Main Existing Datasets for Open Brain Research on Humans

Baptiste Couvy-Duchesne, Simona Bottani, Etienne Camenen, Fang Fang, Mulusew Fikere, Juliana Gonzalez-Astudillo, Joshua Harvey, Ravi Hassanaly, Irfahan Kassam, Penelope Lind, et al.

Chapter 24

Main Existing Datasets for Open Brain Research on Humans

Baptiste Couvy-Duchesne, Simona Bottani, Etienne Camenen, Fang Fang, Mulusew Fikere, Juliana Gonzalez-Astudillo, Joshua Harvey, Ravi Hassanaly, Irfahan Kassam, Penelope A. Lind, Qianwei Liu, Yi Lu, Marta Nabais, Thibault Rolland, Julia Sidorenko, Lachlan Strike, and Margie Wright

Abstract

Recent advances in technology have made possible to quantify fine-grained individual differences at many levels, such as genetic, genomics, organ level, behavior, and clinical. The wealth of data becoming available raises great promises for research on brain disorders as well as normal brain function, to name a few, systematic and agnostic study of disease risk factors (e.g., genetic variants, brain regions), the use of natural experiments (e.g., evaluate the effect of a genetic variant in a human population), and unveiling disease mechanisms across several biological levels (e.g., genetics, cellular gene expression, organ structure and function). However, this data revolution raises many challenges such as data sharing and management, the need for novel analysis methods and software, storage, and computing.

Here, we sought to provide an overview of some of the main existing human datasets, all accessible to researchers. Our list is far from being exhaustive, and our objective is to publicize data sharing initiatives and help researchers find new data sources.

Key words Genetic, Methylation, Gene expression, Brain MRI, PET, EEG/MEG, Omics, Electronic health records, Wearables

1 Aims

We sought to provide an overview and short description of some of the main existing human datasets accessible to researchers. We hope this chapter will help publicize them as well as encourage the sharing of datasets for open science. As much as possible, we tried to provide practical aspects, such as data type, file size, sample demographics, study design, as well as links toward data use/transfer agreements. We hope this can help researchers study larger and more diverse data, in order to advance scientific discovery and improve reproducibility.

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_24,

This chapter does not aim to provide an exhaustive list of the dataset and data types currently available. In addition, the interested readers may refer to the complementary chapters that focus on data processing, feature extraction, and existing methods for their analyses.

2 Introduction

The availability of data used in research is one of the cornerstones of open science, which contributes to improving the quality, reproducibility, and impact of the findings. In addition, data sharing increases openness and transparent and collaborative scientific practices. The global push for open science is exemplified by the recent publication of UNESCO guidelines [1], the engagement of many research institutions, and the requirements of some scientific journals to make data available upon publication. Finally, the sharing and re-use of data also maximizes the return on investment of the agencies (e.g., states, charities, associations) that fund the data collection.

In light of this, our chapter aims at providing a broad (albeit partial) overview of some of the human datasets publicly available to researchers. To assist researchers and data managers, we first describe the different file formats and the size of the different data types (*see* Table 1). As many of these data are high-dimensional, the size of the data can cause storage and computational challenges, which need to be anticipated before download and analysis. Of note, some datasets cannot be downloaded or analyzed outside of a dedicated system/server. This is the case of the UK Biobank (UKB) exome and whole genome sequencing, whose sheer size has led to the creation of a dedicated Research Analysis Platform, accessible (at some cost) by UKB-approved researchers. In addition, the Swedish registry data is only accessible via national dedicated servers due to the extreme sensitive nature of the data.

This chapter breaks down into sections that focus on each data type, although the same dataset may be mentioned in several sections. Beyond a practical writing advantage (each author or group of authors contributed a section), this also reflects the fact that most datasets are organized around a central data type. For example, the ADNI (Alzheimer's disease Neuroimaging Initiative) focuses on brain imaging and later included genotyping information. Another example is the UKB, which released genotyping data of the 500 K participants in 2017, is now collecting brain MRI (as well as cardiac and abdominal MRI, whole-body DXA, and carotid ultrasound), and has recently made available sequencing data. The different sections also discuss and present the specific data sharing tools and portals (e.g., LONI for brain imaging, GTEx for gene expression) or organization of the different fields (e.g., consortia in

Table 1
Overview of data types and sample size

Data	Subtype	File type/extension	Description	Approx. size of one sample
Clinical	Self-reports	Text		~100 KB
	EHR	Text		~100 KB
Neuroimaging	MRI	NIfTI (.nii/.nii.gz) or DICOM (.dcm)	3D image (.nii) 2D slices (.dcm) that compose the 3D volume	T1w ~50 MB T2 FLAIR ~30 MB SWI ~30 MB DWI ~400 MB (highly variable depending on sequence) fMRI ~500 MB (100 MB per minute of acquisition) ~1 GB after processing
	PET	NIfTI (.nii / .nii.gz) or DICOM (.dcm)	3D image (.nii) 2D slices (.dcm) that compose the 3D volume	~15 MB
	EEG	.edf, .gdf, .eeg, .csv, .mat	Raw EEG signal (.egg, .gdf, .edf – file formats), processed/formatted data (.csv, .mat)	~15 MB ~50 MB
	MEG	.fif, .bin, .csv	Raw MEG signal (.fif, .bin) processed/formatted data (.csv, .mat)	
Genetics	Twin/family sample	Text	Phenotypes and pedigree information	~10 KB
	Genotyping	.bgen, .bed (.bim + . bed + .fam)	Binary files	~5 MB
	GWAS summary statistics	Text	Effect size, test statistic, p-value, effect allele from association testing	Not individual-level data, ~500 MB for a genome-wide association summary
	Exome sequencing	.cram, .bcf, .vcf	Each variant site with multiple sequence quality metrics and trained machine learning filters to identify and exclude inconsistencies. Followed by initial QC with various parameters and thresholds	~1 M
	Whole genome sequencing	.cram, .bcf, .vcf	Curated using TOPMed variant calling algorithm. All freezes are fully processed	~5 MB

