Longitudinal Structural Brain Changes in Bipolar Disorder: A Multicenter Neuroimaging Study of 1,232 Individuals by the ENIGMA Bipolar Disorder Working Group

Supplement 1

Supplemental Methods and Materials

Participating sites and cohort characteristics

All studies used standard diagnostic instruments, including SCID, MINI, and/or DIGS. Most studies (N=13) included bipolar type I (BDI). One study included only bipolar type II (BDII). Seven included both BDI and BDII disorders, and 5 included only BDI. Two studies were first manic episodes studies. One study was a medication trial. One study provided only patient data, and one study provided only HC data. Most studies did not exclude comorbidities or specific medication use. Hence, the total sample, collected from 9 countries and 4 continents, represents an ecologically valid and a generalizable representation of BD patients in clinical treatment.

The time elapsed between scans ranged from 0.5 – 9.8 years (mean 2.6, SD 1.6 years), which was accounted for in statistical analyses by computing yearly change rates. We excluded scans with an inter-scan interval below 0.5 years not only to allow sufficient time to capture recurrent mood episodes and associated brain alterations, but also to minimize potential inaccuracies in yearly change computations for shorter time intervals. Further, this study included participants who were scanned at least twice.

MRI processing

The protocol used was adapted from the ENIGMA plasticity Working Group (1), which was slightly modified to allow for individual time point processing and inspection. The standardized protocol is publicly available online (http://enigma.ini.usc.edu/protocols/) in order to foster open science and replication. FreeSurfer (2-6) was used on-site to perform cortical reconstructions and subcortical segmentations at each imaging time point. All images were first processed cross-sectionally and then with the longitudinal stream implemented in FreeSurfer v5.1 or higher (7). Specifically, an unbiased within-subject template space and image was created to overcome asymmetry related

processing bias using inverse consistent registration methods. Following processing steps were initialized with common information from the within-subject template, which increase accuracy and statistical power (7). FreeSurfer versions varied across but not within sites (see Table S4 for versions applied). Although previous analyses from the ENIGMA-BD Working Group showed that scanner field strength, voxel volume, and FreeSurfer version did not significantly influence effect size estimates (8), we controlled for 'imaging site' to account for any potential site-related variations, as typically done in multi center studies.

After longitudinal processing, quality control was performed on a region of interest (ROI) level for each time point image aided by a visual inspection guideline that included pass/fail segmentation examples. ROIs failing quality inspection were excluded from subsequent analyses.

Coding of demographic and clinical variables

Age, BMI, age of onset, and number of mood episodes were continuous variables. Smoking status referred to current status at scan day and was a binary variable coding for non-smoker (0) or smoker (1). The following medication types were investigated: lithium, antipsychotic, antidepressant, and antiepileptic drugs. Medication use variables were binary coding for using (1) or not using (0) the respective medication type. Psychiatric comorbidity was coded as binary for having (1) or not having (0) a certain comorbid diagnosis. The same applied for history of psychotic symptoms. Psychiatric comorbidities investigated were general anxiety disorder (GAD), obsessive compulsive disorder (OCD), attention deficit hyperactivity disorder (ADHD), posttraumatic stress disorder (PTSD), panic disorder, (social) phobia, and eating disorders. Alcohol and substance use variables were 3-level factors indicating "no risk for abuse/dependence", "abuse", or "dependence". One center provided Alcohol Use Disorders Identification Test (AUDIT) data, which was converted and assigned to one of the aforementioned categories using the AUDIT score ranges: 1-7 = no risk, 8-14 = abuse, >15 = dependence. Current mood state was a 4-level factor coding for euthymic, depressed, manic, or mixed mood state at scan day. Ethnicity was coded as "White", "Black", "Asian", or "other". As different sites provided different measures for education (e.g., years versus categories), for all centers, education level

was grouped into the following categories (1: <= 10 years, 2: 11-13 years, 3: 14-16 years, 4: >16 years). Bipolar subtype was a 3-level factor coding for bipolar "type 1" (BDI), "type 2" (BDII), or "NOS/other".

Sensitivity analysis testing for potential confounders

Main analysis. We tested whether the observed group differences in yearly change rates were affected by demographic or clinical variables (listed in Table 1) including including BMI, intracranial volume (ICV), education level, smoking status, medication use, and psychiatric comorbidity (at each TP) by entering those variables separately as additional covariates in the main model. Multi-level factors were added as random factors. Variables were tested one at a time to increase sample size in sensitivity tests, as not all centers/participants provided the respective data tested. For comorbidity data, only data at TP1 was tested as covariate since almost 50% of patients' comorbidity data was not available at TP2. Similarly, smoking information was available for only 50% of participants. Instead of using it as covariate, we tested the potential effects of smoking by comparing yearly change rates between known smokers and non-smokers in the combined cohort while correcting for case-control status. Furthermore, we repeated the main analysis when excluding individuals with a specific psychiatric comorbidity (at each TP) and only in people self-identified 'white' ethnic background to test for potential effects by ancestry. To test if the results were affected by a larger age range in HC and the presence of adolescents, we further matched groups on age range by excluding individuals younger than 18 and older than 75 years. Within patients, we also correlated yearly change rates with age of disease onset and compared change rates between bipolar subtypes BDI and BDII. Results of sensitivity analyses are provided in supplementary Data S1.

Moreover, we compared change rates between patients with and without a history of psychotic symptoms, and tested for the effects of mood state on yearly change rates. Finally, since lithium use has been associated with increased cortical volume (9-12) and antipsychotic drug use with cortical decline (13) we explored potential medication effects on yearly change rates by comparing patients using and not using a specific medication type.

Correlations between change rates and manic episodes between time points. The correlation analyses were repeated after regressing out the effects of age, sex, imaging-site, and/or the number of depressive episodes between time points to test if the observed correlations hold when correcting for these factors (Data S2). Given the observed effects on change rates of FGA use and history of psychosis in patients, we also tested if the correlations between thickness changes and number of mood episodes remained when controlling for these factors. We also repeated these correlation analyses when excluding outliers in Figures S10-S11 and when excluding the SBP Stockholm cohort, which previously showed associations between cortical decline and manic episodes (14) to ensure that the correlations observed were not driven by this center. As it is yet unclear whether or not the occurrence of a manic episode precedes cortical decline, we further re-ran correlation analyses after excluding the STOP-EM cohort, which was a first episode mania cohort, and after excluding both the Stockholm and the STOP-EM cohort. The second first episode cohort (FEMS Melbourne) was not included in these correlation analyses as no information on the exact number of mood episodes was available. Given the previous findings that cortical decline related to mania in BDI and hypomania in BDII (14, 15), we further explored the associations between yearly change rates and manic episodes within BDI as well as hypomanic episodes in BDII. The total sample size for correlation analyses was n=230, but varied slightly depending on mood episode measure and brain region under investigation (Data S2).

Supplemental Results

Effects of demographic and clinical variables (sensitivity tests)

The main results were not affected when controlling for ethnicity, BMI, ICV, medication use, or psychiatric comorbidities; nor did results change when excluding individuals with specific comorbidities, reported alcohol or drug use (see Data S1 for results of all sensitivity tests).

The statistical significance of group differences weakened for changes in fusiform (p=0.038) and parahippocampal thickness (p=0.090) when controlling for education. Of note, however, education may relate to the phenotype of interest (16). Further, education was not a significant predictor in the model ($0.218 < p_{education} < 0.482$) and sample size was

reduced when controlling for it. Results corrected for education should hence be treated with caution. Further, there were no differences in yearly change rates between smokers and non-smokers, and the results remained when only analyzing individuals with "white" ethnic background.

Antiepileptic drug (AE) use at TP2 significantly predicted right lateral ventricle change rates ($p_{\_AE} < 0.005$), and controlling for AE use weakened group differences ($p_{\_left\_ventr}=0.032$ and $p_{\_right\_ventr} = 0.085$). First generation antipsychotic drug (FGA) use at TP1 significantly predicted right fusiform thickness changes ($p_{FGA} = 0.009$) and ventricular volume changes (left and right: $p_{FGA} <0.001$), but controlling for FGA did not affect group differences. Second generation antipsychotic (SGA) use at TP2 significantly predicted both left ($p_{\_SGA} = 0.007$) and right ($p_{\_SGA} = 0.0039$) lateral ventricle change rates, and group differences were slightly less significant when controlling for it ($p = 0.027$ for left ventricle; $p=0.033$ for right ventricle). However, the use of SGA or AE at either time point was not associated with fusiform/parahippocampal thickness or bilateral ventricle volume change rates. FGA at TP1 was associated with larger increases in bilateral ventricular volume ($p_{left}=0.004$, $p_{right}=0.014$) and faster decrease in right fusiform thickness ($p<0.001$) in patients (note only 14 patients used FGA). Similarly, history of psychosis (at TP1) was related to faster decline in right parahippocampal thickness ($p=0.035$). There were no differences associated with the use of lithium or antidepressants at either TP, nor with the use of FGA at TP2. Furthermore, there were no differences related to bipolar subtype, with the exception that BDI patients showed a decline in right parahippocampal thickness, whereas BDII patients showed thickness increases in the same region (BDI versus BDII mean difference: $p=0.010$; Data S1). Neither age of onset nor mood states did significantly predict yearly change rates. See Supplementary Data S1 for statistical details of all sensitivity tests.

Differential change rate vs. mood episode correlations in BDI and BDII

In exploratory analyses, we observed differential correlations in BD subtypes. Whereas correlations with manic episodes in BDI were mainly observed in prefrontal regions, correlations with hypomanic episodes in BDII were observed mainly in lingual and temporal regions (results shown in Data S2).

Table S1: Descriptive statistics of included samples. Abbreviations: BDI (bipolar disorder type 1), BDII (bipolar disorder type 2), AE (antiepileptics), FGA (first generation antipsychotics), SGA (second generation antipsychotics), AD (antidepressants), ME (manic episodes), DE (depressive episodes), HME (hypomanic episodes), MixE (mixed episodes) between time points, TP1 (baseline), TP2 (follow-up).

