Supplementary Information

Table of Contents

Supplementary Notes

Supplementary Note 1: Statistical Modeling Approaches Classification 2
Supplementary Note 2: Search Strategy and Full Search Strings 3
Supplementary Note 3: Site Effect Handling Classification 4
Supplementary Note 4: NORMA Framework Definitions and Scoring Guidance 6

Supplementary Tables

Supplementary Table 1: Studies excluded at full-text screening with reasons 14
Supplementary Table 2: PRISMA 2020 checklist 17

Supplementary Figures

Supplementary Fig. 1: Site handling distribution by modeling family 20

Supplementary Files (Available as separate downloads)

Supplementary Data 1: per-study detail (Excel)
Supplementary Data 2: reference cohort details (Excel)
Supplementary Data 3: per-study NORMA assessments (Excel)
Supplementary Data 4: NORMA editable form (Word)

Supplementary Notes

Supplementary Note 1: Statistical Modeling Approaches Classification

We grouped studies into five modeling categories based on their distributional output and methodological characteristics. The labels reported in each study (Supplementary Data 1) were mapped to these categories as follows.

(1) Linear and generalized linear models (n=10). Classical regression frameworks that parameterize the conditional mean: linear mixed model (LMM); polynomial regression (POLY); generalized linear model (GLM); and ordinary least squares regression (OLSR).

(2) Distributional regression models (n=28). Models that allow distributional parameters beyond the mean (e.g., scale and shape) to vary with covariates, as well as quantile- and smoother-based frameworks: generalized additive models for location, scale, and shape (GAMLSS); generalized additive model (GAM); generalized additive mixed model (GAMM); quantile regression (QUANTREG); locally estimated scatterplot smoothing (LOESS); and multivariate fractional polynomial regression (MFPR).

(3) Bayesian regression models (n=34). Bayesian linear regression (BLR); warped Bayesian linear regression (WBLR); and hierarchical Bayesian regression (HBR).

(4) Gaussian process-based models (n=29). Gaussian process regression (GPR).

(5) Deep learning and other methods (n=21). Autoencoder variants, including standard autoencoder (AE), conditional variational autoencoder (ConVAE), multimodal variational autoencoder (mmVAE), multimodal subject-invariant variational autoencoder (mmSIVAE), adversarial autoencoder (AAE), and vector-quantized variational autoencoder (VQ-VAE); generative pre-trained transformer (GPT); person-based similarity index (PBSI); multivariate archetypal analysis (MAA); network-based normative model (NWK); nonparametric interval estimation (NIE); functional component-of-individual networks (FUNCOIN); and nearest-neighbor normativity (N3). Studies with unknown or unreported modeling methods (UNKNOWN/NR) were also assigned to this category.

Where multiple modeling approaches were evaluated within the same study, we classified each study according to the primary model emphasized in the main results. If no clear primary model could be identified, we assigned the study to the more flexible or normative modeling-specific category according to the following hierarchy: deep learning/other > Gaussian process models > Bayesian regression > distributional regression > linear/GLM.

Supplementary Note 2: Search Strategy and Full Search Strings

We searched five databases between April 25 and April 28, 2025. The following search terms were used for each database with filters applied for English language, human studies, and publication date from January 1, 2005 onward.

PubMed (April, 25, 2025):

("Normative model*" OR "Normative range" OR "Normative trajector*" OR "Canonical model*" OR "Reference model*" OR "Reference trajector*" OR "Reference growth curve" OR "Reference Chart*" OR "Life course trajector*" OR "Lifespan trajector*" OR "Lifespan growth" OR "Lifespan change*" OR "Developmental trajector*" OR "Aging trajector*")

MEDLINE via Ovid (April, 25, 2025):

(normative model* OR normative range OR normative trajector* OR canonical model* OR reference model* OR reference trajector* OR reference growth curve OR reference chart* OR life course trajector* OR lifespan trajector* OR lifespan growth OR lifespan change* OR developmental trajector* OR aging trajector*).mp.

Embase via Ovid (April, 25, 2025):

(normative model* OR normative range OR normative trajector* OR canonical model* OR reference model* OR reference trajector* OR reference growth curve OR reference chart* OR life course trajector* OR lifespan trajector* OR lifespan growth OR lifespan change* OR developmental trajector* OR aging trajector*).mp.

PsycINFO (April, 28, 2025):

"(TI (""Normative model*"" OR ""Normative range"" OR ""Normative trajector*"" OR ""Canonical model*"" OR ""Reference model*"" OR ""Reference trajector*"" OR ""Reference growth curve"" OR ""Reference chart*"" OR ""Life course trajector*"" OR ""Lifespan trajector*"" OR ""Lifespan growth"" OR ""Lifespan change*"" OR ""Developmental trajector*"" OR ""Aging trajector*""))
OR
(AB (""Normative model*"" OR ""Normative range"" OR ""Normative trajector*"" OR ""Canonical model*"" OR ""Reference model*"" OR ""Reference trajector*"" OR ""Reference growth curve"" OR ""Reference chart*"" OR ""Life course trajector*"" OR ""Lifespan trajector*"" OR ""Lifespan growth"" OR ""Lifespan change*"" OR ""Developmental trajector*"" OR ""Aging trajector*""))"

