

Corresponding author(s): Julio E. Villalón-Reina, Paul M. Thompson

Last updated by author(s): 4/3/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No custom software was used to acquire the data (MRI scans of human brains).

Data analysis The code for analysis can be found here: <https://github.com/villalonreina/ENIGMA-DTI-Normative-Modeling>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The estimated normative models' data generated in this study have been deposited in the PCNportal5 database under accession code [<https://pcnportal.dccn.nl/>]. Source data are provided with this paper.

The data generated in this study, such as source data files containing summary statistics underlying each figure (e.g., group means, standard errors, regression coefficients, model outputs, and aggregated metrics) are provided in the Supplementary Information and Source Data files.

We fully support the journal's commitment to transparency and reproducibility. However, several datasets used in this study (ABCD, UK Biobank, ADNI) are governed by Data Use Agreements that prohibit redistribution of individual participant-level data, including publication as supplementary .csv files and/or sharing on online repositories (e.g. Github).

Researchers wishing to reproduce analyses using the underlying individual-level data may obtain access directly through the respective data access procedures:

- ABCD Study: <https://nda.nih.gov/abcd>
- UK Biobank: <https://www.ukbiobank.ac.uk/enable-your-research>
- ADNI: <https://adni.loni.usc.edu/data-samples/access-data/>

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

We used biological sex as a covariate in the regression models for training the normative models. This information was collected by each of the studies that made the data available (HCP, UKBB, ABCD, etc.). A list of all the datasets included and their biological sex can be found in the Methods section. References to each study are all cited.

Reporting on race, ethnicity, or other socially relevant groupings

We did not use race, ethnicity or socioeconomic status in our statistical analyses.

Population characteristics

We used age as a covariate in the regression models when training and testing the normative models. We used datasets that are population based, such as ABCD and the UKBB, and datasets that had participants with brain disease and healthy controls. For the training phase we used only the healthy controls for the latter datasets. In the testing phase we used participants diagnosed with Mild Cognitive Impairment, Alzheimer's Disease and 22q11.2 Deletion Syndrome.

Recruitment

Participants were recruited by other studies, initiatives and projects (i.e., UKBB, HCP, ABCD, ADNI, etc) that have made their data available for public use under custom data user agreements.

Ethics oversight

Each of the studies, initiatives and projects (i.e., UKBB, HCP, ABCD, ADNI, etc.) have done ethics oversight and sought approval in their corresponding institutions.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Suppose we have a real-valued random variable $N(t)$ on the interval $t = 0$ to 100 , and N is distributed with mean $\mu(t)$ and variance $s(t)$. To determine the smallest number of training samples (K) needed to bound the Kolmogorov-Smirnov (K-S) statistic of $N(t)$ to be no greater than 0.01 everywhere, with 95% confidence, we can use the properties of the K-S test.

The K-S statistic (D_n) for a sample size (n) is given by:

$$D_n = \sup_t |F_n(t) - F(t)|$$

where $F_n(t)$ is the empirical cumulative distribution function (ECDF) and $F(t)$ is the true cumulative distribution function (CDF).

For a 95% confidence level, the critical value (d) for the K-S statistic can be approximated using:

$$d = \frac{1.36}{\sqrt{n}}$$

To ensure that the K-S statistic is no greater than 0.01 , we set ($D_n \leq 0.01$). Therefore, we need:

$$\frac{1.36}{\sqrt{n}} \leq 0.01$$

Solving for (n):

$$\sqrt{n} \geq \frac{1.36}{0.01}$$

$$\sqrt{n} \geq 136$$

$$n \geq 136^2$$

$$n \geq 18496$$

Thus, the smallest number of training samples (K) needed is approximately 18,496 to bound the K-S statistic to be no greater than 0.01 everywhere, with 95% confidence. This calculation is independent of the specific forms of $(\mu(t))$ and $(\sigma(t))$ because the K-S test is non-parametric and does not rely on the underlying distribution's parameters. However, it assumes that the distribution of $(N(t))$ is continuous and that the sample size is sufficiently large to approximate the distribution accurately.

Additionally, in our estimated percentiles, we have a non-constant (stationary) variance that depends, to some extent, on age. Thus, if we want to enforce an upper bound on the fitted percentile error everywhere, we need more samples. This was recently tested in a paper by Bozek et al., (<https://doi.org/10.1016/j.neuroimage.2023.119864>) who concluded that increasing the sample size up to N=50,000 in these conditions would increase the accuracy of the percentile estimation.