(continued)

Table 1
(continued)

Data	Subtype	File type/extension	Description	Approx. size of one sample
Genomics	Methylation	.idats	Raw binary intensity file (two per sample, one for green and red channels)	~16 MB (450 K) ~27 MB (EPIC)
		.txt or csv.	Processed, normalized, and filtered beta matrix	~8 MB (450 K) ~14 MB (EPIC)
	Expression	.bed (.bim + .bed + .fam)	Fully processed, filtered, and normalized gene expression matrices (in BED format) for each tissue	~1 MB per individual
		Text	Covariates used in eQTL analysis. Includes genotyping principal components and PEER factors	~6 KB
Smartphone and sensors	*Actigraphy*	Varies with product (e.g., GENEActiv uses .bin)	Raw accelerator data	~500 MB (GENEActiv, 14 days of recording)

Size of one sample may vary based on the technology used to generate the data (e.g., MRI resolution, genotyping chip)

genetics). Every time, we have tried to include the largest dataset (s) available, as well as the commonly used ones, although the selection may be subjective and reflect the authors' specific interests (e.g., age or disease groups).

All datasets are listed in a single table (*see* Table 2), which includes information about country of origin, design (e.g., cross-sectional, longitudinal, clinical, or population sample), and age range of the participants. Unless specified, the datasets presented include male and female participants, although the proportion may differ depending on the recruitment strategy and disease of interest. In addition, the table lists (and details) the different data types that have been collected on the participants. We have only focused on a handful of data types: genetic data (including twin/family samples, genotyping, and exome and whole genome sequencing), genomics (methylation and gene expression), brain imaging (MRI and PET), EEG/MEG, electronic health records (hospital data and national registry), as well as wearable and sensor data. However, we have included additional columns "Other omics" and "Specificities" that list other types of data being collected, such as proteomics, metabolomics, microRNA, single-cell sequencing, microbiome, and non-brain imaging.

Our main table (*see* Table 2) also includes the URLs to the dataset websites and data transfer/agreement. From our experience, data access can take between an hour and up to a few months. The agreements almost always require a review of the project and to acknowledge the data collection team and funding sources (e.g., under the form of a byline, a paragraph in the acknowledgment, and more rarely co-authorships). Standard restrictions of use include that the data cannot be redistributed and that the users do not attempt to identify participants. Specific clauses are often added depending on the nature of data and the specific laws and regulations of the countries it originates from.

There is a growing scientific and ethical discussion about the representativity of the datasets being used in research. Researchers should be aware of the biases present in some datasets (e.g., "healthy bias" in the UKB [2]), which should be taken into account in study design (e.g., analysis of diverse ancestry being collected in genetics [3]), when reporting results [2, 4] and evaluating algorithms [5, 6]. Overall, our (selected) list exemplifies the need for datasets from under-represented countries or groups of individuals (e.g., disease, age, ancestry, socioeconomic status) [7, 8]. Our main table (*see* Table 2) will be accessible online, in a user-friendly, searchable version. Finally, we will also make this table collaborative (via GitHub https://github.com/baptisteCD/MainExistingDatasets) in order to grow this resource beyond this book chapter.

Table 2
Description of a selection of the main human datasets available for research

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
Alzheimer's Disease Neuroimaging Initiative (ADNI)	2215	55–90	USA and Canada	Longitudinal (up to 9 years of follow-up, ongoing)	Neurological (Alzheimer's), cognition, lumbar puncture	MRI (T1w, T2, DWI, rsfMRI) PET (^{18}F-FDG, FBB, AV45, PiB)	NA	Genotyping, WGS	Methylation (subset of 653 individuals—3 time points)	NA	Transcriptomics, CSF proteomics, metabolomics, lipidomics	http://adni.loni.usc.edu/data-samples/access-data/	[11, 191]	Four waves (ADNI1, 2, GO, 3) with different inclusion and protocols
Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL)	726	60+	Australia	Longitudinal (up to 6 years of follow-up)	Neurological (Alzheimer's), cognition, lumbar puncture	MRI (T1w, T2, DWI, rsfMRI) PET (PiB, AV45, Flute)	NA	Genotyping	Methylation	ActiGraph activity (10% of sample)	NA	https://ida.loni.usc.edu/collaboration/access/appLicense.jsp https://aibl.csiro.au/	[12]	Neuroimaging and selected clinical data available via the LONI platform Full sample ($N \sim 1200$), including extended clinical, genotyping, or methylation available via application to CSIRO
Open Access Series of Imaging Studies v3 (OASIS3)	1096	42–95	USA	Longitudinal (up to 12 years of follow-up)	Neurological (Alzheimer's), cognition	MRI (T1w, T2w, FLAIR, ASL, SWI, time of flight, rsfMRI, and DWI) PET (PIB, AV45, FDG)	NA	NA	NA	NA	NA	https://www.oasis-brains.org https://www.oasis-brains.org/#access	[13]	Retrospective dataset from imaging projects collected by WUSTL Knight ADRC over 30 years Two other (non-independent) datasets available: OASIS1 and OASIS2
Adolescent Brain Cognitive Development (ABCD)	~11,878	9–12	USA	Longitudinal (up to 2 years of follow-up, ongoing)	Self- and parental rating. Substance use, mental health (psychiatry), cognition, physical health	MRI (T1w, T2, rsfMRI, tfMRI)	NA	Genotyping, pedigree (twinning)		iPAD tasks and testing	NA	https://abcdstudy.org https://nda.nih.gov/abcd/request-access	[16, 17]	Objective of 10 years of follow-up
Enhancing NeuroImaging Genetics through Meta-analysis (ENIGMA)	>50,000 indiv. incl.: 9572 (SCZ), 6503 (BD), 10,105 (MDD), 1868 (PTSD), 3240 (SUD), 3665 (OCD), 4180 (ADHD), 18,605 (life span)	3–90 (no restriction)	Worldwide (43+ countries)	Cross-sectional and longitudinal	Psychiatry, neurology, addiction, suicidality, brain injury, HIV, antisocial behavior	T1w, DWI, rsfMRI, tfMRI	Resting-state EEG	Genotyping, CNVs, pedigree (twinning)	Methylation	NA	NA	https://enigma.ini.usc.edu/join/ http://enigma.ini.usc.edu/research/download-enigma-gwas-results/	Neuroimaging projects [27 [112, 114]	Consortium organized around genetics and/or disease/trait working groups as well as non-clinical working groups with focus on sex, healthy aging, plasticity, etc. Imaging and genetic protocols and genome-wide association statistics are available for download on the ENIGMA website