Scopus (April, 28, 2025):

TITLE-ABS-KEY ("normative model*" OR "normative range" OR "normative trajector*" OR "canonical model*" OR "reference model*" OR "reference trajector*" OR "reference growth curve" OR "reference chart*" OR "life course trajector*" OR "lifespan trajector*" OR "lifespan growth" OR "lifespan change*" OR "developmental trajector*" OR "aging trajector*")

Supplementary Note 3: Site Effect Handling Classification

We classified site effect handling strategies into four categories based on how site-related variability was addressed in relation to the normative model. For each study, we identified the primary normative model from which individual deviation scores were computed and reported in the main analysis, and classification was based on site handling within this primary model. Site-inclusive specifications that appeared only in sensitivity or benchmarking analyses, but were not used to derive the main results, were not counted toward the Modeled category.

(1) Single-site (N/A): Studies using data from a single acquisition site, where site effects are not applicable. When a study applied a pre-existing normative model trained on a single-site reference dataset (e.g., Cam-CAN) to a multi-site clinical cohort, classification was determined by how site effects in the clinical cohort were handled, not by the reference cohort.

(2) None reported: Multi-site studies in which site effects were not addressed either prior to or within the normative model. This category includes studies that made no reference to site effects, as well as studies that applied post-hoc adjustments to deviation scores (for example, centering deviation scores by within-site control means, or linear recalibration of z-scores) after the normative model had already been fitted and applied. This category also includes studies in which site or scanner variability was addressed only at the evaluation stage, for example by restricting the reference distribution to subgroups of controls matched on site, scanner, or protocol when computing individual deviation scores. Such post-hoc or inference-stage adjustments were not treated as harmonization because they did not alter the input data or the model structure.

(3) Harmonized: Site or scanner effects were addressed through explicit correction procedures applied to the input data prior to normative model fitting. This includes statistical batch-removal methods such as ComBat, NeuroCombat, and CovBat, as well as signal-level harmonization approaches specific to particular modalities (for example, RISH-based harmonization of diffusion-weighted signals). Harmonization was defined by its timing (pre-fitting) and its target (input features or raw signals), irrespective of the specific algorithm used. Studies were also classified as Harmonized when the primary normative model did not include site as a covariate but authors recommended or reported pre-fitting harmonization as the preferred site-handling workflow, even if site-inclusive models were additionally evaluated.

(4) Modeled: Site effects were explicitly incorporated within the primary normative model, such that site entered the estimation of the reference distribution used to derive deviation scores. Implementations included site as a random effect in hierarchical Bayesian regression, site as a covariate in Gaussian process regression or quantile regression, and analogous approaches in other modeling families. Classification as Modeled required that the primary model reported in the main analysis (from which deviation scores were computed) contained site as a parameter. Studies that evaluated site-inclusive models only as sensitivity or benchmarking analyses were classified according to their primary implementation.

When studies employed multiple site effect handling strategies (for example, applying data harmonization prior to model fitting and including site as a covariate within the normative model), a single-category assignment was required for the field-level summary (Supplementary Fig. 1). Classification was determined by the primary normative model from which deviation scores were derived for the main analysis, applying the operational hierarchy Modeled > Harmonized > None > N/A to that primary implementation. When the primary model contained site as a parameter, the

study was classified as Modeled regardless of whether additional pre-fitting harmonization was also applied. When the primary model did not contain site as a parameter but pre-fitting harmonization was applied to input data, the study was classified as Harmonized. Site-inclusive models evaluated only as sensitivity or benchmarking analyses did not alter the primary classification. This hierarchy was adopted for classification purposes only and does not imply methodological superiority of any approach; the appropriate strategy depends on study-specific design considerations, including sample size, number of sites, and the flexibility of the normative model.

Supplementary Note 4: NORMA Framework Definitions and Scoring Guidance

Development and structure

No existing quality assessment tool addresses the methodological requirements specific to normative modeling. We adapted criteria from the NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies[1], complemented by relevant items from the CLAIM checklist (Checklist for Artificial Intelligence in Medical Imaging)[2] for modeling-specific aspects such as data partitioning, reproducibility, and performance reporting. Both instruments were specified in our registered protocol. One item (CA-6, deviation score interpretation) was developed for this review and has no counterpart in either source tool, since neither addresses the interpretation of individual-level deviation scores from a normative reference distribution. The resulting 18-item NORMA checklist is organized into three domains: common elements (CE, 1 item), model development and validation (NM, 11 items covering reference cohort, modeling approach, validation, and reproducibility), and clinical application (CA, 6 items covering clinical cohort characterization and deviation interpretation). NORMA was developed during data extraction by consolidating and reorganizing these source tools into a single structured checklist adapted for normative modeling. NORMA itself was not pre-registered and was not developed through a formal consensus process. We present it as a starting point for structured discussion of reporting practices in normative modeling, not as a finalized standard.

