

Normative modeling in neuroimaging across psychiatric and neurological disorders: a systematic review and the NORMA reporting checklist

Naoki Takamatsu

Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan <https://orcid.org/0000-0003-1704-6411>

Issei Ueda

Department of Child Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; Medical Data Science Course, Faculty of Medicine, Shinshu University, Japan

Yoshito Saito

Department of Psychiatry, The University of Melbourne, Carlton, Victoria, Australia

Yusuke Takahashi

Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Takahide Etani

Keio University Hospital, Tokyo, Japan <https://orcid.org/0000-0003-0027-2551>

Shunsuke Tamura

Department of Psychiatry, Division of Clinical Neuroscience, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan

Koki Takahashi

Keio University, Tokyo, Japan

Amaki Tsukazaki

Graduate School of Medicine, Keio University, Tokyo, Japan

Shuhei Shibukawa

Department of Child Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; Department of Radiological Technology, Faculty of Health Science, Juntendo University, Tokyo, Japan <https://orcid.org/0000-0001-8638-7522>

Andrew Zalesky

Department of Psychiatry and Department of Biomedical Engineering, The University of Melbourne, Melbourne, Victoria, Australia <https://orcid.org/0000-0003-2298-9908>

André F. Marquand

Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands; Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands <https://orcid.org/0000-0001-5903-203X>

Yoji Hirano

Department of Psychiatry, Division of Clinical Neuroscience, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan

Shinsuke Koike

Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; Department of Child Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; Center for Evolutionary Cognitive Sciences, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan; International Research Center for Neurointelligence, The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Tokyo, Japan <https://orcid.org/0000-0002-3375-236X>

Shinichiro Nakajima

shinichiro.1.nakajima@gmail.com

Department of Neuropsychiatry, Keio University School of Medicine, Tokyo 160-8582, Japan <https://orcid.org/0000-0003-4229-1910>

Analysis

Keywords:

Posted Date: July 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10398741/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: Yes there is potential Competing Interest. S.N. has received research support, manuscript fees or speaker's honoraria from Asahi Quality & Innovations, Ltd., Teijin Pharma, Sumitomo Pharma, Meiji Seika Pharma, Otsuka, PDR pharma, and MSD within the past three years. The remaining authors declare no competing interests.

Abstract

Case-control comparisons of group means in psychiatric and neurological neuroimaging cannot quantify individual deviation from population norms. Normative modeling, akin to pediatric growth charts, locates individuals within a reference distribution to yield individualized deviation scores. We reviewed 122 normative-modeling studies across neuroimaging and neurophysiological measures. Structural MRI accounted for 71% of studies. Statistical methodology has shifted from Gaussian process regression toward Bayesian and distributional approaches. Reference cohort sizes spanned 51–77,157 (median 642). The 15 cohorts of 10,000 or more drew most often on UK Biobank (73%), whether directly or within multi-site aggregations, and 20% provided self-reported race or ethnicity. External validation was reported in 8% of applicable studies, and a single model accounted for 38% of pre-trained applications. Methodological refinement has outpaced the empirical foundations and reporting transparency required for clinical translation. To address this, we introduce NORMA, an 18-item reporting checklist for model construction, validation, and clinical application.

Main

Psychiatric neuroimaging has been built on the case-control paradigm¹. Over three decades, studies comparing group averages between patients and healthy controls have produced extensive findings, and large-scale consortia such as ENIGMA (Enhancing Neuroimaging Genetics through Meta-Analysis) have extended this approach through coordinated mega-analyses, identifying regional volume reductions across major psychiatric diagnoses². Pooling thousands of scans was expected to reduce the sampling biases of individual studies and to yield reliable, diagnosis-specific markers. The clinical yield, however, has been limited. Effect sizes are modest: hippocampal volume reduction in schizophrenia, one of the largest effects identified by ENIGMA, yields a Cohen's *d* of approximately -0.46, with the distributions largely overlapping between patient and control groups². Cross-disorder comparisons further suggest that these modest effects are not even diagnosis-specific; subcortical volume effect-size profiles for schizophrenia, bipolar disorder, and major depressive disorder correlate at 0.95 to 0.98³. These comparisons rest on an aggregation of individual case-control studies, each subject to sampling biases from recruitment procedures and site-specific inclusion criteria⁴, and mega-analytic pooling propagates rather than corrects such biases. What these studies share is the analytic approach itself, a case-control design with linear covariate adjustment. The structural convergence observed across diagnoses may therefore reflect shared methodological constraints as much as shared neurobiology.

Case-control designs are built to compare groups, not individuals. They cannot say where a given individual falls relative to the population^{4,5}, and the diagnoses they contrast are defined by clinical presentation rather than a shared biological cause. Within a single diagnosis, genome-wide association studies have identified hundreds of risk loci for schizophrenia⁶, and across diagnoses nearly three-quarters of identified loci are pleiotropic, shared across eight major psychiatric disorders⁷. The same holds in brain structure, where a meta-analysis across six psychiatric disorders found gray matter loss that was shared rather than diagnosis-specific⁸. Similar biology can therefore underlie different diagnoses, and the same diagnosis can arise from different biology. Developmental psychopathology has long recognized this pattern, in which distinct pathophysiological processes produce similar clinical presentations and shared risk factors lead to divergent outcomes, under the terms equifinality and multifinality⁹, framed more recently as a many-to-many mapping between biology and diagnosis¹⁰. Dimensional frameworks such as the Research Domain Criteria (RDoC) and the Hierarchical Taxonomy of Psychopathology (HiTOP) have emerged in response, reframing psychopathology as continuous dimensions that cut across disorders^{11,12}. They do not by themselves quantify how far a given individual's measurements deviate from a normative reference. This mismatch between biology and clinical phenotype means that within-group biological variance can overwhelm between-group differences, and the group average is then not a representative summary but a composite of heterogeneous individual profiles. Brain structure also changes nonlinearly across the lifespan¹³, yet case-control comparisons typically assume that linear covariate adjustment for age controls for developmental and aging effects. When the age and sex composition of patient and control groups differs systematically, linear adjustment cannot disentangle disease effects from demographic confounds operating along a nonlinear trajectory.

Whereas case-control designs compare group means, normative modeling characterizes how each individual deviates from population expectations. The approach draws on a long tradition in pediatric medicine; growth charts have been used for over a century to track height and weight against population-derived centiles¹⁴. The clinician does not ask whether one group of children is taller than another, but whether this particular child falls within the expected range for their age and sex. Normative models apply the same logic to the brain^{5,15} (Fig. 1). A reference model is constructed from a normative cohort, estimating the distribution of a neuroimaging measure, its mean and its variance, as a function of covariates such as age and sex. Each individual is then positioned within this distribution and assigned a deviation score, a *z*-score or centile that reflects the individual's position in the population, not just their distance from the mean^{13,16,17}. These models use flexible nonlinear functions to fit the relationship between age and brain structure, accommodating developmental and aging trajectories that linear covariate adjustments of case-control designs cannot handle. The same logic applies whether the clinical question concerns a psychiatric or a neurological disorder.

Normative models applied across these disorders point to substantial heterogeneity in individual deviation profiles, even among patients who share the same diagnosis. Among 322 patients with schizophrenia, 79% showed an infra-normal deviation somewhere in the brain, only modestly more than the 59% of healthy individuals, and no single region or tract deviated in more than 18% of patients, against about 5% in healthy individuals¹⁸. A transdiagnostic study of six disorders showed a similar trend, with at most 7% of any diagnostic group deviating at any single region, against 3% in healthy individuals¹⁹. Because these deviations fall in different locations, the group mean at a given region is a composite of a deviating minority and a mostly normal majority rather than a shared abnormality. Excluding extreme deviators from case-control comparisons of cortical thickness in autism spectrum disorder (ASD) accordingly reduced the significant group differences by half²⁰. The heterogeneity also extends beyond the spatial distribution of deviations. When positive and negative deviations coexist within the same diagnostic group, the group average masks abnormalities in both directions, and in ASD nearly 28% of patients showed structural deviations opposite in sign to those of the rest²¹.