ChiNese Brain PET Template (CNPET)	116	27–81	China	Cross-sectional	Healthy subjects	PET (^{18}F-FDG)	NA	NA	NA	NA	NA	https://www.nitrc.org/doi/landing_page.php?table=groups&id=1486&doi= https://www.nitrc.org/projects/cnpet	[39]	Dataset made to build a Chinese-specific SPM PET brain template
Centre for Integrated Molecular Brain Imaging (CIMBI)	~2000	17–93	Denmark	Cross-sectional	Mental and physical state, personality, and background Neuropsychological measures (memory, language, etc.)	MRI (T1w, T2w, DWI, fMRI) PET (^{11}C-5-HT)	NA	Genetic polymorphisms relevant for the 5-HT system	NA	NA	NA	https://www.cimbi.dk/index.php	[38]	Data sharing only within the European Union (EU). Visiting access for non-EU researchers
Motor Imagery dataset from Cho et al. [192]	52	24.8 SD = 3.86 years old	South Korea	Cross-sectional	Healthy subjects	NA	EEG (64 channels); real hand movements, motor imagination hand movement tasks	NA	NA	NA	NA	http://moabb.neurotechx.com/docs/generated/moabb.datasets.Cho2017.html#moabb.datasets.Cho2017	[192]	BCI experiments
EEG Alpha Waves dataset	20	19–44 years old	France	Cross-sectional	Healthy subjects	NA	EEG (16 channels); resting state, eyes open/closed	NA	NA	NA	NA	https://zenodo.org/record/2605110#.YTeT5Z4zZhE	[193]	Alpha waves dataset. BCI experiments
Multitaper spectra recorded during GABAergic anesthetic unconsciousness	55		USA	Cross-sectional	Healthy participants and patients receiving an anesthesia care in an operating room context	NA	EEG (64 for healthy volunteers, 6 for patients under anesthesia)	NA	NA	NA	NA	https://physionet.org/content/eeg-power-anesthesia/1.0.0/	[194]	Patients under anesthesia during surgery
A multi-subject, multi-modal human neuroimaging dataset	19 (16 for MEG and EEG, 3 for fMRI)	23–37 years old	UK	Longitudinal (2 visits over 3 months)	Healthy subjects	fMRI	MEG, EEG (70 channels); resting state, sensory–motor tasks	NA	NA	NA	NA	https://legacy.openfmri.org/dataset/ds000117/	[195]	OpenfMRI database, accession number ds000117
Motor imagery, uncued classifier application	10	26–46 years old	Germany	Cross-sectional	Healthy subjects	NA	EEG (59 channels); hands, feet, and tongue motor imagination tasks	NA	NA	NA	NA	http://www.bbci.de/competition/iv/desc_1.html	[196]	Data used for a BCI competition. Also include simulated/synthetic data
BCI Competition 2008—Graz data set A	9	NA	Austria	Cross-sectional	Healthy subjects	NA	EEG (22 channels); hands, feet, and tongue motor imagination tasks	NA	NA	NA	NA	http://www.bbci.de/competition/iv/desc_2a.pdf	[197]	Data used for a BCI competition