Source items: mapping to NIH and CLAIM

The 17 adapted NORMA items draw from 5 items in the NIH Quality Assessment Tool and 14 items in the CLAIM checklist. The number of source items used (19) exceeds the number of adapted NORMA items (17) for two reasons. Some NORMA items consolidate content addressed by overlapping source items. Some content applies separately to the reference cohort and to the clinical cohort and is captured by paired NORMA items that draw on the same source. For example, NM-1 (reference cohort selection) and CA-1 (clinical cohort selection) together draw on NIH items 2 and 4 and CLAIM items 8 and 35, all of which address eligibility, exclusion, and sample-size reporting. NIH item 14 (confounders) is the source for both NM-2 and CA-2, because covariate handling is reported once for the reference cohort and once for the clinical cohort.

NIH source items used: item 1 (research objectives), item 2 (study population), item 4 (inclusion and exclusion), item 11 (outcome measures), item 14 (confounders).

CLAIM source items used: item 7 (data sources), item 8 (inclusion and exclusion criteria), item 9 (data pre-processing), item 13 (image acquisition protocol), item 19 (how data were assigned to partitions), item 22 (detailed description of model), item 25 (details of training approach), item 28 (metrics of model performance), item 32 (evaluation on internal data), item 33 (testing on external data), item 35 (numbers of patients or examinations included and excluded), item 36 (demographic and clinical characteristics of cases in each partition), item 37 (performance metrics and measures of statistical uncertainty), item 43 (statement about the availability of software, trained model, and data).

NORMA item	Source
CE-1	NIH item 1
NM-1	NIH items 2, 4; CLAIM items 8, 35
NM-2	NIH item 14
NM-3	CLAIM item 7
NM-4	CLAIM item 13

NM-5	CLAIM item 9
NM-6	CLAIM items 19, 32
NM-7	CLAIM item 33
NM-8	CLAIM items 22, 25, 43
NM-9	CLAIM items 28, 37
NM-10	CLAIM item 36
NM-11	CLAIM items 22, 43
CA-1	NIH items 2, 4; CLAIM items 8, 35
CA-2	NIH item 14
CA-3	CLAIM item 7
CA-4	NIH item 2; CLAIM items 35, 36
CA-5	NIH item 11
CA-6	Author-developed

Scoring rules

Each item was scored as Y (fully reported), P (partially reported), N (not reported or unclear), or NA (not applicable). For items with multiple sub-criteria, Y was assigned when all sub-criteria were met, N when none were met, and P for intermediate cases. NA was assigned in two situations. First, the item is conditional on a study characteristic that does not apply; for example, NM-4 (image acquisition protocol) is NA when only secondary data are used. Second, the item is logically contingent on a prior NA; for example, NM-10 (partition characteristics) is NA when NM-6 indicates that no data partitioning was performed.

Several items branch on two operational distinctions defined in the Methods. *New normative model* versus *pre-trained normative model* refers to whether the study estimated model parameters on a reference cohort in the present study or applied previously published parameters without re-estimation. *Primary data collection* versus *secondary data* refers to whether the imaging data were newly acquired by the study team or sourced from an existing dataset. This distinction applies separately to the reference cohort and the clinical cohort.

Item definitions and assessment criteria

CE. Common elements

CE-1. Research objectives. Are the study objectives stated?

Y requires:

- Research hypotheses or questions are stated.
- The purpose of applying normative modeling is described (e.g., diagnostic support, personalized medicine, characterization of individual heterogeneity).

NM. Normative modeling

NM-1. Reference cohort selection. Are eligibility criteria for the reference cohort and image quality standards described?

For studies building a new normative model, Y requires:

- Eligibility criteria for the reference cohort are described. For open datasets, the dataset paper or official page is cited. For newly collected data, inclusion and exclusion criteria are provided.
- Image quality and data completeness standards are specified (e.g., motion artifacts, missing data).
- Final reference-cohort sample size and reasons for exclusion are reported.

For studies using a pre-trained normative model, Y requires:

- The original publication describing the reference cohort is cited.
- If the model is recalibrated or adapted (e.g., transfer learning, local re-fitting), eligibility criteria for the local sample are described.
- Image quality and data completeness standards are specified (e.g., motion artifacts, missing data).
- Final sample size and reasons for exclusion are reported.

Scoring guidance. The unit of evaluation is the reference cohort, the healthy population on which the normative model is built. This must not be conflated with calibration controls used to adapt a pre-trained model, or with healthy or patient subgroups inside the clinical cohort. For dataset citation, a DOI or official URL in the body or Supplement is sufficient; a generic URL listed only in the Acknowledgements is not. For image quality, both the exclusion target and the verification procedure must be identifiable; phrasing limited to "subjects passing quality control were included" is insufficient. For sample size, the final reference-cohort N together with exclusion criteria (e.g., Euler > 5, motion threshold, FreeSurfer failure) is sufficient; per-reason breakdowns or pre-exclusion totals are not required.

NM-2. Reference cohort covariate handling. Are covariates handled appropriately when building the normative model?