Across the normative-modeling literature, models differ in statistical approach, sample composition, harmonization strategy, and validation procedure. This variation can make findings difficult to compare across studies. Two recent reviews have addressed conceptual and practical aspects of the approach. One outlined the methodological promise and challenges of normative modeling for quantifying individual risk, focusing on model assumptions, uncertainty estimation, and practical recommendations for model building²². The other examined the broader transition from case-control to normative approaches, with particular attention to multi-site harmonization and the integration of distributed data sources⁴. Neither provided a systematic evaluation of how normative models are being implemented in practice. Here we present a systematic review of normative modeling in neuroimaging across psychiatric and neurological applications, evaluating how normative models are constructed, validated, and applied to clinical populations. It does not rank model classes or synthesize clinical findings. From this evidence we develop the NORMA (NORMative Modeling Assessment) checklist, an 18-item instrument spanning three domains: common elements (CE), model development and validation (NM), and clinical application (CA) (Table 1). NORMA is intended to make the decisions behind each study explicit and comparable across reports.

Results

Methodological characteristics

Database screening identified 122 studies meeting inclusion criteria (Fig. 2). Reports excluded at full-text review are listed with reasons in Supplementary Table 1. Some met more than one criterion and are shown under a single primary reason. The number of studies increased from a single study in 2018 and annual counts below 10 through 2021 to between 14 and 41 per year from 2022 to 2024. Per-study methodological characteristics are provided in Supplementary Data 1, and the joint distribution across modality, modeling method, model build, and clinical focus is shown in Fig. 3. Structural MRI (sMRI) was used in 87 studies (71%), functional MRI (fMRI) in 26 (21%), and diffusion MRI (dMRI) in nine (7%). Ten studies (8%) used other measures, including EEG or MEG (five) and PET (three). Eight combined more than one modality. Because each multimodal study is counted under every modality it used, these counts exceed the number of studies using each modality alone. Within the 87 studies that used sMRI, T1-weighted images were the dominant contrast (86 studies), while T2-weighted images were a primary input in two, one of which used both. Most studies built new normative models (105, 86%). Sixteen (13%) applied pre-trained models, and one combined both approaches. Psychiatric disorders were the focus in 85 studies (70%) and neurological disorders in 28 (23%), and both in 9 (7%).

Statistical approaches

The five categories appeared in the literature in sequence: GPR in 2018, linear and deep learning approaches in 2019, distributional regression in 2020, and several variants of Bayesian regression in 2022. Distributional regression lets the scale and shape of the distribution, not only its mean, depend on covariates (Supplementary Note 1). Bayesian regression approaches led the annual count in 2023 and 2024 (17 of 41 studies in 2024). In 2025 through April, distributional and Bayesian regression were nearly even (9 and 8 of 32). Distributional regression grew in parallel, from one study in 2020 to ten in 2024 and nine through early 2025. It matched GPR in 2023 (4 each) and exceeded it in 2024 (10 vs 8) and 2025 (9 vs 6). Across the four years since Bayesian regression entered the field, Bayesian and distributional regression together accounted for 60 of 106 studies (57%).

Site effect handling

Of the 122 studies, 119 used multi-site data (Supplementary Fig. 1). Site or scanner effects were modeled within the normative model in 50 studies (41%), removed through harmonization before fitting in 28 (23%), and left unadjusted in 41 (34%). The remaining three were single-site. Modeling families differed in whether they addressed site at all: Bayesian regression studies did so in 94% and distributional regression studies in 78%, while the linear, GPR, and deep/other families addressed it less often than either of these two families. Among the studies that did address site, the chosen strategy differed as well. Distributional regression studies split about evenly between the two site strategies, pre-fitting harmonization and within-model adjustment, whereas Bayesian regression studies relied almost exclusively on the latter (88%, 30 of 34). At the other extreme, only 11% (3 of 27) of GPR studies did the same. Pre-trained models modeled site more often than newly built ones (62% vs 37%).

Reference cohorts

Reference cohort characteristics are summarized in Supplementary Data 2. Of 122 included studies, 16 (13%) applied a pre-trained normative model from previous studies, 105 built reference cohorts directly, and one combined a pre-trained sMRI model with a newly built fMRI reference cohort²³. Cohort-size analyses are restricted to these 105 directly-built cohorts (Fig. 4a). Cohort sizes ranged from 51 to 77,157, with a median of 642 (IQR 281–1,978). Most studies (65, 62%) used cohorts of fewer than 1,000 participants, while 15 used cohorts of 10,000 or larger. Among the 15 largest-cohort studies, 11 drew on UK Biobank and 7 on at least one US-based dataset; nine were multi-site aggregations, five incorporating the Adolescent Brain Cognitive Development (ABCD) study. Only one was anchored outside the UK or USA²⁴. Race or ethnicity composition was reported quantitatively in 3 of these 15 (20%), where the cohorts were 64 to 97% White²⁵⁻²⁷, while a fourth described its sample's ancestry bias only qualitatively¹³.

The application of pre-trained models concentrated on a small number of source models (Fig. 4b). Four sources accounted for 14 of the 16 pre-trained applications (88%)^{13,16,17,28}; the most-used source alone accounted for 6 (38%)¹⁷. The other two applications each used a different source model.

Demographic reporting was uneven across the new-build subset. Age mean was reported in 75 of 105 studies (71%) and age range in 66 (63%). Together, 92 (88%) reported at least one of the two, while 13 (12%) reported neither. Sex composition was reported in 85 (81%). Of these, 25 (29%) had distributions outside the 40 to 60% female range, in some cases by design.

Age coverage at the developmental and old-age extremes was sparse. Of the 66 studies that reported numerical age bounds, 13 (20%) included participants under age 5; 26 (39%) extended to age 70 or above. Six studies (9%) spanned both extremes within a single reference cohort. In the 69 studies that reported both age mean and SD, the mean $\pm$ 2SD approximation extended below age 5 in 19 cohorts (28%), reached age 70 or above in 19 (28%), and reached age 80 or above in 11 (16%).

Validation strategies

Validation reporting was uneven across the 106 studies that built or extended a normative model (Fig. 5a). External validation, whether a fitted model generalizes to an independent healthy cohort not used in fitting, was reported in 8% of applicable studies (NM-7), the largest gap among validation items. Internal validation through a held-out partition or cross-validation was reported in 64% (NM-6), and model performance metrics on the reference cohort in 50% (NM-9). Of these, central-tendency accuracy was reported in 77% but a calibration statistic of the deviation-score distribution in only 30% (16 of 53). Where a held-out partition was designated (NM-10, 56 applicable studies), demographic characteristics were reported fully in 39% and at least partially in 59%.

Clinical applications

Clinical-application items were also only partially reported (Fig. 5b). Inclusion criteria for the clinical cohort were reported in some form in 99% of studies and fully in 52% (CA-1). Sample sizes, demographics, and clinical characteristics of the clinical cohort were reported in all 122 studies and fully in 66% (CA-4). Assessment measures and procedures were reported in 82% of studies and fully specified in 52% (CA-5). Deviation interpretation, how a study characterized the clinical significance of its deviation scores, was reported in all 122 studies and fully developed in 59% (CA-6). The largest gap was clinical covariate handling beyond age and sex, with 29% of studies adjusting for or stratifying on factors such as medication, comorbidity, illness severity, or treatment status (CA-2).

