Breaking the Norm: Population-Scale Normative Modeling of Brain Structure in Depression and Anxiety

Julius Wiegert[1,2], Sebastián Marty-Lombardi[1,2], Jailan Oweda[1,2], Esra Lenz[1,2], Peter Ahnert[3], Klaus Berger[4], Hermann Brenner[5, 6], Josef Frank[7], Hans J. Grabe[8], Karin Halina Greiser[9], Johanna Klinger-König[8], André Karch[4], Michael Leitzmann[10], Claudia Meinke-Franze[11], Rafael Mikolajczyk[12,13], Frauke Nees[14,15,16], Thoralf Niendorf[17], Oliver Sander[18], Carsten Oliver Schmidt[11], Steffi G. Riedel-Heller[19], Kerstin Ritter[20,21,22], Annette Peters[23,24,25], Tobias Pischon[26,27,28], Stephanie Witt[7,29], Johannes Nitsche[1,2], Joonas Naamanka[1,2,30], Sebastian Volkmer[1,2], Antonia Mai[1,2], Amrou Abas[1,2], Xiuzhi Li[1,2], Andreas Meyer-Lindenberg[2], Tobias Gradinger[1,2], Fabian Streit[1,2,29], Urs Braun[1,2], and Emanuel Schwarz*[1,2,29]

[1]Hector Institute for Artificial Intelligence in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[3]Universität Leipzig, Institute for Medical Informatics, Statistics and Epidemiology

[4]Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany

[5]German Cancer Research Center (DKFZ), Heidelberg, Germany

[6]Network Aging Research, Heidelberg University, Heidelberg, Germany

[7]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[8]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany

[9]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[10]Department of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

[11]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

[12]Institute for Medical Epidemiology, Biometrics, and Informatics, Medical Faculty of the Martin Luther University Halle-Wittenberg, Halle (Saale)

[13]German Center for Mental Health (DZPG), partner site Halle-Jena-Magdeburg, Halle (Saale), Germany
[14]Institute of Medical Psychology and Medical Sociology, University Medical Center Schleswig-Holstein, Kiel University, Kiel, Germany
[15]Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[16]Institute of Medical Psychology, Faculty of Medicine, Ludwig-Maximilians-Universität München, Munich, Germany
[17]Berlin Ultrahigh Field Facility (B.U.F.F.), Max Delbrueck Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany
[18]Clinic for Rheumatology and Hiller Research Center, University Hospital, Heinrich-Heine-University Düsseldorf, Germany
[19]Institute of Social Medicine, Occupational Health and Public Health (ISAP), Leipzig University, Leipzig, Germany
[20]Department of Machine Learning, Hertie Institute for AI in Brain Health, University of Tübingen, Germany
[21]Department of Psychiatry and Neurosciences, Charité – Universitätsmedizin Berlin (corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Berlin, Germany
[22]Tübingen AI Center, Tübingen, Germany
[23]Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Neuherberg, Germany
[24]Chair of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Munich, Germany
[25]German Center for Mental Health (DZPG), partner site Munich, Munich, Germany
[26]Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Molecular Epidemiology Research Group, Berlin, Germany
[27]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[28]Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Biobank Technology Platform, Berlin, Germany
[29]German Center for Mental Health (DZPG), Partner Site Mannheim - Heidelberg - Ulm, Germany
[30]SleepWell Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland

September 2025

Abstract

We applied deep normative modeling to structural MRI data from two large cohorts (German National Cohort, $N \approx 29{,}000$ and UK Biobank, $N \approx 25{,}000$) to characterize individual-level brain deviations along symptom dimensions of depression, anxiety, and alcohol use. Each brain was embedded into a 256-dimensional latent space, allowing us to quantify both the magnitude and direction of deviation from a normative reference trained on the non/low-symptomatic subpopulation. Deviation magnitude increased with symptom severity, and directional patterns separated mood-anxiety and alcohol-use tendencies. These deviation axes generalized across cohorts and supported individual-level classification of symptomatic group membership, especially at higher symptom levels. Combining deviations with polygenic risk scores improved classification performance, particularly for depressive and anxiety measures, indicating complementary contributions of imaging and genetics. Our findings demonstrate that structural brain deviations reflect meaningful, continuous variation in affective and behavioral symptoms.

Contact: emanuel.schwarz@zi-mannheim.de

[*] Corresponding author

Introduction

Major depressive disorder (MDD) and generalized anxiety disorder (GAD) are some of the most prevalent psychiatric diseases worldwide (MDD 8% and GAD 3% 1-year prevalence), significantly impacting individuals and society and thus making them a major contributor to the global burden of disease [1, 2, 3].

MDD is an affective disorder that is marked by impacts on multiple psychological domains, including mood (e.g., persistent feelings of hopelessness), motivation (e.g., chronic loss of energy), general interest in life (e.g., anhedonia), and cognition (e.g., impaired concentration) [4]. It often manifests in somatic symptoms such as sleep disturbances (e.g., insomnia), changes in body weight (e.g., unintentional weight loss), and altered psychomotor activity (e.g., agitation). MDD is associated with suicidal ideation, and about 2–8% of inpatients with the disorder die by suicide [5].

GAD is an anxiety disorder (ANX) characterized by excessive and persistent anxiety and worry lasting more than six months, typically concerning more than one of everyday domains such as health, work, or finances [6]. These worries are difficult to control and are accompanied by symptoms affecting psychomotor activity (e.g., restlessness), motivation (e.g., easy fatigability), cognition (e.g., difficulty concentrating), and sleep (e.g., disrupted sleep), as well as increased muscle tension and heightened irritability. Both MDD and ANX show moderate heritability (twin-based estimates of 30–40% [7, 8]), and large-scale GWAS have identified hundreds of common variants of small effect [9, 8]. These findings enable the construction of polygenic risk scores (PRS), which can be applied to independent cohorts to capture a fraction of genetic liability, although their predictive utility in psychiatric disorders remains modest [10].

While MDD and GAD are currently defined as distinct disease entities, they share many commonalities. First, there is substantial overlap in typically affected domains (e.g., cognition and sleep). Second, the comorbidity between MDD and GAD is estimated to be as high as 26% for individuals with a principal diagnosis of GAD also receiving a secondary diagnosis of MDD [11]. Third, common psychopharmacological treatments overlap considerably, with selective serotonin reuptake inhibitors (SSRIs) being first-line treatments for both disorders. Furthermore, the substantial genetic overlap—estimated at a correlation of 1.0 in females and 0.74 in males [12]—together with evidence for similarities in typical patterns of neural function and structure [13, 14], supports the hypothesis of shared biological underpinnings between the two disorders. Those commonalities complicate their disentanglement and have led to discussions about whether ANX and MDD share the same neurobiological mechanisms [13].

Both disorders exhibit substantial heterogeneity at the individual level, including variation in symptom manifestation, age of onset, and number of episodes [15, 16]. This heterogeneity is captured by the common classification of the diseases, the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) [17]. While diagnostic systems like the DSM-5 use categorical criteria based on thresholds, the underlying vulnerability model and symptom distribution in the population are continuous [18]. Apart from general disease criteria and differential diagnostic exclusion criteria, for MDD, the DSM-5 lists nine symptoms/affected domains (reflected in the PHQ-9 questionnaire items), of which at least five need to be present/affected. Two individuals diagnosed with MDD may share only a single symptom or affected domain. For GAD, a diagnosis requires the presence of at least three out of six possible symptoms, which similarly permits considerable variation in symptom profiles across individuals. Furthermore, the severity of each symptom can differ markedly between symptomatic subjects, adding another layer of variability [19, 20].

This clinical heterogeneity is mirrored on the biological level, particularly in attempts to predict MDD using brain imaging. A recent benchmark study using the ENIGMA MDD dataset (N = 5,365) demonstrated that predicting MDD from structural magnetic resonance imaging (sMRI) data on brain region level remains highly challenging [21]. Machine learning models achieved approximately 63% balanced accuracy under ideal conditions contrasting MDD patients to healthy controls (HCs), which dropped to near chance levels (52%) after harmonization for site effects. Stratification by clinical variables (e.g., age of onset, antidepressant use) did not improve classification. These findings underscore that current structural brain features lack a sufficiently consistent signal for reliable individual-level prediction of MDD in a classification setting, at least using the methodology employed in this study.

Implicitly, such conventional supervised learning approaches rely on the assumption that individuals sharing the same diagnosis (e.g., MDD or GAD) exhibit underlying biological similarity and that diagnostically distinct groups are meaningfully separable. These methods also fail to fully exploit the disproportionate availability of imaging data from healthy individuals in large population cohorts, compared to the more limited data available for symptomatic subjects with specific diagnoses.

Normative modeling has gained attraction [22, 23, 24], as it models individual-level characteristics in relation to a reference model, typically inferred using data of a healthy population [22]. This allows quantifying an individual's deviation from the learned notion of normality and may thus aid in deciphering biological and clinical heterogeneity. Normative modeling has been successfully applied to a plethora of psychiatric disorders, including autism spectrum disorder [25], schizophrenia [26], and dementia [27, 28]. In MDD, normative models of the brain-functional connectome have pointed to distinct neurophysiological subtypes and revealed substantial inter-subject variability in functional connectivity deviations [29, 30]. Brain-structural deviations found in MDD via normative modeling have further been found to improve diagnostic classification, predict treatment response, and occur in early- and late-onset cases [31, 32].

Many earlier normative models have relied on univariate or region-based measures, modeled independently, often reducing deviations to scalar scores per brain region [22]. Such approaches may be less suited to capturing subtle, distributed, or nonlinear patterns of brain variation that are thought to play a role in psychiatric disorders. Modern AI methods have shifted brain MRI analysis from hand-crafted features to end-to-end models that operate on minimally processed whole-image volumes [33]. This enables the extraction of distributed, disease-relevant patterns directly from raw spatial data. At the same time, population-scale sMRI datasets like the German National Cohort (NAKO, $N \approx 30\,000$) [34] and the UK Biobank (UKB, $N \approx 50\,000$) [35] became available. Our framework addresses the described research gaps by encoding whole-brain voxel-wise data into a nonlinear latent space and modeling deviations as both magnitude and direction within this embedding. This allows detection of transdiagnostic dimensions and fine-grained individual heterogeneity beyond the reach of conventional normative models. Seizing this opportunity, we implemented a novel normative modeling framework that leverages deep learning-based feature extraction from whole-brain sMRI. The NAKO cohort served as the discovery dataset, while the UKB was used for external validation. We pursued four main aims: (i) to develop a deep normative model based on sMRI data to characterize brain variation in individuals with symptoms related to MDD and GAD; (ii) to quantify the magnitude of individual deviations from healthy normative variation among participants with MDD- and GAD-related symptoms; (iii) to assess the directionality and specificity of these deviations by comparing them to those observed in individuals with alcohol use disorder (AUD)-related symptoms, included as a positive control group given its well-documented structural brain alterations [36, 37]; and (iv) to evaluate the predictive utility of individual deviations and their associations with genetic risk.

For our main analysis, we operationalized MDD, GAD, and AUD using questionnaire-based sum scores—PHQ-9 [38], GAD-7 [20], and AUDIT-C [39]. We selected these three scores to ensure the availability of ordinal severity measures for each condition and because they are available in both the NAKO and UKB cohorts. To further substantiate our findings, we conducted additional analyses using alternative operationalizations, including self-reported physician diagnoses and lifetime diagnoses obtained from hospital records and clinical interviews.

Our approach provides a novel, data-driven framework for dissecting the heterogeneity of complex mental health conditions, specifically MDD and GAD, and supports the development of more personalized diagnostic and therapeutic strategies tailored to these highly overlapping and heterogeneous disorders.

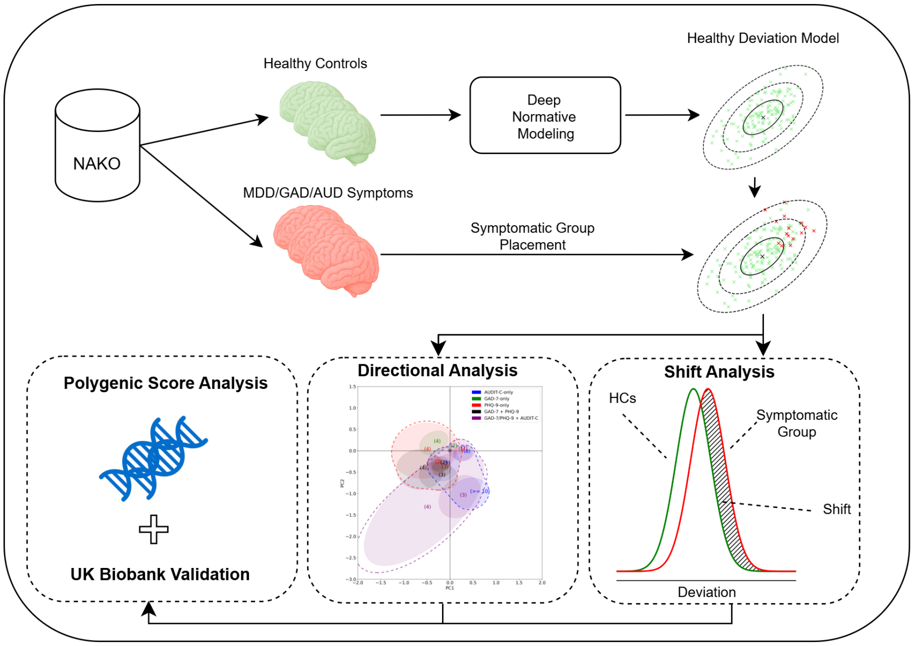

Figure 1: The overall workflow. We began by training a deep normative model on healthy controls (HCs) from the NAKO cohort (German National Cohort) to learn a low-dimensional representation of normative brain variation. This model defined a normative distribution in the learned embedding space, against which we contrasted symptomatic group data, including symptoms of major depressive disorder (MDD), anxiety (ANX), and alcohol use disorder (AUD), in two main analyses. The shift analysis quantified whether symptomatic subjects' deviations exceeded the natural heterogeneity of the HC population. The directional analysis characterized how symptomatic subjects deviated from the normative manifold by visualizing the directionality of case emergence relative to HC variability. Finally, we performed external validation on the UKB cohort and further tested whether case-specific deviations were associated with genetic risk for the corresponding phenotype.

Results

Sample Characteristics

We used two large-scale, population-based cohorts: the NAKO [40, 41], a nationwide German initiative providing extensive health and imaging data with a particular emphasis on the causes and early stages of common chronic diseases, and the UKB [35], a UK-based resource offering comprehensive health, genetic, and imaging information. After filtering out participants with missing questionnaire data on the PHQ-9, GAD-7, or AUDIT-C, as well as those with neurodegenerative diseases, we retained 29,357 sMRI samples from NAKO and 24,838 from UKB for our analyses (see Supplementary Tables 1–6 for details).

Participants were stratified into symptom severity classes based on established cutoff scores on self-report questionnaires, reflecting current symptom burden of MDD, GAD, and AUD rather than formal diagnosis. For depressive symptoms (PHQ-9), scores of 0–4 indicated none to minimal symptoms, 5–9 mild, 10–14 moderate, 15–19 moderately severe, and 20–27 severe symptoms [38]. For anxiety symptoms (GAD-7), scores of 0–4 were considered minimal, 5–9 mild, 10–14 moderate, and 15–21 severe [20]. Alcohol-related risk (AUDIT-C) was categorized as low (0–4), increasing (5–7), higher risk (8–10), and possible dependence (11-12) [39]. Due to pragmatic sample size considerations, we further split individuals with scores above 7 into separate groups: 8, 9, and ≥ 10. This approach allowed us to retain some resolution in symptom severity while ensuring sufficiently large group sizes for statistical power.

