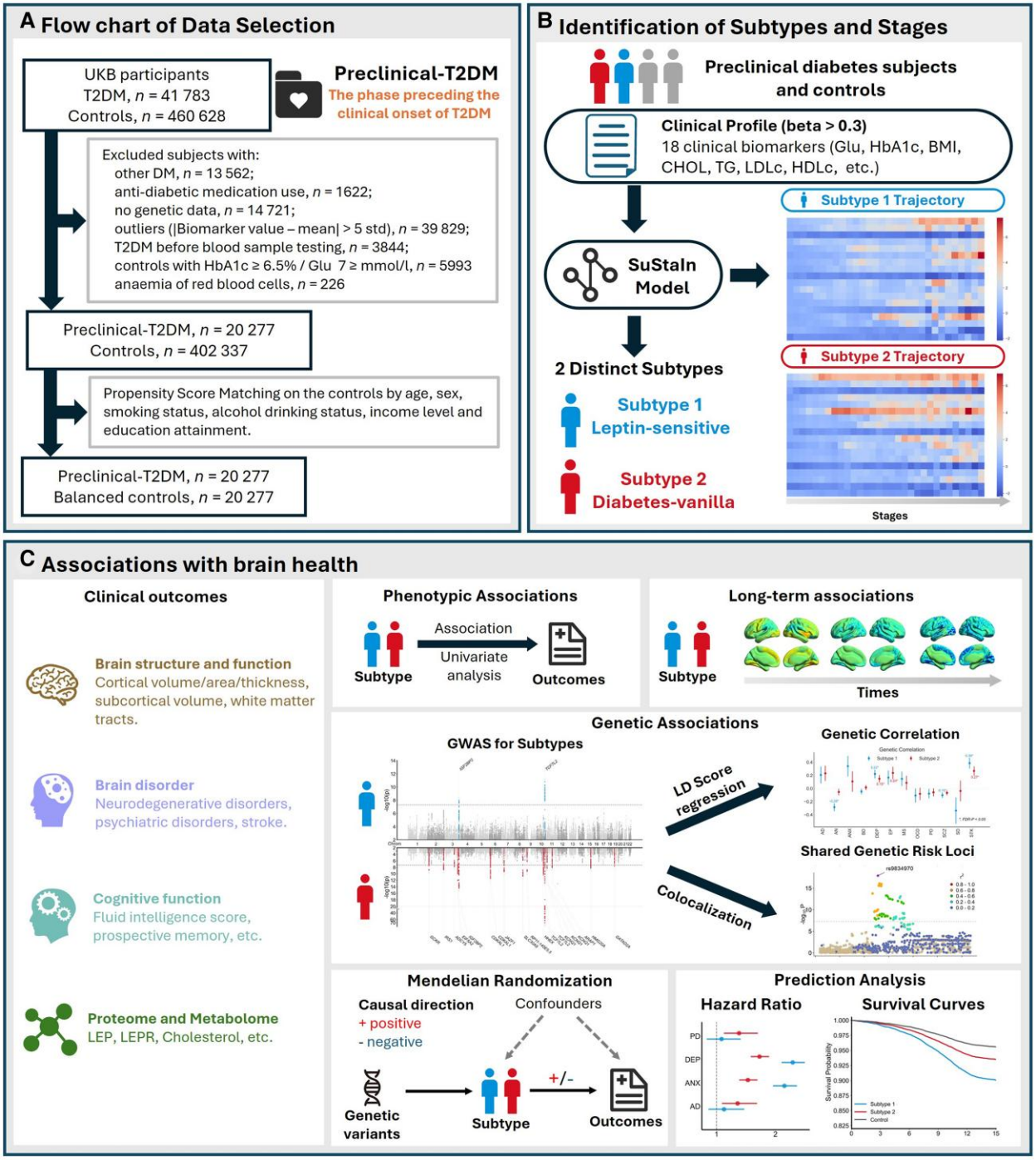

Figure 1 Overview of the study design. (**A**) Flow chart depicting the inclusion and exclusion criteria for data selection from the UK Biobank (UKB). We selected subjects with preclinical-type-2 diabetes mellitus (T2DM), defined as the phase preceding the onset of T2DM, along with a corresponding balanced control group from the UKB for analysis. (**B**) Application of the SuStaIn model to identify subtypes and stages of preclinical-T2DM. The model identified two subtypes: Subtype 1 (S1, leptin-sensitive) and Subtype 2 (S2, diabetes-vanilla) with two distinct progression trajectories, using 18 preclinical-T2DM-associated clinical biomarkers. (**C**) Associations with brain health. We examined the associations of the identified subtypes with brain health using multiple analytical methods, including phenotypic associations, long-term imaging-derived phenotype association analysis, genetic associations, Mendelian randomization and prediction analysis. AD = Alzheimer's disease; ANX = anxiety disorder; BMI = body mass index; CHOL = cholesterol; DEP = depression; Glu = glucose; HbA1c = haemoglobin A1c; HDLc = high-density lipoprotein cholesterol; LD = linkage disequilibrium; LDLc = low-density lipoprotein cholesterol; LEP = leptin; LEP = leptin receptor; PD = Parkinson's disease; TG = triglycerides.

Given the computational demands of the SuStaIn model and insight from prior research suggesting an optimal number of features of approximately 20 for computational feasibility,[8,10] we implemented a cutoff criterion to select the most important biomarkers for SuStaIn modelling after data imputation (Supplementary material, 'Methods' section). This process identified 18 pivotal biomarkers, consisting of haemoglobin A1c (HbA1c), body mass index (BMI), high-density lipoprotein cholesterol, triglyceride-glucose

index, high light scatter reticulocyte count, glucose, reticulocyte count, immature reticulocyte fraction, apolipoprotein A, alanine aminotransferase, triglycerides, urate, white blood cell count, cholesterol, C-reactive protein, lymphocyte count, low-density lipoprotein cholesterol (LDLc) and vitamin D. For detailed information on the selected biomarkers, please refer to Supplementary Table 2.

Identification of subtypes and stages for preclinical-type 2 diabetes mellitus

To unveil the diverse manifestations of preclinical-T2DM and explore its interplay with brain health, we utilized the SuStaIn model to categorize participants with preclinical-T2DM into distinct subtypes and stages of disease progression. We applied the linear z-scored SuStaIn model,[15] integrating the 18 selected clinical biomarkers for subtype and stage identification (Supplementary material, 'Methods' section).

Furthermore, to ascertain the optimal number of preclinical-T2DM subtypes, we employed a 10-fold cross-validation approach, ranging from one to five subtypes. Performance evaluation was conducted using the Cross-Validation Information Criterion (Supplementary material, 'Methods' section).[11,15] Therefore, considering the two-subtype model for its optimal balance of simplicity and explanatory power, we chose to interpret the findings based on the two-subtype model.

Phenotypic associations between preclinical-type 2 diabetes mellitus subtypes and brain health

We examined the phenotypic associations between preclinical-T2DM subtypes and brain health. Our analysis included a broad spectrum of phenotypic measures, consisting of brain structure and function, brain disorders, cognitive functions and metabolic and proteomic profiles (Supplementary material, 'Methods' section).

GWAS on preclinical-type 2 diabetes mellitus subtypes

We conducted GWASs using genotyped and imputed data from the UKB on both subtypes of preclinical-T2DM. Genome-wide genotyping data was performed on all UKB participants using the UK Biobank Axiom Array, followed by imputation using the Haplotype Reference Consortium and UK10K as reference panels (GRCh37 assembly).[16] The analyses were stratified into two comparative sets: S1 ($n = 7942$) versus controls ($n = 20\,277$) and S2 ($n = 9439$) versus the same control group ($n = 20\,277$). After stringent quality control (Supplementary material, 'Methods' section), the GWAS analyses were performed using PLINK software (v.1.90 beta), with adjustments for covariates including sex, age, smoking status, alcohol drinking status, income level, education attainment and the first ten principal components to address potential population stratification effects. The genome-wide significance threshold was set to 5.0×10^{-8}. We employed the Functional Mapping and Annotation (FUMA)[17] platform to annotate the results of the GWAS (Supplementary material, 'Methods' section).

Genetic correlation and colocalization of preclinical-type 2 diabetes mellitus subtypes and brain health

To evaluate the genetic associations between preclinical-T2DM subtypes and brain health, we employed two genetic analysis techniques: genetic correlation and colocalization analysis, utilizing publicly available GWAS results (Supplementary Table 11). Firstly, we utilized linkage disequilibrium score regression (LDSC)[18] to measure the genetic correlation between the T2DM subtypes and outcomes. Only high-quality single nucleotide polymorphisms (SNPs) documented in the HapMap3 dataset were utilized for estimation, with the LD score derived from the 1KGp3 EUR panel employed for LDSC analysis.

Furthermore, to determine whether the preclinical-T2DM subtypes share a common causal variant with brain disorders, we conducted colocalization analysis using the R package 'coloc' (Supplementary material, 'Methods' section).[19-21] In accordance with established conventions,[22] variants with a posterior probability of LDSC > 0.75 were considered colocalized variants (indicating shared causal variants) for preclinical-T2DM subtypes and brain disorders.

Mendelian randomization analyses

To further explore the causal relationship between preclinical-T2DM subtypes and brain disorders, we conducted two-sample Mendelian randomization (MR) analyses using the R package 'TwoSampleMR' (Supplementary material, 'Methods' section).

Prediction analyses

To evaluate the predictive power of these subtypes for disease progression, we examined the efficacy and applicability of preclinical-T2DM subtypes as predictive clinical indicators for brain disorders as outcomes. We further investigated whether incorporating subtype information could enhance risk prediction of brain disorders (Supplementary material, 'Methods' section).

Results

Identification of robust subtype and stages for preclinical-type 2 diabetes mellitus

Two distinct subtypes of preclinical-T2DM were identified with 18 biomarkers of 40 544 individuals in UKB using the SuStaIn algorithm (Fig. 2A). We selected preclinical-T2DM individuals, defined as the individuals at the cohort baseline who will have T2DM at follow-up based on the inclusion and exclusion criteria (Fig. 1A), resulting in 20 277 preclinical-T2DM and 20 277 propensity score matched controls. Due to the computation requirements, we selected 18 of 62 biomarkers from the baseline UKB dataset due to their significance for preclinical-T2DM, as determined by larger effect sizes in univariable logistic regression (Fig. 2A and Table 1). Using a ten-fold cross-validation approach, we determined the most robust result, revealing two distinct subtypes, Subtype 1 (S1) and Subtype 2 (S2), with 36 subtype-specific stages (S1, $n = 7\,942$, mean age 59 years, 52% female; and S2, $n = 9\,439$, mean age 60 years, 37% female). Notably, we excluded 2896 individuals assigned as being in stage 0, as stage 0 indicates that none of the biomarkers have reached the z-score threshold. This distinction highlights two distinct clinical and pathophysiological trajectories for preclinical-T2DM (Fig. 2A and B). Moreover, cross-validation demonstrated a high consistency in the identification of subtypes for each preclinical-T2DM participant, with the majority of subjects (96.16% on average) consistently assigned to the same group across validation folds (Supplementary Fig. 2). These findings further confirm the stability and reproductivity of the results obtained from SuStaIn.

