https://doi.org/10.1038/s41562-025-02186-6

Diverse and large-scale brain data in child development research

Liang Zhang & Gui Xue

Check for updates

Postnatal brain development is important for individual and societal outcomes. We need large-scale cohort studies from diverse populations to generate generalizable insights into the factors that affect children's brain development. Here we discuss the contribution of the Chinese Child Brain Development project.

CREDIT: KLAUS VEDFELT/DIGITALVISION/GETTY

Forty-three per cent of children under the age of five in low- and middle-income countries (about 250 million children) risk not reaching their developmental potential. This has important implications, including projected financial losses of approximately 26% lower annual earnings in adulthood. A recent survey found a 17.5% prevalence of mental disorders among Chinese children[1], a figure that is expected to rise owing to the COVID-19 pandemic. Addressing these global challenges requires urgent scientific insights into the brains and cognitive development of children.

The human brain is immature at birth and it undergoes substantial development postnatally, which enables it to adapt to the environment but also introduces risks of developmental issues. Ensuring a child's intellectual and socioemotional development is crucial for individual families; this also benefits society as a whole, as the cognitive skills of the population have been linked to individual earnings and economic growth[2].

Big data help to answer big questions

Capturing how the brain changes during development and learning is extremely challenging, because these changes unfold over extended periods of time and occur at multiple, interacting levels that include the social environment, behaviours, cognition, brain structure and function, and genetics. These changes interact and affect various domains such as mathematics, reading, cognitive abilities and socioemotional functions. Large representative data are essential for addressing these multifaceted relationships and obtaining unbiased, reproducible results.

The expanded search space. The brain, genes and the environment involve high-dimensional data. For instance, a typical structural scan of the human brain at a 1 × 1 × 1 resolution would produce several million voxels. A whole-genome scan yields several million single-nucleotide polymorphisms. Similarly, environmental exposures (the exposome) represent accumulated exposures that vary in space and time, which results in large datasets. The sheer number of statistical tests necessitates not only considerable computational power but also rigorous analytical methods to avoid false-positive results. A large sample size is required to meet the stringent corrections needed for multiple comparisons.

The prolonged developmental trajectory. Each stage of human brain development is marked by distinct psychosocial and neurobiological changes, which are influenced by genetic and environmental factors. For instance, middle childhood is characterized by ongoing brain maturation and rapid cognitive development, whereas adolescence features marked hormonal changes and substantial synaptic pruning processes (that is, selective loss of unused neuronal connections). Researchers need cross-sectional studies with large sample sizes, longitudinal follow-ups or a combination of these approaches (for example, accelerated longitudinal studies) to effectively characterize the developmental trajectory and identify age-specific factors. In any case, the required sample size is considerably larger than is required for studies focused solely on adults.

Sample bias and generalization. It is crucial to consider diversity both within and across cultures for findings to be accurate and generalizable. For example, owing to the predominance of studies that involve individuals of European descent, polygenic risk scores derived from genome-wide association studies predict individual risk more accurately in European than in non-European individuals[3]. Similarly, in psychological research, failing to account for individuals' experiences and behaviours within their cultural contexts can undermine the accuracy and generalizability of findings related to diverse psychological phenomena, such as well-being and academic outcomes[4]. The same applies to brain development: we need representative and random sampling.

To ensure representative sampling, large overall sample sizes are necessary to provide sufficient statistical power for studying minority

Table 1 | Examples of large child brain development studies

Study name	Year started	Locations	Number of participants (age range in years)	Design
ABCD	2018	USA	11,875 (9–10)	Longitudinal
cVEDA	2016	India	14,000 (0–25)[a]	Cross-sectional
dHCP	2014	UK	1,500 (20–44 post-conceptional weeks)	Cross-sectional
Generation R	2002	Netherlands	9,778 (prenatal to adulthood)	Longitudinal
HCP-D	2016	USA	1,350 (5–21)	Cross-sectional
HBCD	2021	USA	7,500 (0–10)	Longitudinal
HBN	2017	USA	10,000 (5–21)	Cross-sectional
IMAGEN	2010	Europe	2,000 (14–22)	Longitudinal
PING	2009	USA	1,493 (3–20)	Cross-sectional
PNC	2009	USA	9,498 (8–21)	Mixed (about 500 longitudinal)
SYS	2003	Canada	1,029 (12–18)	Longitudinal
YCS	2015	Netherlands	2,500 (prenatal to 7) 1,850 (8–16)	Longitudinal