HCs were defined using harmonized thresholds across the three instruments in both cohorts. Specifically, individuals scoring below 5 on both PHQ-9 and GAD-7, and below 8 on AUDIT-C, were included. Additional exclusion criteria for HCs were applied to remove participants with indicators of psychiatric conditions: in NAKO, this included lifetime diagnoses assessed via the MINI interview [42] and self-reported diagnoses related to MDD, ANX, or panic disorder (PD); in UKB, exclusions were based on matching the respective ICD-10 codes from hospital records as well as self-reported diagnoses.

The final group distributions and demographics used in subsequent analyses are summarized in Table 1 (NAKO) and Table 2 (UKB).

Group	**Criterion**	**Count**	**Age** $\mu(\pm\sigma)$	**Females (%)**
HC	–	15 941	48 (±12)	39
AUDIT-C (8)	AUDIT-C sum $= 8$	633	49 (±13)	15
AUDIT-C (9)	AUDIT-C sum $= 9$	249	49 (±13)	10
AUDIT-C ≥ 10	AUDIT-C sum ≥ 10	111	51 (±13)	11
MDD (mild)	PHQ-9 sum $\in [5, 9]$	6 829	51 (±11)	53
MDD (moderate)	PHQ-9 sum $\in [10, 14]$	1 464	46 (±12)	56
MDD (moderately severe)	PHQ-9 sum $\in [15, 19]$	423	45 (±12)	55
MDD (severe)	PHQ-9 sum ≥ 20	181	47 (±12)	55
MDD (MINI)	MINI Diagnosis MDD	4 752	59 (±11)	54
MDD (Diagnosis)	Doctor's Diagnosis	3 636	50 (±11)	60
GAD (mild)	GAD-7 sum $\in [5, 9]$	5 566	47 (±12)	54
GAD (moderate)	GAD-7 sum $\in [10, 14]$	1 058	46 (±12)	56
GAD (severe)	GAD-7 sum ≥ 15	320	46 (±12)	59
ANX/PD (Diagnosis)	Doctor's Diagnosis	1 716	49 (±12)	59

Table 1: Diagnosis counts and demographics in the NAKO cohort (German National Cohort). Abbreviations: HC, healthy controls; AUDIT-C, Alcohol Use Disorders Identification Test–Consumption; PHQ-9, Patient Health Questionnaire-9; MDD, major depressive disorder; MINI, MINI International Neuropsychiatric Interview; GAD-7, Generalized Anxiety Disorder-7; GAD, generalized anxiety disorder; ANX/PD, anxiety or panic disorder.

Group	**Criterion**	**Count**	**Age** $\mu(\pm\sigma)$	**Females (%)**
HC	–	13 582	65 (±7)	53
AUDIT-C (8)	AUDIT-C sum $= 8$	845	63 (±7)	32
AUDIT-C (9)	AUDIT-C sum $= 9$	611	62 (±7)	25
AUDIT-C ≥ 10	AUDIT-C sum ≥ 10	631	63 (±7)	22
AUD	ICD-10	192	64 (±8)	25
MDD (mild)	PHQ-9 sum $\in [5, 9]$	3 357	61 (±7)	63
MDD (moderate)	PHQ-9 sum $\in [10, 14]$	770	61 (±8)	62
MDD (moderately severe)	PHQ-9 sum $\in [15, 19]$	267	60 (±7)	66
MDD (severe)	PHQ-9 sum ≥ 20	132	58 (±6)	70
MDD (Diagnosis)	ICD-10	754	62 (±8)	65
MDD (Diagnosis)	Doctor's Diagnosis	3 918	63 (±7)	66
GAD (mild)	GAD-7 sum $\in [5, 9]$	2 927	61 (±8)	66
GAD (moderate)	GAD-7 sum $\in [10, 14]$	559	61 (±8)	65
GAD (severe)	GAD-7 sum $\in [15, 21]$	276	59 (±7)	67
GAD	ICD-10	19	63 (±7)	66
ANX (Diagnosis)	ICD-10	654	63 (±7)	65
ANX (Diagnosis)	Doctor's Diagnosis	2 646	63 (±7)	63

Table 2: Diagnosis counts and demographics in the UKB cohort (UK Biobank). Abbreviations: HC, healthy controls; AUDIT-C, Alcohol Use Disorders Identification Test–Consumption; AUD, alcohol use disorder; PHQ-9, Patient Health Questionnaire-9; MDD, major depressive disorder; GAD-7, Generalized Anxiety Disorder-7; GAD, generalized anxiety disorder; ANX, anxiety disorder; ICD-10, International Classification of Diseases, 10th Revision.

Shift Analysis

Our deep normative model was constructed on sMRI data of the 15,941 NAKO HCs. Specifically, we used whole-brain voxelwise T1-weighted MRI grey matter maps, spatially normalized to a common template and downsampled to an 80×80×80 voxel grid. Our model compressed each image into a 256-dimensional latent embedding. In this space, we quantify the deviations for symptomatic groups from the HC centroid.

Figure 2A depicts the distribution of deviations observed in severely symptomatic PHQ-9, GAD-7, and AUDIT-C groups compared to an independent holdout group of 1,594 NAKO HCs (10 %), illustrating how symptomatic groups diverge from normative variation. **Figure 2B** shows that the magnitude of deviation is associated with the likelihood of belonging to at least moderately symptomatic groups. We quantified the proportion of each group whose deviations exceeded the range observed among HCs (**Figure 3A**, henceforth referred to as "shift"). At the symptomatic group level, participants with higher symptom severity levels for MDD and GAD (encoded on a scale from 1 [mild] to 4 [severe]) exhibited significantly larger shifts. Spearman rank correlations on the group level between symptom severity and shift revealed strong positive associations when jointly analyzing the PHQ-9 and GAD-7 groups ($\rho = 0.92$, $p = 0.006$), i.e., levels 1 to 4 for PHQ-9 and 2 to 4 for GAD-7 and their corresponding shift values.

The group with the highest shift was alcohol-related, with a shift of over 20% in the top AUDIT-C (≥ 10) group (**Figure 3 A**). As depicted in more detail in the Supplementary Figure 3, we observed that recent, severe symptoms of depression and anxiety exhibited greater deviations compared to lifetime diagnoses captured by the lifetime-related interviews, such as the MINI interview and self-reported clinical diagnoses by a physician. Exact (corrected) p-values for all groups are provided in Supplementary Tables 10 and 11. Finally, we present the results of the sex-stratified analysis in Supplementary Figure 4, which demonstrate that the trends persist.

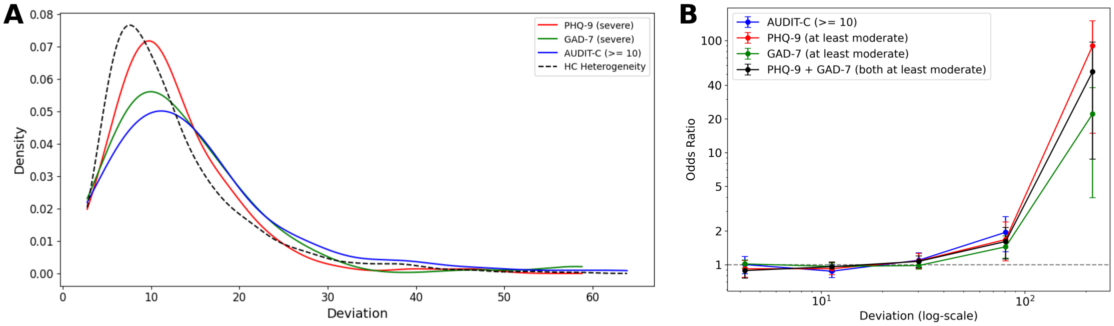

Figure 2: **A**: Density estimates of deviations from the norm for three symptomatic groups. Deviations are computed based on our deep normative pipeline. The dashed black line represents the natural heterogeneity among independent healthy controls (HCs) set aside for evaluation, serving as the normative reference. The areas where symptomatic group distributions exceed the HC distribution, particularly at higher deviation values, form the basis for computing the fraction of symptomatic group deviation not captured by the natural variability in HCs (the *shift*). **B**: Relationship between deviation from the normative distribution and the OR of having Patient Health Questionnaire-9 (PHQ-9) $\geq$ 10, Generalized Anxiety Disorder-7 (GAD-7) $\geq$ 10, and an Alcohol Use Disorders Identification Test–Consumption (AUDIT-C) $\geq$ 10. Deviation scores were stratified into 5 bins with edges defined on a base-10 logarithmic scale, spanning the minimum to maximum observed patient distances. This ensured relatively finer resolution in the distribution tail where case enrichment was expected. A strong increase in odds ratio (OR) with increasing deviation illustrates that individuals further from the norm are substantially more likely to be symptomatic. Error bars denote 95% confidence intervals obtained via 1000 bootstrap resamples. In the highest deviation bin (215.26), there was 1 HC, 0 AUDIT-C samples, 6 PHQ-9 samples, 7 GAD-7 samples, and 6 PHQ-9 + GAD-7 samples.

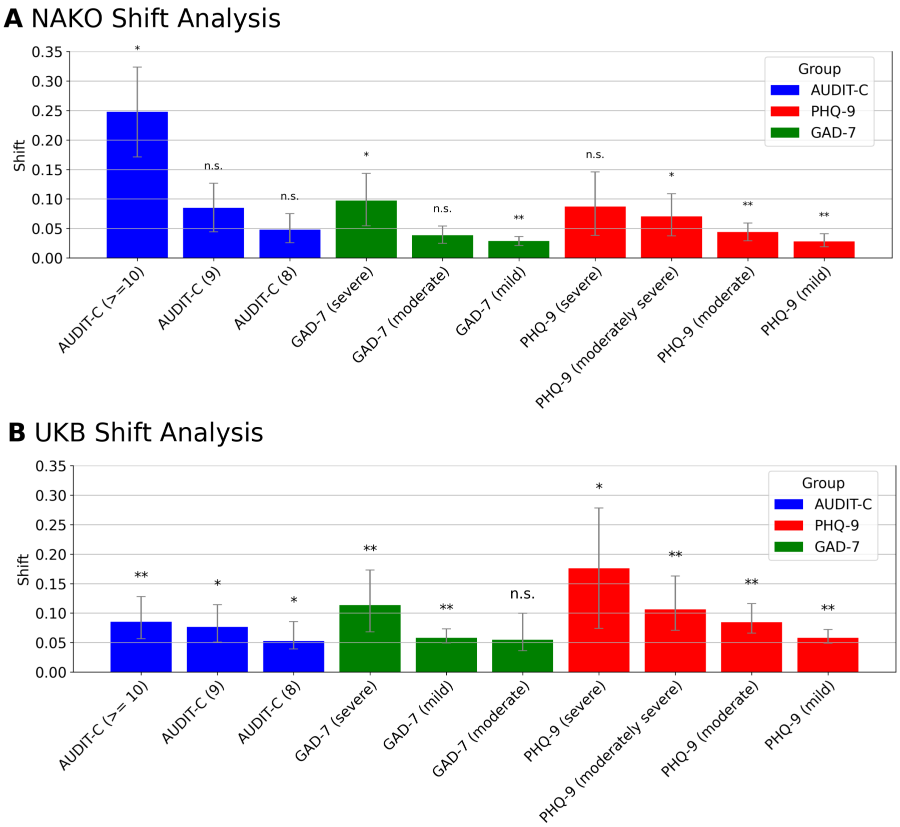

Figure 3: Fraction of symptomatic group deviations unexplained by healthy control variability (*shift*) for all symptomatic groups for NAKO cohort (German National Cohort) (**A**) and the UKB cohort (UK Biobank) (**B**). To assess statistical significance, deviation scores were modeled as the dependent variable in a linear regression with group membership as the main predictor, adjusting for sex, age, age-squared, and sex–age interactions. Asterisks indicate the level of statistical significance after multiple testing correction using Benjamini-Hochberg correction [43]: * $p_{FDR} < 0.05$, ** $p_{FDR} < 0.01$. Error bars indicate 95% confidence intervals obtained from 1000 bootstrap samples. Symptomatic groups are defined by AUDIT-C (Alcohol Use Disorders Identification Test–Consumption), PHQ-9 (Patient Health Questionnaire-9), and GAD-7 (Generalized Anxiety Disorder-7) thresholds.

Directional Analysis

We assessed directional similarity between L2-normalized 256-dimensional deviation vectors by computing mean pairwise cosine similarity within and across mutually exclusive symptom severity groups based on PHQ-9, GAD-7, and AUDIT-C scores. For each symptom severity stratum within these scales, we first computed the median deviation vector, and the similarity analysis was performed on these median vectors rather than on individual participant vectors. Cosine similarity was chosen because it captures the alignment of deviation vectors in the embedding space independently of their magnitude, thereby isolating directional differences. The resulting groupings were defined without overlap, ensuring that similarities were evaluated in the absence of comorbid symptoms. The combined mood-anxiety category (PHQ-9 and GAD-7 strata) showed higher within-group similarity (mean = 0.316) than similarity with the AUD group (mean = 0.212), yielding a difference of 0.104 (one-sided permutation test (10,000 permutations) for within > cross similarity: $p = 0.0063$). This indicates that mood and anxiety disorders share more similar deviation directions and are more distinct from AUD in this deviation-vector space. In contrast, when comparing the mutually exclusive PHQ-9 and GAD-7 strata, pooled within-class similarity (0.269) was lower than cross-class similarity between the two strata (0.322; difference $= -0.053$, $p = 0.748$), suggesting that depressive and anxiety symptom deviation vectors are intermixed rather than forming separable subclusters.

To visualize the directional deviations, we use principal component analysis (PCA). We again computed group representing median deviation vectors in the representation space learned by our normative model, and assessed whether specific symptom profiles share common structural signatures and how these relate to one another and the normative population (**Figure 4A**).