Cohort	SBP		Barcelona		Milano		Sydney		FOR2107-Marburg	
	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases
N	46	53	13	13	0	18	53	0	206	24
age	40.9 + 14.8	41.5 + 14.0	38.7 + 9.0	40.7 + 13.0	-	47.9 + 12.0	22.1 + 4.1	-	35.6 + 13.6	41.5 + 10.0
females	23	34	7	10	-	11	28	-	129	14
time between scans	5.6 + 0.4	6.0 + 0.9	0.8 + 0.3	0.8 + 0.3	-	1.0 + 0.5	2.1 + 0.1	-	2.1 + 0.27	2.0 + 0.26
ME between TPs	-	2.9 + 16.7	-	n.a.	-	0.1 + 0.2	-	-	-	0.3 + 0.7
HME between TPs	-	1.7 + 5.1	-	n.a.	-	0.1 + 0.2	-	-	-	n.a.
DPE between TPs	-	3.7 + 6.3	-	n.a.	-	0.2 + 0.4	-	-	-	1.2 + 2.2
MixE between TPs	-	0.3 + 1.6	-	n.a.	-	0.1 + 0.2	-	-	-	n.a.
BDI (TP1; TP2)	-	28; 28	-	13; 13	-	12; 12	-	-	-	18; 18
BDII (TP1; TP2)	-	25; 25	-	0; 0	-	6; 6	-	-	-	6; 6
Lithium (TP1; TP2)	-	29; 25	-	5; 6	-	8; 14	-	-	-	6; n.a.
AE (TP1; TP2)	-	17; 14	-	7; 8	-	10; 9	-	-	-	10; n.a.
FGA (TP1; TP2)	-	0; 0	-	4; 2	-	2; 1	-	-	-	1; n.a.
SGA (TP1; TP2)	-	10; 12	-	6; 7	-	4; 3	-	-	-	10; n.a.
AD (TP1; TP2)	-	22; 23	-	1; 1	-	11; 11	-	-	-	8; n.a.

Cohort	FOR2107-Muenster		MNC		Singapore		NUI Galway		FEMS Melbourne	
	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases
N	82	9	104	20	41	18	28	9	20	26
age	29.1 + 11.4	39.8 + 15.4	38.5 + 13.1	41.2 + 12.0	34.0 + 10.8	33.3 + 8.6	33.5 + 8.8	26.4 + 6.5	21.6 + 2.4	21.2 + 2.3
females	52	1	52	8	20	10	14	4	12	7
time between scans	2.2 + 0.2	2.2 + 0.2	2.3 + 0.2	2.3 + 0.3	5.4 + 1.4	3.6 + 1.6	3.2 + 0.9	3.2 + 1.0	1.1 + 0.1	1.1 + 0.3
ME between	-	0.2 + 0.4	-	0.6 + 0.9	-	1.1 + 0.8	-	n.a.; n.a.	-	n.a.; n.a.
HME between	-	n.a.	-	n.a.	-	n.a.	-	n.a.; n.a.	-	n.a.; n.a.
DPE between	-	0.9 + 0.8	-	1.3 + 2.5	-	0.7 + 0.8	-	n.a.; n.a.	-	n.a.; n.a.
MixE between	-	n.a.	-	n.a.	-	n.a.	-	n.a.; n.a.	-	n.a.; n.a.
BDI (TP1; TP2)	-	2; 3	-	15; 11	-	18; 18	-	9; 9	-	21; 21
BDII (TP1; TP2)	-	7; 6	-	5; 8	-	0; 0	-	0; 0	-	0; 0
Lithium (TP1; TP2)	-	1; n.a.	-	8; 6	-	6; 4	-	9; 1	-	26; 14
AE (TP1; TP2)	-	2; n.a.	-	6; 9	-	11; 12	-	1; 1	-	0; 0
FGA (TP1; TP2)	-	0; n.a.	-	2; 1	-	4; 3	-	0; 0	-	0; 0
SGA (TP1; TP2)	-	4; n.a.	-	13; 11	-	12; 13	-	9; 5	-	26; 12
AD (TP1; TP2)	-	5; n.a.	-	16; 11	-	3; 5	-	1; 4	-	0; 1

Cohort	Halifax		Oslo TOP		STOP-EM		Oslo Malt	
	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases
N	23	39	293	19	36	48	33	29
age	32.3 + 19.0	51.1 + 11.5	57.3 + 15.0	27.2 + 7.6	23.3 + 5.0	22.8 + 4.2	32.5 + 9.4	33.3 + 6.8
females	15	26	177	12	20	26	18	20
time between scans	4.7 + 1.2	4.0 + 2.2	1.6 + 0.5	1.0 + 0.2	3.1 + 2.4	4.3 + 2.5	2.4 + 0.1	2.3 + 0.5
ME between	-	0.1 + 0.3	-	0.1 + 0.3	-	0.6 + 1.0	-	n.a.
HME between	-	0.1 + 0.4	-	0.1 + 0.3	-	0.5 + 0.9	-	4.8 + 5.9
DPE between	-	0.7 + 1.7	-	0.8 + 0.9	-	1.5 + 1.9	-	3.0 + 2.8
MixE between	-	n.a.; n.a.	-	0	-	0.1 + 0.7	-	n.a.
BDI (TP1; TP2)	-	27; 28	-	3; 2	-	48; 48	-	0; 0
BDII (TP1; TP2)	-	12; 11	-	15; 12	-	0; 0	-	29; 29
Lithium (TP1; TP2)	-	18; 20	-	3; 0	-	16; 16	-	2; 2
AE (TP1; TP2)	-	19; 22	-	3; 1	-	2; 2	-	12; 23
FGA (TP1; TP2)	-	0; 0	-	0; 0	-	1; 2	-	1; 1
SGA (TP1; TP2)	-	11; 16	-	7; 2	-	35; 20	-	2; 6
AD (TP1; TP2)	-	14; 11	-	7; 0	-	3; 13	-	10; 13

Table S2: Diagnostic instruments used to obtain diagnostic and clinical information.

Nr.	Site	Diagnostic instruments used to obtain diagnostic and clinical information (incl. mood episodes)	Method for obtaining medication information
1	SBP	Structured Clinical Interview for DSM-IV. Clinical records.	Detailed clinical interview and review of clinical notes
2	FIDMAG-Barcelona	Structured Clinical Interview for DSM-IV and Research Diagnostic Criteria (RDC).	Detailed clinical interview and review of case notes.
3	Milano OSR	Structured Clinical Interview for DSM-IV	Detailed clinical interview
4	Sydney	Structured Clinical Interview for DSM-IV. (Diagnostic Interview for Genetic Studies (DIGS) (for 22-30 year-olds); the Kiddie-Schedule for Affective Disorders and Schizophrenia for School-Aged Children – Present and Lifetime Version (K-SADS-BP) (for 12-21 year-olds)).	Detailed clinical interview and review of case notes
5	FOR2107-Marburg	Structured Clinical Interview for DSM-IV for Axis I Diagnoses	Self-report and hospital records
6	FOR2107-Muenster	Structured Clinical Interview for DSM-IV for Axis I Diagnoses	Self-report and hospital records
7	MNC	Structured Clinical Interview for DSM-IV for Axis I Diagnoses	Self-report and hospital records
8	Singapore	Structured Clinical Interview for DSM-IV for Axis I Diagnoses	Detailed clinical interview and review of case notes
9	NUI Galway	Structured Clinical Interview for DSMIV-TR-Patient Edition for patients and SCID_NP for controls	Detailed clinical interview outlining dose and duration of all psychotropic medication, supplemented by clinical notes where necessary.
10	FEMS Melbourne	Patients meeting DSM-IV-TR criteria for bipolar disorder on a structured clinical interview (SCID-P) were included.	Detailed clinical interview and Medication Adherence rating Scale (MARS).
11	Halifax	Structured Clinical Interview for DSM-IV for Axis I Diagnoses; (Halifax): Participants were recruited from patients followed up at a specialized Mood Disorders Program at Dalhousie University, Halifax, NS. The Program is a tertiary care clinic providing consultation services to family physicians and community psychiatrists and following up patients with BD. The diagnostic interviews were performed by pairs of clinicians, according to the Schedule for Affective Disorders and Schizophrenia, Lifetime version (SADS-L) and diagnoses were made according to DSM-IV criteria.	Questionnaire with self and interviewer reporting, in part using validated instruments; (Halifax): Patients had regular follow ups at the clinic, including monitoring of Li levels at least twice per year. Furthermore, we established illness course and treatment response to Li using NIMH life charts (NIMH-LCMTM)
12	Oslo TOP	The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I)	Clinical interviews and medical charts
13	STOP-EM	Mini International Neuropsychiatric Interview (MINI), DSM-IV criteria version 5.0.	Detailed Clinical Interview
14	Oslo Malt	Mini International Neuropsychiatric Interview (MINI), DSM-IV criteria version 5.0.	Stanley Foundation Network Entry Questionnaire (NEQ).

Table S3: Exclusion criteria for each cohort.

Nr.	Site	Exclusion criteria for study enrolment
1	SBP	Patients: younger than 18 years, not in euthymic state. Controls: any psychiatric axis I or axis II disorder or neurological conditions, family history of schizophrenia or bipolar disorder in first-degree relatives, drug or alcohol abuse.
2	FIDMAG-Barcelona	All patients with bipolar disorder were right-handed. Exclusion criteria were age younger than 18 or older than 65 years, history of neurological disease or brain trauma, and alcohol/substance abuse in the 12 months prior to participation. Patients were also required to have a current IQ in the normal range (>70). All patients were diagnosed using DSM-IV and Research Diagnostic Criteria (RDC), based on a detailed clinical interview and review of case notes. All healthy controls met the same exclusion criteria as the patients, and they were interviewed and excluded if they reported a history of mental illness and/or treatment with psychotropic medication other than non-regular use of benzodiazepines or similar drugs for insomnia. They were also questioned about family history of mental illness and excluded if a first-degree relative had experienced symptoms consistent with major psychiatric disorder and/or had received any form of in- or outpatient psychiatric care. The healthy controls were selected to be matched with the patients on demographic variables and on premorbid IQ.
3	Milano OSR	Exclusion criteria were age younger than 18; the presence of other diagnoses on Axis I; the presence of pregnancy; history of epilepsy or major medical, neurological disorders or brain trauma; history of drug or alcohol abuse or dependency. Patients were also required to have a current IQ in the normal range (>70). No patient had received electroconvulsive therapy within 6 months prior to study enrolment.
4	Sydney	Control (CN) participants were defined as those who did not have a first-degree relative with either BD I or II, recurrent major depressive disorder (MDD), schizophrenia, schizoaffective disorder, recurrent substance abuse or any past psychiatric hospitalization. Additionally, they did not have a second-degree relative with a history of psychosis or who had been hospitalized for a mood disorder. All subjects were aged between 12 and 30 years. For those aged between 12 and 21 an adapted version of the Schedule for Affective Disorders and Schizophrenia for School-Age Children – Present and Lifetime Version (K-SADS-BP) was developed specifically for use in the US-Australia collaborative study of young people at genetic risk for BD. For participants aged between 22 and 30 the DIGS (Version 4) was used to measure the current and lifetime presence of axis I DSM-IV disorders.
5	FOR2107-Marburg	Inclusion criteria: age 18-65 years; patients were diagnosed of bipolar I disorder by SCID-Interview, currently depressed, (hypo) manic or remitted. Exclusion criteria all: any MRI contraindications; any neurological abnormalities. Exclusion criteria controls: any current or former psychiatric disorder; Exclusion criteria patients: substance dependence or current benzodiazepine treatment (wash out of at least three half-lives before study participation)
6	FOR2107-Muenster	Inclusion criteria: age 18-65 years; patients were diagnosed of bipolar I disorder by SCID-Interview, currently depressed, (hypo)manic or remitted. Exclusion criteria all: any MRI contraindications; any neurological abnormalities. Exclusion criteria controls: any current or former psychiatric disorder; Exclusion criteria patients: substance dependence or current benzodiazepine treatment (wash out of at least three half-lives before study participation)
7	MNC	Inclusion criteria: age 18-65 years; patients were diagnosed of bipolar I disorder by SCID-Interview, currently depressed, (hypo)manic or remitted. Exclusion criteria all: any MRI contraindications; any neurological abnormalities. Exclusion criteria controls: any current or former psychiatric disorder; Exclusion criteria patients: substance dependence or current benzodiazepine treatment (wash out of at least three half-lives before study participation)
8	Singapore	Inclusion: 1) DSM IV diagnosis of Bipolar Disorder (Patients), 2) age: 21-65 3), English speaking, 4) provision of informed written consent. Exclusion criteria: 1) History of significant head injury, 2) significant neurological diseases (such as epilepsy, cerebrovascular accident) or medical illnesses, 4) significant DSM IV alcohol or substance use or dependence, 6) Contraindications to MRI (e.g., pacemaker, orbital foreign body, recent surgery/procedure with metallic devices/implants deployed), 7) pregnant women, 8) claustrophobia.
9	NUI Galway	Inclusion criteria: DSM-IV diagnosis of bipolar disorder (patients); age >18 and <60. Exclusion criteria: history of neurological illness (comorbid); lifetime DSM-IV axis 1 disorder or family history of psychotic or affective disorder in first- or second-degree relatives (controls); history of substance and/or alcohol misuse in the past year; learning disability; recent oral steroid use.