Reporting completeness

Across the 18 items of the NORMA checklist (Table 1), reporting followed a consistent pattern (Table 2; Fig. 5c). The most basic descriptive items were reported almost universally: research objectives (CE-1) and clinical data sources (CA-3) in every study, and reference cohort data sources (NM-3) in 97%. The items needed to evaluate or reuse a model were reported far less completely. Reference cohort selection (NM-1), the statistical approach (NM-8), and clinical inclusion criteria (CA-1) appeared in almost every study but were fully specified in only about half (51%, 51%, and 52%), and reproducibility resources (NM-11) in 51%.

The per-study matrix shows that these gaps run across the field instead of concentrating in a few studies (Fig. 6). Three or more applicable items were unreported in 73 of 122 studies (60%), and the widest gaps, external validation (NM-7) and clinical covariate handling (CA-2), form the densest columns.

Discussion

This systematic review identified 122 normative-modeling studies in psychiatric and neurological neuroimaging, a literature that has grown from a single study in 2018 to 41 in 2024. Alongside this growth, the methods in use have diversified. The predominant modality remains sMRI (71% of studies), followed by fMRI (21%) and dMRI (7%), with extension to neurophysiological and molecular measures. Statistical approaches have shifted from GPR and linear models toward Bayesian and distributional regression. As multi-site data have become the norm, site effects are increasingly modeled within the normative model itself, particularly in Bayesian/HBR implementations, so that the uncertainty in an individual's deviation estimate can be separated into population, site, and individual components. This decomposition is largely specific to current normative models. The data these models rest on, however, have advanced less than the methods themselves. Most studies build a new reference cohort instead of reusing or extending an existing model, external validation was reported in only 8% of applicable studies, and clinical cohorts are often incompletely described. Beneath these gaps lie two basic, unresolved issues, both concerning the study samples: which population the reference cohort represents, and how the clinical cohort is defined. These two preconditions, not further methodological refinement, now determine whether normative modeling can support the individual-level inference it was developed to provide.

Of the two, the reference cohort is the more basic because every deviation score is defined relative to that cohort. The reference distribution is stable only when its outer percentiles, where extreme deviations fall, are estimated from enough data. In our review, the 105 newly built reference cohorts had a bimodal size distribution, with a median of 642 participants, 62% below 1,000, and only 15 at 10,000 or more. Whether that suffices has no fixed answer, since the size needed depends on the model and the measure, and no minimum has been established^{4,22}. Even so, stable outer percentiles consistently require samples in the thousands. Brain-morphometry work found overall predictive accuracy plateauing only near 3,000, with reference samples to date small to modest, between 145 and 870, close to our own median¹⁶, while a functional-connectivity model found its main components stable only around 2,000 and finer ones unstable even at 24,000²⁹. A simulation comparing linear and nonlinear models pointed the same way for structural measures, with the outer percentiles of hippocampal volume needing samples in the several thousands and their bias barely improving even at 50,000³⁰. Most of the cohorts we reviewed fall below the sizes these analyses indicate, so their most extreme deviation scores are the least reliable. That is the first reason individual-level interpretation rests on the reference cohort rather than on the modeling applied to it.

A cohort large enough to yield a stable estimate of the distribution is not, for its size alone, representative of the patients it is applied to. The largest cohorts were demographically narrow. Of the 15 with 10,000 participants or more, most drew on UK Biobank, and several also on ABCD, both Western cohorts. Race or ethnicity was rarely reported, and the few cohorts that did so were largely White²⁵⁻²⁷. Several of these cohorts, including the most widely used lifespan reference models, acknowledge that they draw on predominantly European-ancestry populations^{13,17}, a profile commonly summarized as Western, Educated, Industrialized, Rich, and Democratic (WEIRD)³¹. One such cohort went further and excluded its smaller ethnic groups because their post-quality-control

numbers were too few to model, trading coverage for stability²⁷. The reviewed studies say as much of their own data, describing cohorts that stay unrepresentative in sex, IQ, and ethnicity despite their size³² and reference models whose European ancestry does not match that of the patients studied³³. Coverage, not size, determines what a deviation score means. For a person the cohort represents, a z-score of -2 marks a position in a known distribution. For a person it does not, the same score may instead reflect the cohort's lack of data for people like them. The number is identical in both cases, so on its own it cannot separate a genuine deviation from a gap in coverage. In widely used structural brain reference models, self-reported race can be predicted from deviation scores with high accuracy, and this bias persists even when race is included in the model³⁴. Reference cohort representativeness therefore determines the generalizability of a deviation score: how far it can be interpreted, and for whom.

A representative reference cohort is necessary but not sufficient for individual-level clinical application. The second precondition is the clinical cohort itself, and how consistently its patients are defined. Full clinical inclusion criteria and assessment measures were reported in only 52% of studies, and clinical covariates beyond age and sex in only 29%. An individual deviation reflects the disorder together with factors such as medication and comorbidity. Because these factors are rarely measured or adjusted for, the disorder cannot be separated from their effects. In one cohort, comorbidities varied too widely across patients to be accounted for³⁵, and another study noted that patients recruited as research volunteers may not reflect the broader clinical population³⁶. Beyond reporting, the inconsistency is also conceptual, since diagnoses are defined more by clinical presentation than by a single biological cause, so that a single label groups heterogeneous individuals, documented across depression^{37,38}, ASD^{39,40}, and schizophrenia⁴¹⁻⁴⁷. The same label may denote different populations across studies. Individual-level clinical use thus turns less on the imaging than on how clearly clinical cohorts are defined and reported, limitations of the field rather than of normative modeling.

The cohorts determine what a deviation score can mean, but its reliability depends on the model that produces it. This turns on two checks in particular, calibration and external validation, both of which are rarely reported. Calibration concerns the whole distribution, its spread and shape as much as its mean, so that a given deviation marks the percentile it claims. Among studies that assessed performance, most reported only average accuracy, and calibration in fewer than a third. A model can be accurate on the mean while still misplacing those percentiles. External validation tests the model on an independent healthy cohort from different sources, not on held-out data from the same source. It is the least reported of all 18 items. Without it, whether a model applies to a new site, scanner, or population is simply assumed, and a deviation may reflect the individual or the model's poor fit to that setting.

These preconditions and checks can be evaluated and built upon only when studies report them in enough detail. Across the NORMA items, the descriptive ones, such as the research objective and data sources, were reported in almost every study. By contrast, the items capturing how the reference cohort was selected, how the model was constructed, and how it could be reproduced were fully reported far less often. Yet these are the items needed to compare, reproduce, or reuse a model. Only one of the 122 studies reported every applicable item in full. NORMA, the checklist we propose, specifies what to report, from building the reference cohort to applying the model, so that these elements are stated explicitly rather than pieced together from each paper's text. When studies meet this standard, their results can build on one another. Otherwise, even sound models remain isolated.

The strength of this review is that it does not merely summarize the diverse methodology of normative modeling but reframes it as a question of the preconditions for clinical translation, asking what must hold before normative models can support individual-level clinical use. Its scope is broad, spanning psychiatric and neurological disorders and modalities from neuroimaging to neurophysiological measures. It places the reference cohort at the center of that account and develops NORMA to make explicit the decisions an adequate normative model and its application require. These contributions come with several limitations. Inclusion was restricted to English-language publications, and studies published after the April 2025 cutoff are not captured. Studies developing normative models without a clinical application, and normative modeling in other medical specialties, fell outside our eligibility criteria, and the methods used there may differ. Methodological heterogeneity precluded a formal meta-analysis, so the synthesis is narrative and characterizes how models are built and reported, not the validity of their findings. A GRADE-style certainty rating therefore did not apply. Transparency was assessed with NORMA, scored item by item against operationalized criteria through team consensus. NORMA may not capture every relevant dimension, but it reflects what is currently feasible.