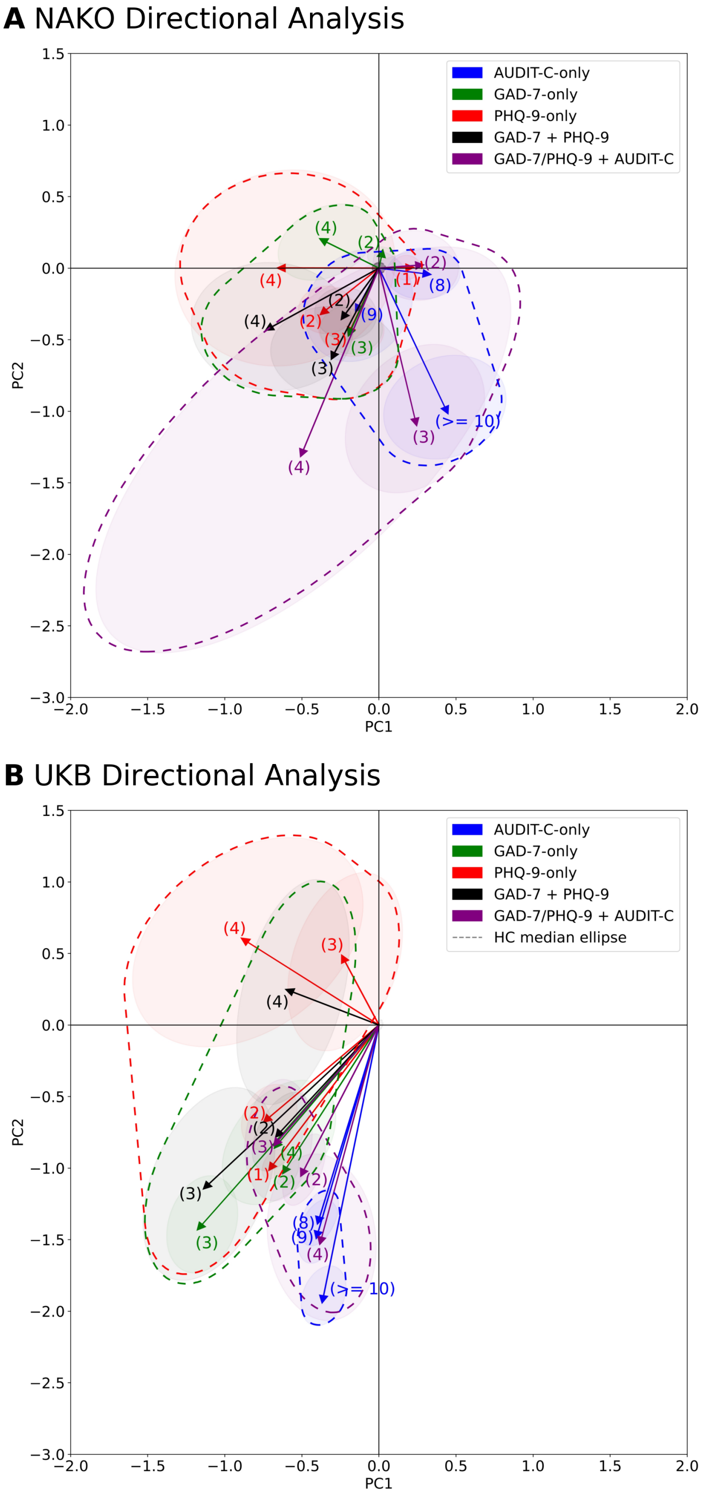

Figure 4: Directional analyses based on principal component analysis PCA projection (two dimensions, 21% explained variance) of normative deviation vectors across diagnostic groups in the NAKO cohort (German National Cohort) (**A**) and the UKB cohort (UK Biobank) (**B**). Shaded ellipses denote the ±1 SD contour of the bootstrap (1,000 resamples) distribution of the group-level geometric median vectors. The dark ellipse around the origin indicates the median deviation among healthy controls (HCs), serving as a reference for interpreting the deviations relative to normative variability. Numbers indicate severity levels, defined by PHQ-9 (Patient Health Questionnaire-9) and GAD-7 (Generalized Anxiety Disorder-7) scores, and the sum for AUDIT-C (Alcohol Use Disorders Identification Test–Consumption). “MDD/GAD + ALC” denotes individuals with either acute MDD or GAD symptoms and elevated alcohol use (AUDIT-C $\geq$ 10). For PHQ-9, levels (1) to (4) correspond to mild, moderate, moderately severe, and severe symptoms, respectively. For GAD-7, levels (2) to (4) correspond to mild, moderate, and severe symptoms. The scattered area includes the ellipses of GAD, PHQ, GAD+PHQ groups (red), AUDIT-C groups (blue) and comorbid groups (purple).

To visualize potential heterogeneity within clinical phenotypes, we further subdivided participants based on comorbidity status. We kept the distinction between depressive and anxiety symptoms, differentiating between individuals with comorbid PHQ-9 and GAD-7 symptom elevation and those with elevations on only one of the two scales. This allowed us to assess whether the resulting three groups (PHQ-9, GAD-7, and GAD-7 + PHQ-9) show deviations in similar or distinct directions in the learned space.

An analogous stratification was introduced for alcohol-related symptoms. Participants with elevated AUDIT-C scores were divided into those with comorbid depressive or anxiety symptoms (i.e., also elevated PHQ-9 or GAD-7 scores) and those with elevated AUDIT-C scores only.

Two broad patterns emerge. First, AUDIT-C–defined groups (blue scattered ellipse) occupy a direction largely distinct from the mood-anxiety symptom axes: their central deviation estimates cluster away from HCs along a subspace different from that traced by PHQ-9 and GAD-7 groups (red and green scattered ellipses). This separation is consistent with at least partially dissociable structural alterations associated with high alcohol consumption versus depressive/anxiety symptoms. Within the alcohol domain, the AUDIT-C strata show a graded displacement of the group vectors with higher thresholds ((8)→(9)→($\geq$10)), suggesting an exposure–response trend. Second, lower symptom-burden PHQ-9 and GAD-7 groups ((1)/(2)) lie close to the HC origin, indicating minimal group-level deviation in this projection, whereas higher symptom levels shift further from the origin. Finally, participants meeting both high AUDIT-C and elevated PHQ-9/GAD-7 criteria cluster in an intermediate region between the "alcohol" and "mood-anxiety" directions, consistent with additive or mixed deviation patterns rather than a purely alcohol- or purely mood-anxiety-like profile.

As detailed in the Supplementary Figure 5, we observed consistent deviation patterns for both self-reported doctor diagnoses of MDD and GAD, as well as lifetime MDD diagnoses assessed via the MINI interview, within the NAKO cohort.

Finally, we assessed whether the directional analysis in the 256-dimensional space was influenced by sex. We compared the cosine similarity between the median deviation vectors of sex-stratified groups and their corresponding joint-sex groups. For women, the mean cosine similarity was 0.682 ± 0.221, and for men it was 0.786 ± 0.127. This reflects a modest difference, with men being, on average, closer to the joint-sex solution. However, as shown in Supplementary Figure 6, this difference disappears after filtering out small groups, suggesting that it likely arises from greater variability in the central tendency estimates of smaller samples.

External Validation: UKB

To evaluate the generalizability of the deviation shifts and directional patterns observed in NAKO, we applied the deep normative modeling model to sMRI data from the UKB. We refit the normative reference distribution on a subset of independent 90% UKB HCs (N=12,223) for the shift analysis, in order to mitigate cohort effects. Using the original NAKO HC reference to evaluate UKB symptomatic groups led to systematically inflated deviation estimates (median 44.24 vs. 11.93 when using the NAKO parameters), likely due to cohort effects. Consequently, within-cohort references were required to ensure a valid comparison of the shift analysis results. Importantly, this adjustment was confined to the shift analysis; all other components of the pipeline remained fully externally validated.

Figure 3B shows that, although PHQ-9–related deviation shifts tend to be larger and AUDIT-C–related shifts tend to be smaller in UKB than in NAKO, the overall pattern remains consistent:

groups with increasing symptom severity exhibit progressively larger deviation shifts.

In the UKB 256-dimensional deviation space, mood-anxiety symptom groups (PHQ-9, GAD-7) again showed higher within-group similarity (mean = 0.712) than similarity with the AUDIT-C groups (mean = 0.623), a mean difference of 0.089 ($p = 0.0413$). Within the mood-anxiety domain, the pooled within-class similarity for MDD and GAD (0.723) exceeded their cross-class similarity (0.632) by 0.091, although—as in NAKO—this difference was not statistically significant ($p = 0.1786$). The primary difference compared to NAKO is that cosine similarities are generally higher in UKB, while the relative pattern of group differences remains consistent.

The projection of UKB participants into the directional PCA deviation space derived from NAKO is presented in **Figure 4B**. Several relative patterns are consistent across the two cohorts. In both, AUDIT-C groups (scattered, blue ellipse) cluster in distinct regions from the mood-anxiety symptom groups. The scattered red and green ellipses indicating mood-anxiety symptomatology (PHQ-9, GAD-7, and their combination) occupy largely the same space in UKB as in NAKO. Participants with comorbid mood-anxiety and AUDIT-C symptom elevation again fall in an intermediate region between the "alcohol" and "mood-anxiety" areas. Within the mood-anxiety domain, the GAD-7–only, PHQ-9–only, and combined groups continue to share large overlapping areas, suggesting that even without identical symptom profiles, their deviation patterns remain similar.

Notable differences are also evident. In UKB, all groups appear shifted relative to the healthy control centroid. Moreover, AUDIT-C and low symptom severity groups ((1)/(2)) are generally positioned farther from the HC centroid than in NAKO, suggesting stronger overall deviation in the alcohol-related domain in the UKB sample.

In Supplementary Figure 3, we show that the directions are largely consistent with ICD-10 and self-reported diagnoses for the respective disorders based on UK hospital records.

We further tested whether the prolonged time interval between MRI acquisition and questionnaire completion in the UKB (median absolute deviation: 742 days) compared to the NAKO (median absolute deviation: 15 days) influenced the results. For this analysis, we compared the lowest 20% and highest 20% of time differences with the overall median deviation vectors per group. Across all groups, the mean pairwise cosine similarity was 0.80 ± 0.16 for the lower 20% and 0.71 ± 0.20 for the upper 80%. As shown in Supplementary Figure 7, this difference can be attributed to unstable central tendency estimates, as the gap between the two solutions diminishes when restricting the analysis to larger group sizes.

Importance of Brain Regions

We examined correlations between the 256 normative deviation dimensions and three symptom scores (PHQ-9, GAD-7, AUDIT-C) to identify which dimensions captured clinically relevant variation. Of the 256 dimensions, 14 were significantly (q-values < 0.05) associated with PHQ-9 scores (5.4%), 4 with GAD-7 (1.6%), and 35 with AUDIT-C (13.7%).

To improve interpretability, we next regressed each symptom-associated deviation dimension onto classical brain regions by associating them with 99 regional grey matter volume (GMV) estimates (full list in Supplementary Table 17) from FreeSurfer [44]. The resulting associations were standardized to facilitate comparability across regions and dimensions. For the raw associations and more details, we refer to the Supplementary Figures 8-13.

Figure 5 depicts the spatial distribution of the strongest regional associations. Effects were predominantly localized to the cerebellum across all three symptom dimensions. AUDIT-C exhibited the most pronounced negative association with cerebellar grey matter, with additional negative associations in the cuneus (Supplementary Figures 11 and 12).

For comparison, we also computed Spearman correlations directly between deconfounded FreeSurfer GMV features and each symptom score. These direct associations diverged substantially from those obtained via the embedding-based normative model (see Supplementary Figures 11 and 13).

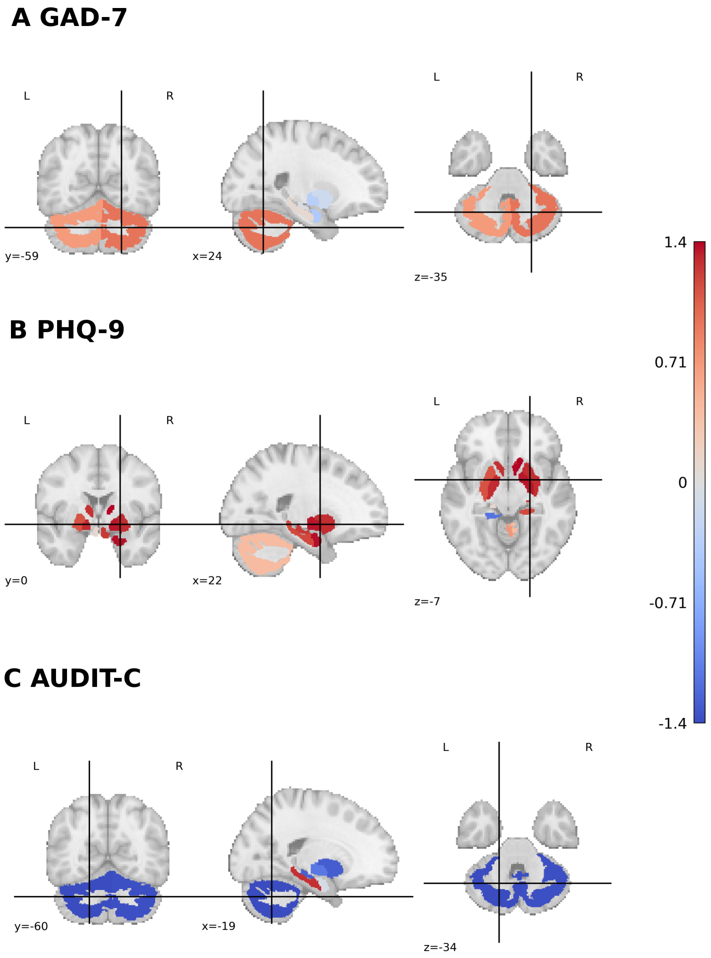

Figure 5: Z-scored associations between brain normative deviations and symptom scores for (A) GAD-7 (Generalized Anxiety Disorder-7), (B) PHQ-9 (Patient Health Questionnaire-9), and (C) AUDIT-C (Alcohol Use Disorders Identification Test–Consumption). Higher values reflect stronger regional associations with individual symptom expression. Associations are z-scored across all symptom scores and brain regions to emphasize relative patterns.

Genetic Associations

We set out to explore whether deviations were associated with polygenic risk for the respective disorders in the UKB cohort. To this end, we computed polygenic risk scores (PRS) for problematic alcohol use (PAU) [45], MDD [9], and ANX [46] using GWAS summary statistics that did not include UKB participants (details in Supplementary Table 9).

We used ANX rather than GAD for two reasons: first, there exist GWAS with substantially greater statistical power for ANX, with a far larger sample size than is available for GAD, leading to more robust PRS estimation; second, although the GAD-7 is nominally a measure of generalized anxiety, it has been shown to capture a broader range of anxiety symptoms beyond GAD [20], making ANX an appropriate genomic proxy. For alcohol use, we selected PAU as it is derived from a GWAS with over one million individuals, providing exceptionally high statistical power, and because it is closely related to AUDIT-C.

We tested three complementary regression models to examine the interplay between genetic risk, brain deviation, and symptoms. First, symptoms were predicted from PRS, deviation, and their interaction. Second, deviation was predicted from PRS. Third, deviation was predicted jointly from PRS and symptoms. All models were adjusted for demographic and genetic covariates (see Methods).

Across all models, no significant interaction effects between PRS and brain deviations on symptom severity were observed. In contrast, the main-effect models revealed significant associations between the genetic risk for MDD and AUD and the deviation dimensions, but not for ANX ($p_{\mathrm{FDR}} < 0.05$). For MDD, 193 dimensions were associated with PRS before adjusting for symptoms and 186 afterward; for AUD, 145 and 153 dimensions were significant, respectively. However, these associations were highly scattered across dimensions, with no discernible spatial or functional clustering, and effect sizes were generally small (Supplementary Tables 14-16).

Classification Performance

We evaluated classification performance using regularized logistic regression applied to the 256-dimensional normative deviations obtained in the UKB cohort. To assess multimodal prediction, we additionally considered the three PRSs from the genetic analyses as features.

To quantify the added value of each modality beyond demographic covariates, we included age, age^2, sex, and sex × age as baseline features in all models. This yielded four model types: confounders only, PRS only (+ confounders), deviations only (+ confounders), and PRS + deviations (+ confounders). All models were evaluated in a 10-fold stratified cross-validation binary classification setting against UKB HCs.

Table 3 summarizes the results. Across all depression-related groups, combining normative deviations with PRS consistently yielded the highest mean balanced accuracy (BACC) and area under the curve (AUC). We observed a general increase in mean AUC and BACC with increasing symptom severity, although the standard deviation of performance also increased. Notably, the performance for moderate depressive symptoms (PHQ-9 (2)) was the same as with ICD-10 MDD diagnoses (BACC: $65 \pm 2\%$).