Y requires:

- Demographic covariates such as age and sex are accounted for.
- For multi-site reference data, the method used to address site effects is specified (e.g., site as a covariate, ComBat-family harmonization, hierarchical modeling).

Scoring guidance. The site-handling method must be identifiable. Phrasing such as "accounting for site differences" without specifying the method is not sufficient.

NM-3. Reference cohort data sources. Are the neuroimaging data sources of the reference cohort described?

For studies building a new normative model, Y requires:

- The data source is specified. For open datasets, the database or study name is provided. For newly collected data, collection sites and time period are provided.

For studies using a pre-trained normative model, Y requires:

- The data source is specified (e.g., dataset or study name, citation of the original publication).

NM-4. Image acquisition protocol. Is the image acquisition protocol described?

For primary data collection, Y requires:

- Imaging modality is specified, including subtype where applicable (e.g., MRI: T1-weighted, diffusion, functional).
- Acquisition parameters are specified (e.g., TR, TE, voxel resolution).
- Equipment specifications are provided (e.g., manufacturer, field strength).

For secondary data, NA.

NM-5. Data preprocessing. Are the preprocessing procedures described?

Preprocessing here refers to all steps used to derive analysis-ready neuroimaging features, including image preprocessing pipelines, quality control and exclusion, outlier handling, feature extraction and aggregation, transformations, and missing-data handling.

For studies that performed preprocessing, Y requires:

- Software or tools are specified (e.g., FreeSurfer, FSL, R or Python).
- Preprocessing steps or the standard pipeline are identified (e.g., recon-all, fMRIPrep).
- Quality control procedures are described.

For studies using preprocessed derivatives from existing projects with no further processing by the study team, NA.

Scoring guidance. The scope of NM-5 is the entire derivation of analysis-ready features, not the image pipeline alone. A check applies when the main tool can be identified; exhaustive listing is not required. For QC procedures, the rule must be identifiable (metric, threshold, manual or automated procedure); "as described previously" without criteria in the present paper is insufficient. NA is restricted to cases in which derived features (e.g., precomputed ROI summaries) are used directly without any additional processing. Such cases are uncommon in practice.

NM-6. Internal validation of the reference cohort. Is a data partitioning strategy used for internal validation?

For studies building a new normative model, Y requires:

- A partitioning strategy is used in which the model is evaluated on samples not used to fit it in that iteration (e.g., hold-out test set, k-fold cross-validation, leave-one-out).

For studies using a pre-trained normative model, NA.

NM-7. External validation of the reference cohort. Is the normative model evaluated on an independent external dataset of healthy individuals?

For studies building a new normative model, Y requires:

- The model is evaluated on an independent external dataset of healthy individuals not used during model development.

For studies using a pre-trained normative model, NA.

Scoring guidance. The evaluation must be on a fully independent healthy cohort. Validation in the clinical cohort does not satisfy NM-7, even when the clinical cohort contains a healthy comparison arm. A check applies when deviation behavior or generalization is assessed in an independent

external healthy cohort. As a worked example, one study[3] applied a normative model trained on healthy reference data to an independent external healthy transfer set for validation.

NM-8. Statistical approach to normative modeling. Is the statistical approach described?

For studies building a new normative model, Y requires:
- The model class is specified (e.g., linear regression, Gaussian process regression, hierarchical Bayesian regression, distributional regression).
- Key model settings are reported.
- Software and libraries used for implementation are specified.

For studies using a pre-trained normative model, NA.

Scoring guidance. Key model settings refer to specific values or types that govern model behavior, stated in the body or Supplement. Examples include GPR kernel type; LOESS or GAM smoothing parameter or degrees of freedom; BLR warping family; Bayesian priors; regularization parameters; or analogous specifications for other model classes. The model name alone, or a general description without identifiable settings, is not sufficient.

NM-9. Model performance. Is model performance quantitatively assessed for the reference cohort?

For studies building a new normative model, Y requires:
- At least one quantitative performance metric is reported:
  - Central tendency (accuracy of the predicted mean; e.g., explained variance or R^2, mean absolute error (MAE), mean squared error (MSE) or root mean squared error (RMSE), standardized mean squared error (SMSE), or the correlation between predicted and observed values).
  - Overall goodness-of-fit (e.g., mean standardized log-loss (MSLL) or log-score, or the Akaike or Bayesian information criterion (AIC or BIC)).
  - Calibration of the modeled distribution (e.g., skewness or kurtosis of the deviation scores, a Shapiro-Wilk test on the residuals, a quantile-quantile (Q-Q) or calibration plot, mean absolute centile error, or coverage of prediction intervals).

For studies using a pre-trained normative model, NA.

Scoring guidance. A single metric of any kind satisfies the item. Central-tendency and overall-fit metrics quantify how well the model predicts on average, whereas calibration metrics quantify whether the modeled distribution, and thus the centiles used to place an individual, fits the data[3-5]. A normality or distribution test counts as calibration only when applied to the model's residuals or deviation scores against a normal distribution, not when comparing cohorts or checking group-comparison assumptions.