Normative modeling has completed a necessary first step, showing that individual deviations from population norms differ from one patient to the next even within a single diagnosis. Whether the approach can move toward clinical application, however, now depends less on the precision of any single study than on two field-level preconditions: how representative the reference cohorts are, and how consistently the clinical cohorts are defined. Refining the methods remains important, but on its own it cannot settle these questions. Doing so will require shared practice across studies in building reference cohorts, reporting analytic choices, and validating models. Until these preconditions are in place, individual-level deviation scores are best interpreted as research measures, not standalone tools for clinical decision-making.

Methods

Protocol and reporting standards

This systematic review followed PRISMA 2020 guidelines⁴⁸ and was prospectively registered (PROSPERO CRD420251015671). The PRISMA checklist is provided in Supplementary Table 2. Four deviations from the registered protocol occurred during analysis: consortium and public-repository datasets were collapsed into a single multi-source category; the raw-versus-preprocessed data distinction was instead captured as a reporting transparency item (NM-5); modified GRADE was not applied due to the absence of pooled effect estimates; and qualitative reporting bias assessment was likewise not applied.

Search strategy and study selection

We searched PubMed, Embase, MEDLINE, PsycINFO, and Scopus for studies published from January 2005 onward, with the last search in April 2025 (full search strings in Supplementary Note 2). After duplicate removal (Fig. 2), two reviewers independently screened the remaining records in Covidence, with

disagreements resolved by discussion or third-reviewer consultation. Studies were included if they applied normative modeling to neuroimaging or neurophysiological data in psychiatric or neurological populations and reported individual-level deviation metrics. For synthesis, studies were grouped by imaging modality, statistical modeling approach, model build, and clinical domain. We excluded non-empirical articles, studies developing normative models without clinical application, non-English articles, and studies that reported only a brain-age gap, which summarizes deviation as a single predicted-age index rather than feature-level deviations from a reference distribution.

Data extraction

Several reviewers extracted and verified study characteristics (imaging modality, statistical framework, model build, reference cohort sample size and age range, demographic composition, site effect handling, validation strategy, and clinical application domain). Part of this work was supported by an AI-based tool, and all final decisions were made by the human reviewers. No clinical outcome measures or effect sizes were extracted, as this review is methodological in focus. We did, however, record each study's clinical population and target disorder, and the clinical domain of NORMA (CA-1 to CA-6) captures how studies characterized their clinical cohorts and interpreted deviation scores. When information was missing or ambiguous, we recorded it as "not reported" instead of imputing values. Discrepancies were resolved by discussion or third-reviewer consultation. Studies were classified as building a new normative model when model fitting was documented within the manuscript, with evidence including the training cohort, covariate specification, fitting procedure, or model performance. Studies that applied previously published model parameters without re-estimation were classified as pre-trained, irrespective of overlap with prior reference cohorts. Borderline cases were resolved through team discussion. The reference cohort and the clinical cohort of each study were also labeled as primary data collection when imaging data were newly acquired by the study team, or as secondary data when sourced from existing datasets. Some included studies were preprints at the time of search and were published during the review. For these, we extracted data and assessed reporting from the version of record where available and cited that version. For the year-by-year counts, each study was dated by its first public release, so the trend reflects when the work entered the field rather than its later publication date.

Statistical modeling approaches were grouped into five categories based on their distributional output and methodological characteristics: linear and generalized linear models; distributional regression (e.g., generalized additive models for location, scale, and shape, GAMLSS, and quantile regression); Bayesian regression (including hierarchical Bayesian formulations); GPR; and deep learning or other methods. Studies employing multiple approaches were classified by their primary model. Site effect handling was similarly grouped into operational categories. Validation strategies were assessed at the reporting level using NORMA items NM-6 and NM-7. Full classification rules are provided in Supplementary Notes 1, 3. Per-study mappings are provided in Supplementary Data 1.

For cohort-size summaries, we excluded the single study that combined a pre-trained component with a newly built reference cohort²³. Studies that fit multiple normative models within the same paper were treated as one observation, regardless of whether the models drew from one or several underlying reference cohorts. The largest training-phase n was used as the representative cohort size. We extracted numerical age bounds where available, including those deducible from inclusion criteria. Age coverage was assessed using two complementary metrics: the reported numerical range, which captures observed extremes, and the mean $\pm$ 2SD interval, which approximates the age window of reliable reference-distribution coverage.

For the 15 studies with reference cohorts of 10,000 or larger, a focused secondary extraction of geographic recruitment region and race or ethnicity reporting status was performed. Two protocol-registered classifications were collapsed during analysis. Consortium and public-repository datasets often blur in practice, especially for the field's most-used sources, and we treated them as a single multi-source category. Raw and preprocessed data similarly proved difficult to separate consistently across studies, since most combined partially preprocessed derivatives with additional in-house processing. The transparency of preprocessing reporting is instead characterized through NORMA item NM-5.

Reporting transparency assessment

Reporting transparency was assessed using the NORMA framework. The registered protocol specified the NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies (<https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools>), complemented by relevant CLAIM checklist⁴⁹ items for modeling-specific aspects. Broad reporting standards for neuroimaging, such as COBIDAS^{50,51}, establish how study design, acquisition, and analysis are documented. NORMA is complementary and specific to normative modeling, addressing decisions such as how the reference cohort is defined and how the model is validated and calibrated. To target the reporting gaps that recurred across the reviewed studies, we built it on the NIH assessment tool and the CLAIM checklist. NORMA itself was not pre-registered. It comprises 18 items: 17 adapted from these sources and one item we developed to capture the interpretation of deviation scores and their clinical significance. The framework is organized into three domains: common elements, model development and validation, and clinical application (Table 1). Two scoring conventions are described below. For NM-7 (external validation), credit was given only when validation used an independent healthy reference cohort from sites or sources distinct from the training cohort. Same-site additional participants and clinical-cohort healthy arms did not satisfy the criterion. For NM-10 (partition characteristics), studies that performed only k-fold cross-validation without designating named held-out partitions were assessed as not applicable. k-fold CV is an internal model-selection procedure, and independent demographic characterization of its iterative partitions is not conventionally reported. In addition, for NM-9 (model performance), we recorded whether each study reported central-tendency, overall goodness-of-fit, or calibration metrics of the modeled distribution⁵²⁻⁵⁴. Full item definitions and scoring guidance are provided in Supplementary Note 4. An editable form is provided in Supplementary Data 4.

Two reviewers independently applied the checklist to 22 studies (18%), without AI support, to calibrate scoring criteria (Gwet's AC2 = 0.67, 95% CI: 0.59–0.75; weighted percent agreement = 85.5%). This calibration informed the final item definitions used throughout the review. For the remaining 100 studies, a custom application submitted each study's full materials (main text and supplementary) with the NORMA checklist to three command-line coding agents (Claude Code with Claude Opus 4.6, Codex CLI with GPT-5.1-Codex-Max, and Gemini CLI with Gemini 3 Pro) and compiled their item-level assessments. A

reviewer then independently assessed each study, using these AI-generated outputs only as a reference, and a second reviewer verified every assessment. Disagreements and low-confidence cases were resolved through discussion, and all final judgments were made by the human reviewers. Per-study assessments are provided in Supplementary Data 3. The framework characterizes only the transparency of methodological reporting. It is not an assessment of risk of bias, methodological quality, or the certainty of evidence. Because NORMA's item definitions were both shaped and applied within this review, the item-level rates are a descriptive application of the framework, not an independent validation, and they reflect the review team's judgment against operationalized criteria, not independent adjudication.

Synthesis

Synthesis was narrative rather than meta-analytic, as the heterogeneity of statistical frameworks and reference cohorts precluded pooled effect estimates. The four GRADE domains are instead captured descriptively by NORMA item-level reporting rates. We examined sample size distributions by modeling approach, age coverage and demographic reporting completeness within the new-build subset, and validation strategies and site effect handling practices across modeling approaches. Analyses were performed in R (version 4.3.2) and Stata MP (version 18). Figures were generated using Python (version 3.12) with matplotlib.