Two distinct metabolic profiles of preclinical-type 2 diabetes mellitus

The two subtypes of preclinical-T2DM exhibited significant differences in clinical biomarkers (Fig. 2D, Table 1 and Supplementary

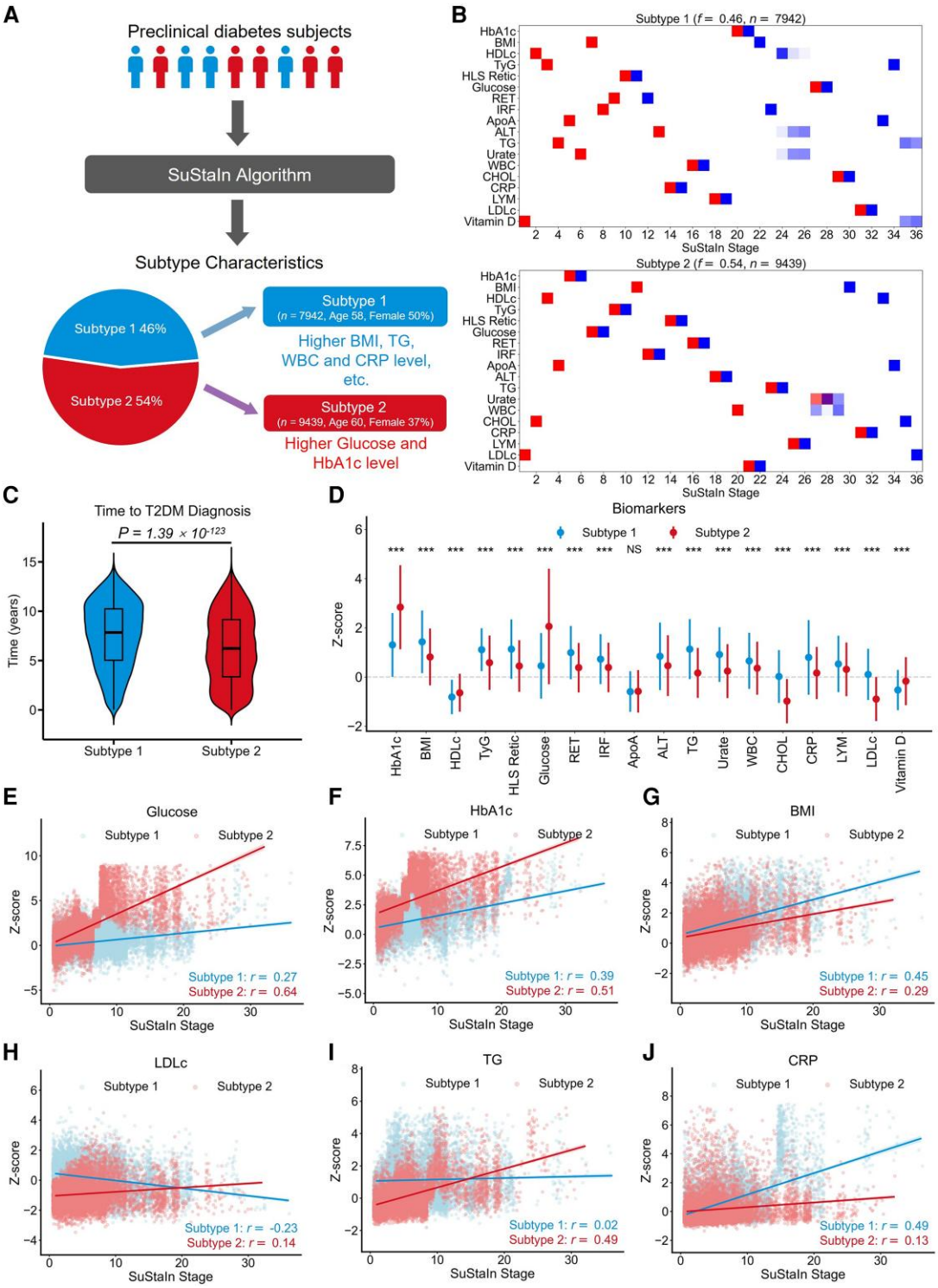

Figure 2 Identification of subtypes and stages of preclinical-type 2 diabetes mellitus. (**A**) An overview of the two distinct preclinical-type 2 diabetes mellitus (T2DM) subtypes identified by the Subtype and Stage Inference (SuStaIn) algorithm. (**B**) Positional variance diagrams of two distinct metabolic trajectories obtained from SuStaIn. The diagrams visualize the cumulative probability of each biomarker reaching a specific z-score threshold. The colours indicate the two z-score thresholds for each biomarker: red indicates a mild biomarker change, blue indicates a severe biomarker change. The colour density represents the proportion of the posterior distribution where events (*y*-axis) occur at specific positions in the sequence (*x*-axis); f represents the proportion of individuals assigned to each phenotype. (**C**) Comparison of time to T2DM diagnosis for the two subtypes. (**D**) Comparison of the mean z-scores of 18 selected clinical biomarkers across the two identified subtypes, where biomarkers were z-scored relative to the control group, adjusting for age, sex, smoking status, alcohol drinking status, income level and educational attainment. A higher z-score indicates a greater deviation from the control group norm. (**E–J**) Progressions of various biomarkers across SuStaIn stages. Six selected biomarkers with remarkably distinct progression in the two subtypes are illustrated. The progressions of other biomarkers are presented in Supplementary Figs 5–16. *r* = Pearson's correlation between biomarkers and SuStaIn stages for each subtype. ALT = alanine aminotransferase; ApoA = apolipoprotein A; BMI = body mass index; CRP = C-reactive protein; CHOL = cholesterol; HbA1c = haemoglobin A1c; HDLc = high-density lipoprotein cholesterol; HLS Retic = high light scatter reticulocyte count; IRF = immature reticulocyte fraction; LDLc = low-density lipoprotein cholesterol; LYM = lymphocyte count; RET = reticulocyte count; TG = triglycerides; TyG = triglyceride-glucose index; WBC = white blood cell count.

Table 1 Basic characteristics and the 18 clinical biomarkers of the two subtypes of preclinical-type 2 diabetes mellitus

Variables	Normal range	S1 ($n = 7942$)	S2 ($n = 9439$)	P-value
Age, years	–	58.35	60.47	1.44×10^{-85}
Sex (female), %	–	49.94	37.07	7.14×10^{-66}
Time to diabetes diagnosis, years	–	7.53	6.27	1.39×10^{-123}
HbA1c, mmol/mol	<42	39.78	**45.28↑**	$<2.23 \times 10^{-308}$
Glucose, mmol/l	3.9–5.6	5.19	**6.11↑**	$<2.23 \times 10^{-308}$
TyG	6.98–10.71	9.26	9.03	1.80×10^{-201}
BMI	18.5–24.9	**33.21↑**	**30.69↑**	4.98×10^{-227}
HDLc, mmol/l	>1.6	**1.18↓**	**1.20↓**	8.42×10^{-11}
LDLc, mmol/l	<3	**3.71↑**	2.82	$<2.23 \times 10^{-308}$
TG, mmol/l	<1.69	**2.79↑**	**1.94↑**	$<2.23 \times 10^{-308}$
CHOL, mmol/l	<5.17	**5.79↑**	4.63	$<2.23 \times 10^{-308}$
ApoA, g/l	1.02–2.0	1.40	1.40	0.34
ALT, U/l	7–56	31.10	27.67	7.02×10^{-61}
HLS Retic, $\times 10^{12}$ cells/l	–	0.027	0.022	$<2.23 \times 10^{-308}$
RET, $\times 10^{12}$ cells/l	–	0.082	0.068	1.07×10^{-275}
IRF	0.16–0.24	0.33↑	0.31↑	2.37×10^{-110}
WBC, $\times 10^{9}$ cells/l	4–11	7.93	7.44	5.23×10^{-69}
CRP, mg/l	<10	4.57	2.80	5.54×10^{-219}
LYM, $\times 10^{9}$ cells/l	–	2.27	2.10	2.62×10^{-57}
Urate, μmol/l	M: 200–420 F: 140–360	373.29	339.56	1.16×10^{-177}
Vitamin D, nmol/l	30–50	38.21	46.54	3.63×10^{-175}

Abnormal values of clinical indexes are marked in bold. Up-arrows indicate values above the normal range and down-arrows indicate values below the normal range. ALT = alanine aminotransferase; ApoA = apolipoprotein A; BMI = body mass index; CHOL = cholesterol; CRP = C-reactive protein; F = female; HbA1c = haemoglobin A1c; HDLc = high-density lipoprotein cholesterol; HLS Retic = high light scatter reticulocyte count; IRF = immature reticulocyte fraction; LDLc = low-density lipoprotein cholesterol; LYM = lymphocyte count; M = male; RET = reticulocyte count; S1/S2 = Subtype 1/2; TG = triglycerides; TyG = triglyceride-glucose index; WBC = white blood cell count.

Fig. 17). Compared with S2, S1 exhibited elevated BMI, total cholesterol, triglycerides, urate and LDLc (Table 1, Fig. 2D, G and H and Supplementary Figs 11, 12, 14 and 17). These heightened biomarkers may suggest a correlation with more severe metabolic dysregulation in lipid metabolism for S1. Additionally, S1 demonstrated higher levels in inflammatory biomarkers, including reticulocyte count, immature reticulocyte fraction, C-reactive protein, white blood cell count and lymphocyte count (Fig. 2D and J and Supplementary Figs 7, 8, 13, 15 and 17), potentially indicating a more pronounced inflammatory activity. Moreover, S1 presented with lower levels of Vitamin D (Fig. 2D and Table 1). Studies have reported that vitamin D supplementation among individuals with prediabetes can mitigate the risk of developing T2DM and facilitate the transition from prediabetes to normoglycaemia.[23] In contrast, S2 was associated with elevated levels of glucose as well as HbA1c, with these increases becoming more pronounced across the SuStaIn stages (Fig. 2E and F and Supplementary Fig. 17). Furthermore, it was observed that individuals in S2 were, on average, closer to T2DM diagnosis, with an estimated average time of 6.5 years to diagnosis, compared with 7.6 years for S1 (Fig. 2C, Table 1 and Supplementary Fig. 4). This observation suggests that although both subtypes were in the early stage of T2DM, S2 individuals were closer to the T2DM diagnosis, and therefore may represent those with more advanced glycaemic abnormalities. These findings underscore the heterogeneity within preclinical-T2DM and suggest that the two subtypes may represent distinct metabolic profiles.