Anxiety-related groups showed similar patterns, although the benefits of incorporating PRS and normative deviations were less consistent than those observed in MDD-related groups.

In alcohol-related groups, models incorporating normative deviations—or the combined approach—outperformed the PRS-only model, with performance increasing alongside severity.

In all groups, the best-performing model included PRS and/or normative deviation features, indicating that both contribute predictive value beyond demographic covariates alone, with mean improvements in AUC of 5.6%, 7.2%, and 3.7% for ANX-, MDD-, and alcohol-related groups, respectively.

To contextualize the predictive value of the deep normative model deviations, we repeated the classification experiments using 99 standard deconfounded FreeSurfer-derived GMV features instead of deviations. Across all groups, the BACC improved moderately by an average of $2 \pm 3\%$ and the AUC by $1 \pm 4\%$ when using deviations. A detailed breakdown of the FreeSurfer baseline results is provided in Supplementary Table 19, with further performance details presented in Supplementary Tables 20 and 21.

Target	**Confounders only**		**PRS only**		**Deviations only**		**PRS + Deviations**	
	BACC (%)	AUC (%)	BACC (%)	AUC (%)	BACC (%)	AUC (%)	BACC (%)	AUC (%)
GAD-7 (2)	58±2	61±2	59±1	**64±1**	58±2	61±2	**60±1**	64±1
GAD-7 (3)	58±6	62±5	**63±3**	**68±3**	61±2	64±3	63±3	67±3
GAD-7 (4)	59±9	62±12	**65±8**	**69±11**	**65±8**	65±6	63±5	68±7
ANX ICD-10	55±3	60±4	62±2	**67±2**	62±2	63±4	**64±3**	67±3
ANX (diag.)	56±2	58±2	58±2	62±3	57±2	60±2	**59±2**	**63±2**
PHQ-9 (1)	58±1	61±1	**60±1**	63±2	59±2	61±1	60±1	**64±2**
PHQ-9 (2)	61±2	65±2	64±2	68±3	63±2	67±2	**65±2**	**70±2**
PHQ-9 (3)	59±6	65±7	61±11	65±14	61±6	69±7	**63±7**	**72±4**
PHQ-9 (4)	63±11	65±26	72±7	68±15	70±7	74±7	**77±8**	**80±7**
MDD ICD-10	57±3	61±3	62±2	68±2	62±3	65±3	**65±2**	**70±2**
MDD (diag.)	58±1	61±1	60±1	64±2	59±1	62±1	**61±1**	**65±1**
AUDIT-C (8)	60±3	65±3	**61±2**	**65±3**	60±3	65±4	60±3	65±3
AUDIT-C (9)	63±3	68±4	**65±4**	69±4	64±4	69±3	65±4	**70±4**
AUDIT-C $\geq$ 10	63±3	68±3	64±3	69±4	**66±3**	72±3	66±3	**73±3**
AUD ICD-10	55±9	56±10	59±7	61±7	**62±4**	63±4	59±6	**64±6**

Table 3: Balanced accuracy (BACC; computed at a decision threshold of 0.5) and area under the receiver operating characteristic curve (AUC) for all classifications. Values are mean ± standard deviation over 10-fold stratified cross-validation, sorted by target. PRS = polygenic risk scores; Deviations = brain structural deviations defined by our normative model. Targets are defined by symptom measures: GAD-7 (Generalized Anxiety Disorder-7), PHQ-9 (Patient Health Questionnaire-9), AUDIT-C (Alcohol Use Disorders Identification Test–Consumption), and ICD-10 diagnosis from hospital records or self-reported physician-reported diagnoses (diag.). Bold values indicate the best classifier for each target; in case of ties, the classifier with the smaller standard deviation is preferred, and if ties remain, the one requiring fewer features is chosen (Occam's razor).

Discussion

Building a novel deep learning approach for normative modeling enabled us to examine the directional deviations of the brain structure of individuals with mental health-relevant symptoms from a large-scale reference population. We observed that individuals with MDD and GAD showed directionally similar deviations that tended to increase with increasing symptom scores, and that were largely specific against those observed with increasing alcohol abuse-related symptoms. These findings could be validated across two independent large-scale cohorts and are consistent with the view that the examined mental health conditions exert a severity-related, dimensional association with brain structure, extending normal physiological variation and aligning with the broadly assumed dimensional nature of mental health. A potential reason for the difference in AUDIT-C–related shifts between NAKO and UKB is the variation in alcohol unit definitions across the cohorts (see Supplementary Section 1.2.1). The brain-structural deviations were associated with polygenic risk for the respective diagnostic constructs, further supporting their biological plausibility. We illustrated how such deviations can be used in multimodal individual-level classifiers that currently have limited predictive power, but that could, in the future, inform personalized medicine applications.

The directional overlap of MDD and GAD deviations and specificity against AUD-related symptoms is consistent with the well-established clinical and biological overlap of MDD and GAD [47], including their high genetic correlation [48, 49] and the considerably lower overlap of either disorder with AUD. As some individuals with AUD show pronounced affective symptoms [50], we anticipated that such individuals would show an intermediate deviation profile. This was supported by our data, supporting the relevance of the directional deviation analysis for capturing the transdiagnostic neurobiological structure, with possible implications for personalized psychiatry. The observed separation between internalizing and externalizing symptom profiles also aligns well with the dimensional structure proposed by the Hierarchical Taxonomy of Psychopathology (HiTOP) [18], which groups MDD and GAD within the internalizing spectrum and AUD within the externalizing domain, although it is important to note that some AUD-related alterations in brain structure may arise from the direct neurotoxic effects of alcohol rather than from an externalizing liability.

In MDD-related groups, the highest classification performance was always but once achieved when combining PRS with brain structural deviations, indicating the strongest complementary effect in our analyses. For anxiety-related groups, the benefits of multimodal integration were less consistent, suggesting weaker complementarity. More precise individual-level prediction appeared possible only at higher levels of symptom severity, particularly for MDD. Nevertheless, the added value of biological measures such as PRS and brain structural deviations remained moderate overall, underscoring the difficulty of predicting psychiatric traits like MDD and anxiety with high accuracy.

Mapping deviations back to classical GMV volumetric features illustrated that AUDIT-C showed the strongest associations with reduced cerebellar GMV. This aligns with prior evidence implicating cerebellar dysfunction in chronic alcohol use [51, 52]. In contrast, associations for MDD and GAD-related symptoms were notably weaker and more spatially diffuse (Supplementary Figures 11 and 12), consistent with prior observations of distributed neural correlates in MDD [53] and psychiatric disorders in general [54]. These findings highlight the ability of the deep learning approach to detect subtle, spatially unconstrained variation by leveraging high-dimensional embeddings, which may be difficult to capture with traditional univariate approaches. Notably, region-level associations derived from the model diverged markedly from direct Spearman correlations between deconfounded GMV and symptom score. This divergence likely reflects the model's capacity to extract non-linear, multivariate effects beyond marginal volume–symptom associations. This capacity is particularly relevant for depressive and anxiety symptoms, where neural alterations may be inherently non-localizable.

Our approach advances normative modeling in several methodological respects. First, by training self-supervised representations directly from MRI images, we avoided feature engineering and preserved distributed structural variance. Second, the normative autoencoder provided a scalable way to capture nonlinear, high-dimensional structure. Finally, by extending deviation analysis to the multivariate embedding space, we were able to examine not only the magnitude of individual divergence but also the directional alignment of deviations across conditions, enabling formal comparison of shared versus distinct neurobiological profiles. Taken together, this framework extends earlier Gaussian process–based or ROI-level normative models by supporting the detection of subtle, spatially unconstrained, and potentially transdiagnostic neurobiological patterns consistent with dimensional models of psychopathology.

From a personalized medicine perspective, our results support the utility of the developed deep normative modeling for parsing the biological architecture of psychiatric phenotypes. By identifying brain-structural effects that vary systematically with symptom severity, our normative modeling approach offers a biologically grounded and data-driven framework that may support future psychiatric risk stratification. The absence of notable deviations in individuals with milder or subclinical symptoms points to the challenges of applying conventional supervised classification approaches successfully to population-based cohorts as compared to those with selected and often severely affected cases that may amplify group differences and reduce heterogeneity [55, 56].

The study has several limitations. First, the UKB cohort includes individuals with a potentially long interval between MRI acquisition and questionnaire completion, which may lead to inconsistencies in associations between brain-structural deviations and clinical data. Although our analyses indicate that the observed directional deviations were robust against this time gap, a residual uncertainty remains. Second, the NAKO and UKB uses different MRI parameters, which may impact the segmentation of whole brain gray matter volumes. However, as our deviation model was replicated across the cohorts, it was likely stable against this source of variation. Third, both cohorts suffer from selection biases towards an underrepresentation of severe mental health problems that are likely amplified in the MRI subsamples [57, 58, 59]. Fourth, our normative modeling was limited to a single data modality, structural MRI, which likely does not capture the full complexity of psychiatric disorders. Finally, the contrastive feature extractor used in our pipeline was trained in a fully unsupervised manner and is thus optimized for general differentiability rather than for highlighting symptom-relevant structural features. Consequently, there may be brain-structural differences related to psychiatric symptoms that are not captured by our model. As our analyses are observational in nature, they cannot establish causality; the observed brain–symptom associations may reflect predisposing liability, illness-related consequences, or environmental influences.

In summary, our deep normative modeling framework reveals consistent deviation profiles across depressive, anxiety, and alcohol-related symptoms, capturing both the magnitude and direction of structural brain divergence from normative patterns. By leveraging high-dimensional embeddings and multivariate deviation analysis, the approach moves beyond traditional feature-based models to detect subtle, spatially unconstrained, and transdiagnostic neurobiological signatures. These findings support the dimensional, transdiagnostic nature of neurobiological changes associated with the investigated conditions and illustrate the potential of advanced normative deep learning as a framework for future applications, including personalized medicine.

Methods

Procedures were approved by Ethics Committee II of the University of Heidelberg Medical Faculty Mannheim (protocol no. 2025-827).

Cohorts - NAKO

The NAKO is a population-based study designed to investigate the causes and early stages of common chronic diseases [34]. The study enrolled 205,000 participants (aged 20–69 years) from 18 centers across Germany, of whom 30,861 underwent MRI at five centers with dedicated facilities [60]. The baseline assessment, conducted between 2014 and 2019, comprised a comprehensive list of examinations, standardized face-to-face medical interviews and touchscreen-based self-report questionnaires, as well as biomaterial collection. The study was approved by the local ethics committees of all participating centers and conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent.

We performed an initial data filtering step by excluding samples with missing questionnaire data (PHQ-9, GAD-7, or AUDIT-C), diagnostic data (MINI or self-reported diagnosis), or confounder data (age, sex, or scan site). Participants with a history of stroke (N = 301) or Parkinson's disease (N = 23) were also excluded to reduce potential bias from neurodegenerative conditions. In total, 1,570 samples were excluded due to missing questionnaire or diagnostic data. After filtering, 15,914 HCs and 13,443 symptomatic participants remained.

Cohorts - UK Biobank (UKB)

The UKB is a population-based cohort comprising approximately 500,000 participants aged 40–69 years at recruitment across 22 study centers, with MRI data collected at four dedicated imaging centers [35]. The UK Biobank obtained ethical approvals from the Northwest Multicenter Research Ethics Committee, the Community Health Index Advisory Group, the Patient Information Advisory Group, and the National Health Service National Research Ethics Service. All participants provided written informed consent. Study procedures were conducted in accordance with the principles of the Declaration of Helsinki. As part of the imaging sub-study, T1-weighted structural brain MRI data are available for 49,279 individuals. A total of 2,589 individuals with neurodegenerative diseases, identified via ICD-10 codes, were initially excluded (details in Supplementary Table 2). After further excluding individuals with missing questionnaire data (PHQ-9, GAD-7, or AUDIT-C) and missing confounder data (sex, age, or total intracranial volume (TiV)), 24,838 samples remained. To evaluate the generalizability of our findings in the NAKO cohort, we used the UKB dataset as a large, external validation cohort, ensuring that the patterns observed in NAKO are reproducible across different populations and imaging protocols.

UK Biobank - Symptomatic Group and Healthy Control Definitions

Since PHQ-9 and GAD-7 scores were available, the same HC criteria were applied as in NAKO. However, as analyzed in detail by Dutt et al. [61], the questionnaires were completed at a time point independent of the scan date. This results in varying temporal deviations between the MRI scan and the corresponding data collection for UKB, with a median absolute deviation of 742 days, whereas for NAKO it is only 15 days. This is an inherent limitation of the UKB cohort, as it may weaken the relationship between the MRI scans and the psychopathological data in the UKB cohort. For UKB, we additionally included ICD-10 diagnoses for MDD, GAD, ANX, and AUD as well as self-reported diagnoses.

Measures

Magnetic Resonance Imaging

MRI is a non-invasive neuroimaging technique that provides high-resolution structural images of the brain by exploiting the magnetic properties of hydrogen nuclei. In psychiatric research, structural T1-weighted MRI is widely used to quantify grey matter anatomy and examine associations with behavior and clinical outcomes [62].

MRI Protocol in NAKO

Structural T1-weighted images were acquired using 3D MPRAGE (TR = 2300ms, TE = 2.98ms, TI = 900ms) on a 3T MRI scanner (Skyra, Siemens, Healthineers, Erlangen, Germany) with 1mm isotropic spatial resolution [63]. We process T1-weighted 3T MRI scans using the Computational Anatomy Toolbox (CAT12, v12.9) [64], which is a voxel-based morphometry tool. The steps here include bias correction, affine registration to the MNI152NLin2009cAsym template space, tissue probability map (TPM)-based segmentation, normalization, modulation, and spatial resampling to extract modulated GMVs with a 1mm isotropic resolution. To enhance computational efficiency, we downsample the images from a 113 × 137 × 113 to an 80 × 80 × 80 voxel grid using spline interpolation. Finally, we standardize all non-zero voxel intensities to harmonize intensity distributions. Gaussian smoothing was introduced during data augmentation in the contrastive learning pipeline.

MRI Protocol in UKB

We use the processed T1-weighted GMV images of the assessment center two provided by the UK Biobank. Structural T1-weighted images were acquired using 3D MPRAGE (TR = 2000ms, TE = 2.01 - 2.03ms, TI = 880ms) on a 3T scanner (Siemens, Siemens, Healthineers, Erlangen, Germany) with 1mm isotropic spatial resolution [65]. The UKB dataset was preprocessed using FSL (v0.1.1) FAST [66], which included bias correction and tissue segmentation. To ensure comparability with the NAKO data, we registered the UKB images to the MNI152NLin2009cAsym template used in the NAKO imaging pipeline, using FSL's FLIRT tool. We then applied the same downsampling and standardization procedures as used for the NAKO data. Since the UKB data was processed using a different pipeline, this approach also enables us to evaluate the robustness of our findings across differing imaging protocols.

Questionnaires

Self-report questionnaires are widely used in psychiatry and mental health research to quantify symptom severity, to screen for disorders, and to monitor treatment response [67]. Their major advantages include low cost, ease of administration, and validated scoring systems that enable standardized measurement across large populations. However, they also have important limitations: they rely on self-reports and thus are subject to recall and social desirability biases, may be influenced by cultural and linguistic factors, and might miss rare symptoms that are not included in predefined diagnostic criteria. Despite these constraints, mental health questionnaires remain essential tools in psychiatric research, epidemiology, and routine care. Below, we describe the three instruments used in this study by giving a short description of the role of the specific clinical questionnaires and their advantages and disadvantages.