NM-10. Reference cohort partition characteristics. Are characteristics of the reference cohort described for each data partition?

For studies building a new normative model with data partitioning, Y requires:
- Sample size for each partition (e.g., training, validation, test) is reported.
- Demographic characteristics for each partition (e.g., age, sex) are described.

For studies using a pre-trained normative model, or for studies in which no partitioning is performed, NA.

Scoring guidance. When NM-6 is judged as no partitioning, NM-10 is automatically NA, not blank or N. When partitioning is present and partitions are fixed (hold-out, split-half, or named training/validation/test sets), N and demographic features for each partition are required for Y. k-fold cross-validation alone, without a fixed independent test set, is NA: the folds are an internal model-selection procedure, and independent demographic characterization of iterative folds is not conventionally reported.

NM-11. Reproducibility resources. Is sufficient information provided to support reproducibility?

For studies building a new normative model, Y requires:

- Software and library versions are specified (e.g., Python 3.8, PCNtoolkit v0.30, R 4.2.0).
- The availability of reproducibility resources (code, trained models, or data) is stated, including access conditions and restrictions.

For studies using a pre-trained normative model, Y requires:

- The source of the normative model is provided (citation, repository URL, or version identifier).

Scoring guidance. For software versions, a check applies when at least one major tool used in normative model construction (e.g., imaging pipeline, analysis environment, modeling toolkit) is specified with a version. Exhaustive listing is not required. For availability, the criterion is whether accessibility is stated, not whether resources are publicly accessible. A check applies when code, trained models, or data are described as open, restricted (e.g., available on request, with an authorization process), or not available with a stated reason.

CA. Clinical application

CA-1. Clinical cohort selection. Are eligibility criteria for the clinical cohort and image quality standards described?

For primary data collection, Y requires:

- Patient eligibility and diagnostic criteria are described (e.g., DSM-5, ICD-10, structured clinical interview).
- Image quality and data completeness standards are specified (e.g., motion artifacts, missing data).
- Final sample size and reasons for exclusion are reported.

For secondary data, Y requires:

- The original publication is cited, and any additional eligibility criteria applied in the present study are described.
- Image quality and data completeness standards are specified (e.g., motion artifacts, missing data).
- Final sample size and reasons for exclusion are reported.

CA-2. Clinical covariate handling. Are clinical covariates accounted for in analyses involving the

clinical cohort?

Y requires:

- Clinical covariates beyond age and sex are addressed in analyses of deviation scores (e.g., medication, illness duration, symptom severity, comorbidities).

Scoring guidance. The check applies when clinical covariates are addressed through adjustment, stratification, or sensitivity analysis. Reporting clinical indices descriptively, or showing correlations between deviation scores and clinical features without adjustment or stratification, does not satisfy the criterion.

CA-3. Clinical cohort data sources. Are the neuroimaging data sources of the clinical cohort described?

For primary data collection, Y requires:

- The data source is specified (e.g., collection sites, time period).

For secondary data, Y requires:

- The data source is specified (e.g., dataset or study name, citation of the original publication).

CA-4. Clinical cohort characteristics. Are the characteristics of the clinical cohort described?

Y requires:

- Sample sizes are reported.
- Demographic characteristics are described (e.g., age, sex).
- Clinical characteristics beyond demographics are described (e.g., medication status, symptom severity, illness duration).

Scoring guidance. Clinical characteristics require both an indicator (scale name or definition) and a summary statistic (e.g., mean ± SD, median [IQR], proportion). The illustrative examples are not mandatory; standardized indices for symptom severity, function, cognition, or behavioral traits (e.g., ADOS) are also acceptable.

CA-5. Clinical assessment measures. Are the clinical assessment measures described?

Y requires:

- Clinical measures are specified (e.g., symptom rating scales, diagnostic criteria with version).
- Assessment procedures are described (e.g., structured interview, self-report, clinician-rated).

Scoring guidance. For clinical measures, a check applies when the indicator is identifiable; symptom-scale names or named diagnostic criteria are sufficient (version is preferred but not mandatory). Diagnostic category labels alone (e.g., ASD, SCZ, MCI) are not sufficient. For assessment procedures, the method of measurement or evaluation must be stated.

CA-6. Deviation score interpretation. Is the interpretation of deviation scores and their clinical significance described?

Y requires:

- The meaning and direction of deviation scores are explained (i.e., what positive and negative values represent, how the magnitude should be interpreted).
- When deviation scores are used to define categories (e.g., extreme, atypical), quantitative thresholds are reported (e.g., $|Z| > 1.96$, outside the 5th to 95th percentile).
- Clinical implications of deviation patterns are discussed (e.g., relationships with symptoms or diagnosis, contrast with group-level findings, limitations of interpretation).