Phenotypic associations of preclinical-type 2 diabetes mellitus subtypes on brain health

We examined the phenotypic association with 12 brain disorders between the two subtypes and the controls to explore the clinical significance of the two preclinical-T2DM subtypes. Our findings revealed that compared with S2, S1 showed a significantly higher risk of psychiatric disorders [Fig. 3; false discovery rate (FDR) adjusted *P*-values for S1 versus S2 <0.05], including anxiety, bipolar, depression and sleep disorders. In addition, the risk for Alzheimer's disease, Parkinson's disease and stroke was slightly increased in S2 compared with S1, but the difference did not reach statistical significance (Fig. 3; FDR adjusted *P*-values for S1 versus S2 >0.05). These associations highlight the intricate relationship between preclinical-T2DM subtypes and brain health, indicating potential shared mechanism-related pathways and comorbidities.

We also investigated the associations with six cognitive functions between the two subtypes and controls. Both subtypes were associated with significant declines across several cognitive aspects, including numeric memory, symbol digit substitution, reaction time, fluid intelligence and reasoning and Trail-Making Test performance compared with controls, although the differences between S1 and S2 were not statistically significant (Fig. 4A–F). Specifically, S1 showed slightly worse performance on executive functioning, with a higher impairment in symbol digit substitution and the Trail-Making Test compared with S2 (Fig. 4B and F). Conversely, S2 showed slightly worse performance on numeric memory and reaction time, involving the ability to memorize and neural processing speed and response (Fig. 4A and C).

We evaluated the proteomic phenotypes with two subtypes of preclinical-T2DM. In the proteomic analysis, we found BGN-specific proteins (BGN) were highly expressed in S2, while brain-specific and immune-specific proteins, OXT and LAT2, were highly expressed in S1 (Fig. 4G). Fibroblast growth factor 21 (FGF21), with the highest expression observed in S1, was upregulated in S1 compared with healthy controls and S2. FGF21, a pivotal player in the regulation of energy balance and glucose as well as lipid homeostasis, has gathered attention as a therapeutic target for T2DM and obesity. Clinical trials utilizing FGF21 analogues and

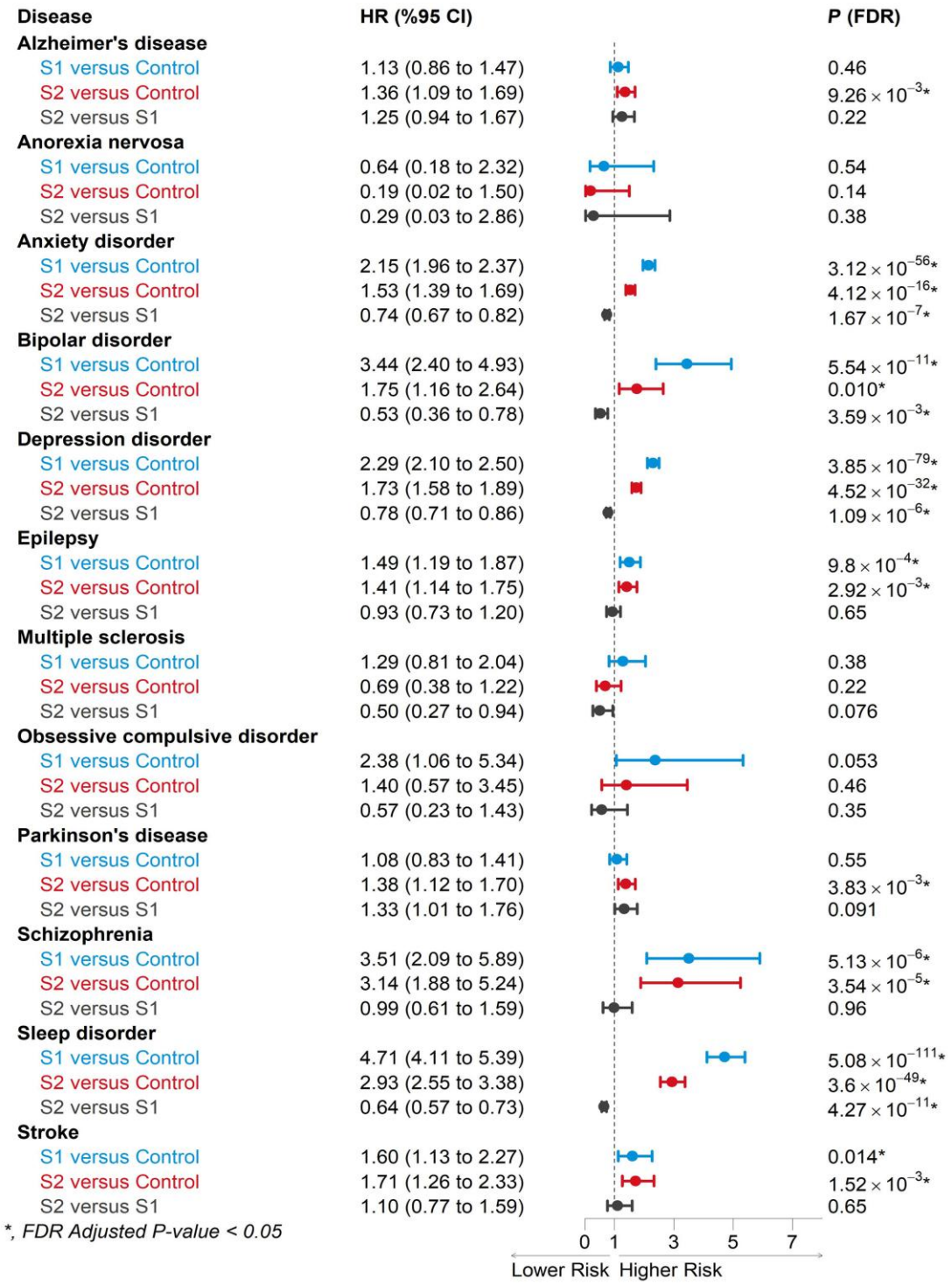

Figure 3 Hazard ratios of the two preclinical-type 2 diabetes mellitus subtypes on the onset of brain disorders. The hazard ratios (HRs) for the two subtypes were estimated by utilizing the Cox proportion hazard model and adjusted by sex, age, smoking status, alcohol drinking status, income level and education attainment. The HR was used to compare the progression rate of the diseases between each subtype and the control group (with the control group as the reference) and also between the two subtypes (with Subtype 1 as the reference). HR >1 indicates an increased risk, while HR <1 indicates a reduced risk, of developing the specific disease. *P*-values were adjusted for multi-testing using false discovery rate (FDR) correction at the threshold of 5% significance level according to the Benjamini-Hochberg procedure. CI = confidence interval; S1 = Subtype 1; S2 = Subtype 2.

mimetics have shown promise in patients with obesity and T2DM,[24] suggesting a potential physiological response of FGF21 to the preclinical status of T2DM. Leptin exhibits a similar expression pattern to FGF21, indicating a potential association with the preclinical state of T2DM (Fig. 4H and I). Conversely, the expression patterns of leptin receptors across the three groups were dramatically opposed, suggesting a possible impairment in leptin receptor signalling (Fig. 4H and I). This observation aligned with studies