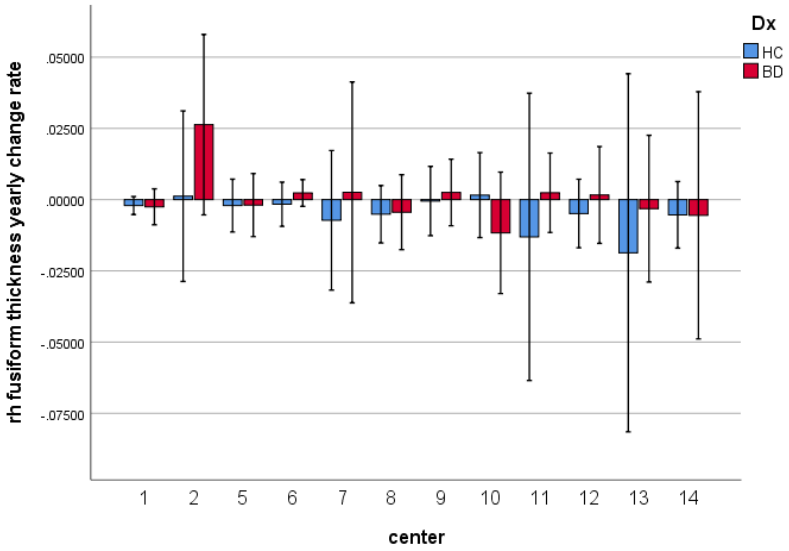
10	FEMS Melbourne	Inclusion Criteria: Meet DSM-IV criteria for mania as part of bipolar I disorder or schizoaffective disorder. YMRS at baseline of at least 20. Not have had a previous treated manic episode. Have the capacity to provide informed consent to the study and comply with study procedures. Be utilising effective contraception if female, sexually active and of childbearing age. Patients will need to have been on quetiapine and lithium as standard therapy for at least 1 month prior to randomisation. Male or female patients age 15 to 25 years Exclusion from the trial includes: Patients with a known or suspected clinically relevant systemic medical disorder. Individuals who are pregnant or lactating. Patients who have had a prior sensitivity or allergy to quetiapine, lithium or their components. Inability to comply with either the requirements of informed consent or the treatment protocol. Non-fluency in English. History of epilepsy. Clinically relevant biochemical or haematological abnormalities at baseline. Patients at immediate risk of self-harm or risk to others. Organic mental disease, including mental retardation (Full scale IQ<70). Use of any of the following cytochrome P450 3A4 inhibitors in the 14 days preceding enrolment including but not limited to: ketoconazole, itraconazole, fluconazole, erythromycin, clarithromycin, troleandomycin, indinavir, nelfinavir, ritonavir, fluvoxamine and saquinavir. Use of any of the following cytochrome P450 inducers in the 14 days preceding enrollment including but not limited to: phenytoin, carbamazepine, barbiturates, rifampin, St. John's Wort, and glucocorticoids. A patient with Diabetes Mellitus (DM) fulfilling one of the following criteria: Unstable DM defined as enrollment glycosylated hemoglobin (HbA1c) >8.5%. Admitted to hospital for treatment of DM or DM related illness in past 12 weeks. Not under physician care for DM Physician responsible for patient's DM care has not indicated that patient's DM is controlled. Physician responsible for patient's DM care has not approved patient's participation in the study. Has not been on the same dose of oral hypoglycaemic drug(s) and/or diet for the 4 weeks prior to randomization. For thiazolidinediones (glitazones) this period should not be less than 8 Weeks. Taking insulin whose daily dose on one occasion in the past 4 weeks has been more than 10% above or below their mean dose in the preceding 4 weeks. Note: If a diabetic patient meets one of these criteria, the patient is to be excluded even if the treating physician believes that the patient is stable and can participate in the study. An absolute neutrophil count (ANC) of £1.5 x 109 per liter.
11	Halifax	Inclusion criteria. The BD patients (both Li and non-Li groups) had to have: (i) a diagnosis of bipolar I or II disorder made by a psychiatrist using the SCID; (ii) at least 10 years of illness; (iii) a history of at least five episodes of illness (including manic, depressive, or mixed episodes); (iv) current Hamilton Depression Rating Scale, 17-item version (HAM-D-17) score < 7; (v) current Young Mania Rating Scale (YMRS) score < 5; (vi) current Clinical Global Impressions Scale–Bipolar (CGI-BP) score < 3; and (vii) a period of euthymia for at least four months prior to scanning, as aside from state- related factors, patients in acute episodes may present with additional difficult to control confounding variables, including recent medication change or substance abuse. The non-Li group had to have less than three months of lifetime Li exposure, more than 24 months prior to the scanning. The Li group had to have a current Li treatment lasting a minimum of 24 months. Exclusion criteria. Individuals from any of the three groups were excluded if they met any of the magnetic resonance imaging (MRI) exclusion criteria or had any serious medical illness (e.g., brain injury, Cushings disease, or conditions treated with corticosteroids). Individuals with BD were excluded if they had: (i) more than one lifetime course of electroconvulsive therapy (ECT) or ECT in the previous 12 months; (ii) comorbid psychiatric disorders, and/or personality disorder; (iii) active substance abuse in the previous 12 months; (iv) significant change in their medication in the previous three months; or (v) current psychotic features or acute suicidality. Individuals from the non-Li group were excluded if they had: (i) Li exposure < 2 years before the scanning; or (ii) lifetime Li exposure of more than three months. The neuropsychiatrically healthy individuals were excluded if they had a personal history of psychiatric disorders. (Halifax) Diabetes Study: The subjects with BD were required to 1) have the diagnosis of bipolar I or II disorder made by a psychiatrist; and 2) be at least 18 years of age. Patients were excluded if they had 1) the diagnosis of organic mood disorder; 2) mood disorder not otherwise specified; or 3) more than one lifetime course of electroconvulsive therapy or electroconvulsive therapy within the last 6 months. The neuropsychiatrically healthy, euglycemic subjects were excluded if they had 1) a personal history of psychiatric disorders; or 2) T2DM. Subjects from any group were excluded if they 1) met any magnetic resonance imaging (MRI) exclusion criteria; 2) suffered from substance abuse in the last 12 months; had a history of 3) neurodegenerative disorders; or 4) cerebrovascular

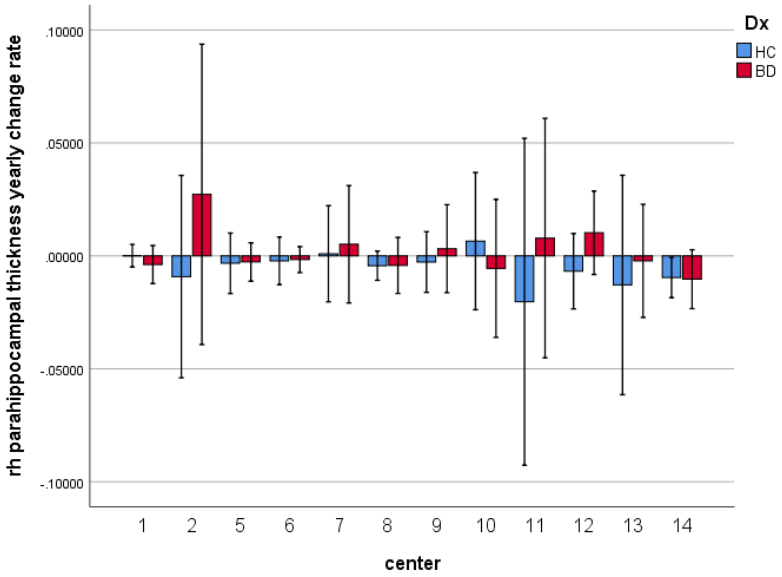
		disease/stroke, as we were interested in the more subtle T2DM-related neuronal changes. Halifax High Risk Study: Families were identified through adult probands with BD, who had participated in 1) previous genetic and high-risk studies for the Halifax sample. Only the offspring from these families, not the probands, were a part of the MRI study. The offspring from BD parents were divided into two subgroups: 1) the Unaffected HR group, which consisted of 50 offspring with no lifetime history of psychiatric disorders. These individuals were at an increased risk for BD because they had one parent affected with a primary mood disorder. 2) The Affected Familial group, which consisted of 36 offspring who met criteria for a lifetime Axis I diagnosis of mood disorders (i.e., a personal history of at least one episode of depression, hypomania, or mania meeting full DSM-IV criteria). When available, we recruited more than one offspring per family. From this study, we provided data only from patients who had a personal history of bipolar disorder.
12	Oslo TOP	All participants aged between 18 and 65 and spoke a Scandinavian language. Participants were excluded if they had or have had severe head injury, IQ below 70, neurological disorder or uncontrolled medical condition that interferes with brain function, metal implants, cardiac pacemaker or other MRI contraindications. Additionally, persons with pathological neuroradiological findings were excluded from the analyzes. Healthy control participants were also excluded if they or any of their first-degree relatives had a lifetime history of a severe psychiatric disorder, or if they had severe substance abuse during the last year.
13	STOP-EM	***Intake Criteria:*** First Episode Mania within 3 months of the episode. We will also include a subgroup who have had recent manic/hypomanic symptoms but may not meet DSM IV criteria for diagnosis. Subjects can be psychotic or non-psychotic, in pure or mixed mania, with or without comorbidity. Age: 14 to 35 years. ***Exclusion Criteria:*** (1) inability to take part in neuropsychological testing, (2) meets the standard criteria for exclusion for magnetic resonance imaging, (3) a previous manic episode diagnosed retrospectively on structured interview or via collateral.
14	Oslo Malt	Inclusion criteria patients: A DSM-IV diagnosis of bipolar disorder type II. Exclusion criteria healthy controls: Controls with previous or current psychiatric illness were excluded from the study. The exclusion criteria for all participants were: A.) age younger than 18 or older than 50 years; B.) previous head injury with loss of consciousness for more than 1 minute; C.) history of neurological or other severe chronic somatic disorder; D.) pregnancy; E.) metallic implants.