Supplementary References

1. National Heart, Lung, and Blood Institute. Study Quality Assessment Tools. https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools (2021).
2. 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). https://doi.org/10.1148/ryai.2020200029
3. Rutherford, S. *et al.* Charting brain growth and aging at high spatial precision. *eLife* **11**, e72904 (2022). https://doi.org/10.7554/eLife.72904
4. Dinga, R. *et al*. Normative modeling of neuroimaging data using generalized additive models of location, scale and shape. Preprint at https://doi.org/10.1101/2021.06.14.448106 (2021).
5. Marquand, A. F., Rutherford, S. & Dinga, R. Fairly evaluating the performance of normative models. *Lancet Digit. Health* **6**, e775 (2024). https://doi.org/10.1016/S2589-7500(24)00200-0

Supplementary Tables

Supplementary Table 1: Studies excluded at full-text screening with reasons

Of 208 full-text articles assessed for eligibility, 86 were excluded. Reasons correspond to the exclusion categories shown in Fig. 2 (PRISMA flow diagram). Studies are grouped by reason and listed alphabetically within each group. A study could meet more than one exclusion criterion.

Study	Reasons for exclusion	DOI
Antoniades 2021	Not population-level	https://doi.org/10.1093/schbul/sbab005
Brouwer 2022	Not population-level	https://doi.org/10.1038/s41593-022-01042-4
Cao 2017	Not population-level	https://doi.org/10.1038/s41598-017-00582-1
Chen 2024	Not population-level	https://doi.org/10.1101/2024.03.02.583073
Gordon 2010	Not population-level	https://doi.org/10.1177/155005941004100404
Gugger 2023	Not population-level	https://doi.org/10.1016/j.nicl.2023.103392
Keehn 2015	Not population-level	https://doi.org/10.1002/aur.1438
Le Scouarnec 2024	Not population-level	https://doi.org/10.1186/s13195-024-01595-5
Lefebvre 2018	Not population-level	https://doi.org/10.3389/fnins.2018.00662
Luo 2025	Not population-level	https://doi.org/10.1016/j.jad.2025.01.025
Siffredi 2021	Not population-level	https://doi.org/10.1101/2021.11.02.21265740
Snyder 2021	Not population-level	https://doi.org/10.1016/j.nicl.2021.102565
Su 2024	Not population-level	https://doi.org/10.3389/fnins.2024.1334508
Yang 2020	Not population-level	https://doi.org/10.1016/j.neuroimage.2019.116227
Yang 2024	Not population-level	https://doi.org/10.1038/s41591-024-03144-x
Zeng 2016	Not population-level	https://doi.org/10.1109/TMI.2016.2538998
Coupé 2019	No individual deviation analysis	https://doi.org/10.1038/s41598-019-39809-8
Coupé 2023	No individual deviation analysis	https://doi.org/10.1002/hbm.26464
Courchesne 2011	No individual deviation analysis	https://doi.org/10.1016/j.brainres.2010.09.101
de Moraes 2022	No individual deviation analysis	https://doi.org/10.3389/fnins.2022.897226
Duan 2024	No individual deviation analysis	https://doi.org/10.7554/eLife.94970
Feldman 2025	No individual deviation analysis	https://doi.org/10.1186/s13229-025-00657-1
Guo 2024	No individual deviation analysis	https://doi.org/10.1038/s41467-024-53550-5
Hu 2024	No individual deviation analysis	https://doi.org/10.1101/2024.08.29.610247
Huberty 2021	No individual deviation analysis	https://doi.org/10.1002/aur.2518
Huizinga 2018	No individual deviation analysis	https://doi.org/10.1016/j.neuroimage.2017.10.040
Neale 2024	No individual deviation analysis	https://doi.org/10.1017/S0033291724002599
Schabdach 2023	No individual deviation	https://doi.org/10.1148/radiol.230096