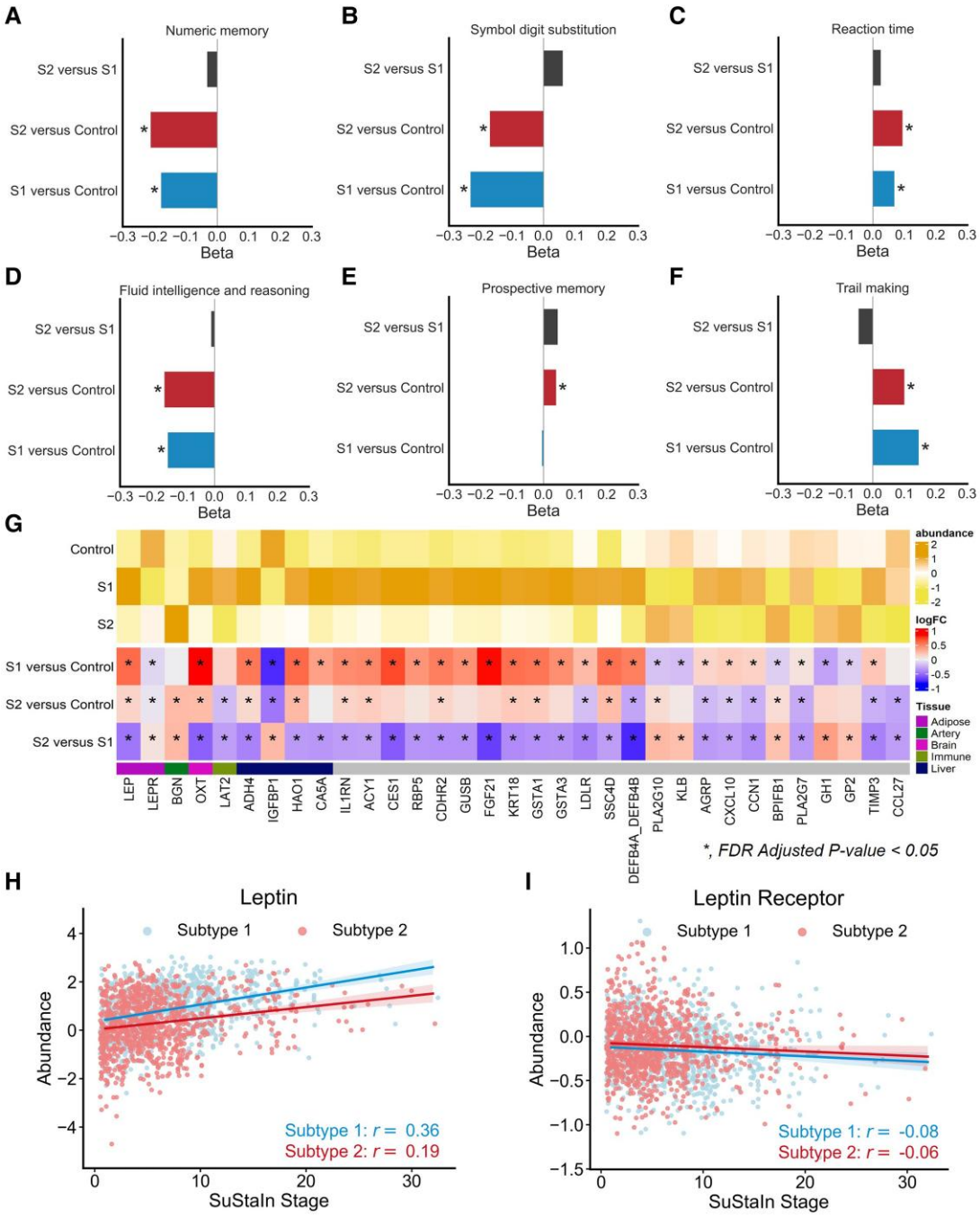

Figure 4 Phenotypic associations with brain health. (**A**–**F**) Phenotypic associations for six cognitive functions between controls and the two subtypes. Associations were measured using univariate analysis, z-scored relative to the control group and adjusted for sex, age, smoking status, alcohol drinking status, income level and educational attainment. *significant association after false discovery rate (FDR) correction (FDR adjusted *P*-value < 0.05). (**G**) Comparison of proteomic expressions between controls and the two subtypes. Log fold-change (logFC) and statistical significance were used to show differences in proteomic expressions across the three groups: Controls, Subtype 1 (S1) and Subtype 2 (S2). Significant differences were assessed using limma with FDR-adjusted *P*-value < 0.05 after adjusting sex and age confounding variables. Thirty-three proteins with logFC of S2 versus S1 >0.2 or <−0.3 are visualized using heat map, and a full list of proteins exhibiting differential abundance is listed in Supplementary material, 'Data' section 9. (**H** and **I**) Association the abundance of leptin and leptin receptor across SuStaIn stages for the two subtypes. SuStain = Subtype and Stage Inference model.

utilizing leptin receptor-deficient db/db mice as models for T2DM.[25] Elevated circulating leptin concentrations, as observed in individuals with obesity, are often attributed to leptin resistance,[26,27] potentially suggesting that S1 is related to leptin resistance. Leptin-resistant syndromes are known to contribute to severe insulin resistance and diabetes.[28-30] Furthermore, insulin-like growth factor binding protein 1 (IGFBP1) as well as β-klotho exhibit decreased expression levels in S1 individuals compared with healthy controls (Fig. 4G). Reduced concentrations of circulating IGFBP1 have been linked to insulin resistance and diabetes,[31] while

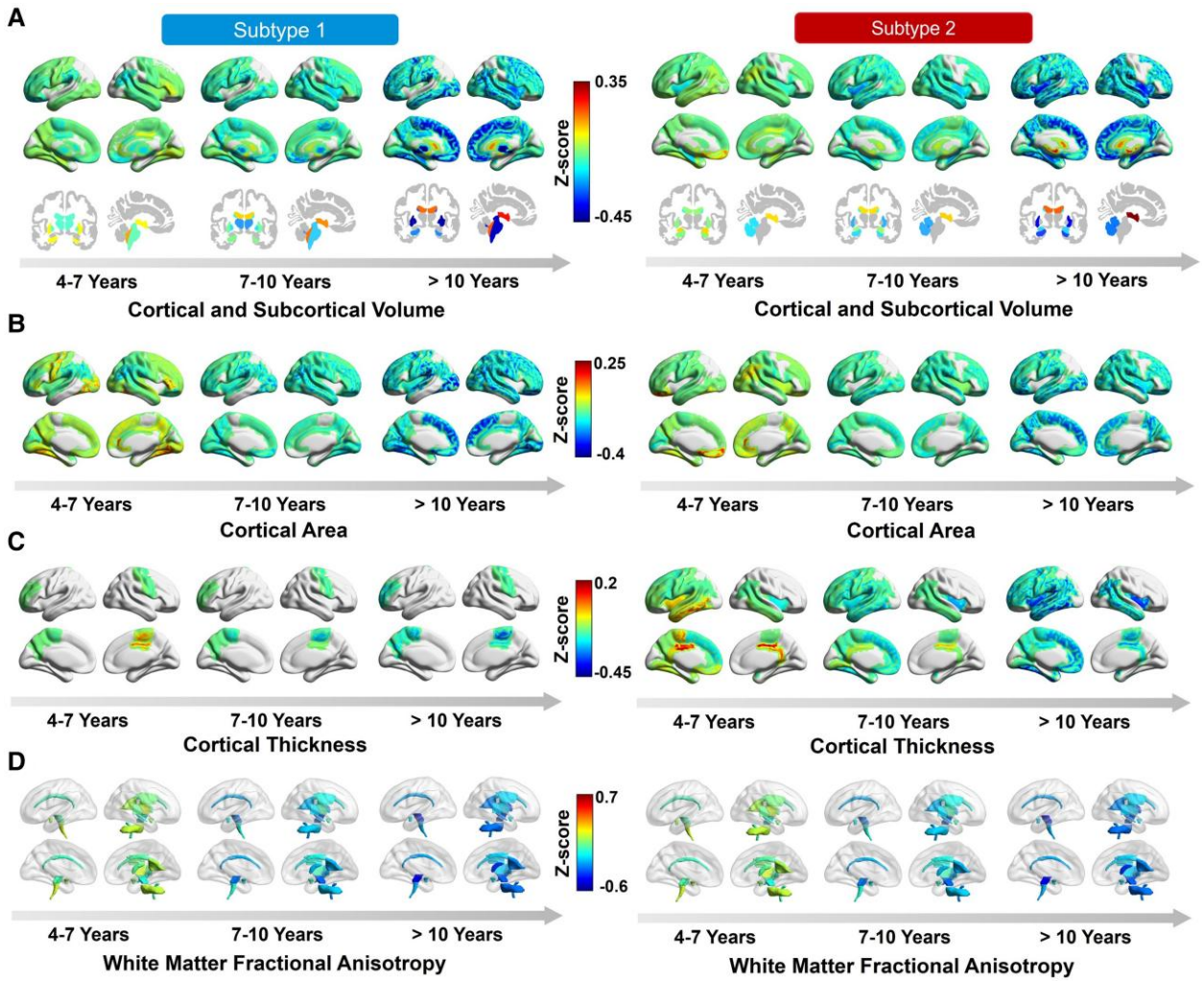

Figure 5 Long-term associations with the brain structures and functions. (**A–D**) Long-term effects of preclinical-type 2 diabetes mellitus subtypes on the selected image derived phenotypes (IDPs) from brain MRIs. Subjects for each subtype were categorized into three groups based on imaging acquisition intervals (4–7 years, 7–10 years and >10 years) to analyse temporal changes in the brain structures and functions. We illustrated the mean z-score of selected IDPs with the most pronounced effects associated with subtypes. Long-term associations with other IDPs were presented in Supplementary Fig. 29. The IDP values were z-scored relative to the control group and adjusted for sex, age, smoking status, alcohol drinking status, income level and educational attainment. Total intracranial volume was also adjusted for cortical and subcortical volumes. The colour map illustrates the mean z-scores of the IDPs for each time interval; red indicates larger IDP values and blue indicates smaller IDP values.

β-klotho, acting as a cell-surface glucose sensor and co-receptor for FGF21, holds promise as a therapeutic target for T2DM by modulating glucose-stimulated insulin release in pancreatic β cells. As the triglyceride to high-density lipoprotein cholesterol (TG: HDLc) ratio and metabolic score for insulin resistance (METS-IR) serve as readily measurable markers of insulin resistance, our findings revealed that S1 exhibited significantly elevated TG: HDLc and METS-IR levels compared with S2 and the healthy control group (Figs 2D and 4G). Significantly higher interleukin-1 receptor agonist (IL-1RA) was observed in S1 compared with S2 and controls. Circulating IL-1RA (encoded by *IL1RN*), an endogenous inhibitor of proinflammatory IL-1β, may be protective against the development of insulin resistance.[32]

Moreover, we evaluated the associations between metabolomic phenotypes and two subtypes of preclinical-T2DM. We observed significantly higher levels of total cholesterol, total triglycerides, total fatty acids, omega-6 fatty acids and total lipids in lipoprotein particles in S1 compared with S2 and the healthy control group (Supplementary Fig. 19). Our findings suggested that these associations tended to be particularly pronounced in S1. Furthermore, we found that the ratios of polyunsaturated fatty acids to total fatty acids, glucose-lactate and glucose were lowest in S2 among the three groups.

Long-term effects of preclinical-type 2 diabetes mellitus subtypes on the brain

The analysis of long-term effects on the brain structure and function across the two subtypes unveiled distinct outcomes (Fig. 5). After adjusting for various covariates, we observed that although the differences in changing rate between the two subtypes were not statistically significant (FDR adjusted *P*-value > 0.05; Supplementary material, 'Data', section 8), both subtypes of preclinical-T2DM were associated with long-term changes in many brain regions. In terms of MRI-based brain structure, S1 exhibited slightly more atrophy from cortical and subcortical volume and cortical area over time, particularly in the putamen, accumbens, thalamus and cerebral white matter (Fig. 5A and B and Supplementary Figs 20 and 21). In contrast,

S2 displayed slightly more atrophy in cortical and subcortical volume and cortical thickness of the fusiform gyrus, superior temporal lobe, superior parietal lobe and middle temporal lobe (Fig. 5A and C and Supplementary Figs 20 and 22). Regarding white matter integrity and microstructural organization, both subtypes showed decreases in mean white matter fractional anisotropy (FA) in regions such as the inferior cerebral peduncle, cerebral peduncle and fornix (Fig. 5D and Supplementary Fig. 23). S1 also demonstrated reductions in the genu of corpus callosum, inferior cerebral peduncle and superior fronto occipital fasciculus, while S2 displayed more reductions in the middle cerebellar peduncle, splenium of corpus callosum and posterior limb of internal capsule (Fig. 5D).