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
BCI Competition 2008—Graz data set B	9	NA	Austria	Cross-sectional	Healthy subjects	NA	EEG (3 channels); hands motor imagination tasks	NA	NA	NA	NA	http://www.bbci.de/competition/iv/desc_2b.pdf	[197]	Data for a BCI competition
EEG dataset from Mumtaz et al. [198]	64 (34 MDD, 30 healthy)	40.3 ± 12.9 years old	Malaysia	Longitudinal (multiple visits to the clinic)	Case control for major depressive disorder	NA	EEG (19 channels)	NA	NA	NA	NA	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171409	[198]	EEG (resting task), MDD based on medical history of patients
EEG data for ADHD/control children	121	7–12 years old	South Korea	Cross-sectional	ADHD (61 children) and healthy controls [61]	NA	EEG (19 channels); visual attention tasks	NA	NA	NA	NA	https://ieee-dataport.org/open-access/eeg-data-adhd-control-children	[199]	Visual cognitive tests on children with ADHD, using videos
EEG and EOG data from Jaramillo-Gonzalez et al. [61]	4	NA	Germany	Longitudinal (2 to 10 visits)	Amyotrophic lateral sclerosis, locked-in state	NA	EEG (116 channels)	NA	NA	NA	NA	https://doi.org/10.6084/m9.figshare.13148762	[200]	Spelling task with eye movement-based answers Electrooculography (EOG)
Temple University Hospital (TUH) EEG Corpus	10,874	1–90	USA	Cross-sectional	Epilepsy, stroke, concussion, healthy	NA	EEG (20–41 channels)	NA	NA	NA	NA	https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml	[201]	Extensive dataset of 30 K clinical EEG recordings collected at TUH from 2002 to today. Epilepsy subset: TUEP dataset
Bern–Barcelona EEG database	5	NA	Spain	Longitudinal	Epilepsy	NA	EEG (64 channels); rest	NA	NA	NA	NA	https://repositori.upf.edu/handle/10230/42829	[202]	Intracranial EEG samples before and after surgery
CHB-MIT Scalp EEG Database	22	1.5–22	USA	Longitudinal	Seizures and intractable seizures	NA	EEG (21 channels); resting state	NA	NA	NA	NA	https://www.physionet.org/content/chbmit/1.0.0/	[203]	Pediatric subjects with intractable seizures
Brain/Neural–Computer Interaction (BNCI) Horizon	2	NA	Austria	Longitudinal	Chronic stroke	NA	EEG (2 channels); eye staring task	NA	NA	NA	NA	http://bnci-horizon-2020.eu/database/data-sets	[204]	The BNCI is an open-source project with several datasets. Stroke is "6. SCP training in stroke (006-2014)"
Queensland Twin Adolescent Brain (QTAB)	422 (baseline)	9–14 (baseline)	Australia	Longitudinal	Population sample. Parental and/or self-report mental health, cognition, and social behavior measures	MRI (T1w, T2w, FLAIR, DWI, rsfMRI, tfMRI, ASL)	NA	Pedigree (twin/sibling), genotyping	NA	Wrist-worn accelerometer	Gut microbiome	https://imaginggenomics.net.au/	[205]	Includes participants from Queensland Twin Registry and Twins Research Australia

Queensland Twin Imaging Study (QTIM)	1200+	12–30	Australia	Cross-sectional	Population sample (as part of BATS). Self-report mental health, cognition, substance use, and personality measures	MRI (T1w, DWI, rsfMRI, tfMRI)	NA	Pedigree (twin/sibling), genotyping	NA	NA	NA	https://imaginggenomics.net.au/	[206]	Include participants from Brisbane adolescent twin study (BATS)
Human Connectome Project, young adults (HCP-YA)	~1200	22–35	USA	Cross-sectional	Population sample. Self-report mental health, cognition, personality, and substance use measures	MRI (T1w, T2w, DWI, rsfMRI, tfMRI)	MEG ($n = 95$)	Pedigree (twin/sibling). Genotyping	NA	NA	NA	https://db.humanconnectome.org	[29]	Expanded to development and aging projects (similar imaging protocols, but extensions do not include twins)
Vietnam Era Twin Study of Aging (VETSA)	1237 (baseline; 500+ with MRI)	51–60 (baseline)	USA	Longitudinal	US veterans, males only	MRI (T1w, DWI, ASL (subset of sample))	NA	Pedigree (twin). Genotyping	NA	NA	NA	https://medschool.ucsd.edu/som/psychiatry/research/VETSA/Researchers/Pages/default.aspx	[83]	Subset of the Vietnam Era Twin Registry, males only
Older Australian Twins Study (OATS)	623 (baseline)	>65	Australia	Longitudinal	Population sample. Self-report medical and mental health and neuropsychological measures	MRI (T1w, DWI, tfMRI); PET	NA	Pedigree (twin/sibling). Genotyping	NA	NA	NA	https://cheba.unsw.edu.au/research-projects/older-australian-twins-study	[84]	Recruited through the Australian Twin Registry
Swedish Twin Registry (STR)	87,000 twin pairs	All ages	Sweden	Longitudinal (since the late 1950s)	NA	NA	NA	Genome-wide single nucleotide polymorphism array genotyping	Methylation is not available in the STR yet but is available in the Swedish adoption/twin study of aging (SATSA), a sub-study of the STR	NA	NA	https://ki.se/en/research/the-swedish-twin-registry	[74]	Twin Registry
UK Biobank (UKB)	~502,000 ~50,000 (imaging subset) ~100,000 (actigraphy subset)	37–73 44–82 (imaging) 40–69 (actigraphy)	UK	Longitudinal	Self-reported and EHR medical history (incl. cancer, neurology, COVID-19)	MRI (T1w, T2w, FLAIR, DWI, SWI, rsfMRI, tfMRI)	NA	Genotyping, exome, WGS, pedigree	NA	Actigraphy	Metabolomics (future release)	https://bbams.ndph.ox.ac.uk/ams/resApplications	[14, 101, 186]	Population-based sample (volunteers), healthy bias, ongoing resource and data collectiontarget MRI sample 100 K, retest 10 K. also available: Cardiac and abdominal MRI, whole-body DXA (dual-energy X-ray absorptiometry),