Only studies with more than 1,000 developmental brain datasets are included. [a]Brain image data were obtained for 1,000 participants aged 10–23 years. ABCD, Adolescent Brain and Cognitive Development; dHCP, Developing Human Connectome Project; HCP-D, The Human Connectome Project Development; HBCD, Healthy Brain and Child Development; HBN, Healthy Brain Network; PING, Paediatric Imaging, Neurocognition, and Genetics; PNC, Philadelphia Neurodevelopmental Cohort; SYS, Sanguenay Youth study; YCS, Youth Cohort Study.

groups or disorders with relatively low occurrence rates (such as autism spectrum disorder). Large sample sizes are indispensable because many prospective studies aim to predict the occurrence of developmental disorders and cannot pre-identify the case group.

The reproducibility crisis. Both brain and behavioural measures often suffer from reliability issues, which contributes to the reproducibility crisis in psychological and neuroimaging research[5]. For instance, it is estimated that approximately 20–30 min of functional scanning is needed to obtain reliable estimates of functional connectivity patterns. Similarly, the test–retest reliability of most cognitive tasks hovers around 0.5, which means that we need to use multiple cognitive tasks that measure the same cognitive construct and longer testing times to achieve reliable results.

Compounding these challenges, the effect sizes in neuroimaging and genetic research are often very small. For example, individual genetic variants account for only up to 1% of the variance in complex traits, including cognitive functions or brain structure and functions[6]. Associations between image-derived phenotypes (such as T1, T2 and diffusion tensor imaging) and non-imaging variables (such as cigarette consumption or cognitive test scores) are also typically minimal[6].

Reliability can be increased by longer testing times and/or greater sample sizes. Continuous efforts have been made to optimize the balance between sample size and testing time to achieve high-reliability levels while minimizing overall costs.

Big developmental brain data needs diversity

To tackle these challenges, many large-scale 'population neuroscience' studies have been conducted worldwide over the past two decades (Table 1).

However, most of the projects in involve participants of European ancestry, which limits the diversity of the sample and the generalizability of the results[7]. For example, the Chinese Human Connectome Project (CHCP), which focuses on adult brain, has revealed an important culture-specific language network[8]. The Born in GuangZhou Cohort Study (BIGCS) reported east Asian-specific genetic associations with maternal total bile acid, gestational weight gain and infant cord blood traits[9].

To address the diversity issue, the Imaging Genetics (IMAGEN) project has teamed up with researchers from India to launch the Consortium on Vulnerability to Externalizing Disorders and Addictions (cVEDA) project, which aims to compare risk constellations and neurobehavioural trajectories for substance misuse and externalizing disorder in developed countries and emerging societies. In September 2021, China initiated the 'Brain Science and Brain-Like Intelligence Technology' project (also known as the China Brain Project (CBP)). One important component of the CBP is the Chinese Child Brain Development (CCBD) project, which aims to understand the environmental and genetic factors that affect brain and cognitive development, academic performance, and psychological well-being.

The CCBD is a large-scale initiative for studying brain and cognitive development in China. The main cohort of the CCBD consists of a cross-sectional sample of more than 5,000 participants aged 6–18 years and a longitudinal sample of more than 20,000 participants aged 6–7 years, who will be tracked until they are 18 years old. Additionally, the CCBD includes a prospective cohort that tracks 10,000 children aged 3–8 years to investigate the causes of Chinese dyslexia, and the tracking of 10,000 children aged 6–12 years from the existing BIGCS[9] to investigate the early predictors of learning difficulties and mental health problems. All participants are based in China. With these large samples, the CCBD expects to substantially improve the reliability and generalization of research findings. In addition, extensive cognitive and behavioural, socioemotional, environmental, EEG, neural and genetic data will be collected. Combined with the inclusion of a broad range of years, the CCBD will greatly expand the scope of questions that can be explored through big brain data. Finally, the CCBD will also incorporate cognitive and environmental interventions to alleviate developmental difficulties. By combining CCBD data and other datasets (as listed in Table 1), researchers could, for example, gain a much deeper understanding of the developmental trajectory, and the effect of urbanization, lifestyle changes and culture, which can guide policies and programmes to enhance cognitive development and educational outcomes. By identifying universal and culture-specific genetic variants, researchers can develop personalized medicine approaches for neurological diseases.