Patient Health Questionnaire-9 (PHQ-9)

The PHQ was developed as part of the larger PRIME-MD project to provide brief, standardized mental health assessments for use in primary care settings [68]. The PHQ-9 consists of nine items that correspond directly to the diagnostic criteria for MDD as defined in the DSM-IV. Each item assesses the frequency of symptoms over the previous two weeks using a four-point Likert scale ranging from 0 (”not at all”) to 3 (”nearly every day”), yielding a total score between 0 and 27. The PHQ-9 has demonstrated strong psychometric properties across diverse populations. Internal consistency is high (Cronbach’s α typically > 0.85), and test–retest reliability is acceptable [69]. A commonly used cutoff of ≥ 10 has shown a sensitivity and specificity of approximately 88% for detecting MDD in primary care samples [68]. In addition to screening, the PHQ-9 is widely used to monitor symptom severity over time due to its brevity, ease of use, and direct mapping onto DSM criteria [69].

Generalized Anxiety Disorder-7 (GAD-7)

The GAD-7 was developed as a brief self-report questionnaire to screen for GAD in primary care settings [20]. It consists of seven items that reflect the core DSM-IV criteria for GAD, assessing symptom frequency over the past two weeks on a 4-point Likert scale from 0 (”not at all”) to 3 (”nearly every day”). In NAKO the applied time frame of GAD-7 was 4 instead of 2 weeks. Total scores range from 0 to 21, with higher scores indicating higher symptom severity. Beyond its original use for detecting GAD, the GAD-7 has also demonstrated utility in capturing broader dimensions of anxiety, including panic, social anxiety, and post-traumatic stress symptoms. Psychometric evaluations have shown excellent internal consistency (Cronbach’s $\alpha = 0.92$) and good test–retest reliability (intraclass correlation $= 0.83$) [20, 70]. A cutoff score of ≥ 10 is commonly used to identify cases with probable GAD, yielding a sensitivity of 89% and specificity of 82% in primary care samples [20]. Its brevity, strong psychometric properties, and alignment with DSM criteria have made the GAD-7 one of the most widely used anxiety screening tools in both clinical and research contexts [71].

Alcohol Use Disorder Identification Test - Consumption (AUDIT-C)

The AUDIT-C is a 3-item screening instrument derived from the 10-item AUDIT, developed by the World Health Organization (WHO) to detect hazardous drinking and potential AUD [39]. The AUDIT-C includes items on drinking frequency, typical quantity, and frequency of binge drinking, and is widely used in both clinical and population-based settings due to its brevity and strong psychometric performance. Each item is scored on a 0–4 scale, yielding a total score from 0 to 12. Psychometric evaluations have demonstrated very good internal consistency (Cronbach’s α up to 0.98) [72]. The screening thresholds that simultaneously maximize sensitivity and specificity are $\geq$ 4 in men (sensitivity 0.86, specificity 0.89) and $\geq$3 in women (sensitivity 0.73, specificity 0.91) [73].

Contrastive Feature Extraction, Deconfounding, and Normative Modeling

Deep normative modeling requires embedding brain MRI images into a suitable representational space that captures variations in brain structure. To achieve this, we use momentum contrast (MoCo) [74] for self-supervised feature extraction, reducing the dimensionality of GMV images to 256 dimensions.

MoCo implements contrastive learning by encouraging similar representations for augmented versions of the same image (positive pairs) while pushing apart representations of other images (negative pairs). It maintains a dynamic dictionary of encoded images (size 8 192) via a momentum encoder, ensuring a rich and diverse set of negative pairs, which is crucial for effective contrastive learning. By decoupling the number of negative pairs from batch size, MoCo remains scalable and efficient. MoCo embeddings were deconfounded for age, age-squared, sex, site, and total intracranial volume using linear residualization to reduce potential confounding effects (Supplementary Figure 2). The deconfounded embeddings were input to the normative model, a fully connected autoencoder trained exclusively on HC data. Deviations of holdout HCs and symptomatic subjects were quantified using the reconstruction errors derived from the autoencoder trained exclusively on HCs, under the assumption that symptomatic subjects—unseen during training—would exhibit larger and more frequent reconstruction errors than HCs.

To account for residual confounding in the deviation vectors, we applied a second deconfounding step by linearly residualizing the deviation vectors produced by our autoencoder, as the latter had reintroduced confounding. This procedure was effective: before residualization, embeddings showed significant associations with age (124 dimensions), age-squared (124), sex (190), and age × sex (188), whereas after residualization, only a single weak association with sex remained (Spearman rank correlation).

For more details on architectural decisions, we refer to the Supplementary Section 2.1 - 2.3.

Shift Analysis

To better understand how symptomatic groups deviate from normative variability, we performed a shift analysis that quantifies the extent to which observed symptomatic deviations put out by the normative autoencoder exceed the natural heterogeneity present among healthy controls. Rather than relying on summary statistics such as means or medians, this approach compares the full distribution of deviations, yielding a holistic, group-level measure of distributional shift relative to HCs.

We then computed the Mahalanobis distance from the healthy reference distribution for each subject, capturing how far their deviation vector lies from the normative range in high-dimensional space. The Mahalanobis parameters (mean and covariance) were estimated using 90% of the HC sample, and the resulting distances were computed for both symptomatic groups and the remaining 10% HC holdout set.

For each diagnostic group, we quantified the degree to which deviations exceeded normative variability by calculating the exponentially weighted area where the symptomatic subject's Mahalanobis distance distribution surpassed that of HCs. This weighting emphasizes larger distances, which likely correspond to more clinically meaningful abnormalities. The resulting metric ranges from 0 to 1 and reflects the proportion of the symptomatic group distribution that cannot be explained by healthy variability alone. For mathematical details we refer to the Supplementary Section 2.4.

Stability Assessment and Significance Testing

For the shift analysis, we used a fixed 10% holdout split of HCs to estimate deviations between symptomatic subjects and HCs in the training cohort. The remaining 90% of HCs were used to estimate the parameters of the Mahalanobis distance (μ and Σ). This setup facilitates consistent comparisons across symptomatic groups by preserving a stable reference set. While the HC train/test split introduces a potential source of variability, the resulting deviation estimates proved robust to resampling. Since the symptomatic group data remain fixed and untouched, any variation can be attributed solely to the HC subset used for modeling. However, we refer to the Supplementary Section 2.5 for empirical validation, which shows that the Mahalanobis distance parameters prove to be stable across different sets of HCs.

To assess the statistical significance of deviation differences between HCs and symptomatic groups, we employed multiple linear regression models. For each symptomatic group, we predicted individual Mahalanobis distances using binary group membership as a predictor, while adjusting for age, age-squared, sex, and sex-age interactions. We chose this approach over permutation testing of the group-level shift because it allows covariate adjustment and yields more stable inference; permutation testing is less suited in this context due to its sensitivity to group size imbalance, which violates the exchangeability assumption. Despite prior deconfounding of our embeddings, we included these covariates to account for residual confounding. To address non-normal residuals, we applied a robust standard error estimator [75]. For outlier-robust inference, significance testing was based on the logarithm of the Mahalanobis distances rather than their raw magnitudes. We assessed the significance of the deviation put out by our normative model as a predictor using type III ANOVA on the fitted models, performed separately for each symptomatic group. We applied BH correction [43] to control the false discovery rate (FDR) across the set of diagnostic tests. This choice was due to the small number of tests, which makes methods estimating the proportion of true null hypotheses unstable.

In contrast, when correlating the sum scores with all 256 deviation dimensions, the number of tests was sufficiently large (3×256) to justify methods estimating the proportion of true null hypotheses. To control for multiple hypothesis tests when correlating the deviation dimensions directly with the questionnaire sum scores while maintaining statistical power, we used the Storey–Tibshirani q-value procedure [76]. As the BH method controls the FDR by rejecting only those null hypotheses that fall beneath a predefined threshold, the q-value procedure corrects the positive FDR (pFDR), assuming an analogous role to the p-values. The pFDR is the expected proportion of false positives among rejected hypotheses, conditioned on at least one rejection:

$$\text{pFDR} = \mathbb{E}\left(\frac{V}{R}\,\middle|\, R > 0\right) \tag{1}$$

where V denotes the number of false positives and R the total number of rejections. By thresholding the pFDR at 0.05, we ensured that the expected proportion of false discoveries among all significant findings does not exceed 5%. The q-value represents the minimum pFDR at which a given test would be deemed significant. We thus accepted all tests with q-values below the specified threshold (e.g., 0.05). We used q-values rather than traditional corrections because they estimate the proportion of true null hypotheses, allowing control of the expected FDR without the excessive conservatism of methods like BH or Bonferroni when performing a large number of tests.

We report “q-values” when using the Storey–Tibshirani procedure and denote BH–adjusted p-values as p_{FDR}. Unless stated otherwise, we control the FDR at $\alpha = 0.05$ for all tests.

Directional Analysis

The previously described shift analysis compares distributions of a single deviation metric (the Mahalanobis distance). To achieve a more granular understanding of how the different diagnosis groups differ from HCs, we further analyzed deviations in the 256-dimensional deviation space learned by our model by omitting the Mahalanobis distance modeling step. Instead of summarizing deviations using a single value, we used the full difference between each subject's original and reconstructed feature representations (the input to the Mahalanobis modeling) as a more granular notion of deviation. This method preserves the original geometry of the deviation vectors (256 dimensions), allowing us to detect distinct patterns associated with different symptoms more precisely than the previous analysis.

We quantified the similarity structure of the embeddings by computing the mean pairwise cosine similarity within and between diagnostic groups. The embeddings were 256-dimensional deviation vectors, obtained after deconfounding and L2-normalization, and were represented by each group's geometric median.

For each group or combination of groups, pairwise similarities were calculated as the dot product between normalized embedding vectors. Within-group similarity was defined as the mean of all unique pairwise similarities among subjects belonging to the same group (upper triangular of the similarity matrix, excluding the diagonal). Cross-group similarity was defined as the mean of all pairwise similarities between subjects in different groups.

To quantify directional differences between similarity distributions, we performed one-sided permutation tests. For two similarity sets, A and B, representing within-group and cross-group pairs respectively, we first computed the observed mean difference:

$$\Delta_{\text{obs}} = \text{mean}(A) - \text{mean}(B).$$

Under the null hypothesis (H_0) that the two sets are drawn from the same distribution, group labels ("within" or "cross") were permuted 10,000 times, preserving the number of observations in each set. For each permutation, the mean difference Δ_{perm} was recomputed, yielding a null distribution of differences. The one-sided p-value was calculated as the proportion of permutations where $\Delta_{\text{perm}} \geq \Delta_{\text{obs}}$. Small p-values therefore indicate evidence that within-group similarity exceeds cross-group similarity, consistent with greater internal homogeneity or separation from the comparison group.

This procedure was applied to (i) mood-anxiety groups (defined based on PHQ-9 and GAD-7 strata) versus AUD-related groups (defined based on AUDIT-C strata) to assess whether mood and anxiety disorders were more internally similar than to AUD, and (ii) MDD-related versus GAD-related to test whether within-diagnosis similarity exceeded cross-diagnosis similarity within the mood-anxiety group.

To visualize the relative positioning of diagnostic groups in the normative embedding space, we performed a multi-step procedure on the 256-dimensional deviation vectors.

After deconfounding, residual vectors were standardized to zero mean and unit variance across dimensions. Principal component analysis (PCA) was then applied to project the standardized deviations into two dimensions for visualization; the first two principal components explained 21% of the variance.

For each diagnostic group, we estimated the group centroid in 2D space using the geometric median, a robust estimator less sensitive to outliers than the arithmetic mean. To assess the stability of centroid locations, we performed 1000 bootstrap resamples per group, each time recomputing the geometric median [77]. From the covariance of the bootstrapped medians, we derived a 1-standard-deviation (1-SD) ellipse characterizing the dispersion of the bootstrapped centroids. These ellipses were plotted in the 2D PCA space after centering on the HC centroid (origin), allowing qualitative assessment of directional shifts between diagnostic groups.

This visualization complemented the quantitative cosine similarity analyses by illustrating the relative displacements in embedding space and the degree of separation between groups. We used PCA for dimensionality reduction because it preserves the global structure of the data; however, the first two components explained only 21% of the variance, which limits the interpretability of fine-scale patterns.

Explaining Normative Deviations

To gain insight into the neuroanatomical factors driving normative deviations, it is essential to evaluate the contribution of original input features of the GMVs to the learned embeddings used in normative modeling. To this end, we implemented a heuristic two-step procedure intended to generate hypotheses about which brain regions may underlie observed deviations. We emphasize that this approach is exploratory in nature and not intended to provide definitive causal inferences.

First, we quantified the relationship between regional GMV measures derived from FreeSurfer and the latent embedding dimensions produced by our model. Second, we linked these embedding dimensions to PHQ-9, GAD-7, and AUDIT-C scores and constructed regional relevance maps that summarize these associations.

For the first step, we used elastic net regression [78] to predict each residualized MoCo embedding dimension from a set of 99 residualized FreeSurfer GMV features. Residualization was performed using linear regression with the same covariates applied to the embeddings—specifically, age, age squared, sex, TiV, and recruitment center. Each deconfounded MoCo embedding dimension was predicted from the deconfounded regional GMV features to identify, via the learned coefficients, which brain regions were associated with which embedding dimensions. Elastic net was chosen for its robustness to multicollinearity and its ability to induce sparsity, facilitating the identification of a stable set of region–embedding associations. The resulting coefficient matrix indexed the contribution of each brain region to each latent dimension.

For the second step, we identified embedding dimensions significantly correlated with the score, using a q-value threshold of 0.05 (Storey's method, $\pi_0 = 0.74$). Each dimension has a set of elastic net–derived coefficients linking it to brain regions. We weighted each region's coefficient by the corresponding dimension–symptom correlation, then summed across all relevant dimensions. This yielded a single relevance score per region for each symptom domain, indicating how strongly structural variability in that region is linked to symptom variation. The scores were standardized for interpretability.

Genetic Association Analysis

To analyze the genetic influence on our brain structural deviations we compute PRSs for MDD (PGC MDD2025 Adams et al. [9]), ANX (Kurki et al. [46] Release R12), and problematic alcohol use (Zhou et al. [45]) in the UKB cohort using the default GenoPred (v2.2.11) pipeline [79]. GenoPred is a standardized framework designed for reproducible PRS computation with extensive quality control included. Within this pipeline, we applied SBayesR [80] for postprocessing the summary statistics. We then correlated the resulting PRS values computed by GenoPred with brain deviation scores and overlaid the correlation maps obtained in the previous step to estimate the proportion of structural deviations potentially influenced by genetic risk for each phenotype. Associations were modeled using multiple linear regression, and multiple testing correction was performed using the BH-correction.

All PRS-based association models included age (squared), sex, sex-age interaction, TiV, and the first five genomic principal components as covariates to adjust for population structure and other potential confounding effects.