Declarations

Acknowledgements

The authors thank colleagues at their respective institutions for helpful discussions during the preparation of this review.

Funding

KS discloses support for the research of this work from the Japan Agency for Medical Research and Development (AMED) [JP24wm0625302]. YH has received grants from AMED [JP24wm0625302], Japan Society for the Promotion of Science (JSPS) KAKENHI [20KK0193, 21H02851, 25K21813] and Japan Science and Technology Agency (JST) K Program grant number [JPMJKP25Y5]. SK has received grants from AMED [JP24wm0625302, JP23tm0524002], JSPS KAKENHI [JP24K02378, JP26H02468], JST Fusion Oriented REsearch for disruptive Science and Technology (FOREST) Program [JPMJFR231Q], and SENSHIN Medical Research Foundation. SN has received grants from AMED [JP24wm0625302, JP24wm0625307] and JSPS KAKENHI [22H03002, 23K24263, 24K02541, 24K10740, 24K02387, 25K10823, 25K10856, 25K21813, 26K02358, 26K10763], Japan Research Foundation for Clinical Pharmacology, Naito Foundation, Takeda Science Foundation, Watanabe Foundation, Osakeno-Kagaku Foundation, and Astellas Foundation within the past three years. YS was supported by the Australian Government Research Training Program Scholarship, the Japan Student Services Organization Scholarship, and the Honjo International Scholarship. YT has received grants from JSPS (26KJ0975). The funders had no role in the design of the review, in the collection, analysis or interpretation of data, in the writing of the manuscript, or in the decision to submit it for publication.

Author contributions

N.T.: Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. I.U.: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – review & editing. Y.S. and Y.T.: Investigation, Methodology, Validation, Visualization, Writing – review & editing. T.E.: Investigation, Methodology, Validation, Writing – review & editing. S.T., K.T., and A.T.: Investigation, Validation, Writing – review & editing. S.S.: Investigation, Methodology, Supervision, Validation, Writing – review & editing. A.Z. and A.M.: Supervision, Writing – review & editing. Y.H.: Methodology, Supervision, Writing – review & editing. S.K. and S.N.: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Competing interests

S.N. has received research support, manuscript fees or speaker's honoraria from Asahi Quality & Innovations, Ltd., Teijin Pharma, Sumitomo Pharma, Meiji Seika Pharma, Otsuka, PDR pharma, and MSD within the past three years. The remaining authors declare no competing interests.

Data availability

All data used for this paper are provided in the main text and Supplementary Information.

References

1. Kapur, S., Phillips, A. G. & Insel, T. R. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* 17, 1174–1179 (2012).
2. van Erp, T. G. et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* 21, 547–553 (2016).
3. Cheon, E. J. et al. Cross disorder comparisons of brain structure in schizophrenia, bipolar disorder, major depressive disorder, and 22q11.2 deletion syndrome: a review of ENIGMA findings. *Psychiatry Clin. Neurosci.* 76, 140–161 (2022).
4. Koike, S., Tanaka, S. C. & Hayashi, T. Beyond case-control study in neuroimaging for psychiatric disorders: harmonizing and utilizing the brain images from multiple sites. *Neurosci. Biobehav. Rev.* 171, 106063 (2025).
5. Marquand, A. F., Rezek, I., Buitelaar, J. & Beckmann, C. F. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol. Psychiatry* 80, 552–561 (2016).

6. Pergola, G., Penzel, N., Sportelli, L. & Bertolino, A. Lessons learned from parsing genetic risk for schizophrenia into biological pathways. *Biol. Psychiatry* 94, 121–130 (2023).
7. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* 179, 1469–1482 (2019).
8. Goodkind, M. et al. Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry* 72, 305–315 (2015).
9. Cicchetti, D. & Rogosch, F. A. Equifinality and multifinality in developmental psychopathology. *Dev. Psychopathol.* 8, 597–600 (1996).
10. Dukart, J., Lotter, L. D., Paquola, C., Eickhoff, S. B. & Schilbach, L. The many-to-many problem of endophenotypes in psychiatry – a biological perspective. *Mol. Psychiatry* 31, 3416–3424 (2026).
11. Insel, T. et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* 167, 748–751 (2010).
12. Kotov, R. et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* 126, 454–477 (2017).
13. Bethlehem, R. A. I. et al. Brain charts for the human lifespan. *Nature* 604, 525–533 (2022).
14. Cole, T. J. The development of growth references and growth charts. *Ann. Hum. Biol.* 39, 382–394 (2012).
15. Marquand, A. F. et al. Conceptualizing mental disorders as deviations from normative functioning. *Mol. Psychiatry* 24, 1415–1424 (2019).
16. Ge, R. et al. Normative modelling of brain morphometry across the lifespan with CentileBrain: algorithm benchmarking and model optimisation. *Lancet Digit. Health* 6, e211–e221 (2024).
17. Rutherford, S. et al. Charting brain growth and aging at high spatial precision. *eLife* 11, e72904 (2022).
18. Lv, J. et al. Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. *Mol. Psychiatry* 26, 3512–3523 (2021).
19. Segal, A. et al. Regional, circuit and network heterogeneity of brain abnormalities in psychiatric disorders. *Nat. Neurosci.* 26, 1613–1629 (2023).
20. Bethlehem, R. A. I. et al. A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Commun. Biol.* 3, 486 (2020).
21. Shan, X. et al. Mapping the heterogeneous brain structural phenotype of autism spectrum disorder using the normative model. *Biol. Psychiatry* 91, 967–976 (2022).
22. Frazza, C., Rutherford, S., Bučková, B. R., Beckmann, C. F. & Marquand, A. F. The promise of quantifying individual risk for brain disorders through normative modeling, a narrative review. *Neurosci. Biobehav. Rev.* 176, 106284 (2025).
23. Rutherford, S. et al. Evidence for embracing normative modeling. *eLife* 12, e85082 (2023).
24. Leenings, R. et al. Judged by your neighbors: a novel framework for personalized assessment of brain structural aging effects in diverse populations. Preprint at medRxiv <https://doi.org/10.1101/2024.12.24.24319598> (2025).
25. Cirstian, R. et al. Lifespan normative models of white matter fractional anisotropy: applications to early psychosis. *Biol. Psychiatry* 99, 559–568 (2025).
26. Frazza, C. et al. Unraveling the link between CNVs, cognition and individual neuroimaging deviation scores from a population-based reference cohort. *Nat. Ment. Health* 2, 1451–1463 (2024).
27. Georgiadis, F. et al. Estimating accelerated retinal decline in mental health disorders through normative modeling. *Schizophr. Bull.* <https://doi.org/10.1093/schbul/sbaf178> (2025).
28. Kia, S. M. et al. Closing the life-cycle of normative modeling using federated hierarchical Bayesian regression. *PLoS One* 17, e0278776 (2022).
29. Kobbersmed, J. R. L., Gohil, C., Marquand, A. & Vidaurre, D. Normative modeling of brain function abnormalities in complex pathology requires a whole-brain approach. *Prog. Neurobiol.* 260, 102906 (2026).
30. Bozek, J., Griffanti, L., Lau, S. & Jenkinson, M. Normative models for neuroimaging markers: impact of model selection, sample size and evaluation criteria. *NeuroImage* 268, 119864 (2023).
31. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83 (2010).
32. Bedford, S. A. et al. Brain-charting autism and attention-deficit/hyperactivity disorder reveals distinct and overlapping neurobiology. *Biol. Psychiatry* 97, 517–530 (2025).
33. Thukral, R. A. et al. Dissecting heterogeneity in functional network connectivity aberrations in antipsychotic medication-naïve first episode psychosis patients – a normative modeling study. Preprint at medRxiv <https://doi.org/10.1101/2024.08.23.24312480> (2024).
34. Rutherford, S. et al. To which reference class do you belong? Measuring racial fairness of reference classes with normative modeling. *Proc. Mach. Learn. Res.* 252, 1–34 (2024).
35. Wolfers, T. et al. Individual differences v. the average patient: mapping the heterogeneity in ADHD using normative models. *Psychol. Med.* 50, 314–323 (2020).
36. Verdi, S. et al. Revealing individual neuroanatomical heterogeneity in Alzheimer disease using neuroanatomical normative modeling. *Neurology* 100, e2442–e2453 (2023).
37. Dinga, R. et al. Evaluating the evidence for biotypes of depression: methodological replication and extension of Drysdale et al. (2017). *NeuroImage Clin.* 22, 101796 (2019).
38. Drysdale, A. T. et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* 23, 28–38 (2017).