GWAS of the two preclinical-type 2 diabetes mellitus subtypes

We investigated the subtype-specific associated SNPs for the two preclinical-T2DM subtypes using GWAS, respectively, identifying two genomic risk loci with two independent lead SNPs for S1 and 15 genomic risk loci with 19 lead SNPs for S2 (Fig. 6A and Supplementary Tables 9 and 10). The SNP-based heritability estimates for the two subtypes were 0.14 and 0.18, respectively (Supplementary Fig. 31). Although the genetic correlation between the two subtypes was 0.80 ($P = 9.37 \times 10^{-35}$), the gene annotation of these lead SNPs revealed an obvious difference between the two subtypes. An S1-associated significant SNP at 3q27.2, rs66513933, is in the intron of *IGF2BP2*, and a high concentration is strongly associated with low T2DM risk.[33] Other risk loci in S1 at 10q25.2-q25.3 with three independent significant SNPs were in the intron of *TCF7L2*, the most potent locus for T2DM.[34] Almost all the genes associated with the lead SNPs of S2 were previously reported to be associated with T2DM, including *IGF2BP2* and *TCF7L2*. For example, *GCKR* is a hepatocyte-specific inhibitor of the glucose-metabolizing enzyme glucokinase[35]; *IRS1* plays a critical role in insulin-signalling pathways[36]; a paralogue of *ELF5A2* is associated with *T2DM*[37]; *CDKAL1* is involved in misfolded insulin, leading to oxidative and endoplasmic reticulum stress in pancreatic β cells[38]; *JAZF1* directly and negatively regulates insulin gene transcription[39]; a loss-of-function of *SLC30A8* protects against T2DM[40]; *HHEX* is repeatedly associated with T2DM[41]; *KCNQ1* is highly associate with the risk of T2DM[42]; ARAP1 is located near risk alleles for T2DM[43]; and *HMG20A* is key to the functional maturity of islet β cells.[44] The consistency between our results and these reported associations indicated the reliably of our GWAS results for the two subtypes. Meanwhile, the large difference between the GWAS of S1 and S2 showed the different genetic sources of the two subtypes, suggesting the rationality of our subtyping results with SuStaIn, which forms a foundation for the genetic association with brain health.

Genetic associations between preclinical-type 2 diabetes mellitus subtypes and brain health

We examined the genetic correlation between the two preclinical-T2DM subtypes and brain health using LDSC. Both subtypes showed significant genetic associations with various brain disorders, including depression disorder and stroke. Notably, S1 exhibited significant genetic associations with anorexia nervosa and schizophrenia, whereas S2 showed significant genetic associations with epilepsy (Fig. 6B). These subtype-specific genetic associations of different diseases reveal different genetic risks in brain disorders. For cognitive traits, we observed strong correlations between fluid intelligence and reasoning for both subtypes, alongside significant correlations between numeric memory and S1 (Supplementary Fig. 32), and between prospective memory and S2.

Next, we identified the shared causal variant between the two subtypes and brain disorders via Bayesian colocalization analyses. Our results revealed that S1 has significant colocalization with bipolar disorder at SNP rs9834970 [Fig. 6C; posterior probability of hypothesis 4 (PPH4) = 0.89]. These findings from genetic associations further indicate the genetic distinctions between the subtypes in relation to different brain disorders, highlighting the importance of considering subtype-specific genetic profiles in understanding the pathogenesis of these complex conditions.

Mendelian randomization for the preclinical-type 2 diabetes mellitus subtypes with brain disorders

In light of the robust associations uncovered through phenotypic and genetic analyses, we expanded our inquiry using two-sample MR analyses to explore the underlying causal link between preclinical-T2DM subtypes and brain disorders. By leveraging instrumental variables (IVs) derived from GWAS summaries of the two subtypes, we identified two IVs for S1 and 20 IVs for S2 (Supplementary Tables 12 and 13). Following correction for multiple testing using an FDR threshold of $P < 0.05$, we discerned significant causal relationships for both subtypes with stroke (Fig. 6D). Notably, several types of MR analysis supported the causal association between S2 and these diseases above, while only one MR analysis, inverse variance weighted, for S1, indicated there still were some differences between the causal relationship identified. Additionally, we conducted an MR Egger intercept test for S2 to assess the presence of horizontal pleiotropy as applicable, which could potentially bias the causal estimates. The results of the test did not indicate significant horizontal pleiotropy (MR Egger $P > 0.05$), suggesting that the MR findings for S2 are robust and unbiased (Supplementary Data 16).

Disease progression prediction using the identified preclinical-type 2 diabetes mellitus subtypes

Finally, we utilized survival analysis to assess the progression of various diseases in relation to the two preclinical-T2DM subtypes. Survival curves illustrated differences in disease progression relative to each subtype (Fig. 7). S1 exhibited faster progression and a higher risk for anxiety disorder, bipolar disorder, depression disorder, and sleep disorder compared with S2 (Fig. 7A–D). Conversely, S2 was associated with slightly faster progression in neurodegenerative disorders, such as Alzheimer's disease and Parkinson's disease, and a slightly higher risk for stroke (Fig. 7E and F and Supplementary Fig. 38). These findings align closely with the phenotypic associations previously identified, further revealing the distinct pathophysiological trajectories of each subtype. Of note, the three survival curves of S1, S2 and healthy controls exhibited an initial rapid decline followed by a slowdown towards the end of the follow-up period concerning most comorbidities (Fig. 7 and Supplementary Figs 33–38). This observation is primarily attributed to the rapid increase in censoring events, suggesting that the observed effects become less pronounced.

Moreover, we evaluated whether the preclinical-T2DM subtypes as biomarkers could enhance the accuracy of disease progression predictions. A baseline model using 18 clinical indicators was enhanced by incorporating information specific to each subtype. The predictive performance was evaluated using the Concordance Index (C-Index). Our results indicated that the inclusion of subtype information could enhance the prediction accuracy for both disease onset and its complications (Supplementary Fig. 39).

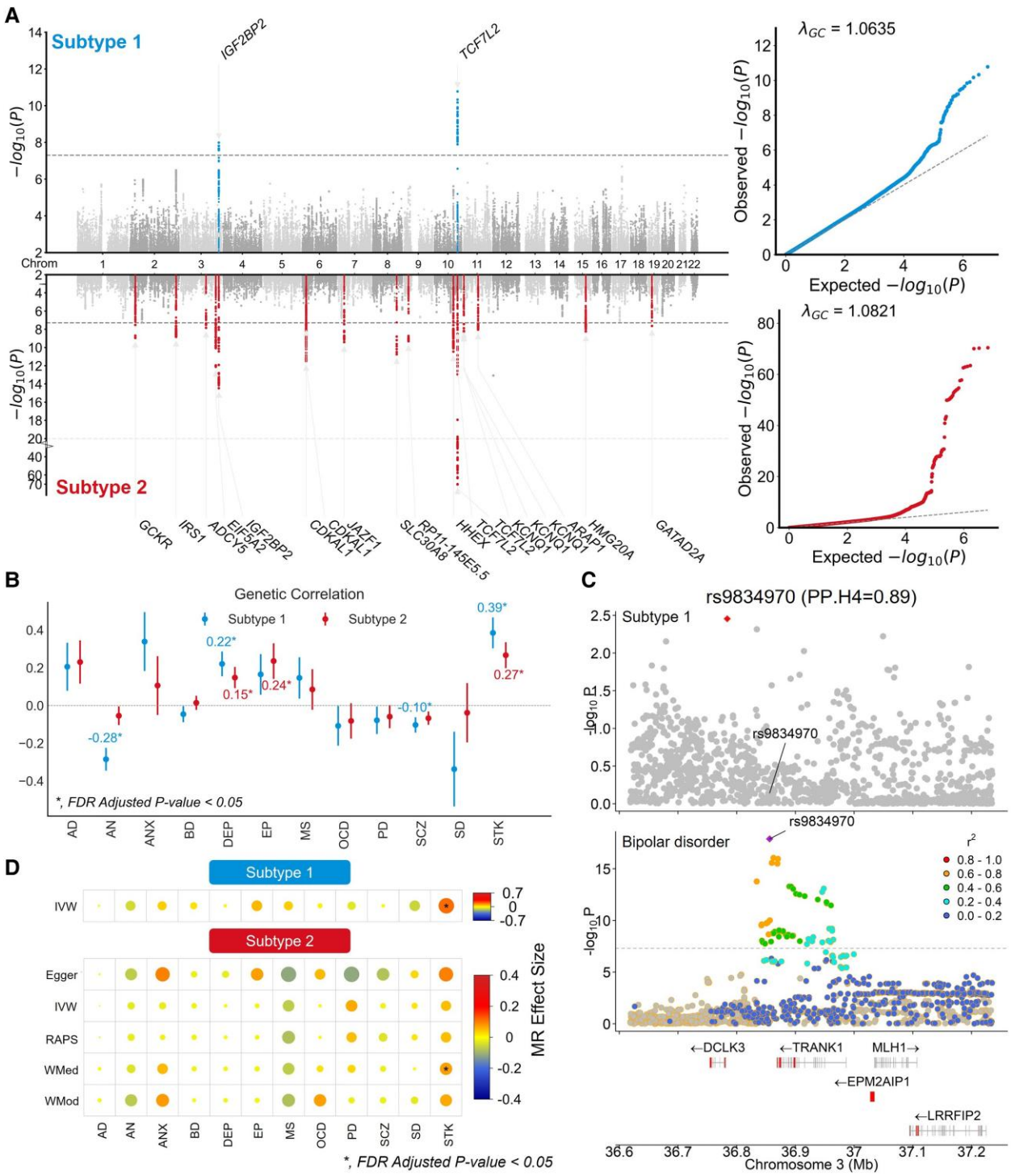

Figure 6 Genetic profiles and associations for the preclinical-type 2 diabetes mellitus subtypes with brain health. (**A**) The Miami and quantile-quantile (Q-Q) plot of genome-wide association study (GWAS) results for the two preclinical-type 2 diabetes mellitus (T2DM) subtypes. The dashed line indicates the significance level ($P < 10^{-08}$). Regions in a sidling window size of 500 kb around the lead single nucleotide polymorphisms (SNPs) are highlighted in the plot. Genes annotated for lead SNPs are marked in each region on the GWAS plot. The genomic control lambda (λ_{GC}) on the Q-Q plot is used to assess the degree of inflation in test statistics due to potential population stratification. A value of λ_{GC} close to 1 indicates no significant bias from population stratification. (**B**) Genetic correlations between subtypes and brain disorders. Results that passed the significance threshold adjusted by the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at the 5% level (FDR adjusted *P*-value < 0.05) are marked in the plot. (**C**) Significant colocalization results between subtypes and diseases (PPH4 > 0.75). (**D**) Genetic causal effects estimated by Mendelian randomization (MR) analyses of subtypes on brain disorders. We employed the inverse variance weighted (IVW) method for Subtype 1 with two SNPs as instrument variables and another four MR methods, MR Egger (Egger), MR-RAPS (RAPS), weighted median (WMed) and weighted mode (WMod) for Subtype 2 with 20 SNPs as instrument variables. Results that passed the significant threshold adjusted by the Benjamini-Hochberg procedure to control the FDR at the 5% level (FDR adjusted *P*-value < 0.05) are marked with asterisks. AD = Alzheimer's disease; AN = anorexia nervosa; ANX = anxiety disorder; BD = bipolar disorder; DEP = depression disorder; EP = epilepsy; MS = multiple sclerosis; OCD = obsessive compulsive disorder; PD = Parkinson's disease; SCZ = schizophrenia; SD = sleep disorder; STK = stroke.