Classification

To quantify the added value of each modality beyond demographic covariates, we included age (mean-centered), age^2, sex, and the interaction term sex × age as baseline features in all models. This resulted in four model types: (i) confounders only, (ii) PRS only plus confounders, (iii) deviations only plus confounders, and (iv) PRS plus deviations plus confounders. Here, "PRS only" refers to models including only polygenic risk score features in addition to the confounders, whereas "deviations only" refers to models including only the 256-dimensional deviation features in addition to the confounders. Classification was performed using logistic regression with elastic net regularization, implemented in `scikit-learn` [81] (version 1.7.0). For each outer training fold, the l_1-ratio was tuned in an inner loop using a single stratified 80/20 split of the training data.

All features for all four models were standardized within each training fold to prevent data leakage. Models were evaluated against HCs in a binary classification setting using stratified 10-fold cross-validation to maintain case–control proportions across folds.

Author Contributions

Julius Wiegert: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft.
Sebastián Marty-Lombardi: Validation (code analysis), Writing – review & editing.
Jailan Oweda: Data curation (imaging preprocessing).
Esra Lenz: Data curation (data extraction).
Antonia Mai, Amrou Abas, Xiuzhi Li: Data curation (imaging preprocessing support and genomic data processing).
Johannes Nitsche: Methodology (statistical analysis design), Writing – review & editing.
Joonas Naamanka, Sebastian Volkmer, Andreas Meyer-Lindenberg: Writing – review & editing.
Tobias Gradinger, Fabian Streit, Urs Braun, Emanuel Schwarz: Conceptualization, Study design, Supervision.
Peter Ahnert, Klaus Berger, Hermann Brenner, Hans J. Grabe, Karin Halina Greiser, Johanna Klinger-König, André Karch, Michael Leitzmann, Claudia Meinke-Franz, Rafael Mikolajczyk, Thoralf Niendorf, Oliver Sander, Carsten Oliver Schmidt, Steffi G. Riedel-Heller, Annette Peters, Tobias Pischon: Resources (NAKO dataset), Writing – review & editing. Frauke Nees, Stephanie Witt, Kerstin Ritter, Josef Frank: Writing – review & editing.
All authors have reviewed and approved the manuscript.

Acknowledgments

The project was conducted with data (NAKO-711) from the German National Cohort (NAKO) (http://www.nako.de/). The NAKO is funded by the Federal Ministry of Education and

Research (BMBF) [project funding reference numbers: 01ER1301A/B/C, 01ER1511D, and 01ER1801A/B/C/D and 01ER2301A/B/C], federal states of Germany and the Helmholtz Association, the participating universities and the institutes of the Leibniz Association. We thank all participants who took part in the NAKO study and the staff of this research initiative. We also thank the participants and scientists involved in making the UK Biobank resource available (http://www.ukbiobank.ac.uk/). This study was conducted under UK Biobank application number 162313. The project was supported by the DZPG (German Centre for Mental Health Research) and by the BMBF (German Ministry of Education and Research) grant 01EE2303E. Fabian Streit is supported by a 2023 NARSAD Young Investigator Grant (#31537) from the Brain & Behavior Research Foundation with support from the Families for Borderline Personality Disorder Research. This work was supported by the Hector foundation II and was endorsed by German Center for Mental Health (DZPG).

Conflicts of Interest

HJG has received travel grants and speakers honoraria from Neuraxpharm, Servier, Indorsia and Janssen Cilag. ES received speaker fees from bfd buchholz-fachinformationsdienst GmbH, Lundbeckfonden, and Janssen-Cilag GmbH, as well as editorial fees from Lundbeckfonden and the Wellcome Trust. AML has received consultancy honoraria from AbbVie, Janssen-Cilag GmbH, Boehringer-Ingelheim, Daimler und Benz Stiftung, Helmut Horten Stiftung, Neurotorium/Lundbeckfonden, Hector Stiftung, Endosane Pharmaceuticals, Elsevier, von Behring-Röntgen-Stiftung, The LOOP Zürich, ECNP, Teva, Medical Research Council/UKRI, Heinrich-Lanz-Stiftung, Johnson & Johnson, Lundbeckfonden, and the Wellcome Trust. He has received lecture honoraria from pro Mente Akademie GmbH, Schön Klinik, Janssen-Cilag, Evangelische Hochschule Ludwigsburg, Landesärztekammer Baden-Württemberg, Klinikum Ingolstadt, PSY (Psychiatrie und Psychotherapie Update Refresher, FOMF), Consorcio Mexicano de Neuropsicofarmacología (MCNP), Universität Klagenfurt, and Universität Norwalk/USA. He has received editorial honoraria (as editor, etc.) from ECNP/Neuroscience Applied and JSPS. He has received authorship honoraria from Beltz Verlag, Thieme Verlag, and Kohlhammer Verlag. He has received project funding from BMBF, DFG, Hector Stiftung, Klaus Tschira Stiftung, and MWK.

Data Availability

Access to and use of NAKO data and biosamples can be obtained via the electronic application portal (`https://transfer.nako.de`). Access to UK Biobank data requires application through the registration and application portal (`http://ukbiobank.ac.uk/register-apply`).

Code Availability

All analysis were conducted in Python (version 3.12). All analysis scripts and the utilized packages are available via GitHub at `https://github.com/wiegertj/DeepNormativeModeling`.

Bibliography

[1] GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9(2):137–150, 2022. doi: 10.1016/S2215-0366(21)00395-3.

[2] Elizabeth Reisinger Walker, Robin E. McGee, and Benjamin G. Druss. Mortality in Mental Disorders and Global Disease Burden Implications: A Systematic Review and Meta-analysis . *JAMA Psychiatry*, 72(4):334–341, 2015. doi: 10.1001/jamapsychiatry.2014.2502.

[3] Yang Wu, Lu Wang, Mengjun Tao, Huiru Cao, Hui Yuan, Mingquan Ye, Xingui Chen, Kai Wang, and Chunyan Zhu. Changing trends in the global burden of mental disorders from 1990 to 2019 and predicted levels in 25 years. *Epidemiology and Psychiatric Sciences*, 32, 2023. doi: 10.1017/S2045796023000756.

[4] Lulu Cui, Shu Li, Siman Wang, Xiafang Wu, Yingyu Liu, Weiyang Yu, Yijun Wang, Yong Tang, Maosheng Xia, and Baoman Li. Major depressive disorder: hypothesis, mechanism, prevention and treatment. *Signal Transduction and Targeted Therapy*, 9(1):30, 2024. doi: 10.1038/s41392-024-01738-y.

[5] Erkki T. Isometsä. Suicides in mood disorders in psychiatric settings in nordic national register–based studies. *Frontiers in Psychiatry*, Volume 11 - 2020, 2020. ISSN 1664-0640. doi: 10.3389/fpsyt.2020.00721.

[6] H U Wittchen and J Hoyer. Generalized anxiety disorder: nature and course. *J Clin Psychiatry*, 62 Suppl 11:15–9; discussion 20–1, 2001.

[7] Ana Maria Fernandez-Pujals, Mark James Adams, Pippa Thomson, Andrew G McKechanie, Douglas H R Blackwood, Blair H Smith, Anna F Dominiczak, Andrew D Morris, Keith Matthews, Archie Campbell, Pamela Linksted, Chris S Haley, Ian J Deary, David J Porteous, Donald J MacIntyre, and Andrew M McIntosh. Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation scotland: Scottish family health study (GS:SFHS). *PLoS One*, 10(11):e0142197, November 2015.

[8] Kirstin L Purves, Jonathan R I Coleman, Sandra M Meier, Christopher Rayner, Katrina A S Davis, Rosa Cheesman, Marie Bækvad-Hansen, Anders D Børglum, Shing Wan Cho, J Jürgen Deckert, Héléna A Gaspar, Jonas Bybjerg-Grauholm, John M Hettema, Matthew Hotopf, David Hougaard, Christopher Hübel, Carol Kan, Andrew M McIntosh, Ole Mors, Preben Bo Mortensen, Merete Nordentoft, Thomas Werge, Kristin K Nicodemus, Manuel Mattheisen, Gerome Breen, and Thalia C Eley. A major role for common genetic variation in anxiety disorders. *Mol Psychiatry*, 25(12):3292–3303, November 2019.

[9] Mark J. Adams, Fabian Streit, Xiangrui Meng, Swapnil Awasthi, Brett N. Adey, Karmel W. Choi, V. Kartik Chundru, Jonathan R.I. Coleman, Bart Ferwerda, Jerome C. Foo, Zachary F. Gerring, Olga Giannakopoulou, Priya Gupta, Alisha S.M. Hall, Arvid Harder, David M. Howard, Christopher Hübel, Alex S.F. Kwong, Daniel F. Levey, Brittany L. Mitchell, Guiyan Ni, Vanessa K. Ota, Oliver Pain, Gita A. Pathak, Eva C. Schulte, Xueyi Shen, Jackson G. Thorp, Alicia Walker, Shuyang Yao, Jian Zeng, Johan Zvrskovec, Dag Aarsland, Ky'Era V. Actkins, Mazda Adli, Esben Agerbo, Mareike Aichholzer, Allison Aiello, Tracy M. Air, Thomas D.

Als, Evelyn Andersson, Till F.M. Andlauer, Volker Arolt, Helga Ask, Julia Bäckman, Sunita Badola, Clive Ballard, Karina Banasik, Nicholas J. Bass, Aartjan T.F. Beekman, Sintia Belangero, Tim B. Bigdeli, Elisabeth B. Binder, Ottar Bjerkeset, Gyda Bjornsdottir, Sigrid Børte, Emma Bränn, Alice Braun, Thorsten Brodersen, Tanja M. Brückl, Søren Brunak, Mie T. Bruun, Margit Burmeister, Pichit Buspavanich, Jonas Bybjerg-Grauholm, Enda M. Byrne, Jianwen Cai, Archie Campbell, Megan L. Campbell, Adrian I. Campos, Enrique Castelao, Jorge Cervilla, Boris Chaumette, Chia-Yen Chen, Hsi-Chung Chen, Zhengming Chen, Sven Cichon, Lucía Colodro-Conde, Anne Corbett, Elizabeth C. Corfield, Baptiste Couvy-Duchesne, Nick Craddock, Udo Dannlowski, Gail Davies, E. J. C. de Geus, Ian J. Deary, Franziska Degenhardt, Abbas Dehghan, J. Raymond DePaulo, Michael Deuschle, Maria Didriksen, Khoa Manh Dinh, Nese Direk, Srdjan Djurovic, Anna R. Docherty, Katharina Domschke, Joseph Dowsett, Ole Kristian Drange, Erin C. Dunn, William Eaton, Gudmundur Einarsson, Thalia C. Eley, Samar S.M. Elsheikh, Jan Engelmann, Michael E. Benros, Christian Erikstrup, Valentina Escott-Price, Chiara Fabbri, Yu Fang, Sarah Finer, Josef Frank, Robert C. Free, Linda Gallo, He Gao, Michael Gill, Maria Gilles, Fernando S. Goes, Scott Douglas Gordon, Jakob Grove, Daniel F. Gudbjartsson, Blanca Gutierrez, Tim Hahn, Lynsey S. Hall, Thomas F. Hansen, Magnus Haraldsson, Catharina A. Hartman, Alexandra Havdahl, Caroline Hayward, Stefanie Heilmann-Heimbach, Stefan Herms, Ian B. Hickie, Henrik Hjalgrim, Jens Hjerling-Leffler, Per Hoffmann, Georg Homuth, Carsten Horn, Jouke-Jan Hottenga, David M. Hougaard, Iiris Hovatta, Qin Qin Huang, Donald Hucks, Floris Huider, Karen A. Hunt, Nicholas S. Ialongo, Marcus Ising, Erkki Isometsä, Rick Jansen, Yunxuan Jiang, Ian Jones, Lisa A. Jones, Lina Jonsson, Masahiro Kanai, Robert Karlsson, Siegfried Kasper, Kenneth S. Kendler, Ronald C. Kessler, Stefan Kloiber, James A. Knowles, Nastassja Koen, Julia Kraft, Henry R. Kranzler, Kristi Krebs, Theodora Kunovac Kallak, Zoltán Kutalik, Elisa Lahtela, Marilyn Lake, Margit Hørup Larsen, Eric J. Lenze, Melissa Lewins, Glyn Lewis, Liming Li, Bochao Danae Lin, Kuang Lin, Penelope A. Lind, Yu-Li Liu, Donald J. MacIntyre, Dean F. MacKinnon, Brion S. Maher, Wolfgang Maier, Victoria S. Marshe, Gabriela A. Martinez-Levy, Koichi Matsuda, Hamdi Mbarek, Peter McGuffin, Sarah E. Medland, Susanne Meinert, Christina Mikkelsen, Susan Mikkelsen, Yuri Milaneschi, Iona Y. Millwood, Esther Molina, Francis M. Mondimore, Preben Bo Mortensen, Benoit H. Mulsant, Joonas Naamanka, Jake M. Najman, Matthias Nauck, Igor Nenadić, Kasper R. Nielsen, Ilja M. Nolt, Merete Nordentoft, Markus M. Nöthen, Mette Nyegaard, Michael C. O'Donovan, Asmundur Oddsson, Adrielle M. Oliveira, Catherine M. Olsen, Hogni Oskarsson, Sisse Rye Ostrowski, Michael J. Owen, Richard Packer, Teemu Palviainen, Pedro M. Pan, Carlos N. Pato, Michele T. Pato, Nancy L. Pedersen, Ole Birger Pedersen, Wouter J. Peyrot, James B. Potash, Martin Preisig, Michael H. Preuss, Jorge A. Quiroz, Miguel E. Renteria, Charles F. Reynolds III, John P. Rice, Saori Sakaue, Marcos L. Santoro, Robert A. Schoevers, Andrew Schork, Thomas G. Schulze, Tabea S. Send, Jianxin Shi, Engilbert Sigurdsson, Kritika Singh, Grant C.B. Sinnamon, Lea Sirignano, Olav B. Smeland, Daniel J. Smith, Tamar Sofer, Erik Sørensen, Sundararajan Srinivasan, Hreinn Stefansson, Kari Stefansson, Peter Straub, Mei-Hsin Su, André Tadic, Henning Teismann, Alexander Teumer, Anita Thapar, Pippa A. Thomson, Lise Wegner Thørner, Apostolia Topaloudi, Shih-Jen Tsai, Ioanna Tzoulaki, George Uhl, André G. Uitterlinden, Henrik Ullum, Daniel Umbricht, Robert J. Ursano, Sandra Van der Auwera, Albert M. van Hemert, Abirami Veluchamy, Alexander Viktorin, Henry Völzke, G. Bragi Walters, Xiaotong Wang, Agaz Wani, Myrna M. Weissman, Jürgen Wellmann, David C. Whiteman, Derek Wildman, Gonneke Willemsen, Alexander T. Williams, Bendik S. Winsvold, Stephanie H. Witt, Ying

Xiong, Lea Zillich, John-Anker Zwart, Twenty-Three, Me Research Team, China Kadoorie Biobank Collaborative Group, Estonian Biobank Research Team, Genes &amp; Health Research Team, HUNT All-In Psychiatry, The BioBank Japan Project, VA Million Veteran Program, Ole A. Andreassen, Bernhard T. Baune, Klaus Berger, Dorret I. Boomsma, Anders D. Børglum, Gerome Breen, Na Cai, Hilary Coon, William E. Copeland, Byron Creese, Carlos S. Cruz-Fuentes, Darina Czamara, Lea K. Davis, Eske M. Derks, Enrico Domenici, Paul Elliott, Andreas J. Forstner, Micha Gawlik, Joel Gelernter, Hans J. Grabe, Steven P. Hamilton, Kristian Hveem, Catherine John, Jaakko Kaprio, Tilo Kircher, Marie-Odile Krebs, Po-Hsiu Kuo, Mikael Landén, Kelli Lehto, Douglas F. Levinson, Qingqin S. Li, Klaus Lieb, Ruth J.F. Loos, Yi Lu, Susanne Lucae, Jurjen J. Luykx, Hermine H.M. Maes, Patrik K. Magnusson, Hilary C. Martin, Nicholas G. Martin, Andrew McQuillin, Christel M. Middeldorp, Lili Milani, Ole Mors, Daniel J. Müller, Bertram Müller-Myhsok, Yukinori Okada, Albertine J. Oldehinkel, Sara A. Paciga, Colin N.A. Palmer, Peristera Paschou, Brenda W.J.H. Penninx, Roy H. Perlis, Roseann E. Peterson, Giorgio Pistis, Renato Polimanti, David J. Porteous, Danielle Posthuma, Jill A. Rabinowitz, Ted Reichborn-Kjennerud, Andreas Reif, Frances Rice, Roland Ricken, Marcella Rietschel, Margarita Rivera, Christian Rück, Giovanni A. Salum, Catherine Schaefer, Srijan Sen, Alessandro Serretti, Alkistis Skalkidou, Jordan W. Smoller, Dan J. Stein, Frederike Stein, Murray B. Stein, Patrick F. Sullivan, Martin Tesli, Thorgeir E. Thorgeirsson, Henning Tiemeier, Nicholas J. Timpson, Monica Uddin, Rudolf Uher, David A. van Heel, Karin J.H. Verweij, Robin G. Walters, Sylvia Wassertheil-Smoller, Jens R. Wendland, Thomas Werge, Aeilko H. Zwinderman, Karoline Kuchenbaecker, Naomi R. Wray, Stephan Ripke, Cathryn M. Lewis, and Andrew M. McIntosh. Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. *Cell*, 188(3):640–652.e9, 2025. doi: 10.1016/j.cell.2024.12.002.