39. Lombardo, M. V., Lai, M. C. & Baron-Cohen, S. Big data approaches to decomposing heterogeneity across the autism spectrum. *Mol. Psychiatry* 24, 1435–1450 (2019).
40. Loth, E. et al. The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Mol. Autism* 8, 24 (2017).
41. Clementz, B. A. et al. Identification of distinct psychosis biotypes using brain-based biomarkers. *Am. J. Psychiatry* 173, 373–384 (2016).
42. Cropley, V. L. et al. Accelerated gray and white matter deterioration with age in schizophrenia. *Am. J. Psychiatry* 174, 286–295 (2017).
43. Gillespie, A. L., Samanaite, R., Mill, J., Egerton, A. & MacCabe, J. H. Is treatment-resistant schizophrenia categorically distinct from treatment-responsive schizophrenia? A systematic review. *BMC Psychiatry* 17, 12 (2017).
44. Howes, O. D. & Kapur, S. A neurobiological hypothesis for the classification of schizophrenia: type A (hyperdopaminergic) and type B (normodopaminergic). *Br. J. Psychiatry* 205, 1–3 (2014).
45. Kirkpatrick, B., Mucci, A. & Galderisi, S. Primary, enduring negative symptoms: an update on research. *Schizophr. Bull.* 43, 730–736 (2017).
46. Mouchlianitis, E., McCutcheon, R. & Howes, O. D. Brain-imaging studies of treatment-resistant schizophrenia: a systematic review. *Lancet Psychiatry* 3, 451–463 (2016).
47. Voineskos, A. N., Jacobs, G. R. & Ameis, S. H. Neuroimaging heterogeneity in psychosis: neurobiological underpinnings and opportunities for prognostic and therapeutic innovation. *Biol. Psychiatry* 88, 95–102 (2020).
48. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71 (2021).
49. Mongan, J., Moy, L. & Kahn, C. E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol. Artif. Intell.* 2, e200029 (2020).
50. Nichols, T. E. et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* 20, 299–303 (2017).
51. Pernet, C. et al. Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research. *Nat. Neurosci.* 23, 1473–1483 (2020).
52. Rutherford, S. et al. The normative modeling framework for computational psychiatry. *Nat. Protoc.* 17, 1711–1734 (2022).
53. Dinga, R. et al. Normative modeling of neuroimaging data using generalized additive models of location, scale and shape. Preprint at bioRxiv <https://doi.org/10.1101/2021.06.14.448106> (2021).
54. Marquand, A. F. et al. Fairly evaluating the performance of normative models. *Lancet Digit. Health* 6, e775 (2024).

Tables

Table 1 | The NORMA reporting checklist for normative modeling studies

Item	Title	Description	
Common elements (CE)			
CE-1	Research objectives	Are the study objectives stated?	A
Model development and validation (NM)			
NM-1	Reference cohort selection	Are eligibility criteria for the reference cohort and image quality standards described?	A
NM-2	Reference cohort covariate handling	Are covariates handled appropriately when building the normative model?	A
NM-3	Reference cohort data sources	Are the neuroimaging data sources of the reference cohort described?	A
NM-4	Image acquisition protocol	Is the image acquisition protocol described?	F
NM-5	Data preprocessing	Are the preprocessing procedures described?	F
NM-6	Internal validation of the reference cohort	Is a data partitioning strategy used for internal validation?	N
NM-7	External validation of the reference cohort	Is the normative model evaluated on an independent external dataset of healthy individuals?	N
NM-8	Statistical approach to normative modeling	Is the statistical approach described?	N
NM-9	Model performance	Is model performance quantitatively assessed for the reference cohort?	N
NM-10	Reference cohort partition characteristics	Are characteristics of the reference cohort described for each data partition?	F
NM-11	Reproducibility resources	Is sufficient information provided to support reproducibility?	A
Clinical application (CA)			
CA-1	Clinical cohort selection	Are eligibility criteria for the clinical cohort and image quality standards described?	A
CA-2	Clinical covariate handling	Are clinical covariates accounted for in analyses involving the clinical cohort?	A
CA-3	Clinical cohort data sources	Are the neuroimaging data sources of the clinical cohort described?	A
CA-4	Clinical cohort characteristics	Are the characteristics of the clinical cohort described?	A
CA-5	Clinical assessment measures	Are the clinical assessment measures described?	A
CA-6	Deviation score interpretation	Is the interpretation of deviation scores and their clinical significance described?	A

Each item is scored Y (fully reported), P (partially reported), N (not reported), or NA (not applicable). Full item definitions and scoring guidance are in Supplementary form in Supplementary Data 4.

Table 2 | Reporting completeness across the 18-item NORMA checklist

Item	Title	Y	P	N	NA	Applicable	Y (%)	Y+P (%)
Common elements (CE)								
CE-1	Research objectives	122	0	0	0	122	100	100
Model development and validation (NM)								
NM-1	Reference cohort selection	62	60	0	0	122	51	100
NM-2	Covariate handling	74	48	0	0	122	61	100
NM-3	Data sources	118	1	3	0	122	97	98
NM-4	Image acquisition	27	0	0	95	27	100	100
NM-5	Preprocessing	81	30	1	10	112	72	99
NM-6	Internal validation (partition)	68	0	38	16	106	64	64
NM-7	External validation	8	0	98	16	106	8	8
NM-8	Statistical approach	54	52	0	16	106	51	100
NM-9	Model performance	53	0	53	16	106	50	50
NM-10	Partition characteristics	22	11	23	66	56	39	59
NM-11	Reproducibility resources	62	48	12	0	122	51	90
Clinical application (CA)								
CA-1	Clinical cohort selection	64	57	1	0	122	52	99
CA-2	Clinical covariate handling	35	0	87	0	122	29	29
CA-3	Clinical data sources	122	0	0	0	122	100	100
CA-4	Clinical characteristics	80	42	0	0	122	66	100
CA-5	Assessment measures	63	37	22	0	122	52	82
CA-6	Deviation interpretation	72	50	0	0	122	59	100

Y, fully reported; P, partial; N, not reported; NA, not applicable. NA: NM-4, secondary data only; NM-5, preprocessed derivatives only; NM-6 to NM-9, pre-trained model; NM-10, pre-trained model or k-fold cross-validation only. Applicable n excludes NA studies; Y (%) and Y+P (%) use this denominator.