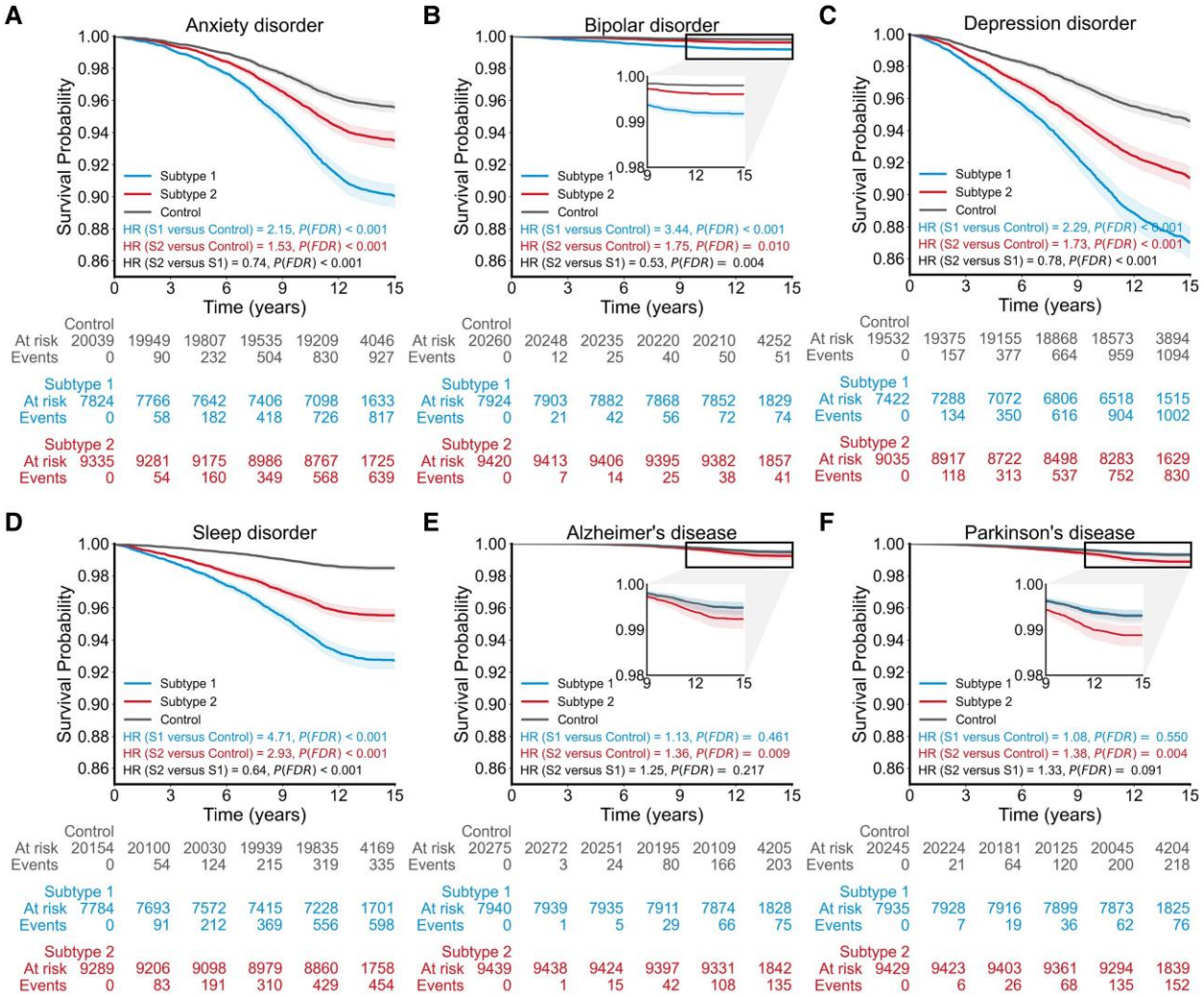

Figure 7 Survival curves for disease progression grouping by two preclinical-type 2 diabetes mellitus subtypes. We illustrate six selected diseases that demonstrated distinct progression rates over the years among the three groups: Subtype 1 (S1), Subtype 2 (S2) and the control group. Cox regression models were used to estimate the survival rates over time for each subtype, accounting for potential covariates such as age and sex, smoking status, alcohol drinking status, income level and education attainment. Survival curves of other diseases are illustrated in Supplementary Figs 32–37. (**A**–**D**) Diseases that showed a faster progression rate in S1 compared with S2 and the control group. (**E** and **F**) Diseases that showed a faster progression rate in S2 compared with S1 and the control group. The tables below the survival curves present the number of subjects currently at risk (not progressed to the specific disease or censored) and the cumulative number of subjects who had an event (progressed to the specific disease) for each group, corresponding to time (years) (x-axis).

Discussion

In this study, we uncovered the distinct subtypes and stages of preclinical-T2DM by using a cohort of 20 277 UK Biobank participants. Utilizing the machine learning algorithm SuStaIn,[8] we demonstrated that the heterogeneous progression of preclinical-T2DM can be delineated by two distinct trajectories. Both subtypes exhibited different illness durations, biomarker profiles, brain health and signatures.

Our analysis confirms the robustness and distinctiveness of the two identified subtypes. To verify the stability of our results, we applied a more stringent cutoff on subtype assignment and excluded subjects with a subtype assignment probability of less than 60%. We observed that the metabolic profiles of the two subtypes remained consistent, with no notable changes observed (Supplementary Table 4). Furthermore, we tested whether the two subtypes could be replicated using only median values of HbA1c and BMI. The results showed that using these two indicators alone failed to capture the differences between the subtypes across other biomarkers (Supplementary Tables 5 and 6). We also showed that there was no significant negative correlation between HbA1c and BMI in the preclinical-T2DM population (Supplementary Fig. 18). These finding further suggests that the two subtypes are not simply defined by 'BMI' or 'glycaemic' profiles but are associated with a variety of biomarkers, reflecting more complex underlying mechanisms.

S1, characterized by a higher leptin and lower leptin receptor phenotype, demonstrates elevated levels of BMI, alanine aminotransferase, LDLc and C-reactive protein. S1 had the highest circulating leptin and lowest circulating leptin receptors among the healthy control and participants with preclinical-T2DM (Fig. 4H and I). Subjects with leptin resistance have also slightly improved LDLc levels in the blood (Fig. 2H), which is revealed in S1. It is worth noting that leptin-deficiency is the main cause of massive obesity because of both hyperphagia and decreased energy expenditure.[45-47] Leptin plays a crucial

role in regulating insulin synthesis and secretion from pancreatic β cells, thereby influencing insulin sensitivity, hepatic glucose production and glucagon levels.[48,49] Previous studies have indicated a correlation between leptin resistance and obesity, abnormal cholesterol levels and heightened risks of psychiatric disorders.[50-52] Our findings not only confirmed these associations but also shed light on the pivotal role of leptin in the initiation and advancement of T2DM. On the contrary, S2 showed higher levels of HbA1c and glucose (Fig. 2D and Table 1). Notably, S2 exhibited a higher proportion of males and a shorter time interval to the onset of T2DM than S1 (Table 1). Additionally, the biomarker trajectories differed between the subtypes (Fig. 2E–J and Supplementary Figs 5–16). For instance, glucose and HbA1c levels escalated more rapidly for S2 than S1 over time (Fig. 2E and F), whereas BMI levels increased more swiftly for S1 than S2 during the progression of both subtypes (Fig. 2G). These factors may be caused by a reduced capacity for insulin utilization in participants with S1. However, it is intriguing to note that the progression curve for the triglyceride-glucose index in S1 initially exceeded that in S2, but the latter rose rapidly and eventually surpassed the former during the progression of preclinical-T2DM (Fig. 2I). This suggested that S1 may not represent a typical insulin-resistant form of T2DM. From a genetic standpoint, S2 manifested more significant genetic associations than S1, with noticeably stronger signals (Fig. 6A). Several identified genetic loci, such as *GCKR* and *IGFI*, have been identified as associated with insulin insensitivity.[53,54] These findings suggested that, while S2 is a T2DM genetic-relevant subtype, S1 is more complex and based on multiple traits that may share common upstream clinical determinants, e.g. leptin resistance. These findings demonstrated the biological plausibility of distinct subtypes of preclinical-T2DM.

We found preclinical-T2DM influenced the structures and functions of the brain. The cumulative damage associated with diabetes can affect brain health, and we propose that the early susceptibility of neurobiological structures to metabolic stresses may facilitate the early onset of this damage. Consequently, these structural changes in the brain may result in functional impairments. To quantify the associations between preclinical-T2DM and brain health, we analysed brain MRI data from 935 individuals in the UKB. A few distinct brain regions were affected by each subtype. For example, participants with S1 exhibited slightly more atrophy in the putamen and accumbens regions, which are primarily associated with emotional regulation and motor control. Atrophy in these regions was associated with higher risks of anxiety and depression disorders.[55,56] In contrast, participants with S2 showed slightly more atrophy in regions such as the fusiform gyrus, superior temporal lobe, superior parietal lobe and middle temporal lobe, which play an important role in language comprehension, visual information processing and memory functions.[57,58] Atrophy in these regions is more closely associated with cognitive decline. Furthermore, the reduction in cortical thickness and white matter function was slightly more pronounced in S2 (Fig. 5C and D and Supplementary Fig. 29). The atrophy pattern observed in S2 may contribute to the slightly increased risk of dementia.[59-61] Our findings suggested that structural changes in the brain could potentially increase the risk of brain disorders. While our study specifically focused on the connections between preclinical-T2DM and the brain, it is also possible to explore the relationships between preclinical-T2DM and other human organs and systems. Conducting a multisystem analysis using biobank-scale data may provide insights into interorgan pathophysiological mechanisms and assist in the prevention of T2DM and the early detection of its effects on human organs.