[10] Ruidong Xiang, Martin Kelemen, Yu Xu, Laura W Harris, Helen Parkinson, Michael Inouye, and Samuel A Lambert. Recent advances in polygenic scores: translation, equitability, methods and FAIR tools. *Genome Medicine*, 16(1):33, February 2024.

[11] T A Brown, L A Campbell, C L Lehman, J R Grisham, and R B Mancill. Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *J Abnorm Psychol*, 110(4):585–599, 2001. doi: 10.1037/0021-843X.110.4.585.

[12] Kenneth S Kendler, Charles O Gardner, Margaret Gatz, and Nancy L Pedersen. The sources of co-morbidity between major depression and generalized anxiety disorder in a swedish national twin sample. *Psychol Med*, 37(3):453–462, November 2006. doi: 10.1017/S0033291706009135.

[13] Joseph Levine, Daniel P. Cole, K. N. Roy Chengappa, and Samuel Gershon M.D. Anxiety disorders and major depression, together or apart. *Depression and Anxiety*, 14(2):94–104, 2001. doi: 10.1002/da.1051.

[14] Zhiyi Chen, Yancheng Tang, Xuerong Liu, Wei Li, Yuanyuan Hu, Bowen Hu, Ting Xu, Rong Zhang, Lei Xia, Jing-Xuan Zhang, Zhibing Xiao, Ji Chen, Zhengzhi Feng, Yuan Zhou, Qinghua He, Jiang Qiu, Xu Lei, Hong Chen, Shaozheng Qin, and Tingyong Feng. Edge-centric connectome-genetic markers of bridging factor to comorbidity between depression and anxiety. *Nature Communications*, 15(1):10560, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-55008-0.

[15] Eiko I Fried and Randolph M Nesse. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR D study. *Journal of affective disorders*, 172:96–102, 2015. doi: 10.1016/j.jad.2014.10.010.

[16] Arijit Nandi, John R. Beard, and Sandro Galea. Epidemiologic heterogeneity of common mood and anxiety disorders over the lifecourse in the general population: a systematic review. *BMC Psychiatry*, 9(1):31, 2009. doi: 10.1186/1471-244X-9-31.

[17] American Psychiatric Association, editor. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*. American Psychiatric Publishing, Washington, DC, USA, 2022. ISBN 978-0-89042-575-6.

[18] Roman Kotov, Robert F Krueger, David Watson, Thomas M Achenbach, Robert R Althoff, R Michael Bagby, Timothy A Brown, William T Carpenter, Avshalom Caspi, Lee Anna Clark, et al. The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, 126(4):454, 2017. doi: 10.1037/abn0000258.

[19] W A van Eeden, A M van Hemert, I V E Carlier, B W Penninx, and E J Giltay. Severity, course trajectory, and within-person variability of individual symptoms in patients with major depressive disorder. *Acta Psychiatr Scand*, 139(2):194–205, 2018. doi: 10.1111/acps.12987.

[20] Robert L Spitzer, Kurt Kroenke, Janet B W Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*, 166(10):1092–1097, 2006. doi: 10.1001/archinte.166.10.1092.

[21] V. Belov, T. Erwin-Grabner, M. Aghajani, et al. Multi-site benchmark classification of major depressive disorder using machine learning on cortical and subcortical measures. *Scientific Reports*, 14:1084, 2024. doi: 10.1038/s41598-023-47934-8.

[22] Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G Ruhe, Christian F Beckmann, and Andre F Marquand. The normative modeling framework for computational psychiatry. *Nat Protoc*, 17(7):1711–1734, 2022. doi: 10.1038/s41596-022-00696-5.

[23] Saige Rutherford, Pieter Barkema, Ivy F Tso, Chandra Sripada, Christian F Beckmann, Henricus G Ruhe, and Andre F Marquand. Evidence for embracing normative modeling. *Elife*, 12, 2023. doi: 10.7554/eLife.85082.

[24] Jelena Bozek, Ludovica Griffanti, Stephan Lau, and Mark Jenkinson. Normative models for neuroimaging markers: Impact of model selection, sample size and evaluation criteria. *NeuroImage*, 268:119864, 2023. doi: 10.1016/j.neuroimage.2023.119864.

[25] Dorothea L. Floris, Thomas Wolfers, Mariam Zabihi, Nathalie E. Holz, Marcel P. Zwiers, and et al. Atypical brain asymmetry in autism—a candidate for clinically meaningful stratification. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(8):802–812, 2021. doi: 10.1016/j.bpsc.2020.08.008.

[26] Thomas Wolfers, Jaroslav Rokicki, Dag Alnaes, Pierre Berthet, Ingrid Agartz, Seyed Mostafa Kia, Tobias Kaufmann, Mariam Zabihi, Torgeir Moberget, Ingrid Melle, Christian F Beckmann, Ole A Andreassen, Andre F Marquand, and Lars T Westlye. Replicating extensive

brain structural heterogeneity in individuals with schizophrenia and bipolar disorder. *Hum Brain Mapp*, 42(8):2546–2555, 2021. doi: 10.1002/hbm.25386.Epub.

[27] Serena Verdi, Andre F Marquand, Jonathan M Schott, and James H Cole. Beyond the average patient: how neuroimaging models can address heterogeneity in dementia. *Brain*, 144(10): 2946–2953, 2021. doi: 10.1093/brain/awab165.

[28] Sayantan Kumar, Philip R O Payne, and Aristeidis Sotiras. Normative modeling using multimodal variational autoencoders to identify abnormal brain volume deviations in Alzheimer's disease. *Proc SPIE Int Soc Opt Eng*, 12465, 2023. doi: 10.1117/12.2654369.

[29] Xiaoyi Sun, Jinrong Sun, Xiaowen Lu, Qiangli Dong, Liang Zhang, Wenxu Wang, Jin Liu, Qing Ma, Xiaoqin Wang, Dongtao Wei, Yuan Chen, Bangshan Liu, Chu-Chung Huang, Yanting Zheng, Yankun Wu, Taolin Chen, Yuqi Cheng, Xiufeng Xu, Qiyong Gong, Tianmei Si, Shijun Qiu, Ching-Po Lin, Jingliang Cheng, Yanqing Tang, Fei Wang, Jiang Qiu, Peng Xie, Lingjiang Li, Yong He, Yong He, Lingjiang Li, Jingliang Cheng, Qiyong Gong, Ching-Po Lin, Jiang Qiu, Shijun Qiu, Tianmei Si, Yanqing Tang, Fei Wang, Peng Xie, Xiufeng Xu, Mingrui Xia, and Mingrui Xia. Mapping neurophysiological subtypes of major depressive disorder using normative models of the functional connectome. *Biological Psychiatry*, 94(12):936–947, 2023. doi: 10.1016/j.biopsych.2023.05.021. Mood Disorder: A Lifespan Perspective on Neurobiology.

[30] Li Sun, Peng Wang, Yuhong Zheng, Jinghua Wang, Jinhui Wang, and Shao-Wei Xue. Dissecting heterogeneity in major depressive disorder via normative model-driven subtyping of functional brain networks. *Journal of Affective Disorders*, 377:1–13, 2025. doi: 10.1016/j.jad.2025.02.033.

[31] Junneng Shao, Jiaolong Qin, Huan Wang, Yurong Sun, Wei Zhang, Xinyi Wang, Ting Wang, Li Xue, Zhijian Yao, and Qing Lu. Capturing the individual deviations from normative models of brain structure for depression diagnosis and treatment. *Biological Psychiatry*, 95(5):403–413, 2024. doi: 10.1016/j.biopsych.2023.08.005.

[32] Guo-Rong Wu and Chris Baeken. Normative modeling analysis reveals corpus callosum volume changes in early and mid-to-late first episode major depression. *Journal of Affective Disorders*, 340:10–16, 2023. doi: 10.1016/j.jad.2023.07.110.

[33] Fabian Eitel, Marc-André Schulz, Moritz Seiler, Henrik Walter, and Kerstin Ritter. Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. *Experimental Neurology*, 339:113608, 2021. ISSN 0014-4886. doi: 10.1016/j.expneurol.2021.113608.

[34] Sabine Schipf, Gina Schöne, Börge Schmidt, Kathrin Günther, Gunthard Stübs, Karin H. Greiser, Fabian Bamberg, Claudia Meinke-Franze, Heiko Becher, Klaus Berger, Hermann Brenner, Stefanie Castell, Antje Damms-Machado, Beate Fischer, Claus-Werner Franzke, Julia Fricke, Sylvia Gastell, Matthias Günther, Wolfgang Hoffmann, Bernd Holleczek, Lina Jaeschke, Annika Jagodzinski, Karl-Heinz Jöckel, Rudolf Kaaks, Hans-Ulrich Kauczor, Yvonne Kemmling, Alexander Kluttig, Lilian Krist, Bärbel Kurth, Oliver Kuß, Nicole Legath, Michael Leitzmann, Wolfgang Lieb, Jakob Linseisen, Markus Löffler, Karin B. Michels, Rafael Mikolajczyk, Iris Pigeot, Ulrich Mueller, Annette Peters, Stefan Rach, Tamara Schikowski, Matthias B. Schulze, Christoph Stallmann, Andreas Stang, Enno Swart, Sabine Waniek, Kerstin Wirkner,

Henry Völzke, Tobias Pischon, and Wolfgang Ahrens. Die Basiserhebung der NAKO Gesundheitsstudie: Teilnahme an den Untersuchungsmodulen, Qualitätssicherung und Nutzung von Sekundärdaten. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 63(3): 254–266, 2020. doi: 10.1007/s00103-020-03093-z.

[35] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. doi: 10.1038/s41586-018-0579-z.

[36] Noora Heikkinen, Eini Niskanen, Mervi Könönen, Tommi Tolmunen, Virve Kekkonen, Petri Kivimäki, Heikki Tanila, Eila Laukkanen, and Ritva Vanninen. Alcohol consumption during adolescence is associated with reduced grey matter volumes. *Addiction*, 112(4):604–613, 2017. doi: 10.1111/add.13697.

[37] Lei Li, Hua Yu, Yihao Liu, Ya-jing Meng, Xiao-jing Li, Chengcheng Zhang, Sugai Liang, Ming-li Li, Wanjun Guo, QiangWang, Wei Deng, Xiaohong Ma, Jeremy Coid, and Tao Li. Lower regional grey matter in alcohol use disorders: evidence from a voxel-based meta-analysis. *BMC Psychiatry*, 21(1):247, 2021. doi: 10.1186/s12888-021-03244-9.

[38] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001. doi: 10.1046/j.1525-1497.2001.016009606.x.

[39] Kristen Bush, Daniel R Kivlahan, Mary B McDonell, Stephan D Fihn, Katharine A Bradley, Ambulatory Care Quality Improvement Project (ACQUIP, et al. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Archives of internal medicine*, 158(16):1789–1795, 1998. doi: 10.1001/archinte.158.16.1789.

[40] Annette Peters, German National Cohort (NAKO) Consortium, Annette Peters, Karin Halina Greiser, Susanne Göttlicher, Wolfgang Ahrens, Maren Albrecht, Fabian Bamberg, Till Bärnighausen, Heiko Becher, et al. Framework and baseline examination of the German National Cohort (NAKO). *European journal of epidemiology*, 37(10):1107–1124, 2022. doi: 10.1007/s10654-022-00890-5.

[41] Fabian Streit, Lea Zillich, Josef Frank, Luca Kleineidam, Michael Wagner, Bernhard T. Baune, and Klaus Berger. Lifetime and current depression in the German National Cohort (NAKO). *The World Journal of Biological Psychiatry*, 24(10):865–880, 2022. doi: 10.1080/15622975.2021.2014152.

[42] Yves Lecrubier, David V Sheehan, Emmanuelle Weiller, Pierre Amorim, Irene Bonora, K Harnett Sheehan, Juris Janavs, and Geoffrey C Dunbar. The mini international neuropsychiatric interview (MINI). a short diagnostic structured interview: reliability and validity according to the CIDI. *European psychiatry*, 12(5):224–231, 1997. doi: 10.1016/S0924-9338(97)83296-8.

[43] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. doi: 0.1111/j.2517-6161.1995.tb02031.x.

[44] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.01.021. 20 YEARS OF fMRI.

[45] Hang Zhou, Rachel L. Kember, Joseph D. Deak, Heng Xu, Sylvanus Toikumo, Kai Yuan, Penelope A. Lind, Leila Farajzadeh, Lu Wang, Alexander S. Hatoum, Jessica Johnson, Hyunjoon Lee, Travis T. Mallard, Jiayi Xu, Keira J. A. Johnston, Emma C. Johnson, Trine Tollerup Nielsen, Marco Galimberti, Cecilia Dao, Daniel F. Levey, Cassie Overstreet, Enda M. Byrne, Nathan A. Gillespie, Scott Gordon, Ian B. Hickie, John B. Whitfield, Ke Xu, Hongyu Zhao, Laura M. Huckins, Lea K. Davis, Sandra Sanchez-Roige, Pamela A. F. Madden, Andrew C. Heath, Sarah E. Medland, Nicholas G. Martin, Tian Ge, Jordan W. Smoller, David M. Hougaard, Anders D. Børglum, Ditte Demontis, John H. Krystal, J. Michael Gaziano, Howard J. Edenberg, Arpana Agrawal, Amy C. Justice, Murray B. Stein, Henry R. Kranzler, Joel Gelernter, and Million Veteran Program. Multi-ancestry study of the genetics of problematic alcohol use in over 1 million individuals. *Nature Medicine*, 29(12):3184–3192, 2023. doi: 10.1038/s41591-023-02653-5.