Table S4: Image acquisition parameters and FreeSurfer software versions used at each site.

Nr.	Site	MRI scanner	Sequence	Flip angle	TR (ms)	TE (ms)	TI (ms)	Voxels (mm)	Gap (mm)	Slices	Direction	FreeSurfer (longitudinal)
1	SBP	1.5 T Signa Excite	3D-SPGR	30	21	6	NA	.7 x .7 x 1.8	1.8	128	Coronal	V6.0
2	FIDMAG-Barcelona	1.5T GE Signa	3D T1-weighted enhanced fast gradient echo (EFGRE3D)	15	2000	3.93	710	.47 x .47 x 1	0	180	Axial	V5.3
3	Milano OSR	3T Gyroscan Intera, Philips	3D T1-weighted enhanced fast field echo	30	25	4.6	NA	.9 x .9 x 1.6	0.8	220	Axial	V6.0
4	Sydney	3T Philips Achieva	3D T1-weighted turbo field echo (TFE)	8	5.5	2.5	NA	1 x 1 x 1	1	180	Sagittal	V5.3
5	FOR2107-Marburg	3T Siemens Magnetom Trio	3D T1-weighted magnetization prepared rapid acquisition on gradient echo (MPRAGE)	9	1900	2.26	900	1 x 1 x 1	0.5	176	Sagittal	V5.3
6	FOR2107-Muenster	3T Siemens PRISMA	3D T1-weighted magnetization prepared rapid acquisition on gradient echo (MPRAGE)	8	2130	2.28	900	1 x 1 x 1	0	192	Sagittal	V5.3
7	MNC	3T Philips Gyroscan Intera	3D Fast gradient echo sequence	9	7.4	3.4	815	.5 x .5 x .5	0	320	Coronal	V5.3
8	Singapore	3T Philips Achieva	3D T1-weighted magnetization prepared rapid acquisition on gradient echo (MPRAGE)	8	7200	3.3	NA	0.9x0.9x0.9	0	180	Axial	V5.3
9	NUI Galway	1.5T Siemens Magnetom	3D T1-weighted magnetization prepared rapid acquisition on gradient echo (MPRAGE)	15	1140	4.38	600	.45 x .45 x .9	0	256	Axial	V5.3
10	FEMS Melbourne	3T Siemens TimTrio (32 channel head coil)	3D T1 Magnetisation-Prepared RApid Gradient-Echo (MPRAGE32)	9	2000	2.24	900	0.9 x 0.9 x 0.9 For 25 subjects: 0.45 x 0.45 x 1.0	0	192	Sagittal	V5.3
11	Halifax	1.5T GE Signa	3D T1-weighted spoiled gradient recalled acquisition in steady state	40	25	5	0	.9375 x .9375 x 1.5	1.5	125	Coronal	V5.3
12	Oslo TOP	3T General Electric Discovery MR750	3D T1-weighted BRAVO sequence	12	8.16	3.18	450	1 x 1 x 1	0	188	Sagittal	V7.1
13	STOP-EM	3.0T Philips Intera	3D T1 Magnetisation-Prepared Rapid Acquisition on Gradient Echo (MPRAGE)	8	1800	3.5	794	1 x 1 x 1	0	180	Axial	V5.1
14	Oslo Malt	3T Philips Achieva	3D T1-weighted turbo field echo (TFE)	7	8.4	2.3	NA	1 x 1 x 1	0	220	Sagittal	V5.3

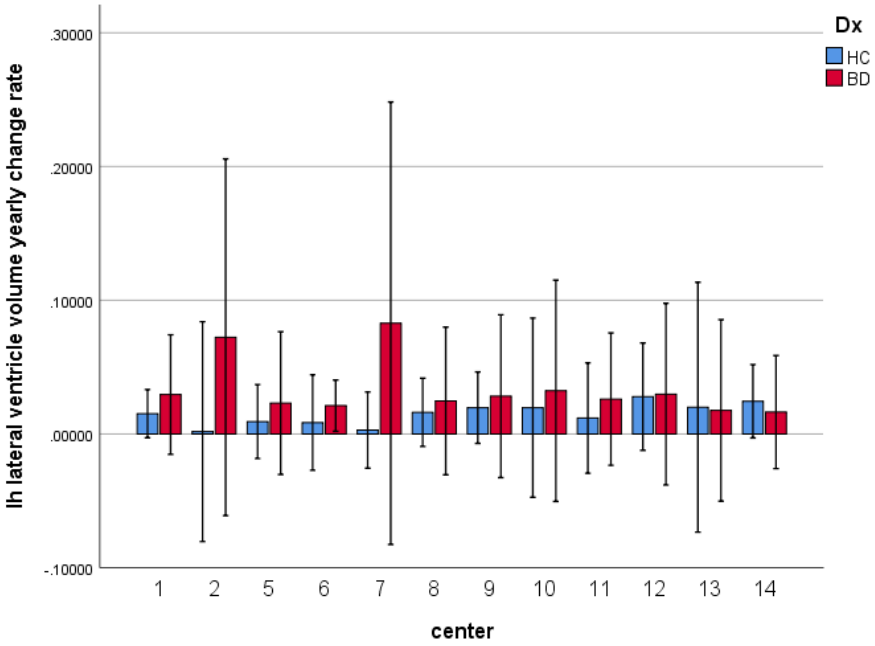
Figures S1-S2: Cortical change rates by center.

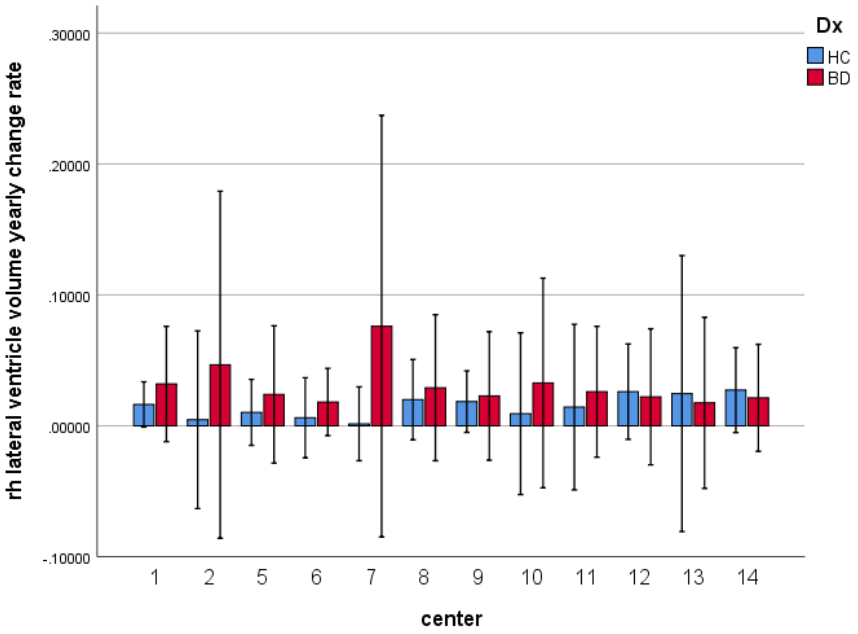




Note: Pattern of group differences were the same and results are still significant when excluding center 2, showing the largest mean of thickness increases in BD (parahippocampal: p=0.012; fusiform: p=0.008).

Figures S3-S4: Ventricular change rates by center.