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
														carotid ultrasound, electrocardiogram, hematological assays, serological antibody response assay, telomere length
Avon Longitudinal Study of Parents and Children (ALSPAC): Accessible Resource for Integrated Epigenomic Studies (ARIES)	2044	Average age: Mothers (antenatal = 28.7, follow-up = 47.5), offspring (birth = 40 weeks, childhood = 7.5, adolescence = 17.1)	UK	Longitudinal (2 time points for mother, 3 for offspring)	Clinical evaluations, obstetric data, cognition, questionnaires	MRI (T1w, DWI, mcDESPOT (subset of offspring cohort))	NA	Genotyping	Methylation		Transcriptomics	http://www.ariesepigenomics.org.uk/ https://github.com/MRCIEU/aries	[207]	General population study (health and development) following 1022 mother–offspring pairs; 2 time points for mother, 3 for offspring
Biobank-based Integrative Omics Studies (The BIOS Consortium)	~4000	18–87	Netherlands	Longitudinal	Clinical information and bioassays (depending on the sub-cohort)	NA	NA	Genotyping, pedigree	Methylation	NA	Transcriptomics, metabolomics	https://www.bbmri.nl/node/24	[208]	Population study including various sub-cohorts, covering differing research designs (LifeLines, Leiden longevity study, Netherlands twin registry, Rotterdam study, CODAM, and the prospective ALS study Netherlands); access via European genome-phenome archive (EGA) and SURFsara high performance computing cloud
Framingham Heart Study	4241	Offspring cohort mean age = 66, third-generation cohort = 45	USA	Cross-sectional (multi-generational)	Extensive clinical evaluations and bioassays, original focus on cardiovascular diseases	MRI (T2w)	NA	Genotyping, WGS	Methylation	NA	Transcriptomics, metabolomics	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000724.v9.p13	[209, 210]	Data collected over two generations of individuals; cardiovascular MRI A subset of individuals are used to identify rare variants influencing brain imaging phenotypes. Data included in NHLBI TOPMed

Women's Health Initiative	2129	50–79	USA	Cross-sectional	Clinical evaluations and bioassays; cardiovascular diseases, women only	NA	NA	Genotyping	Methylation	NA	Transcriptomics, metabolomics, miRNA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001335.v2.p3	[211]	Women only
Sporadic ALS Australia Systems Genomics Consortium (SALSA-SGC)	1395	Predicted mean age (from DNA methylation) 60.4 (controls), 62.9 (cases)	Australia	Cross-sectional	Case control of amyotrophic lateral sclerosis	NA	NA	Genotyping	Methylation	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002068.v1.p1	[212]	
System Genomics of Parkinson's Disease (SGPD)	2333	Predicted mean age (from DNA methylation), mean 70.05; range 26–93	Australia and New Zealand	Cross-sectional	Case control of Parkinson's disease	NA	NA	Genotyping	Methylation	NA	NA	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145361	[213]	
Genetics of DNA Methylation Consortium (goDMC)	32,851	0–91	Worldwide	Cross-sectional	Multiple diseases, but also healthy aging cohorts	NA	NA	Genotyping	Methylation	NA	NA	http://www.godmc.org.uk/cohorts.html	[79]	Consortium gathering 38 independent studies; data access to be obtained from each subsample
Psychiatric Genomics Consortium (PGC)	By 2025, ~2.5 million cases of psychiatric disorders	All ages	Worldwide	Cross-sectional	Psychiatric disorders, substance use disorders, and neurology: Major depressive disorder, cannabis use disorder, alcohol use disorder, schizophrenia, anorexia, bipolar, ADHD, Alzheimer's	NA	NA	Genotyping and sequencing	Expression and methylation data available in some working groups	NA	NA	https://www.med.unc.edu/pgc/shared-methods/open-source-philosophy/	[214, 215]	Consortium organized around disease/trait working groups
The genetic epidemiology of asthma in Costa Rica	4347	6–12	Costa Rica	Cross-sectional	Asthma cases and controls	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA295246	[216]	A subset of individuals from this sample/accession is used in Pharmacogenetic drug response in racially diverse children with asthma. Data included in NHLBI TOPMed
Coronary Artery Risk Development in Young Adults (CARDIA)	3425	18–30	USA	Longitudinal	Coronary Artery Risk	NA	NA	WGS	NA	NA	NA	https://anvilproject.org/ncpi/data/studies/phs001612	[217]	A subset of longitudinal dataset of 3087 self-identified Black and White participants from the CARDIA study were used to study multi-ethnic polygenic risk score

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
														associated with hypertension prevalence and progression
Genetic Epidemiology Network of Arteriopathy (GENOA)	1854	>60	USA	Longitudinal	Elucidate the genetics of target organ complications of hypertension	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001345.v3.p1	[217]	Study was used to study multi-ethnic polygenic risk score associated with hypertension prevalence and progression
Hispanic Community Health Study/ Study of Latinos (HCHS/SOL)	8093	18–74	USA	Longitudinal	A multicenter prospective cohort study for asthma patients	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001395.v1.p1	[217]	Study was used to study multi-ethnic polygenic risk score associated with hypertension prevalence and progression
Women's Health Initiative (WHI)	11,357	>65	USA	Longitudinal	Women's Health Initiative cohort involved study on ischemic stroke, 900 cases of hemorrhagic stroke	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000200.v12.p3	[217]	Study was used to study multi-ethnic polygenic risk score associated with hypertension prevalence and progression
Atherosclerosis Risk in Communities (ARIC)	8975	45–64	USA	Cross-sectional/ longitudinal	Red blood cell phenotype	NA	NA	WGS	NA	NA	NA	https://anvilproject.org/data/studies/phs001211	[218]	WGS association study of red blood cell phenotypes, GWAS statistics available. Data included in NHLBI TOPMed
Rare Variants for Hypertension in Taiwan Chinese (THRV)	2159	>35	Taiwan/China, Japan	Longitudinal	Insulin-resistant cases and controls	NA	NA	WGS	NA	NA	NA	https://anvilproject.org/ncpi/data/studies/phs001387	[219]	Clustering and heritability of insulin resistance in Chinese and Japanese hypertensive families. Data included in NHLBI TOPMed
My Life, Our Future initiative (MLOF)	7482	>18	USA	Cross-sectional	Hemophilia cases and controls	NA	NA	WGS	NA	NA	NAs	https://athn.org/what-we-do/national-projects/mlof-research-repository.html	[220]	Summary statistics, different types of DNA variants detected in hemophilia. Data included in NHLBI TOPMed