It is widely recognized that individuals exhibit worsening symptoms and face an increased risk of T2DM onset and its complications over the course of the illness. Of note, the observed increases in glucose and HbA1c levels were consistent with an increased risk of stroke. This finding aligned with existing knowledge regarding the crucial role of glucose in cerebrovascular functions and its causal link with elevated risks of vascular diseases. In addition, both subtypes were correlated with a decline in cognitive function, a trend mirrored in alterations to brain structure (Fig. 4A–F). Studies have indicated that the probability of T2DM patients experiencing a decline in cognitive abilities is 1.5–2.0 times higher than that of non-diabetic individuals.[58,62,63] However, despite the occurrence of cognitive function impairment during the preclinical stage of T2DM and the irreversible damage to the brain upon diabetes progression, the adverse effects of preclinical-T2DM on cognitive function are often underestimated, even when clinical symptoms are mild or asymptomatic. Furthermore, it is intriguing to note that both subtypes showed differential risks of T2DM and its complications (Figs 3 and 7 and Supplementary Figs 33–38). For instance, participants with S1 exhibited higher hazard ratios than those with S2 for psychiatric disorders, including depression disorder, anxiety disorder, bipolar disorder and sleep disorder. This supports the emerging notion that the presence of leptin may serve as a potential indicator of neurotransmitter alterations, subsequently impacting the psychiatric status of individuals with preclinical-T2DM S1. Intriguingly, colocalization analysis revealed that genes involved in loci associated with S1 were correlated with bipolar disorder (Fig. 6C). These findings may shed light on why metabolic exposure generally accelerates brain tissue loss in conditions such as bipolar disorder, depression disorder or other psychiatric disorders, despite being a complex neurobiological process. Conversely, compared with S1, participants with S2 were associated with a slightly increased risk of developing neurodegenerative disorders such as Alzheimer's disease and Parkinson's disease. Over the past decade, accumulating evidence has suggested a positive association between T2DM and dementia.[64-67] Our results indicated that this link may exist within a subgroup of preclinical-T2DM patients (S2). For individuals with S1, there were no discernible differences between them and healthy controls regarding the risks of Alzheimer's disease and Parkinson's disease (Fig. 7E and F). Although MR analyses suggested no causal associations between Alzheimer's disease/Parkinson's disease and preclinical-T2DM (Fig. 6D), our efforts to identify disease subtypes and their associations with brain disorders were conceptual and predominantly relied on accumulating evidence regarding the progression of preclinical-T2DM and its impacts on the brain, which share biological mechanisms with brain disorders. This could potentially aid in the development of cost-effective health promotion strategies tailored to this extensive and vulnerable population.

The identification of distinct subtypes of preclinical-T2DM opens avenues for personalized disease screening and prevention, ultimately leading to improved patient care and outcomes. It is crucial to acknowledge that the subtypes and stages we identified help to delineate the heterogeneity of preclinical-T2DM and link them to specific treatments, suggesting that predicting clinical outcomes could benefit from stratification based on the biological subtypes of preclinical-T2DM. To this end, the development of a classifier cluster comprising specific subgroups corresponding to each subtype demonstrates enhanced performance in predicting the onset of T2DM and its complications compared with the model based solely on clinical information. Each subtype exhibited unique clinical characteristics and impacts on the brain, underscoring the importance of tailored approaches in disease management. Previous studies have also suggested the potential benefits of disease risk prediction for certain phenotypes of patients based on specific phenotypic or

genetic features.[9,68,69] While further investigation is warranted on the biological mechanisms of the progression of preclinical-T2DM from the interorgan perspective, factors such as increases in HbA1c and glucose levels, as well as leptin resistance, have consistently shown associations with disease onset and brain health. Embracing a perspective of stratified prediction models may unveil the underlying progressive heterogeneity of the disease and facilitate the adoption of more individualized treatment approaches in clinical practice, which holds promise for optimizing patient care, enhancing treatment efficacy and ultimately mitigating the burden of T2DM and its complications on individuals and healthcare systems alike.

The potential clinical impact of our study is multifaceted. Broadly, it aids in dissecting the heterogeneity of preclinical-T2DM into more defined metabolic subtypes, with implications for downstream tasks. For instance, establishing robust preclinical-T2DM subtypes can enhance the accuracy of individualized disease diagnosis and prognosis. Furthermore, modelling preclinical-T2DM heterogeneity offers novel patient stratification and treatment assessment tools for future clinical trials, which are particularly crucial, given the mixed results and clinical limitations of glucose treatments. Recognizing that assessing treatment responses within more homogeneous patient subgroups can significantly enhance the efficacy of clinical trials, our findings suggest that preclinical-T2DM subtyping and staging could improve the ability to identify significant clinical characteristics of T2DM and the differential connections with neurodegenerative diseases and psychiatric disorders, which might otherwise be diluted in case-control comparisons due to underlying heterogeneity. Lastly, the identified subtypes, being both phenotypically and genetically relevant to the brain, serve as reliable prognostic biomarkers, thereby facilitating the risk prediction of brain disorders.

There are strengths as well as limitations to this study. Firstly, while the SuStaIn algorithm offers estimates of preclinical-T2DM trajectories based on cross-sectional clinical indexes, it is crucial to validate these findings using longitudinal data to authenticate the disease progressions over time. Secondly, we identified two distinct subtypes from the UKB dataset, each exhibiting different clinical phenotypes, neuroanatomical signatures and clinical outcomes. Validation using independent discovery and replication cohorts would bolster the reliability of these identified subtypes. With the availability of large-scale cohorts in studies, there is potential to replicate and validate our findings, particularly in contextualizing the proposed subtypes with the brain connectivity, cytoarchitecture, metabolism, neurotransmitter receptors and transporters, gene expression and cognitive function. Thirdly, although our study indicated a connection between preclinical-T2DM S1 and psychiatric disorders, the causality remains unclear and deserves further investigation. Individuals with psychiatric disorders may exhibit abnormal metabolic levels, potentially due to the poor lifestyle factors associated with the disorder.[70,71] Additionally, these conditions may not have been diagnosed at the time of blood collection, potentially confounding the interpretation of the relationship between preclinical T2DM and psychiatric disorders. Finally, while clinical biomarkers and genetics can influence the progression of individuals within the preclinical stage of T2DM, the risk factors characterizing the clusters identified in our study were also shaped by behavioural, environmental and dietary determinants, as well as the use or non-use of medications that lower risk factor levels. Future research integrating these determinants with clinical data is necessary to understand their contributions to the prevalence and trends in preclinical-T2DM subtypes and their impact on brain health.

In summary, our study identified two distinct and stable subtypes of preclinical-T2DM based on cross-sectional clinical data encompassing 18 clinical indexes. These subtypes demonstrated varied associations with brain health. Untangling the underlying mechanisms of abnormal glucose and lipid metabolism implicated in neurodegenerative diseases and psychiatric disorders requires further fundamental and experimental research. Nevertheless, the results clearly demonstrate the critical necessity of monitoring and addressing the brain health needs of individuals in the preclinical stage of T2DM.

Data availability

The data used and generated in this study are provided in the Supplementary material. The phenotypic, genotypic, proteomic and metabolic data used in the study that support subtype and stage modelling and association analyses were obtained from the UK Biobank under application number 89757. Access to the UK Biobank data is available to all researchers with approval (https://www.ukbiobank.ac.uk/enable-your-research/register). Statistical details of the GWAS datasets and download links for all the datasets are available in Supplementary Table 11.

The original source code for the implementation of the SuStaIn algorithm is available on the GitHub repository at https://github.com/ucl-pond/pySuStaIn. The source codes pertaining to this study and data analysis in this manuscript are provided at https://github.com/ZJU-BMI/disease-progression-preclinical-T2DM.

Acknowledgements

The UK Biobank resource was used under application number 89757.

Funding

This work was partially supported by the National Key Research and Development Program of China under Grant No. 2022YFF1202400 and the National Nature Science Foundation of China under Grant No. 82272129.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Sun H, Saeedi P, Karuranga S, et al. IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract*. 2022;183:109119.
2. Kumar A, Gangwar R, Zargar AA, Kumar R, Sharma A. Prevalence of diabetes in India: A review of IDF diabetes atlas 10th edition. *Curr Diabetes Rev*. 2024;20:e130423215752.
3. Galicia-Garcia U, Benito-Vicente A, Jebari S, et al. Pathophysiology of type 2 diabetes mellitus. *Int J Mol Sci*. 2020;21:6275.
4. Garcia-Serrano AM, Duarte JMN. Brain metabolism alterations in type 2 diabetes: What did we learn from diet-induced diabetes models? *Front Neurosci*. 2020;14:229.