[46] Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Wei Zhou, Sanni Ruotsalainen, Kalle Pärn, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Vincent Llorens, Javier Gracia-Tabuenca, Harri Siirtola, Kadri Reis, Abdelrahman G. Elnahas, Benjamin Sun, Christopher N. Foley, Katriina Aalto-Setälä, Kaur Alasoo, Mikko Arvas, Kirsi Auro, Shameek Biswas, Argyro Bizaki-Vallaskangas, Olli Carpen, Chia-Yen Chen, Oluwaseun A. Dada, Zhihao Ding, Margaret G. Ehm, Kari Eklund, Martti Färkkilä, Hilary Finucane, Andrea Ganna, Awaisa Ghazal, Robert R. Graham, Eric M. Green, Antti Hakanen, Marco Hautalahti, Åsa K. Hedman, Mikko Hiltunen, Reetta Hinttala, Iiris Hovatta, Xinli Hu, Adriana Huertas-Vazquez, Laura Huilaja, Julie Hunkapiller, Howard Jacob, Jan-Nygaard Jensen, Heikki Joensuu, Sally John, Valtteri Julkunen, Marc Jung, Juhani Junttila, Kai Kaarniranta, Mika Kähönen, Risto Kajanne, Lila Kallio, Reetta Kälviäinen, Jaakko Kaprio, Nurlan Kerimov, Johannes Kettunen, Elina Kilpeläinen, Terhi Kilpi, Katherine Klinger, Veli-Matti Kosma, Teijo Kuopio, Venla Kurra, Triin Laisk, Jari Laukkanen, Nathan Lawless, Aoxing Liu, Simonne Longerich, Reedik Mägi, Johanna Mäkelä, Antti Mäkitie, Anders Malarstig, Arto Mannermaa, Joseph Maranville, Athena Matakidou, Tuomo Meretoja, Sahar V. Mozaffari, Mari E. K. Niemi, Marianna Niemi, Teemu Niiranen, Christopher J. O´Donnell, Ma´en Obeidat, George Okafo, Hanna M. Ollila, Antti Palomäki, Tuula Palotie, Jukka Partanen, Dirk S. Paul, Margit Pelkonen, Rion K. Pendergrass, Slavé Petrovski, Anne Pitkäranta, Adam Platt, David Pulford, Eero Punkka, Pirkko Pussinen, Neha Raghavan, Fedik Rahimov, Deepak Rajpal, Nicole A. Renaud, Bridget Riley-Gillis, Rodosthenis Rodosthenous, Elmo Saarentaus, Aino Salminen, Eveliina Salminen, Veikko Salomaa, Johanna Schleutker, Raisa Serpi, Huei-yi Shen, Richard Siegel, Kaisa Silander, Sanna Siltanen, Sirpa Soini, Hilkka Soininen, Jae Hoon Sul, Ioanna Tachmazidou, Kaisa Tasanen, Pentti Tienari, Sanna Toppila-Salmi, Taru Tukiainen, Tiinamaija Tuomi, Joni A. Turunen, Jacob C. Ulirsch, Felix Vaura, Petri Virolainen, Jeffrey Waring, Dawn Waterworth, Robert Yang, Mari Nelis, Anu Reigo, Andres Metspalu, Lili Milani, Tõnu Esko, Caroline Fox, Aki S. Havulinna, Markus Perola, Samuli Ripatti, Anu Jalanko, Tarja Laitinen, Tomi P. Mäkelä, Robert Plenge, Mark McCarthy, Heiko Runz, Mark J. Daly, Aarno Palotie, and FinnGen. FinnGen provides

genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, 2023. doi: 10.1038/s41586-022-05473-8.

[47] Lisa Sindermann, Ronny Redlich, Nils Opel, Joscha Böhnlein, Udo Dannlowski, and Elisabeth Johanna Leehr. Systematic transdiagnostic review of magnetic-resonance imaging results: Depression, anxiety disorders and their co-occurrence. *Journal of Psychiatric Research*, 142: 226–239, 2021. ISSN 0022-3956. doi: 10.1016/j.jpsychires.2021.07.022.

[48] Liwei Mei, Yan Gao, Min Chen, Xiao Zhang, Weihua Yue, Dai Zhang, and Hao Yu. Overlapping common genetic architecture between major depressive disorders and anxiety and stress-related disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 113:110450, 2022. doi: 10.1016/j.pnpbp.2021.110450.

[49] John M. Hettema. What is the genetic relationship between anxiety and depression? *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 148C(2):140–146, 2008. doi: 10.1002/ajmg.c.30171.

[50] Alvaro Castillo-Carniglia, Katherine M Keyes, Deborah S Hasin, and Magdalena Cerdá. Psychiatric comorbidities in alcohol use disorder. *Lancet Psychiatry*, 6(12):1068–1080, 2019. doi: 10.1016/S2215-0366(19)30222-6.

[51] Edith V. Sullivan, Natalie M. Zahr, Manojkumar Saranathan, Kilian M. Pohl, and Adolf Pfefferbaum. Convergence of three parcellation approaches demonstrating cerebellar lobule volume deficits in alcohol use disorder. *NeuroImage: Clinical*, 24:101974, 2019. doi: 10.1016/j.nicl.2019.101974.

[52] Ansgar Torvik and Sverre Torp. The prevalence of alcoholic cerebellar atrophy: A morphometric and histological study of an autopsy material. *Journal of the Neurological Sciences*, 75(1): 43–51, 1986. ISSN 0022-510X. doi: 10.1016/0022-510X(86)90049-3.

[53] Nils R. Winter, Ramona Leenings, Jan Ernsting, Kelvin Sarink, Lukas Fisch, Daniel Emden, Julian Blanke, Janik Goltermann, Nils Opel, Carlotta Barkhau, Susanne Meinert, Katharina Dohm, Jonathan Repple, Marco Mauritz, Marius Gruber, Elisabeth J. Leehr, Dominik Grotegerd, Ronny Redlich, Andreas Jansen, Igor Nenadic, Markus M. Nöthen, Andreas Forstner, Marcella Rietschel, Joachim Groß, Jochen Bauer, Walter Heindel, Till Andlauer, Simon B. Eickhoff, Tilo Kircher, Udo Dannlowski, and Tim Hahn. Quantifying deviations of brain structure and function in major depressive disorder across neuroimaging modalities. *JAMA Psychiatry*, 79(9):879–888, 09 2022. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2022.1780.

[54] Ashlea Segal, Linden Parkes, Kevin Aquino, Seyed Mostafa Kia, Thomas Wolfers, Barbara Franke, Martine Hoogman, Christian F. Beckmann, Lars T. Westlye, Ole A. Andreassen, Andrew Zalesky, Ben J. Harrison, Christopher G. Davey, Carles Soriano-Mas, Narcís Cardoner, Jeggan Tiego, Murat Yücel, Leah Braganza, Chao Suo, Michael Berk, Sue Cotton, Mark A. Bellgrove, Andre F. Marquand, and Alex Fornito. Regional, circuit and network heterogeneity of brain abnormalities in psychiatric disorders. *Nature Neuroscience*, 26(9):1613–1629, 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01404-6.

[55] Zachary B Millman, James M Gold, Vijay A Mittal, and Jason Schiffman. The critical need for help-seeking controls in clinical high-risk research. *Clinical Psychological Science*, 7(6): 1171–1189, 2019. doi: 10.1177/2167702619855660.

[56] Kenneth S. Kendler. The super-normal control group in psychiatric genetics: Possible artifactual evidence for coaggregation. *Psychiatric Genetics*, 1(2), 1990. doi: 10.1097/00041444-199001020-00005.

[57] Stefan Rach, Matthias Sand, Achim Reineke, Heiko Becher, Karin Halina Greiser, Kathrin Wolf, Kerstin Wirkner, Carsten Oliver Schmidt, Sabine Schipf, Karl-Heinz Jöckel, Lilian Krist, Wolfgang Ahrens, Hermann Brenner, Stefanie Castell, Sylvia Gastell, Volker Harth, Bernd Holleczek, Till Ittermann, Stefan Janisch-Fabian, André Karch, Thomas Keil, Carolina J. Klett-Tammen, Alexander Kluttig, Oliver Kuß, Michael Leitzmann, Wolfgang Lieb, Claudia Meinke-Franze, Karin B. Michels, Rafael Mikolajczyk, Ilais Moreno Velásquez, Nadia Obi, Cara Övermöhle, Annette Peters, Tobias Pischon, Susanne Rospleszcz, Börge Schmidt, Matthias B. Schulze, Andreas Stang, Henning Teismann, Christine Töpfer, Robert Wolff, and Kathrin Günther. The baseline examinations of the german national cohort (NAKO): recruitment protocol, response, and weighting. *European Journal of Epidemiology*, 40(4):475–489, 2025. ISSN 1573-7284. doi: 10.1007/s10654-025-01219-8.

[58] Katrina A S Davis, Jonathan R I Coleman, Mark Adams, Naomi Allen, Gerome Breen, Breda Cullen, Chris Dickens, Elaine Fox, Nick Graham, Jo Holliday, Louise M Howard, Ann John, William Lee, Rose McCabe, Andrew McIntosh, Robert Pearsall, Daniel J Smith, Cathie Sudlow, Joey Ward, Stan Zammit, and Matthew Hotopf. Mental health in UK biobank - development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. *BJPsych Open*, 6(2), 2020.

[59] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and Health-Related characteristics of UK biobank participants with those of the general population. *Am J Epidemiol*, 186(9):1026–1034, 2017.

[60] Fabian Bamberg, Christopher L. Schlett, Svenja Caspers, Steffen Ringhof, Matthias Günther, Jochen G. Hirsch, Julia Rüdebusch, Pavlína Miklánková, Nora Bittner, Christiane Jockwitz, Michael Forsting, Norbert Hosten, Rudolph Kaaks, Hans-Ulrich Kauczor, Thomas Kroenke, Thoralf Niendorf, Annette Peters, Tobias Pischon, Andreas Stang, Klaus Berger, and Henry Völzke. Baseline MRI Examination in the NAKO Health Study. *Dtsch Arztebl International*, 121(18):587–593, 2024. doi: 10.3238/arztebl.m2024.0151.

[61] Rosie K Dutt, Kayla Hannon, Ty O Easley, Joseph C Griffis, Wei Zhang, and Janine D Bijsterbosch. Mental health in the UK biobank: A roadmap to self-report measures and neuroimaging correlates. *Hum Brain Mapp*, 43(2):816–832, 2021. doi: 10.1002/hbm.25690.

[62] Shergill S. and Maitra R. Neuroimaging. In *Essential Neuroscience for Psychiatrists*, pages 147–173. Cambridge University Press, Cambridge, 2025. doi: 10.1017/9781911623083.006.

[63] Fabian Bamberg, Hans-Ulrich Kauczor, Sabine Weckbach, Christopher L. Schlett, Michael Forsting, Susanne C. Ladd, Karin Halina Greiser, Marc-André Weber, Jeanette Schulz-Menger, Thoralf Niendorf, Tobias Pischon, Svenja Caspers, Katrin Amunts, Klaus Berger, Robin Bülow, Norbert Hosten, Katrin Hegenscheid, Thomas Kröncke, Jakob Linseisen, Matthias Günther, Jochen G. Hirsch, Alexander Köhn, Thomas Hendel, Heinz-Erich Wichmann, Börge Schmidt, Karl-Heinz Jöckel, Wolfgang Hoffmann, Rudolf Kaaks, Maximilian F. Reiser, and Henry and

Völzke. Whole-body MR imaging in the German National Cohort: Rationale, design, and technical background. *Radiology*, 277(1):206–220, 2015. doi: 10.1148/radiol.2015142272. PMID: 25989618.

[64] Christian Gaser, Robert Dahnke, Paul M Thompson, Florian Kurth, Eileen Luders, and the Alzheimer's Disease Neuroimaging Initiative. CAT: a computational anatomy toolbox for the analysis of structural MRI data. *GigaScience*, 13:giae049, 2024. ISSN 2047-217X. doi: 10.1093/gigascience/giae049.

[65] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K. Bangerter, Jesper L.R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N. Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul McCarthy, Christopher Rorden, Alessandro Daducci, Daniel C. Alexander, Hui Zhang, Iulius Dragonu, Paul M. Matthews, Karla L. Miller, and Stephen M. Smith. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166:400–424, 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2017.10.034.

[66] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. FSL. *NeuroImage*, 62(2):782–790, 2012. doi: 10.1016/j.neuroimage.2011.09.015.

[67] Constantina Demetriou, Bilge Uzun Ozer, and Cecilia A. Essau. Self-report questionnaires. In Robin L. Cautin and Scott O. Lilienfeld, editors, *The Encyclopedia of Clinical Psychology*. John Wiley & Sons, Inc., 2015. doi: 10.1002/9781118625392.wbecp507.

[68] Kurt Kroenke and Robert L. Spitzer. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9):509–515, 2002. doi: 10.3928/0048-5713-20020901-06.

[69] Bernd Löwe, Robert L. Spitzer, Kerstin Gräfe, Kurt Kroenke, Andrea Quenter, Stephan Zipfel, and Wolfgang Herzog. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *Journal of Affective Disorders*, 78(2):131–140, 2004. doi: 10.1016/S0165-0327(02)00237-9.

[70] Bernd Löwe, Oliver Decker, Stefanie Müller, Elmar Brähler, Dieter Schellberg, Wolfgang Herzog, and Philipp Y. Herzberg. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. *Medical Care*, 46(3):266–274, 2008. doi: 10.1097/MLR.0b013e318160d093.

[71] Faye Plummer, Laura Manea, Dominic Trepel, and Dean McMillan. Screening for anxiety disorders with the GAD-7 and GAD-2: A systematic review and diagnostic meta-analysis. *General Hospital Psychiatry*, 39:24–31, 2016. doi: 10.1016/j.genhosppsych.2015.11.005.

[72] Yoneatsu Osaki, Aro Ino, Sachio Matsushita, Susumu Higuchi, Yoko Kondo, and Aya Kinjo. Reliability and validity of the alcohol use disorders identification test - consumption in screening for adults with alcohol use disorders and risky drinking in Japan. *Asian Pacific Journal of Cancer Prevention*, 15(16):6571–6574, 2014. doi: 10.7314/APJCP.2014.15.16.6571.

[73] Katharine A. Bradley, Anna F. DeBenedetti, Robert J. Volk, Emily C. Williams, Danielle Frank, and Daniel R. Kivlahan. Audit-c as a brief screen for alcohol misuse in primary care. *Alcoholism: Clinical and Experimental Research*, 31(7):1208–1217, 2007. doi: 10.1111/j.1530-0277.2007.00403.x.

[74] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. doi: 10.1109/CVPR42600.2020.00975.

[75] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000. ISSN 00031305, 15372731.

[76] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002. doi: 10.1111/1467-9868.00346.

[77] E. Weiszfeld and Frank Plastria. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 167(1):7–41, March 2009. doi: 10.1007/s10479-008-0352-z.

[78] Hui Zou and Trevor Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x.

[79] Oliver Pain, Ammar Al-Chalabi, and Cathryn M Lewis. The GenoPred pipeline: a comprehensive and scalable pipeline for polygenic scoring. *Bioinformatics*, 40(10):btae551, 2024. doi: 10.1093/bioinformatics/btae551.

[80] Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature communications*, 10 (1):5086, 2019. doi: 10.1038/s41467-019-12653-0.

[81] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.