	analysis	
Srivastava 2012	No individual deviation analysis	https://doi.org/10.1002/hbm.21206
Truelove-Hill 2020	No individual deviation analysis	https://doi.org/10.1523/JNEUROSCI.2092-19.2019
Wang 2021	No individual deviation analysis	https://doi.org/10.1016/j.pnpbp.2020.110044
Wang 2024	No individual deviation analysis	https://doi.org/10.1038/s41398-024-02803-4
Bashyam 2020	Brain age gap only	https://doi.org/10.1093/brain/awaa160
Chen 2022	Brain age gap only	https://doi.org/10.1016/j.nicl.2022.103003
Cropley 2021	Brain age gap only	https://doi.org/10.1016/j.bpsc.2020.07.014
Eavani 2018	Brain age gap only	https://doi.org/10.1016/j.neurobiolaging.2018.06.013
Ecker 2017	Brain age gap only	https://doi.org/10.1001/jamapsychiatry.2016.3990
Fjell 2014	Brain age gap only	https://doi.org/10.1093/cercor/bhs379
Gutierrez Becker 2018	Brain age gap only	https://doi.org/10.1016/j.neuroimage.2018.03.075
Habes 2021	Brain age gap only	https://doi.org/10.1002/alz.12178
Huang 2022	Brain age gap only	https://doi.org/10.3389/fnagi.2022.823502
Kelly 2022	Brain age gap only	https://doi.org/10.1016/j.neuroimage.2021.118828
Seitz 2021	Brain age gap only	https://doi.org/10.1093/cercor/bhaa220
Sun 2024	Brain age gap only	https://doi.org/10.1038/s41467-024-54657-5
Tian 2023	Brain age gap only	https://doi.org/10.1038/s41591-023-02296-6
Tunç 2019	Brain age gap only	https://doi.org/10.1186/s13229-019-0301-5
Wang 2021	Brain age gap only	https://doi.org/10.1002/hbm.25660
Zabihi 2021	Brain age gap only	https://doi.org/10.1101/2021.03.10.434856
Zhu 2023	Brain age gap only	https://doi.org/10.1038/s41398-023-02379-5
Verdi 2021	Review/other study types	https://doi.org/10.1093/brain/awab165
Walhovd 2018	Review/other study types	https://doi.org/10.1016/j.eurpsy.2017.12.006
van Nederpelt 2025	Review/other study types	https://doi.org/10.1212/WNL.0000000000213618
AlHaddad 2022	No clinical population	https://doi.org/10.1101/2022.06.02.493125
Drakulich 2021	No clinical population	https://doi.org/10.1016/j.neuroimage.2021.117974
Eisma 2024	No clinical population	https://doi.org/10.1186/s12987-024-00525-9
Fawaz 2023	No clinical population	https://doi.org/10.1101/2023.10.16.562598
Fjell 2020	No clinical population	https://doi.org/10.1093/sleep/zsz280
Jeon 2024	No clinical population	https://doi.org/10.3390/brainsci14040401
Kwon 2022	No clinical population	https://doi.org/10.1101/2022.03.31.486476
Lawrence 2021	No clinical population	https://doi.org/10.1007/s11682-021-00548-y
Lotter 2023	No clinical population	https://doi.org/10.1101/2023.05.05.539537
Nomi 2017	No clinical population	https://doi.org/10.1523/JNEUROSCI.3408-16.2017
Peterson 2018	No clinical population	https://doi.org/10.3171/2017.10.PEDS17141
Qian 2025	No clinical population	https://doi.org/10.1007/s00259-025-07211-4
Reynolds 2019	No clinical population	https://doi.org/10.1016/j.neuroimage.2019.04.004

Robinson 2024	No clinical population	https://doi.org/10.1162/imag_a_00247
Romero 2021	No clinical population	https://doi.org/10.1002/hbm.25293
Schilling 2023	No clinical population	https://doi.org/10.1101/2023.09.25.559330
Sun 2023	No clinical population	https://doi.org/10.1101/2023.09.12.557193
Yu 2024	No clinical population	https://doi.org/10.1007/s11357-024-01220-1
Bick 2015	No neuropsychiatric application	https://doi.org/10.1001/jamapediatrics.2014.3212
Blok 2022	No neuropsychiatric application	https://doi.org/10.3389/fpsyt.2022.846201
Corrigan 2024	No neuropsychiatric application	https://doi.org/10.1073/pnas.2403200121
Cromb 2024	No neuropsychiatric application	https://doi.org/10.1093/braincomms/fcae356
Deng 2024	No neuropsychiatric application	https://doi.org/10.1101/2024.11.29.626043
Fraza 2022	No neuropsychiatric application	https://doi.org/10.1101/2022.08.23.505049
Gaiser 2024	No neuropsychiatric application	https://doi.org/10.1038/s41467-024-46398-2
Ho 2025	No neuropsychiatric application	https://doi.org/10.1101/2025.01.05.631276
Kjelkenes 2022	No neuropsychiatric application	https://doi.org/10.1016/j.dcn.2022.101173
Mandal 2024	No neuropsychiatric application	https://doi.org/10.1101/2024.09.24.24314251
Marquand 2016	No neuropsychiatric application	https://doi.org/10.1016/j.biopsych.2015.12.023
Nárai 2024	No neuropsychiatric application	https://doi.org/10.1007/s10802-024-01190-0
Thompson 2021	No neuropsychiatric application	https://doi.org/10.1016/j.clinph.2021.05.025
Wang 2022	No neuropsychiatric application	https://doi.org/10.1016/j.bbi.2021.12.008
Wong 2023	No neuropsychiatric application	https://doi.org/10.1038/s41380-022-01908-w
de Boer 2024	No neuropsychiatric application	https://doi.org/10.1162/imag_a_00132