Genetic Epidemiology of COPD (COPDGene)	19,996	45–80	USA	Cross-sectional	Pulmonary functions	NA	NA	WGS	NA	NA	Gene expression (eQTL) and methylation (mQTL): eQTLs in 48 tissues from GTEx v7	https://anvilproject.org/data/studies/phs001211	[221]	Multi-omic data from GTEx and TOPMed identify potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Cardiovascular Health Study (CHS)	4877	>65	USA	Longitudinal	*Cardiovascular health*	NA		WGS	NA	NA	NA	https://ega-archive.org/studies/phs001368 DOI:https://doi.org/10.1038/s41467-020-18334-7	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Cleveland Family Study (CFS)	3576	>65	USA	Longitudinal	Epidemiological data on genetic and non-genetic risk factors for sleep disordered breathing	NA		WGS	NA	NA	NA	https://ega-archive.org/studies/phs000954 DOI:https://doi.org/10.1038/s41467-020-18334-7	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Framingham Heart Study (FHS)	4241	<65	USA	Longitudinal	Assess risk of cardiovascular disease study	NA		WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000724.v9.p13	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Jackson Heart Study (JHS)	3596	>55	USA	Longitudinal	Assess cardiovascular disease	NA		WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000964.v5.p1	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Multi-Ethnic Study of Atherosclerosis (MESA)	6814	45–84	USA	Longitudinal	Assess cardiovascular disease	NA		WGS	NA	NA	NA	https://www.omicsdi.org/dataset/dbgap/phs001416	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Boston Early-Onset COPD (EOCOPD)	80	<53	USA	Longitudinal	Chronic obstructive pulmonary disease (COPD)	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000946.v5.p1	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed
Genetic Epidemiology of COPD (COPDGene)	10,647	45–80	USA	Longitudinal	Chronic obstructive pulmonary disease (COPD)	NA	NA	WGS	NA	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000951.v5.p5	[221]	Potential molecular mechanisms underlying 4 of the 22 novel loci. Data included in NHLBI TOPMed

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
Genotype-Tissue Expression (GTEx)	948 individuals, 17,382 samples (across 54 tissues)	20–70	USA	Cross-sectional	General medical history, HIV status, potential exposures, medical history at the time of death, death circumstances, serology results, smoking status	NA	NA	WGS	Gene expression (bulk RNA-seq)	NA	NA	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/dataset.cgi?study_id=phs000424.v8.p2&phv=169091&phd=3910&pha=&pht=2742&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1	[222]	Postmortem samples
Developmental Genotype-Tissue Expression (dGTEx)	120 individuals, 3600 samples (across 30 tissues)	0–18	USA	Cross-sectional	General medical history	NA	NA	WGS	Gene expression (bulk RNA-seq)	NA	NA	https://dgtex.org/	NA	Expected data release by late 2022 Four developmental time points: Postnatal (0–2 years), early childhood (2–8 years), pre-pubertal (8–12.5 years), and post-pubertal (12.5–18 years)
Greater-Paris Area Hospital (AP-HP)	Several millions ~130 K patients with brain MRI (~200 K images)	All ages	France	Longitudinal	ICD10 codes for diagnoses, medical reports, and a wide range of physiological variables	All medical imaging modalities and all organs: MRI, PET, SPECT, CT, ultrasounds	NA	NA	NA	NA	NA	https://eds.aphp.fr/	[155]	All patients visiting greater-Paris area hospital (AP-HP)the figures correspond to a specific project approved on the clinical data warehouse (CDW)
Swedish Total Population Register (TPR)	Nationwide coverage (N = 9,775,572 on April 30, 2015)	All ages	Sweden	Longitudinal (since 1968)	NA	NA	NA	NA	NA	NA	NA	www.scb.se/vara-tjanster/bestall-data-och-statistik/bestalla-mikrodata/vilka-mikrodata-finns/individregister/registret-over-totalbefolkningen-rtb/	[158]	Demographic information