BOX 1

A call for action

To fully realize the potential of big brain development data in addressing fundamental question in children's brain and cognitive development and to inform policy, education and healthcare, we recommend:

- Open science and collaboration. To increase the reproductivity and generalization of results, the global brain research community should embrace open science practices such as preregistered reports and governments (including funding agencies) should encourage data sharing and international collaboration.
- Deep phenotyping studies. In conjunction with big cohort studies, funding agencies should also support follow-up, more focused deep phenotyping projects such as the Simons Simplex Collection for autism spectrum disorder[10]. This new type of big data could complement big cohort studies and provide more detailed characterization of developmental subtypes, uncover their protective and risk factors, and inform precise interventions.
- Manipulation and intervention. To provide causal evidence and develop effective interventions, the brain development research community must reach out to the educational, clinical and industry sectors for collaboration and the sharing of resources, expertise and data. This collaborative effort should uphold recognized standards of scientific rigour and integrity (one example in medicine development is the International Council on Harmonisation (ICH) Good Clinical Practice (GCP) Guidelines) to enhance research quality and ensure robust and reliable findings.

Conclusions and future directions

In summary, big data in developmental brain research offers invaluable insights into brain development and the various risk and protective factors, which has important implications for policy, education and healthcare. However, realizing this potential requires acknowledging and addressing the challenges of big data projects. One obvious challenge is the considerable financial investment and extensive coordination and management involved in these projects. Professional and responsible teamwork are invaluable for the success of these projects. In addition, several scientific challenges also need to be addressed. First, even large-scale studies might not provide sufficient or representative samples. Second, large-scale studies often compromise on the detailed characterization of nuanced phenotypes. Finally, many of these studies are observational, and their correlational findings – despite being longitudinal – should not be mistaken for causal relationships. To address these challenges, we invite the government, funding agencies, global research community, and educational, clinical and industry sectors to take up our call for action (Box 1) and commit to the discovery of robust, precise and translational findings.

Liang Zhang [1] **& Gui Xue** [1,2] ✉

[1]State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China. [2]Chinese Institute for Brain Research, Beijing, China.
✉e-mail: gxue@bnu.edu.cn

Published online: 10 April 2025

References

1. Li, F. et al. *J. Child Psychol. Psychiatry* **63**, 34–46 (2022).
2. Hanushek, E. A. & Woessmann, L. *J. Econ. Lit.* **46**, 607–668 (2008).
3. Martin, A. R. et al. *Nat. Genet.* **51**, 584–591 (2019).
4. Brady, L. M., Fryberg, S. A. & Shoda, Y. *Proc. Natl Acad. Sci. USA* **115**, 11406–11413 (2018).
5. Poldrack, R. A. et al. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
6. Smith, S. M. & Nichols, T. E. *Neuron* **97**, 263–268 (2018).
7. He, Y. & Martin, A. R. *Nat. Hum. Behav.* **8**, 197–200 (2024).
8. Ge, J. et al. *Nat. Neurosci.* **26**, 163–172 (2023).
9. Huang, S. et al. *Nature* **626**, 565–573 (2024).
10. Fischbach, G. D. & Lord, C. *Neuron* **68**, 192–195 (2010).

Acknowledgements

The corresponding author is supported by the STI 2030 - Major Projects (2021ZD0200500) and the National Natural Science Foundation of China (32330039 and 32441110)

Competing interests

The authors declare no competing interests.