Figures

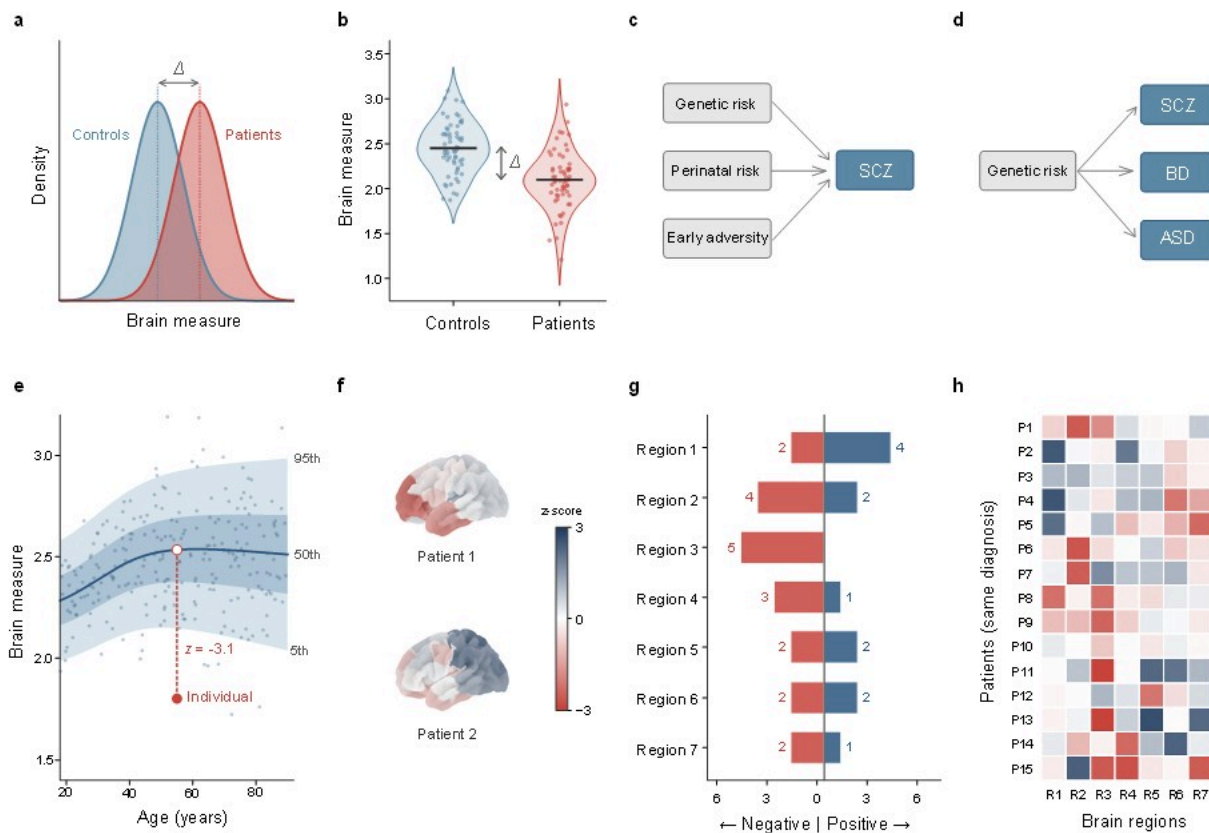


Figure 1

Why normative modeling. **a**, Case-control designs test for a mean difference (Δ) between patients and controls. **b**, Individual values overlap substantially between groups; group-level effects are modest. **c**, Equifinality: distinct biological pathways can converge on the same diagnosis. **d**, Multifinality: a shared risk factor can lead to divergent diagnoses. **e**, Normative modeling places each individual within a reference distribution conditioned on age, yielding a z-score relative to the expected range. **f**, Two patients with the same diagnosis show distinct individual deviation profiles across cortical regions. **g**, Per-region count of patients with extreme negative (red) or positive (blue) deviation ($|z| > 1.96$); opposing signs cancel under group averaging. **h**, Within-diagnosis patient-by-region z-score matrix; individual profiles do not overlap.

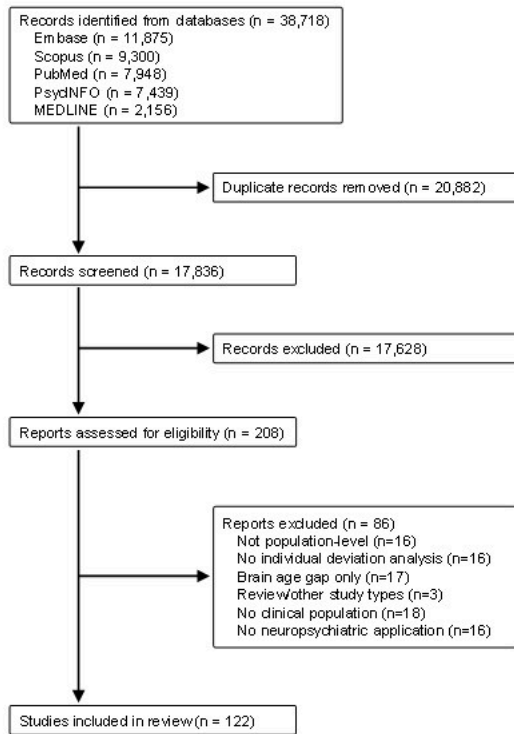


Figure 2

Study selection. PRISMA 2020 flow diagram across five databases.

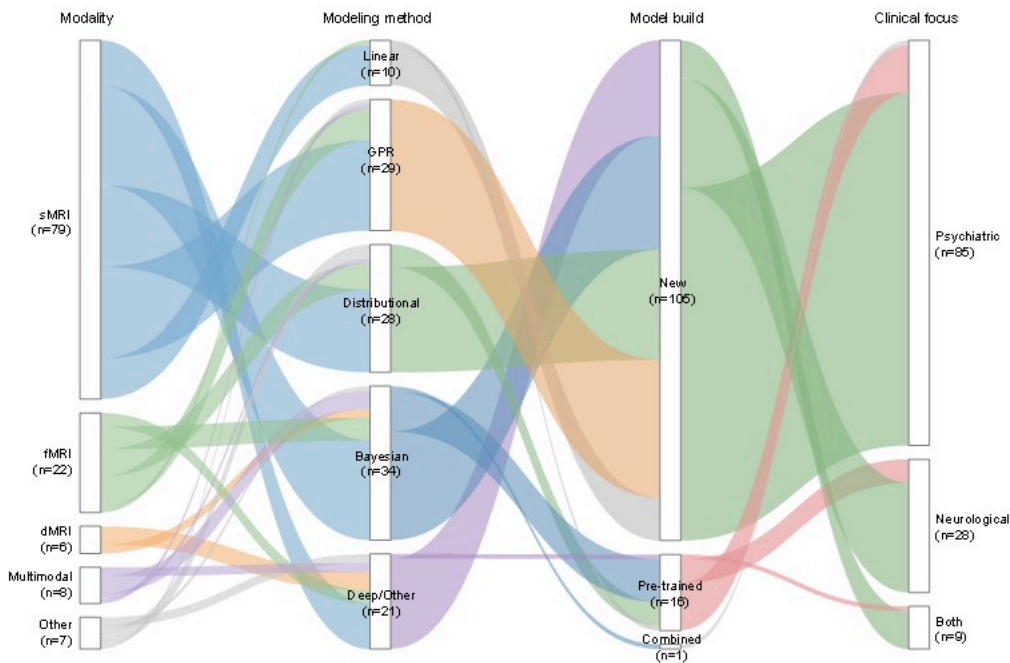


Figure 3

Methodological characteristics of included studies across four dimensions. Sankey diagram showing the distribution of 122 studies across imaging modality, modeling method, model build, and clinical focus. Flow width is proportional to the number of studies. Modeling methods are ordered by approximate complexity and date of introduction (Linear models at top, Deep/Other at bottom). Ribbons are colored by source modality, modeling method, and model build category in successive segments. Bayesian regression was the most common modeling method (n=34), and most studies built new models from primary data (n=105) rather than applying pre-trained models (n=16) or combining the two (n=1).