Swedish Longitudinal Integrated Database for Health Insurance and Labour Market Studies (LISA)	All individuals ≥16 years in Sweden	≥16 years (15 since 2010)	Sweden	Longitudinal (since 1990)	NA	NA	NA	NA	NA	NA	www.scb.se/lisa	[159]	Demographic and socioeconomic information
Swedish Multi-Generation Register (MGR)	Over 11 million	All ages	Sweden	Longitudinal (since 1961)	NA	NA	NA	Pedigree	NA	NA	www.scb.se/en/finding-statistics/statistics-by-subject-area/other/other/other-publications-non-statistical/pong/publications/multi-generation-register-2016/	[160]	Family relation
Swedish National Patient Register (NPR)	Nationwide coverage	All ages	Sweden	Longitudinal (since 1964)	Nationwide inpatient care since 1987 and outpatient care since 2001	NA	NA	NA	NA	NA	www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/the-national-patient-register/	[161]	Clinical diagnoses from inpatient and outpatient care
Swedish Cancer Register (SCR)	Around 60,000 malignant cases are included annually (statistics in 2020)	All ages	Sweden	Longitudinal (since 1958)	Clinical diagnoses of cancer	NA	NA	NA	NA	NA	www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/swedish-cancer-register/	[165]	
Swedish Medical Birth Register (MBR)	Around 80,000–120,000 deliveries annually	Infants at birth	Sweden	Longitudinal (since 1973)	Yes	NA	NA	NA	NA	NA	https://www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/the-swedish-medical-birth-register/	[167]	Register of medical birth
Swedish Causes of Death Register (CDR)	Nearly 100,000 deaths annually	All ages	Sweden	Longitudinal (since 1952)	Cause of death	NA	NA	NA	NA	NA	https://www.socialstyrelsen.se/statistik-och-data/register/alla-register/dodsorsaksregistret/	[170]	Register of cause of death
Swedish Prescribed Drug Register (PDR)	More than 100 million records annually	All ages	Sweden	Longitudinal (since July 2005)	Prescribed drug(s)	NA	NA	NA	NA	NA	https://www.socialstyrelsen.se/en/statistics-and-data/registers/register-information/the-swedish-prescribed-drug-register/	[172]	Register of prescribed drug

(continued)

Table 2
(continued)

Sample	N	Age range	Country	Cross-sectional/ longitudinal	Clinical	Neuroimaging (MRI, PET MRI)	EEG/MEG	Genetics (genotyping, exome, WGS, twins)	Genomics	Smartphones and sensors	Other omics	Website/Data Transfer Agreement	Reference article(s)	Specificities
Swedish Dementia Registry (SDR)	More than 100,000	27–103 (between 2007 and 2012)	Sweden	Longitudinal (since 2007)	Dementia	MRI, PET, CT	EEG	NA	NA	NA	NA	www.svedem.se	[176]	Register of dementia
Swedish Neuro-Register (SNR)	Around 16,000 multiple sclerosis, over 1500 Parkinson's disease, and over 600 myasthenia gravis (numbers from 2015)	All age	Sweden	Longitudinal (since 2004)	Motor neuron disease, MS	MRI, PET, and CT	MEG	Genotyping	NA	NA	NA	www.neuroreg.se	[177]	Register of MND (especially MS)
Swedish Stroke Register (Riks-Stroke)	Around 29, 000 cases (around 21,000 stroke and 8000 TIA) annually (statistics in 2020)	Mean age 75 for stroke; mean age 74 for TIA	Sweden	Longitudinal (since 1994)	Stroke and transient ischemic attack (TIA)	MRI and CT scan	NA	NA	NA	NA	NA	www.riksstroke.org/	[179]	Register of stroke and TIA; includes electrocardiograms
Brisbane Adolescent Twin Sample (BATS)	4000	>11	Australia	Cross-sectional and longitudinal	Population sample. Self-report mental health, cognition, substance use, and personality measures	NA	EEG (15 channels, eyes closed resting; N ~ 1000)	Pedigree (twin/ sibling), genotyping	NA	Wrist-worn accelerometer (N ~ 130)	NA	https:// imaginggenomics. net.au/	[79, 223–225]	Also known as the Brisbane Longitudinal Twin Study (BLTS). Includes participants from the Queensland Twin Registry
mPower	~8000	>18	USA	Longitudinal	Parkinson's disease (subsample self-identified professional diagnosis)	NA	NA	NA	NA	iPhone application	NA	https://parkinsonmpower.org/team https://www.synapse.org/#!Synapse:syn4993293/wiki/247860	[226]	Sample size varies across surveys and tasks completed

We hope this overview could be useful to the readers wanting to replicate findings, maximize sample size and statistical power, develop and apply methods that utilize multi-level data, or even select the most relevant dataset to tackle a research question. We also hope this encourages the collection of new data shared with the community while ensuring interoperability with the existing datasets.

3 Neuroimaging

3.1 Magnetic Resonance Imaging (MRI)

Brain magnetic resonance images are 3D images that measure brain structure (T1w, T2w, FLAIR, DWI, SWI) or function (fMRI). The different MRI sequences (or modalities) can characterize different aspects of the brain. For example, T1w and T2w offer the maximal contrast between tissue types (white matter, gray matter, and cerebrospinal fluid), which can yield structural/shape/volume measurements. They can also be used in conjunction with an injection of a contrast agent (e.g., gadolinium) for detecting and characterizing various types of lesions. FLAIR is also useful for detecting a wide range of lesions (e.g., multiple sclerosis, leukoaraiosis, etc.). SWI focuses on the neurovascular system, while DWI allows measuring the integrity of the white matter tracts. Functional MRI measures BOLD (blood oxygen level dependent) signal, which is thought to measure dynamic oxygen consumption in the different brain regions. Of note, fMRI consists of a series of 3D images acquired over time (typically 5–10 min).

Brain MRI is available as a series of DICOM files (brain slices, traditional format of the MRI machines) or a single NIfTI (single 3D image) format (*see* Table 1). The two formats are roughly equivalent, and most image processing pipelines allow both data sources as input. MR images are composed of voxels (3D pixel), and their size (e.g., 1 × 1 × 1 mm) corresponds to the image resolution.