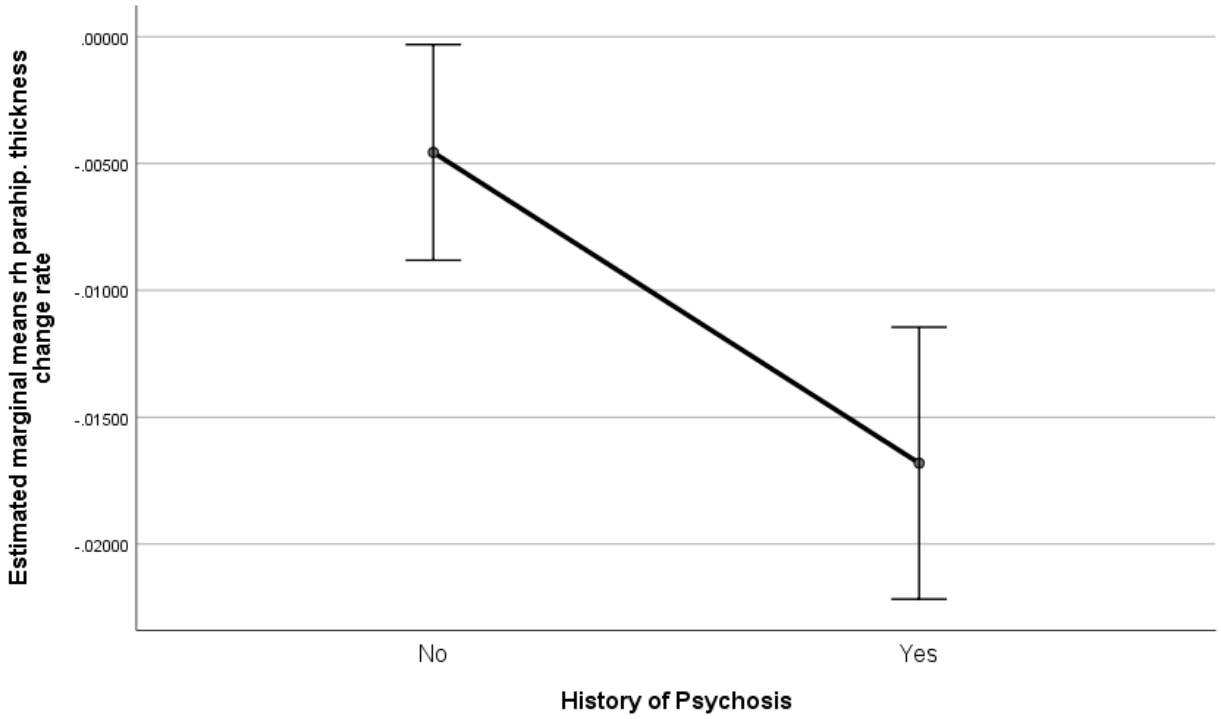


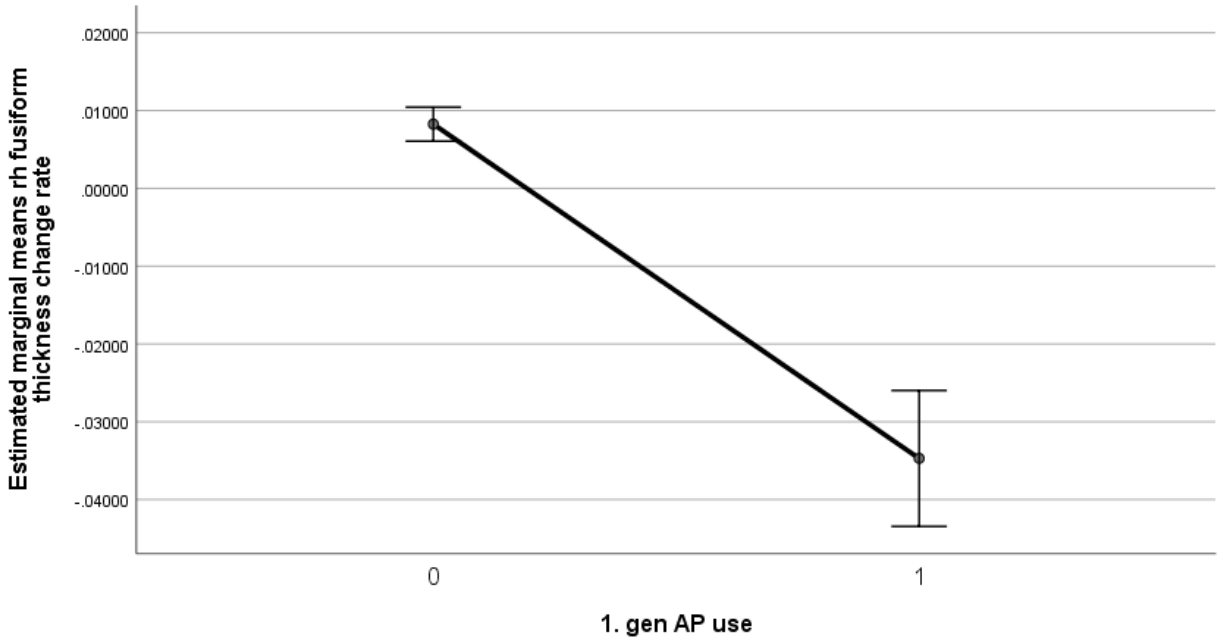


Note: Pattern of group differences were the same and results are still significant when excluding center 7, showing the largest mean of thickness increases in BD (left ventricle: p=0.007; right ventricle: p=0.020) and when excluding both center 2 and 7 (left ventricle: p=0.036; right ventricle: p=0.047).

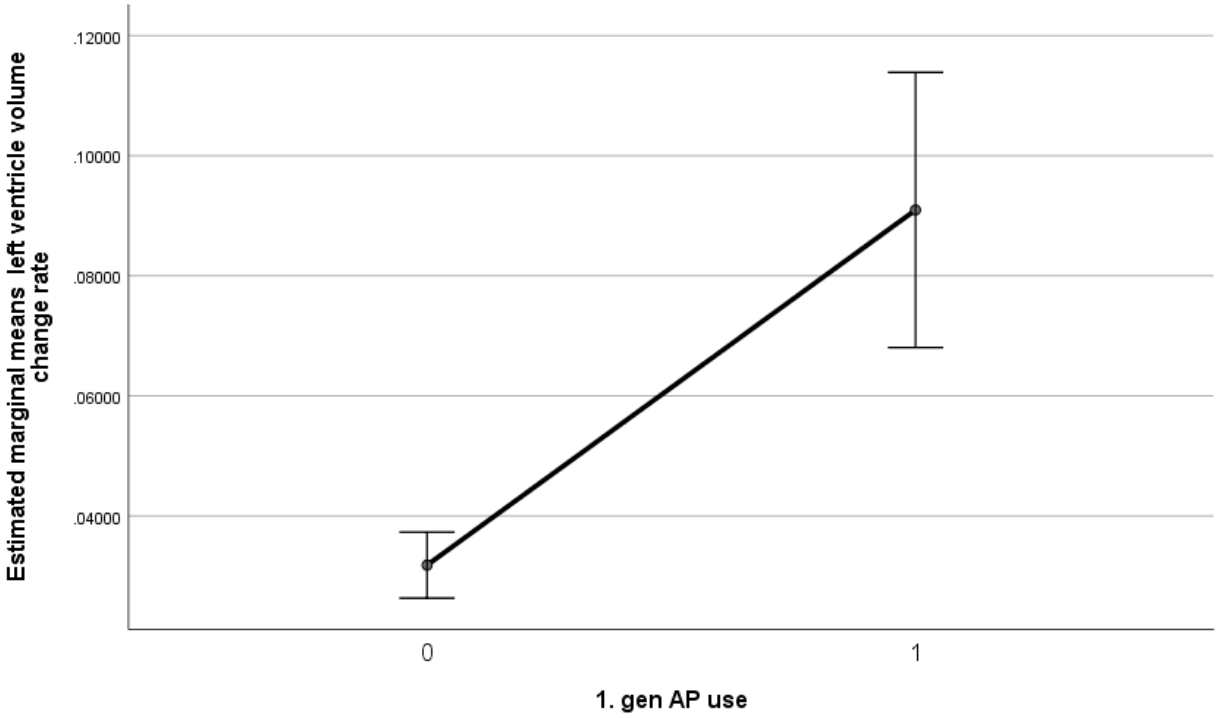
Leave-one-site-out analysis:
In addition to removing visual outliers indicated by Figure S1-S4, we have performed a leave-one-site-out analysis where we repeated the main analysis after excluding each site one at a time. This was done for changes in bilateral ventricle volumes, right parahippocampus, and right fusiform cortex. Results of these tests indicate that our results are not driven by individual sites and support the overall robustness of our findings (Data S1).

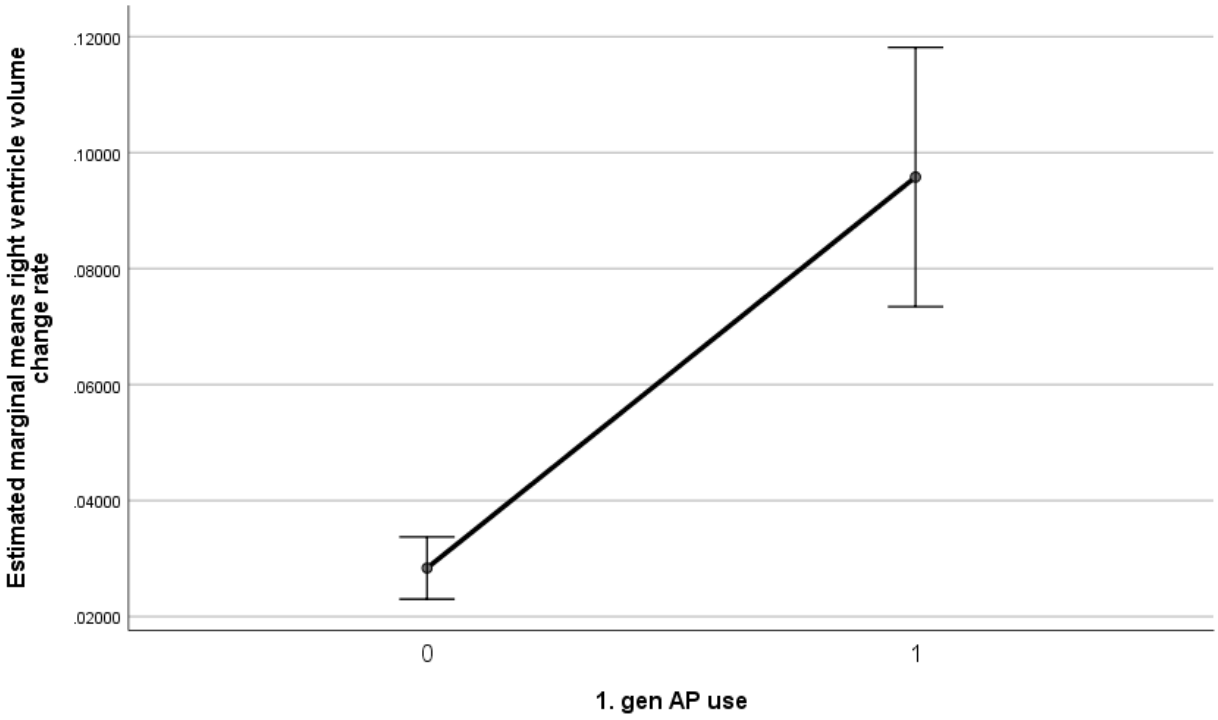
Figures S5-S9: Change rate comparisons within BD patients showing the effects of history of psychosis, first generation antipsychotic drug use, and bipolar subtype. Corresponding statistical results can be found in **Data S1.**