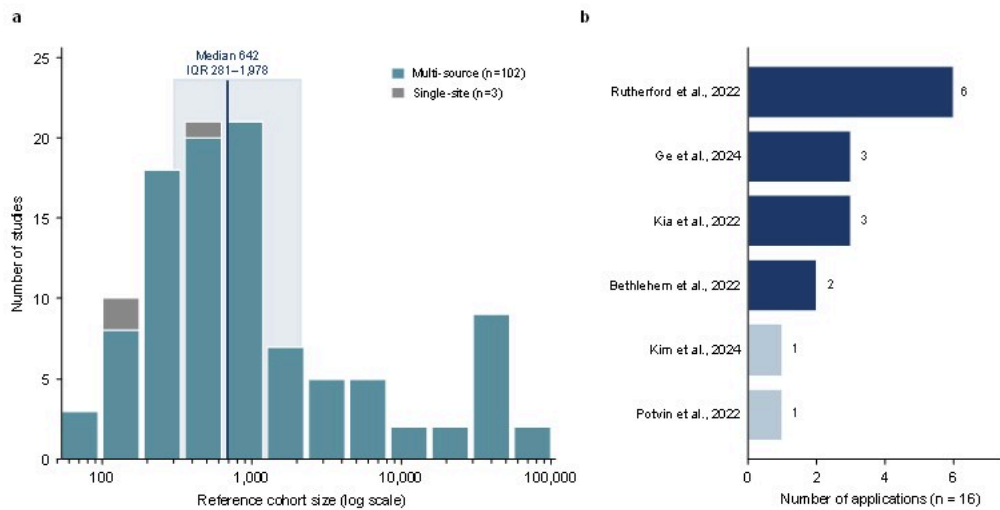


Figure 4

Reference cohort characteristics. **a**, Cohort size distribution for the 105 studies that built new reference cohorts (3 single-site, 102 multi-source). The vertical line shows the median (642) and the shaded region the interquartile range (281–1,978). **b**, Source models reused by the 16 studies that applied a pre-trained normative model. Four sources account for 14 of 16 applications (88%, deep navy).

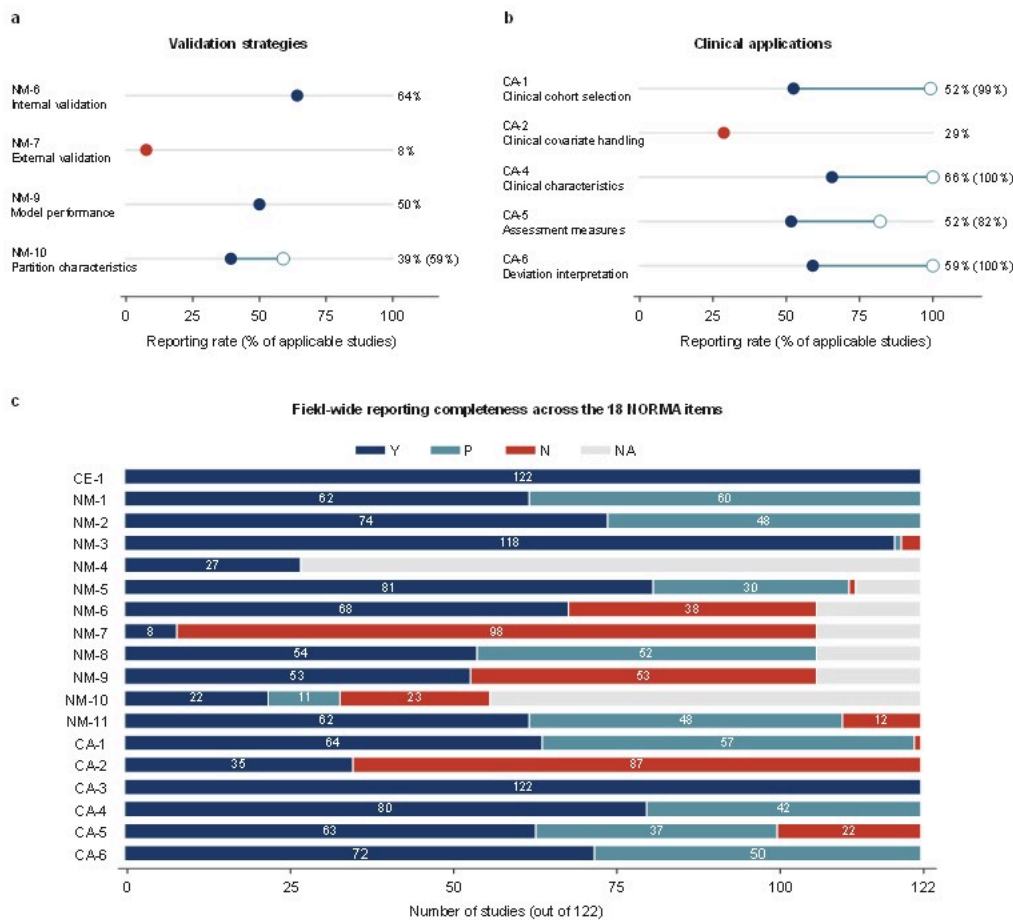


Figure 5

NORMA item-level reporting completeness across 18 items. **a**, Reporting rates for four validation items. Filled markers show full reporting (Y) and open rings show full plus partial (Y+P). Red marks severe field-wide gaps (Y < 30%). **b**, Reporting rates for five clinical-application items, same convention. CA-3 (clinical data sources) is omitted because all 122 studies reported it. **c**, Stacked bar showing the count of studies in each reporting category (Y, fully reported; P, partial; N, not reported; NA, not applicable) for all 18 NORMA items in checklist order. Per-study assessments are provided in Supplementary Data 3.

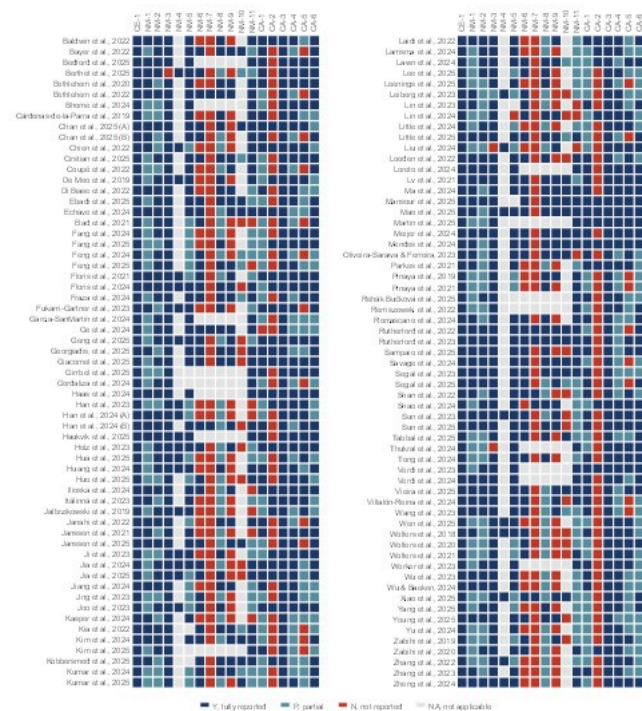


Figure 6

Per-study NORMA reporting matrix. Item-level NORMA assessments for each of the 122 included studies. Rows: studies in alphabetical order, split across two columns. Cells indicate reporting category: Y, fully reported; P, partial; N, not reported; NA, not applicable per item-specific rules. The matrix represents the present review team's judgment against operationalized criteria (Methods).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryData1PerStudyCharacteristics.xlsx](#)
- [SupplementaryData2ReferenceCohortCharacteristics.xlsx](#)
- [SupplementaryData3PerStudyNORMAAssessments.xlsx](#)
- [SupplementaryData4NORMAReportingChecklist.docx](#)
- [Supplementaryinformation.pdf](#)