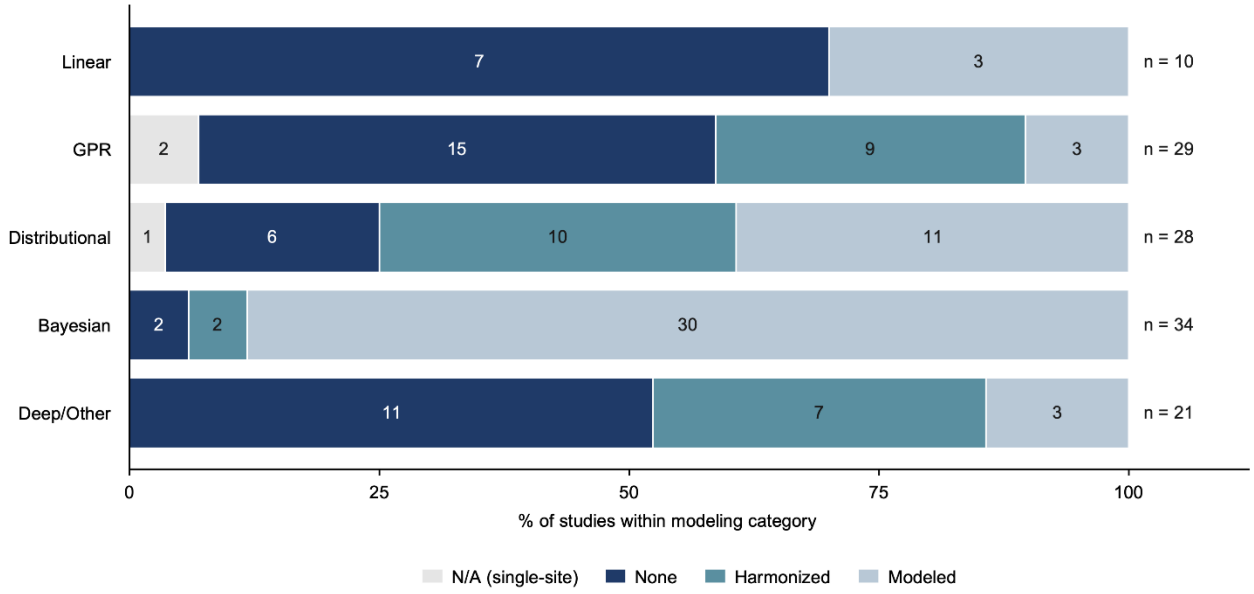
Reason categories: Not population-level, normative models constructed from within-subject longitudinal data or individual-level distributions rather than from a population-level reference cohort. No individual deviation, normative models built without computing or reporting individual-level deviation scores. Brain age gap only, brain-based regression to predict chronological age without deriving deviation metrics from normative trajectories of specific brain measures. Review/other, review articles, methodological concept papers, or project descriptions without population-level normative analysis. No clinical population, normative modeling applied only to healthy participants, with no clinical comparison group. No neuropsychiatric, populations without a clinically diagnosed psychiatric or neurological disorder, including healthy development, environmental exposures, non-neuropsychiatric medical conditions, subclinical symptoms, or methodological demonstrations.

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Main (¶1–3)
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Main (¶5)
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods (Search strategy and study selection)
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods (Search strategy and study selection)
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Methods (Search strategy and study selection); Supplementary Note 1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods (Search strategy and Study Selection)
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods (Data extraction)
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Methods (Reporting transparency assessment); Supplementary Data 1–3
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Methods (Data extraction); Supplementary Data 1, 2
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods (Reporting transparency assessment); Table 1; Supplementary Note 4
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	N/A
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods (Data extraction; Synthesis)
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods (Data extraction; Synthesis)
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods (Synthesis); Tables 2;

Section and Topic	Item #	Checklist item	Location where item is reported
			Figures 3–6
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	N/A
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods (Synthesis); Results (Methodological characteristics; Site effect handling; Reference cohorts; Validation strategies)
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	N/A
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods (Reporting transparency assessment; Synthesis)
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	N/A - narrative synthesis (not meta-analytic)
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results (Methodological characteristics, ¶1); Figure 2
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Supplementary Table 1
Study characteristics	17	Cite each included study and present its characteristics.	Results (Methodological characteristics; Statistical approaches; Site effect handling); Figure 3; Supplementary Data 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Results (Reporting transparency); Table 2; Figure 5; Figure 6; Supplementary Data 3
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	N/A
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results (Methodological characteristics; Reporting transparency); Tables 2, Figure 3
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Results (throughout); Tables 2; Figures 3–6
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results (Methodological

Section and Topic	Item #	Checklist item	Location where item is reported
			characteristics; Site effect handling; Reference cohorts; Validation strategies; Clinical applications); Figure 4; Supplementary Fig. 1
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	N/A
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	N/A
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	N/A
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion (¶1–5)
	23b	Discuss any limitations of the evidence included in the review.	Discussion (¶7)
	23c	Discuss any limitations of the review processes used.	Discussion (¶7)
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion (¶6, 8)
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Methods (Protocol and reporting standards)
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	PROSPERO (CRD420251015671)
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Methods (Protocol and reporting standards; Data extraction; Synthesis)
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Funding
Competing interests	26	Declare any competing interests of review authors.	Competing interests
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability; Supplementary Data 1–4; Supplementary Note 4

Supplementary Figures



Supplementary Fig. 1: Site handling distribution by modeling family. Strategy distribution within each modeling category (n = 122 studies). Segment widths represent within-category percentages. Counts are shown within segments; row totals appear to the right. Strategies: N/A (single-site); None (multi-site without correction); Harmonized (input data harmonization prior to model fitting); Modeled (site as a parameter within the normative model). Classification rules are detailed in Supplementary Note 3. GPR, Gaussian process regression.