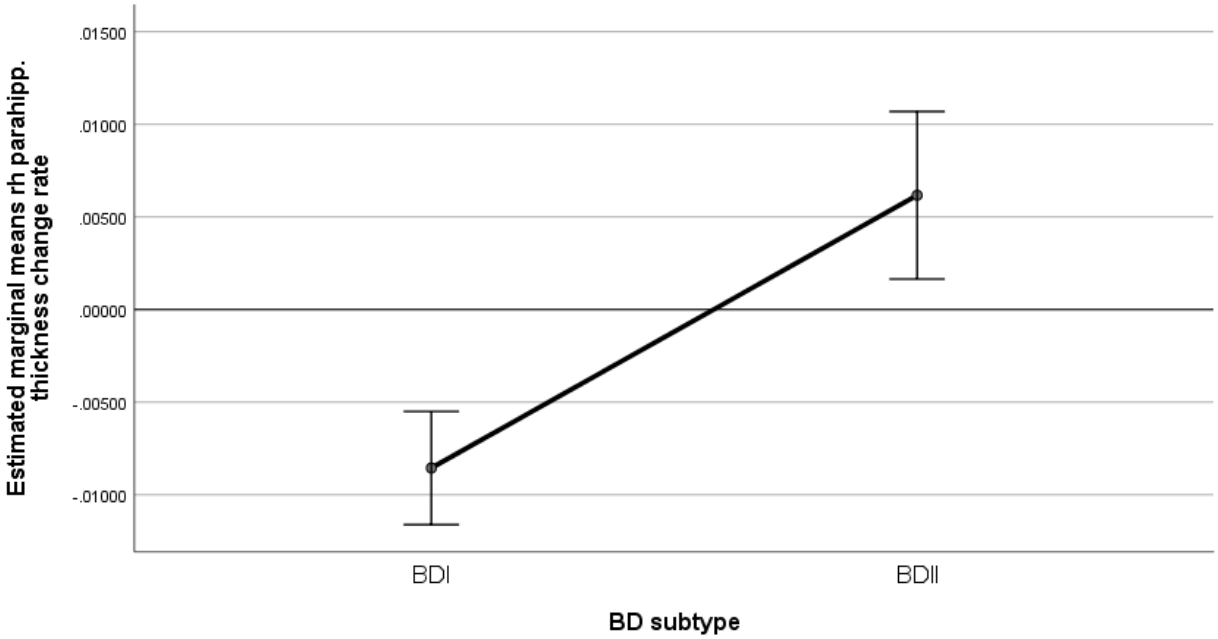


0: not using; 1: using



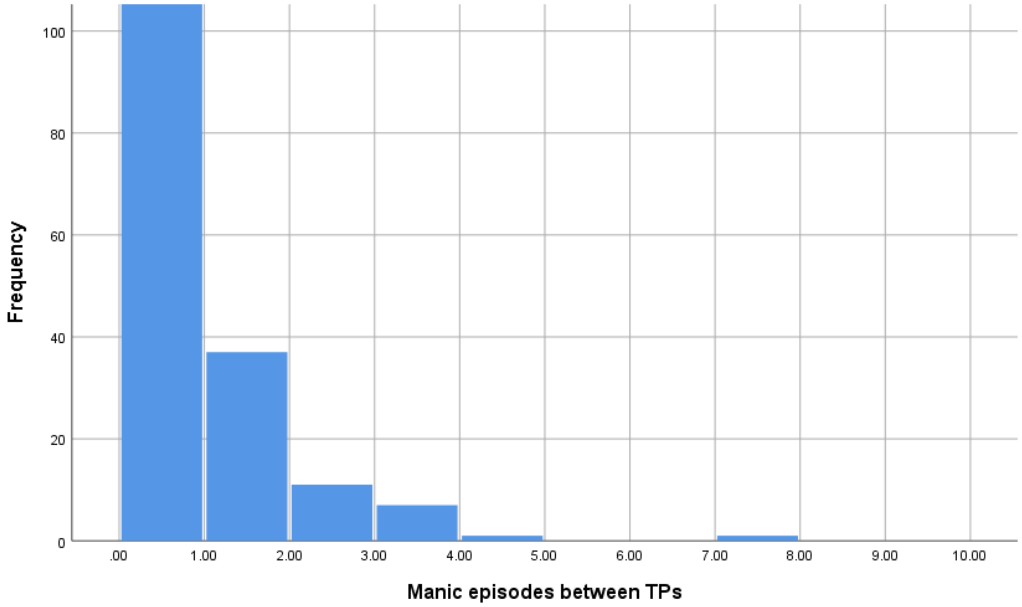


0: not using; 1: using


Estimated marginal means rh parahipp. thickness change rate
.01500
.01000
.00500
.00000
-.00500
-.01000
BDI
BDII
BD subtype

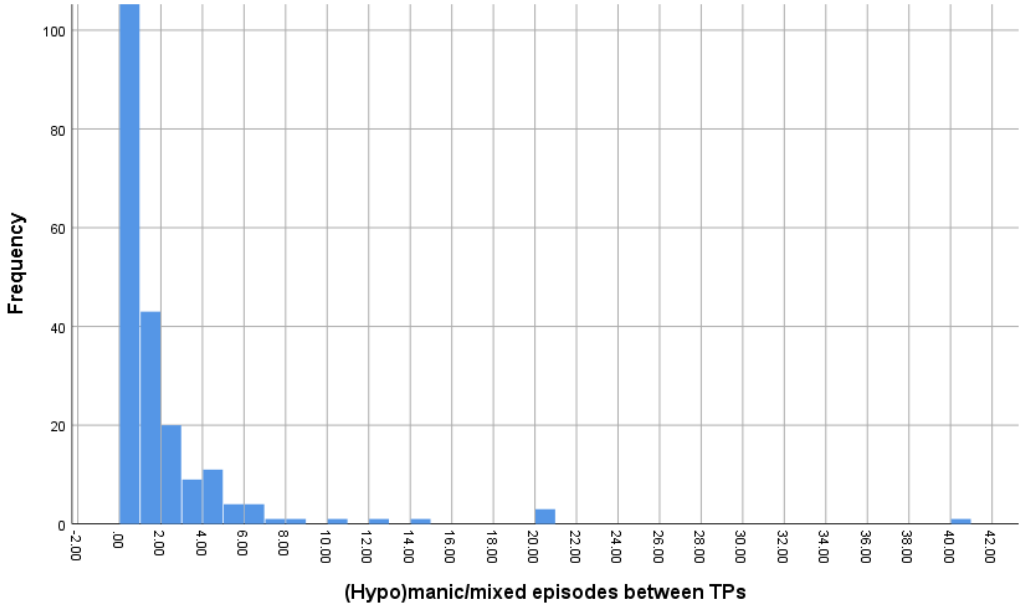
Distributions of mood episodes between time points (TPs):

Figure S10



Frequency-axis is not fully displayed for better visualization. The maximum frequency value at zero episodes was 162. One outlier reported over 100 mood episodes between time points (not shown in this graph). There was no indication that justified exclusion of this person from the correlation analysis. However, the results did not change when excluding the participant (**Data S2**).

Figure S11



Frequency-axis is not fully displayed for better visualization. The maximum frequency value at zero episodes was 152. One outlier with over 100 reported mood episodes between time points is not shown in this graph. The correlation results did not change when excluding this participant (**Data S2**).

Table S5: Intercorrelations between ventricular and subcortical volume change rates. These analyses were exploratory, hence, not corrected for multiple testing.

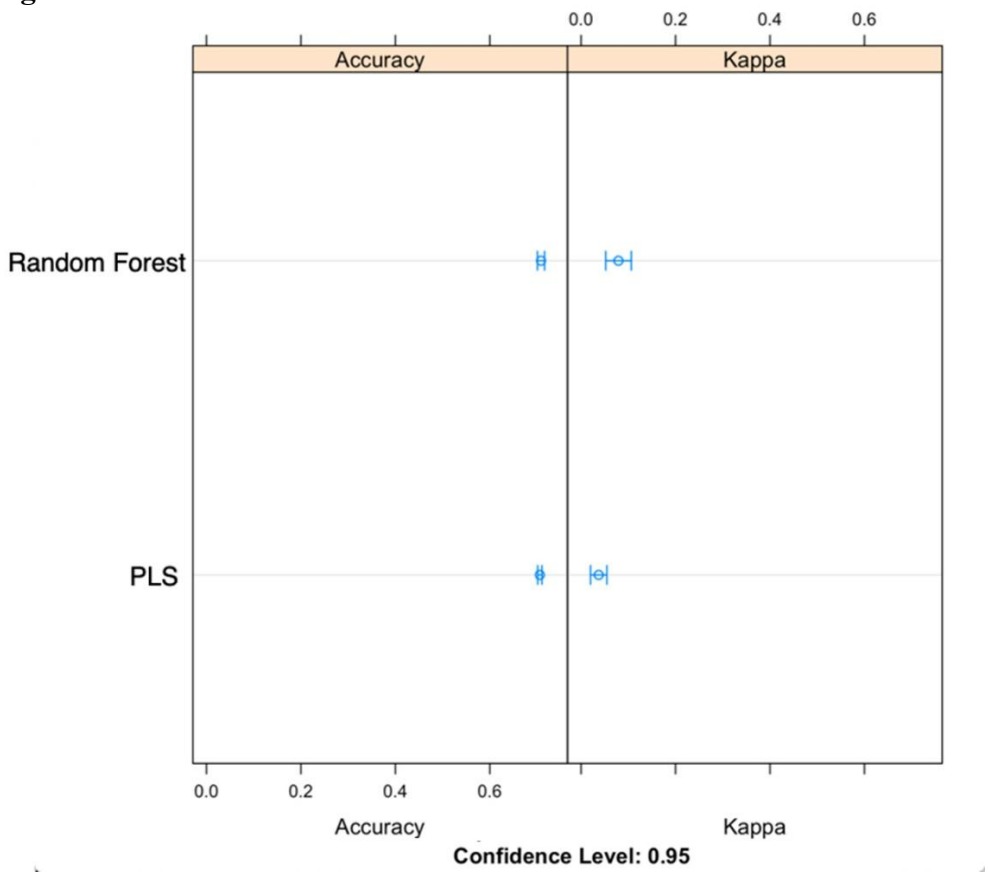
Change rate in region	result	Left Lateral Ventricle	Right Lateral Ventricle
Left Thalamus	r	-0.517	-0.522
	p-value	<0.001	<0.001
	N	297	297
Left Caudate	r	-0.384	-0.384
	p-value	<0.001	<0.001
	N	320	320
Left Putamen	r	-0.209	-0.209
	p-value	<0.001	<0.001
	N	298	298
Left Pallidum	r	-0.164	-0.126
	p-value	0.004	0.027
	N	309	309
Left Hippocampus	r	-0.358	-0.383
	p-value	<0.001	<0.001
	N	309	309
Left Amygdala	r	-0.152	-0.145
	p-value	0.007	0.010
	N	317	317
Left Accumbens	r	0.077	0.064
	p-value	0.168	0.254
	N	319	319
Right Thalamus	r	-0.611	-0.639
	p-value	<0.001	<0.001
	N	302	302
Right Caudate	r	-0.262	-0.288
	p-value	<0.001	<0.001
	N	319	319
Right Putamen	r	-0.309	-0.308
	p-value	<0.001	<0.001
	N	299	299
Right Pallidum	r	0.017	0.056
	p-value	0.765	0.319
	N	314	314
Right Hippocampus	r	-0.431	-0.45
	p-value	<0.001	<0.001
	N	311	311
Right Amygdala	r	-0.229	-0.251
	p-value	<0.001	<0.001
	N	316	316