Data exclusions	All data exclusions were pre-established. For instance, we did manual quality control of the brain scans for part of the data and excluded scans with bad quality. For the other part of the data we used an anomaly detection algorithm to identify potential outliers that were further checked manually. For some datasets, explicitly, for PPMI, ADNI, HBN and OASIS3, the exclusion criteria were pre-established, meaning that only healthy controls were included in the training of the normative models. All cases in those studies were excluded from training. for testing we used the controls for ADNI and OASIS3. The same logic applied for the UCLA and the NIMHANS datasets.
Replication	We ran the training and the testing of the normative models 10 times, each time splitting the entire dataset randomly by 80% for training and 20% testing. This is similar to cross-validation. Additionally we used data from 19 different datasets from different countries in order to ensure the reproducibility across diverse datasets. Also, as part of the review process we ran the same experiments with fully manually QCed data, with data harmonized with ComBat and in all cases the results were identical, as shown in the responses to the reviewers. Across experimental repetitions and across data settings (manually QCed, ComBat harmonized) the replications were successful.
Randomization	We added covariates age and sex for all regression analyses. We did randomize the train test splits (see above and Methods section) for each of the 10 experimental runs. We could not randomize between cases and controls because of the design of the study that inherently assumes that the diagnosis of the participants must be known beforehand.
Blinding	Blinding was not relevant for this study. For the normative modeling training we used datasets of general population (UKBB, ABCD, etc) and the healthy control subjects of clinical datasets (ADNI, PPMI, OASIS), for which we needed to know the diagnosis of the participants. The testing of the normative models on patients with Alzheimer's disease or 22q deletion syndrome also needed the information of the participants' diagnosis "a priori".

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Clinical data
☐	☐ Dual use research of concern
☐	☐ Plants

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See ICLAC register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

Palaeontology and Archaeology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>
☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	<i>For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.</i>
Wild animals	<i>Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.</i>
Reporting on sex	<i>Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.</i>
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>
Outcomes	<i>Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.</i>

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No Yes

- ☒ ☐ Public health
- ☒ ☐ National security
- ☒ ☐ Crops and/or livestock
- ☒ ☐ Ecosystems
- ☒ ☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No Yes

- ☒ ☐ Demonstrate how to render a vaccine ineffective
- ☒ ☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
- ☒ ☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
- ☒ ☐ Increase transmissibility of a pathogen
- ☒ ☐ Alter the host range of a pathogen
- ☒ ☐ Enable evasion of diagnostic/detection modalities
- ☒ ☐ Enable the weaponization of a biological agent or toxin
- ☒ ☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Since the study is not an fMRI one, this does not apply.

Design specifications

Since the study is not an fMRI one, this does not apply.

Behavioral performance measures

Since the study is not an fMRI one, this does not apply.

Acquisition

Imaging type(s)

Diffusion MRI

Field strength

1.5 Tesla, 3 Tesla, 4 Tesla

Sequence & imaging parameters

Data from multiple sites/studies were used. All parameters are available in supplementary Table S1.

Area of acquisition

Whole brain scans for all subjects.

Diffusion MRI



Used



Not used

Parameters Data from multiple sites were used. All parameters are available in supplementary Table S1.

Preprocessing

Preprocessing software

Details are described in the Methods section, subsection 2, and in the Supplementary Methods Table 1. In general we used three software tools: FSL (<https://fsl.fmrib.ox.ac.uk/fsl/docs/#/>), DIPY (<https://dipy.org/>) and MrTrix (<https://www.mrtrix.org/>). Tortoise (<https://tortoise.nibib.nih.gov/tortoise>) was also used in a few instances. Some datasets were preprocessed by the laboratories that acquired and released the MRI scans. An example is the ABCD dataset for which one laboratory is in charge of preprocessing the entire dataset and releasing it to the public (<https://abcdstudy.org/>). In those cases we refer the reader to the corresponding manuscript describing the preprocessing.

Normalization

All data was registered to the ENIGMA-DTI template linearly and non-linearly using ANTs (<https://stnava.github.io/ANTs>)

Normalization template

The ENIGMA-DTI template, available at <https://enigma.ini.usc.edu/protocols/dti-protocols/>

Noise and artifact removal

We used denoising algorithms (LPCA, MPPCA, Patch2Self) available in DIPY, FSL and/or MrTrix. Gibbs removal algorithms

were also used, available in DIPY an MrTrix.

Volume censoring

Subjects with more than 30% of the volumes with uncorrected artifacts were removed from the study. This was performed for the datasets that were manually QCed.

Statistical modeling & inference

Model type and settings

Hierarchical Bayesian Regression. Hyperpriors and priors are set to follow a Gaussian distribution. Specific parameters are presented in Kia et al., 2022 (<https://doi.org/10.1371/journal.pone.0278776>).

Effect(s) tested

This does not apply to our study. We extracted Z-scores as a metric of deviation from the reference model.

Specify type of analysis: ☐ Whole brain ☒ ROI-based ☐ Both

Anatomical location(s)

Automated analysis of the white matter following the ENIGMA-DTI protocol: <https://enigma.ini.usc.edu/protocols/dti-protocols/>. This protocol extracts average white matter DTI metrics from the deep white matter.

Statistic type for inference

(See [Eklund et al. 2016](#))

Normative modeling involves predictive inference and probabilistic regression. With hierarchical Bayesian regression we created reference models against which individual data points were compared. With predictive inference we mapped the variations across large diffusion MRI datasets to identify typical patterns, establish deviations and make predictions at the individual level.

Correction

We used the False Discovery Rate (FDR) to correct for multiple comparisons when making inferences about individual deviations from the reference models (see Methods section)

Models & analysis

n/a | Involved in the study

- ☐ ☐ Functional and/or effective connectivity
☐ ☐ Graph analysis
☐ ☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Independent variables: age, sex and scanner protocol.