5. Xourafa G, Korbmacher M, Roden M. Inter-organ crosstalk during development and progression of type 2 diabetes mellitus. *Nat Rev Endocrinol*. 2024;20:27-49.
6. Zheng R, Xu Y, Li M, et al. Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults. *Cell Rep Med*. 2023;4:100958.
7. Young AL, Oxtoby NP, Garbarino S, et al. Data-driven modelling of neurodegenerative disease progression: Thinking outside the black box. *Nat Rev Neurosci*. 2024;25:111-130.
8. Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nat Commun*. 2018;9:4273.
9. Tao P, Conarello S, Wyche TP, et al. Metabolomics and lipidomics analyses aid model classification of type 2 diabetes in non-human primates. *Metabolites*. 2024;14:159.
10. Jiang Y, Li W, Li J, et al. Identification of four biotypes in temporal lobe epilepsy via machine learning on brain images. *Nat Commun*. 2024;15:2221.
11. Chen D, Wang X, Voon V, et al. Neurophysiological stratification of major depressive disorder by distinct trajectories. *Nat Mental Health*. 2023;1:863-875.
12. Jiang Y, Wang J, Zhou E, et al. Neuroimaging biomarkers define neurophysiological subtypes with distinct trajectories in schizophrenia. *Nat Mental Health*. 2023;1:186-199.
13. Xiao F, Caciagli L, Wandschneider B, et al. Identification of different MRI atrophy progression trajectories in epilepsy by subtype and stage inference. *Brain*. 2023;146:4702-4716.
14. Cholerton B, Baker LD, Montine TJ, Craft S. Type 2 diabetes, cognition, and dementia in older adults: Toward a precision health approach. *Diabetes Spectr*. 2016;29:210-219.
15. Aksman LM, Wijeratne PA, Oxtoby NP, et al. Pysustain: A Python implementation of the subtype and stage inference algorithm. *SoftwareX*. 2021;16:100811.
16. Bycroft C, Freeman C, Petkova D, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562:203-209.
17. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8:1826.
18. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47:291-295.
19. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. 2014;10:e1004383.
20. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol*. 2020;82:1273-1300.
21. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet*. 2021;17:e1009440.
22. Zheng J, Haberland V, Baird D, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat Genet*. 2020;52:1122-1131.
23. Zhang Y, Tan H, Tang J, et al. Effects of vitamin D supplementation on prevention of type 2 diabetes in patients with prediabetes: A systematic review and meta-analysis. *Diabetes Care*. 2020;43:1650-1658.
24. Geng L, Lam KSL, Xu A. The therapeutic potential of FGF21 in metabolic diseases: From bench to clinic. *Nat Rev Endocrinol*. 2020;16:654-667.
25. Suriano F, Vieira-Silva S, Falony G, et al. Novel insights into the genetically obese (ob/ob) and diabetic (db/db) mice: Two sides of the same coin. *Microbiome*. 2021;9:147.
26. Bungau S, Behl T, Tit DM, et al. Interactions between leptin and insulin resistance in patients with prediabetes, with and without NAFLD. *Exp Ther Med*. 2020;20:197.
27. Mazor R, Friedmann-Morvinski D, Alsaigh T, et al. Cleavage of the leptin receptor by matrix metalloproteinase-2 promotes leptin resistance and obesity in mice. *Sci Transl Med*. 2018;10:eaah6324.
28. Friedman JM. Leptin and the endocrine control of energy balance. *Nat Metab*. 2019;1:754-764.
29. Ahima RS, Flier JS. Leptin. *Annu Rev Physiol*. 2000;62:413-437.
30. Friedman J. The long road to leptin. *J Clin Invest*. 2016;126:4727-4734.
31. Rajwani A, Ezzat V, Smith J, et al. Increasing circulating IGFBP1 levels improves insulin sensitivity, promotes nitric oxide production, lowers blood pressure, and protects against atherosclerosis. *Diabetes*. 2012;61:915-924.
32. Herder C, Nuotio ML, Shah S, et al. Genetic determinants of circulating interleukin-1 receptor antagonist levels and their association with glycemic traits. *Diabetes*. 2014;63:4343-4359.
33. Wittenbecher C, Ouni M, Kuxhaus O, et al. Insulin-like growth factor binding protein 2 (IGFBP-2) and the risk of developing type 2 diabetes. *Diabetes*. 2019;68:188-197.
34. Del Bosque-Plata L, Martínez-Martínez E, Espinoza-Camacho MÁ, Gragnoli C. The role of TCF7L2 in type 2 diabetes. *Diabetes*. 2021;70:1220-1228.
35. Fernandes Silva L, Vangipurapu J, Kuulasmaa T, Laakso M. An intronic variant in the GCKR gene is associated with multiple lipids. *Sci Rep*. 2019;9:10240.
36. Kovacs P, Hanson RL, Lee YH, et al. The role of insulin receptor substrate-1 gene (IRS1) in type 2 diabetes in Pima Indians. *Diabetes*. 2003;52:3005-3009.
37. Mastracci TL, Colvin SC, Padgett LR, Mirmira RG. Hypusinated eIF5A is expressed in the pancreas and spleen of individuals with type 1 and type 2 diabetes. *PLoS One*. 2020;15:e0230627.
38. Ghosh C, Das N, Saha S, Kundu T, Sircar D, Roy P. Involvement of Cdkal1 in the etiology of type 2 diabetes mellitus and microvascular diabetic complications: A review. *J Diabetes Metab Disord*. 2022;21:991-1001.
39. Kobiita A, Godbersen S, Araldi E, et al. The diabetes gene JAZF1 is essential for the homeostatic control of ribosome biogenesis and function in metabolic stress. *Cell Rep*. 2020;32:107846.
40. Flannick J, Thorleifsson G, Beer NL, et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet*. 2014;46:357-363.
41. Zhang J, McKenna LB, Bogue CW, Kaestner KH. The diabetes gene Hhex maintains δ-cell differentiation and islet function. *Genes Dev*. 2014;28:829-834.
42. Erfani T, Sarhangi N, Afshari M, Abbasi D, Meybodi HRA, Hasanzad M. KCNQ1 common genetic variant and type 2 diabetes mellitus risk. *J Diabetes Metab Disord*. 2020;19:47-51.
43. Li L, Xu L, Wen S, Yang Y, Li X, Fan Q. The effect of lncRNA-ARAP1-AS2/ARAP1 on high glucose-induced cytoskeleton rearrangement and epithelial-mesenchymal transition in human renal tubular epithelial cells. *J Cell Physiol*. 2020;235:5787-5795.
44. Mellado-Gil JM, Fuente-Martín E, Lorenzo PI, et al. The type 2 diabetes-associated HMG20A gene is mandatory for islet beta cell functional maturity. *Cell Death Dis*. 2018;9:279.
45. Xu J, Bartolome CL, Low CS, et al. Genetic identification of leptin neural circuits in energy and glucose homeostases. *Nature*. 2018;556:505-509.
46. Clément K, van den Akker E, Argente J, et al. Efficacy and safety of setmelanotide, an MC4R agonist, in individuals with severe obesity due to LEPR or POMC deficiency: Single-arm, open-label,

multicentre, phase 3 trials. *Lancet Diabetes Endocrinol*. 2020;8: 960-970.
47. Takahashi K, Yamada T, Hosaka S, et al. Inter-organ insulin-leptin signal crosstalk from the liver enhances survival during food shortages. *Cell Rep*. 2023;42:112415.
48. Pereira S, Cline DL, Glavas MM, Covey SD, Kieffer TJ. Tissue-specific effects of Leptin on glucose and lipid metabolism. *Endocr Rev*. 2021;42:1-28.
49. Duquenne M, Folgueira C, Bourouh C, et al. Leptin brain entry via a tanycytic LepR-EGFR shuttle controls lipid metabolism and pancreas function. *Nat Metab*. 2021;3:1071-1090.
50. Changchien TC, Tai CM, Huang CK, Chien CC, Yen YC. Psychiatric symptoms and leptin in obese patients who were bariatric surgery candidates. *Neuropsychiatr Dis Treat*. 2015;11:2153-2158.
51. Dallner OS, Marinis JM, Lu YH, et al. Dysregulation of a long noncoding RNA reduces leptin leading to a leptin-responsive form of obesity. *Nat Med*. 2019;25:507-516.
52. Zeng Q, Song J, Sun X, et al. A negative feedback loop between TET2 and leptin in adipocyte regulates body weight. *Nat Commun*. 2024;15(1):2825.
53. Kimura M, Iguchi T, Iwasawa K, et al. En masse organoid phenotyping informs metabolic-associated genetic susceptibility to NASH. *Cell*. 2022;185:4216-4232.e16.
54. Barbieri M, Bonafè M, Franceschi C, Paolisso G. Insulin/IGF-I-signaling pathway: An evolutionarily conserved mechanism of longevity from yeast to humans. *Am J Physiol Endocrinol Metab*. 2003;285:E1064-E1071.
55. Talati A, van Dijk MT, Pan L, et al. Putamen structure and function in familial risk for depression: A multimodal imaging study. *Biol Psychiatry*. 2022;92:932-941.
56. Lin H, Bruchmann M, Straube T. Altered putamen activation for social comparison-related feedback in social anxiety disorder: A pilot study. *Neuropsychobiology*. 2023;82:359-372.
57. Xie W, Bainbridge WA, Inati SK, Baker CI, Zaghloul KA. Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nat Hum Behav*. 2020;4:937-948.
58. Ennis GE, Saelzler U, Umpierrez GE, Moffat SD. Prediabetes and working memory in older adults. *Brain Neurosci Adv*. 2020;4: 2398212820961725.
59. Migliaccio R, Cacciamani F. The temporal lobe in typical and atypical Alzheimer disease. *Handb Clin Neurol*. 2022;187:449-466.
60. Ma D, Fetahu IS, Wang M, et al. The fusiform gyrus exhibits an epigenetic signature for Alzheimer's disease. *Clin Epigenetics*. 2020;12:129.
61. Vogel JW, Young AL, Oxtoby NP, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. 2021;27:871-881.
62. Parashar A, Mehta V, Malairaman U. Type 2 diabetes mellitus is associated with social recognition memory deficit and altered dopaminergic neurotransmission in the amygdala. *Ann Neurosci*. 2018;24:212-220.
63. Koekkoek PS, Kappelle LJ, van den Berg E, Rutten GE, Biessels GJ. Cognitive function in patients with diabetes mellitus: Guidance for daily care. *Lancet Neurol*. 2015;14:329-340.
64. Biessels GJ, Despa F. Cognitive decline and dementia in diabetes mellitus: Mechanisms and clinical implications. *Nat Rev Endocrinol*. 2018;14:591-604.
65. Ehtewish H, Arredouani A, El-Agnaf O. Diagnostic, prognostic, and mechanistic biomarkers of diabetes mellitus-associated cognitive decline. *Int J Mol Sci*. 2022;23:6144.
66. Du H, Meng X, Yao Y, Xu J. The mechanism and efficacy of GLP-1 receptor agonists in the treatment of Alzheimer's disease. *Front Endocrinol (Lausanne)*. 2022;13:1033479.
67. Zhou Y, Dong J, Song J, Lvy C, Zhang Y. Efficacy of glucose metabolism-related indexes on the risk and severity of Alzheimer's disease: A meta-analysis. *J Alzheimers Dis*. 2023;93: 1291-1306.
68. Zhao B, Li T, Fan Z, et al. Heart-brain connections: Phenotypic and genetic insights from magnetic resonance images. *Science*. 2023;380:abn6598.
69. Zhang J, Zhan J, Jin J, et al. An ensemble penalized regression method for multi-ancestry polygenic risk prediction. *Nat Commun*. 2024;15:3238.
70. Penninx BWJH, Lange SMM. Metabolic syndrome in psychiatric patients: Overview, mechanisms, and implications. *Dialogues Clin Neurosci*. 2018;20:63-73.
71. Chaput JP, McHill AW, Cox RC, et al. The role of insufficient sleep and circadian misalignment in obesity. *Nat Rev Endocrinol*. 2023; 19:82-97.