Right Accumbens	r	-0.198	-0.2
	p-value	<0.001	<0.001
	N	320	320

Multivariate and machine learning classification

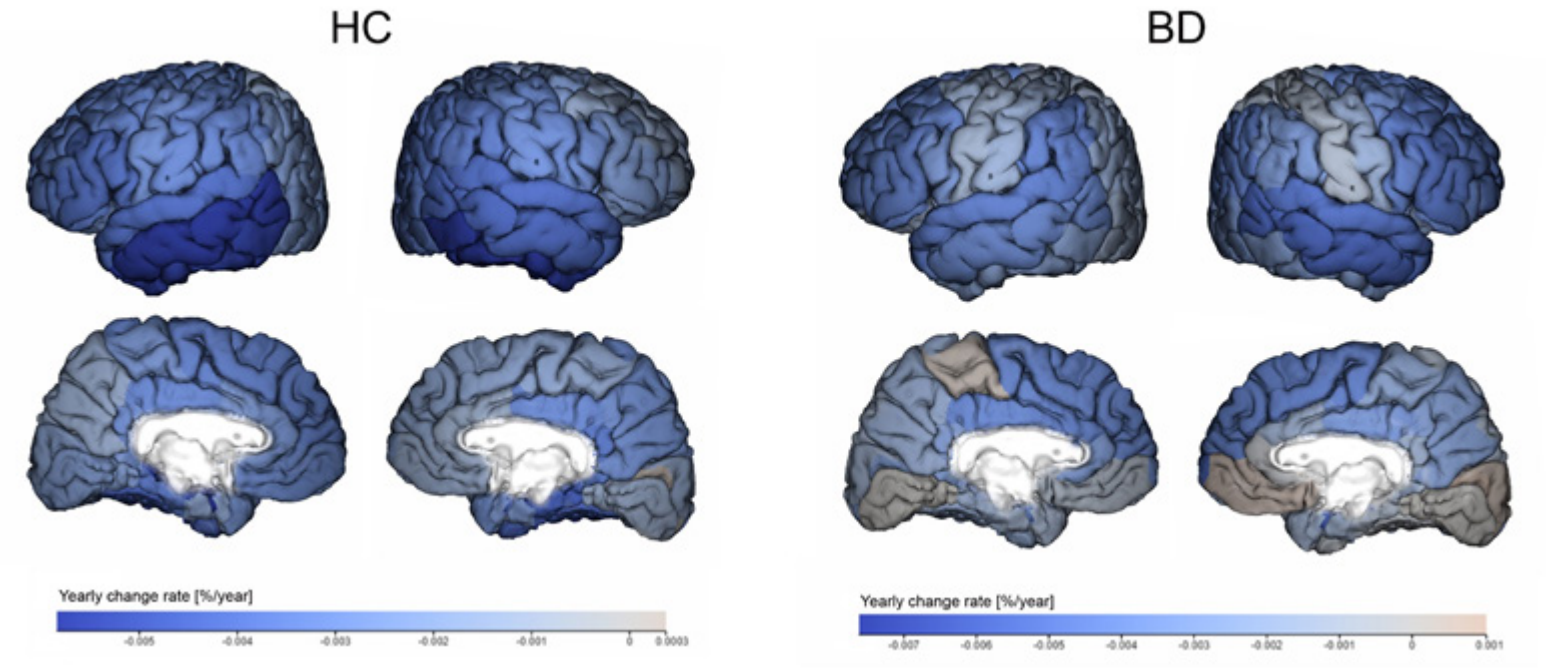
Based on results presented in Table S5, we tested whether multivariate classification methods (PLS and Random Forest) could distinguish between BD patients and controls based on regional change rate data. We first trained a Partial Least Squares Classifier as one of the most simplistic and interpretable linear projective algorithms and then benchmarked it against a non-linear classifier (Random Forest), using an ensemble of multiple (n=500) decision trees trained on random feature subsets combined by a majority vote. Both classifiers failed to provide reasonable accuracy levels when unbalanced classes are taken into account (Acc~0.7, Kappa~0.1), indicating that HC and BD could not reliably be classified based on the brain change measures investigated in this study.

Figure S12:



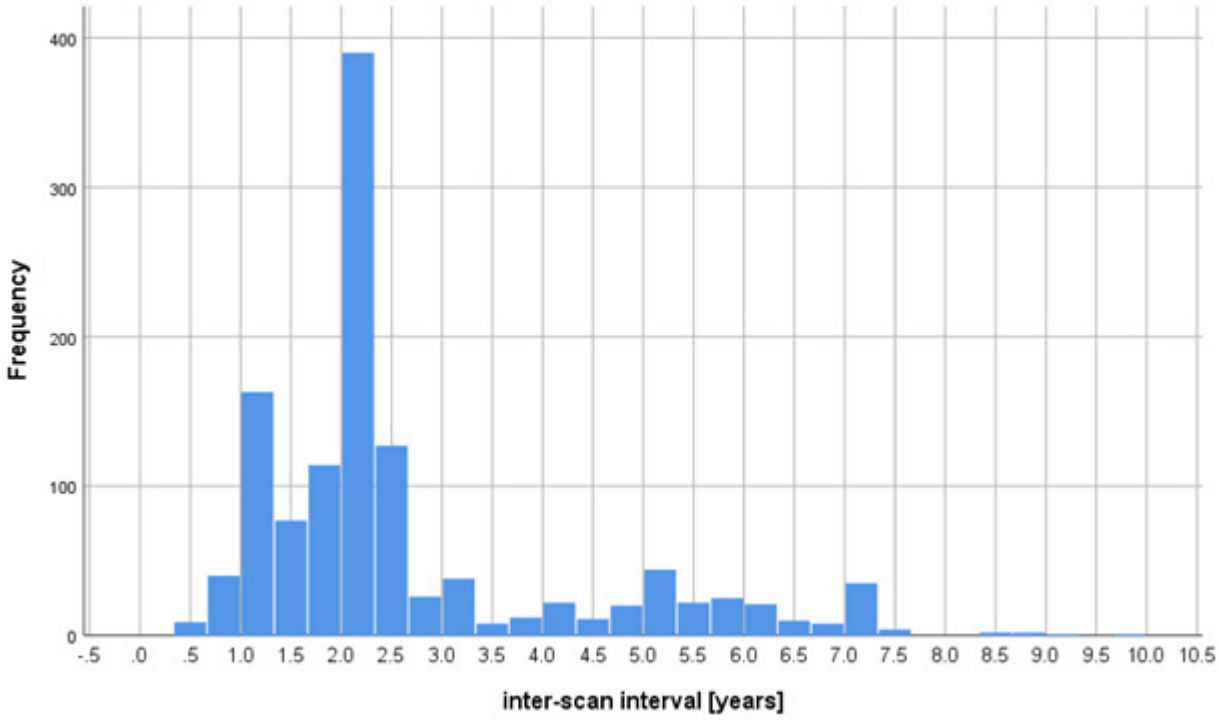
Accuracy and Kappa plotted for both multivariate classification algorithms tested: Random Forest (top) and Partial least squares (PLS, bottom).

Figure S13:



Yearly change rates within each group. Raw (uncorrected) means of yearly change rates of cortical thickness for each ROI and each group mapped into brain space. The HC group (left) displays negative rates (cortical thinning; cold colors) across the whole brain. While BD patients (right) show similar decreases in thickness in some areas, they also show rates close to zero (no changes) or positive rates (thickening; warm colors) in other areas. Numerical values as well as means and SD for other brain phenotypes (area and subcortical volumes) are provided in Data S3. Results for group comparisons are presented in the main text (Figure 1 and 3).

Figure S14:



Histogram plot of interscan interval in the combined cohort.

Table S6: Cohort characteristics: demographics and clinical variables.

Variable	BD	HC	BD vs HC (p-value)	data available* (n)
N (total = 1232)	307	925	-	-
Sex (F/M), n (%)	172(56%)/135(44%)	539(58%)/386(42%)	0.505	1232
Ethnicity (White/Black/Asian/Other), n (%)	204(83%)/1(0.4%)/26(10.6%)/14(6%)	536 (90%)/0(0%)/54(9%)/3(1%)	0.005 (w/other)	838
Age at TP1, years, mean ± SD [min; max]	35 ± 14 [16; 75]	41 ± 18 [15; 85]	<0.001	1232
time between TP1 and TP2, mean ± SD	3.3 ± 2.2	2.4 ± 1.3	<0.001	1232
Education (1/2/3/4), n (%); mean ± SD	41(16%)/69(28%)/103(41%)/38(15%); 2.5 ± 0.9	47(5%)/244(28%)/271(31%)/314(36%); 3.0 ± 0.9	<0.001; <0.001)	1127
BMI at TP1 [TP2], mean ± SD	26.4 ± 5.8 [27.1 ± 5.4]	24.4 ± 4.1 [24.8 ± 4.1]	<0.001 [<0.001]	970 [988]
Smokers at TP1 [TP2], n (%)	39 (28%) [29 (31%)]	65 (10%) [51 (12%)]	<0.001 [<0.001]	793 [517]
ICV, liter, at TP1 [TP2], mean ± SD	1.493 ± 0.140 [1.493 ± 0.139]	1.496 ± 0.180 [1.496 ± 0.180]	0.807 [0.782]	
Age of onset, years, mean ± SD	22 ± 9	-	-	
Subdiagnosis (BD1/BD2/NOS) at TP1 [TP2], n (%)	202 (66%)/99(32%)/6(2%) [199 (66%)/97 (32%)/6(2%)]	-	-	307 [302]
Mood episodes				
Mood state (euthymic/depressed/manic/mixed) at TP1 [TP2], n (%)	185(70%)/54(20%)/24(9%)/2(1%) [177(86%)/23(11%)/7(3%)/0(0%)]	-	-	265 [207]
Number of hypomanic episodes at TP1, mean ± SD (range)	9 ± 16 (0-100)	-	-	166
Number of hypomanic episodes between TP1 and TP2, mean ± SD (range)	1.4 ± 4 (0-30)	-	-	160
Experienced hypomanic episode between TP1 and TP2, n (%)	54 (31%)	-	-	175
Number of manic episodes at TP1, mean ± SD (range)	2 ± 4 (0-40)	-	-	240
Number of manic episodes between TP1 and TP2, mean ± SD (range)	1 ± 8 (0-111)	-	-	203
Experienced manic episode between TP1 and TP2, n (%)	59 (27%)	-	-	216
Number of depressive episodes at TP1, mean ± SD (range)	9 ± 14 (0-90)	-	-	238
Number of depressive episodes between TP1 and TP2, mean ± SD (range)	1.9 ± 3.5 (0-30)	-	-	234
Experienced depressive episode between TP1 and TP2, n (%)	145 (58%)	-	-	250
Number of mixed episodes at TP1, mean ± SD (range)	0.3 ± 1.9 (0-15)	-	-	117
Number of mixed episodes between TP1 and TP2, mean ± SD (range)	0.2 ± 1.1 (0-10)	-	-	105
Experienced mixed episode between TP1 and TP2, n (%)	6 (6%)	-	-	106
Medication use				
Lithium at TP1 [TP2], n (%)	120 (60%) [94 (36%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	1227 [1187]
Antiepileptics at TP1 [TP2], n (%)	90(30%) [90 (34%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	1230 [1188]
Antipsychotics (1. gen.) at TP1 [TP2], n (%)	12(4%) [7 (3%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	1228 [1187]
Antipsychotics (2. gen.) at TP1 [TP2], n (%)	145(48%) [104 (40%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	1177 [1137]
Antidepressants at TP1 [TP2], n (%)	90(30%) [82 (31%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	1229 [1188]

Comorbidity and substance use				
Alcohol use (abuse/dependence) at TP1 [TP2], n (%)	18 (6%)/16(6%) [14 (8%)/11(6%)]	36 (4%)/2(0.2%) [17(12%)/0(0%)]	<0.001 [0.006]	1102 [324]
Substance use (abuse/dependence) at TP1 [TP2], n (%)	22 (9%)/11(4%) [2 (0.4%)/19(5%)]	0 (0%)/0(0%) [0 (0%)/0(0%)]	<0.001 [<0.001]	997 [569]
ADHD at TP1 [TP2], n (%)	17 (7%) [15 (9%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	939 [518]
OCD at TP1 [TP2], n (%)	6 (1%) [6 (3.5%)]	0 (0%) [0 (0%)]	<0.001 [0.001]	941 [518]
GAD at TP1 [TP2], n (%)	19 (7%) [8 (5%)]	1 (0.1%) [1 (0.3%)]	<0.001 [0.001]	941 [518]
PTSD at TP1 [TP2], n (%)	3 (1%) [2 (1%)]	0 (0%) [0 (0%)]	0.021 [0.107]	942 [518]
Panic disorder at TP1 [TP2], n (%)	19 (7%) [5 (3%)]	0 (0%) [0 (0%)]	<0.001 [0.004]	941 [518]
Phobia at TP1 [TP2], n (%)	14 (5%) [13 (8%)]	0 (0%) [0 (0%)]	<0.001 [<0.001]	942 [518]
Eating disorder at TP1 [TP2], n (%)	6 (2%) [2 (1%)]	0 (0%) [0 (0%)]	<0.001 [0.107]	942 [516]
History of psychotic symptoms at TP1 [TP2], n (%)	86 (46%) [56 (41%)]	1 (0.3%) [1 (0.2%)]	<0.001 [<0.001]	532 [771]

Demographics and clinical variables. Means +/- SD or number of participants are listed for both groups and both timepoints. Data at T2 is listed in parenthesis. Abbreviations: BDI: bipolar type 1, BDII: bipolar type 2, ME: manic episodes, DE: depressive episodes, HME: hypomanic episodes, MXE: mixed episodes, AD: antidepressants, AE: antiepileptics, AP: antipsychotics, CS: central stimulants, GAD: general anxiety disorder, OCD: obsessive compulsive disorder, ADHD: attention deficit hyperactivity disorder, PTSD: posttraumatic stress disorder. *Percentages are given in relation to available data.

Table S7: Correlations between regional change rates and mood episodes between time points.

Change rate in region	Manic episodes	(Hypo)manic and mixed episodes
left lingual	r=-0.26, p<0.001, n=198	r=-0.24, p<0.001, n=228
right *pars orbitalis*	n.s.	r=-0.24, p<0.001, n=230
left *pars opercularis*	n.s.	r=-0.23, p<0.001, n=229
right caudal anterior cingulate	n.s.	r=-0.23, p<0.001, n=230
left caudal middle frontal	n.s.	r=-0.23, p<0.001, n=227
right paracentral	n.s.	r=-0.22, p=0.001, n=229
left rostral middle frontal	n.s.	r=-0.22, p=0.001, n=230
right *pars triangularis*	n.s.	r=-0.22, p=0.001, n=229
right *pars opercularis*	n.s.	r=-0.22, p=0.001, n=230
left isthmus cingulate	n.s.	r=-0.21, p=0.001, n=228
left superior frontal	n.s.	r=-0.22, p=0.001, n=228
left transverse temporal	n.s.	r=-0.21, p=0.001, n=229
left frontal pole	r=-0.23, p=0.001, n=198	n.s.

Statistical results of (Spearman's r) correlations between regional change rates and the number of manic or the combined number of manic, hypomanic, and mixed episodes are listed. Only results for regions that were significant after correction for multiple testing are shown. See Data S2 for detailed results and results in other brain areas. n.s.: not significant after adjusting for multiple testing. Overall, only negative associations, and no correlations with surface area or subcortical volumes were observed. Anatomical locations of the regions listed are shown in **Figure 5**.

Study limitations

Although structural brain changes are often interpreted as neuronal gain or loss, the imaging method we used cannot reveal what biological mechanisms underlie the observed MRI-derived brain changes (69). Further, potential signal drifts of MRI scanners are a common problem in longitudinal studies. Here, each center used the same scanner for baseline and follow-up investigation, and a potential signal drift would similarly affect both BD and HC within-site data. This makes it unlikely that possible scanner drifts would explain the observed group differences.

Groups differed in inter-scan interval, but this was accounted for by the use of annualized change rates. BD patients were younger than HC, and age-related brain changes may be of larger magnitude in older people (70, 71). However, age did not correlate with cortical change rates in our study; age only correlated positively with change rates of ventricular volumes in HC, but not in BD. In addition, age was used as covariate, accounting for individual age-related variation in change rates, and results obtained from sensitivity analyses in age-range matched adults did not change our conclusions (Data S1). If age difference nevertheless would have affected our results, we would expect group differences in ventricular volume changes to be even more pronounced if groups were of same age. However, whether and how longitudinal brain changes in BD depend on age remains to be investigated in future studies.

We analyzed two-time point MRI data. Change rates for patients that were lost to follow remain therefore unknown. Future studies are warranted to investigate such subgroups and to include additional timepoint data, which may improve change rate accuracy and can potentially detect non-linear relationships between brain changes and time. It is also possible that even larger sample sizes will have power to detect smaller effects; with the present sample, we had 98% power to detect group differences with an effect size of 0.25.

Although there was no indication that our results were affected by demographic or clinical variables, the results of the respective follow-up tests should be interpreted with caution as they were performed on subsamples; how cortical changes or the number of mood episodes relate to medication effects can be better addressed using refined between time point cumulative medication use data and in randomized clinical trials. However, this is it not feasible with long-term studies. Moreover, it is challenging to accurately assess the number of mood episodes, especially in cases that did

not require hospitalizations or rely on self-report. Also, how the number of hospitalizations as well as the duration of mood episodes relate to longitudinal brain changes in BD remains to be investigated in future studies.

Although the ROI approach provides better comparability to previous studies that used the same brain parcellation method, analyses with higher regional resolution, e.g., voxel-wise or surface-based vertex-wise analyses, could potentially reveal focal cortical variations that remained undetected at the ROI level.

Finally, although the investigated cohort is ecologically valid and represents typical clinical cohorts of BD patients, our findings do not allow conclusions about brain changes that occur in the natural course of BD if untreated, due to obvious is ethical concerns. Also, the high percentage of 'white' identified participants warrants future investigations in different populations. Although we attempted to parse patient groups with potential differential brain trajectories, such as those that experienced frequent manic episodes, refined data-driven analyses aimed at the identification of other potential subpopulations in even larger samples are warranted.

Supplemental References:

1. Brouwer RM, Panizzon MS, Glahn DC Genetic influences on individual differences in longitudinal changes in global and subcortical brain volumes: Results of the ENIGMA plasticity working group. 2017; 38: 4444-4458.

2. Dale AM, Fischl B, Sereno MI Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*. 1999; 9: 179-194.

3. Fischl B, Sereno MI, Dale AM Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*. 1999; 9: 195-207.

4. Fischl B, Dale AM Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc Natl Acad Sci U S A*. 2000; 97: 11050-11055.

5. Fischl B, van der Kouwe A, Destrieux C, Halgren E, Segonne F, Salat DH, et al. Automatically parcellating the human cerebral cortex. *Cereb Cortex*. 2004; 14: 11-22.

6. Fischl B, Salat DH, van der Kouwe AJ, Makris N, Segonne F, Quinn BT, et al. Sequence-independent segmentation of magnetic resonance images. *Neuroimage*. 2004; 23 Suppl 1: S69-84.

7. Reuter M, Schmansky NJ, Rosas HD, Fischl B Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*. 2012; 61: 1402-1418.

8. Hibar DP, Westlye LT, van Erp TG, Rasmussen J, Leonardo CD, Faskowitz J, et al. Subcortical volumetric abnormalities in bipolar disorder. *Mol Psychiatry*. 2016; 21: 1710-1716.

9. Hajek T, Bauer M, Simhandl C, Rybakowski J, O'Donovan C, Pfennig A, et al. Neuroprotective effect of lithium on hippocampal volumes in bipolar disorder independent of long-term treatment response. *Psychol Med*. 2014; 44: 507-517.

10. Cousins DA, Aribisala B, Nicol Ferrier I, Blamire AM Lithium, gray matter, and magnetic resonance imaging signal. *Biol Psychiatry*. 2013; 73: 652-657.

11. Monkul ES, Matsuo K, Nicoletti MA, Dierschke N, Hatch JP, Dalwani M, et al. Prefrontal gray matter increases in healthy individuals after lithium treatment: a voxel-based morphometry study. *Neurosci Lett*. 2007; 429: 7-11.

12. Sun YR, Herrmann N, Scott CJM, Black SE, Khan MM, Lanctot KL Global grey matter volume in adult bipolar patients with and without lithium treatment: A meta-analysis. *J Affect Disord*. 2018; 225: 599-606.

13. Ho BC, Andreasen NC, Ziebell S, Pierson R, Magnotta V Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. *Arch Gen Psychiatry*. 2011; 68: 128-137.

14. Abé C, Liberg B, Song J, Bergen SE, Petrovic P, Ekman CJ, et al. Longitudinal Cortical Thickness Changes in Bipolar Disorder and the Relationship to Genetic Risk, Mania, and Lithium Use. *Biological Psychiatry*. 2020; 87: 271-281.

15. Abé C, Ekman CJ, Sellgren C, Petrovic P, Ingvar M, Landen M Manic episodes are related to changes in frontal cortex: a longitudinal neuroimaging study of bipolar disorder 1. *Brain*. 2015; 138: 3440-3448.

16. Glahn DC, Bearden CE, Bowden CL, Soares JC Reduced educational attainment in bipolar disorder. *J Affect Disord*. 2006; 92: 309-312.