In practice, most MR images are archived and shared via web-based applications and more rarely using specific software (e.g., UKB). The two major web platforms are XNAT (eXtensible Neuroimaging Archive Toolkit) [9], an open-source platform developed by the Neuroinformatics Research Group of the Washington University School of Medicine (Missouri, (1, 2)), and IDA (Image and Data Archive) created by the Laboratory of NeuroImaging of the University of South California (LONI, https://loni.usc.edu/). Of note, XNAT also allows to perform some image processing [9].

The neuroimaging community has developed BIDS (Brain Imaging Data Structure), a standard for MR image organization to accommodate multimodal acquisitions and facilitate processing.

In practice, few datasets come in BIDS format, and tools have been developed to assist with download and conversion (e.g., https://clinica.run) [10].

We have listed a handful of datasets (*see* Table 1), which is far from being exhaustive but aims at summarizing some of the largest and/or most used samples. Our selection aims at presenting diverse and complementary samples in terms of age range, populations, and country of origin.

First, we have described three clinical elderly samples from the USA and Australia, with a focus on Alzheimer's disease and cognitive disorders. The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2004 and funded by a partnership between private companies, foundations, the National Institute of Health, and the National Institute for Aging. ADNI is a longitudinal study, with data collected across 63 sites in the USA and Canada. To date, four phases of the study have been funded, which makes ADNI one of the largest clinical neuroimaging samples to study Alzheimer's disease and cognitive impairment in aging. ADNI collected a wide range of clinical, neuropsychological, cognitive scales as well as biomarkers, in addition to multimodal imaging and genotyping data [11]. Sites contribute data to the LONI, which is automatically shared with approved researchers without embargo. The breadth of data available and its accessibility have made ADNI one of the most used neuroimaging samples, with more than 1000 scientific articles published to date.

The Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) started in 2006 and has since recruited about 1100 participants over 60 years of age, who have been followed over several years (*see* Table 1) [12]. AIBL collected data across the different Australian states and, similar to ADNI, consisted in an in-depth assessment of individual cognition, clinical status, genetics, genomics, as well as multimodal brain imaging [12]. In 2010, AIBL partnered with ADNI to release the AIBL imaging subset and selected clinical data via the LONI platform. Having the same MRI protocols and similar fields collected, AIBL represents a great addition to the ADNI study, by boosting statistical power or allowing for replication. The full clinical information as well as genetics, genomics, and wearable data (actigraphy watches) are not available via the LONI and require a direct application to the Commonwealth Scientific and Industrial Research Organisation (CSIRO) (*see* Table 1).

The Open Access Series of Imaging Studies v3 (OASIS3) is another longitudinal sample comprising almost 1100 adult participants (*see* Table 1) [13]. Its main focus is around aging and neurological disorders, and the application/approval process is extremely fast (typically a couple of days). OASIS3 is hosted on XNAT and is the third dataset to be made available by the Washington University in Saint Louis (WUSTL) Knight Alzheimer's Disease Research

Center (ADRC), although the three datasets are not independent and cannot be analyzed together. Contrary to ADNI and AIBL, OASIS3 is a retrospective study that aggregates several research studies conducted by the WUSTL over the past 30 years. As a result, the data collected may vary from one individual to the next, with a variable time window between visits. In that sense, OASIS3 resembles data from clinical practice, with individual specific care/assessment pathways.

The UK Biobank (UKB) imaging study [14] is the largest brain imaging study to date, with around 50,000 individuals already imaged (target of 100,000). The imaging wave complements the wealth of data already collected in the previous waves (*see* Table 1; see also Subheading 5 for a description of the full dataset). Considering the sheer size of the data, the biobank shares raw and processed images as well as structured data (measurements of regions of interest) [15]. Data is accessible upon request by all bona fide researchers, with certified profiles. Data access requires payment of a fee, which only aims to cover the biobank functioning costs. The UKB has developed proprietary tools for a secure download and data management (https://biobank.ndph.ox.ac.uk/showcase/download.cgi).

The Adolescent Brain Cognitive Development (ABCD) is an ongoing longitudinal study of younger individuals, recruited aged 9–10 years and who will be followed over a decade [16, 17]. The ABCD focuses on cognition, behavior, and physical and mental health (e.g., substance use, autism, ADHD) of adolescents. It includes self- and parental rating of the adolescents as well as a description of the familial environment [17]. ABCD data is hosted on the NIMH data archive and requires obtaining and maintaining an NDA Data Use Certification, which requires action from a signing official (SO) from the researcher institution, as defined in the NIH eRA Commons (https://era.nih.gov/files/eRA-Commons-Roles-10-2019.pdf).

The Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) disease working groups have stemmed from the ENIGMA genetics project (see Subheading 5.3) to perform worldwide neuroimaging studies for a wide range of disorders (e.g., major depressive disorder [18], attention-deficit hyperactivity disorder [19], autism [20], post-traumatic stress disorder [21], obsessive–compulsive disorder [22], substance dependence [23], schizophrenia [24], bipolar disorder [25]) as well as traits of interest (e.g., sex, healthy variation [26]); see [27] for a review. Each working group may conduct simultaneously several research projects, proposed and led by its members. Each site of the working group choses the project(s) they contribute to and performs the analyses. Of note, most ENIGMA working groups still rely on a meta-analytic framework, even if recent projects (e.g., machine learning) now require sharing data onto a central server. Interested

researchers can contribute new data and propose analyses or new image processing pipelines to the different working groups. The ENIGMA samples typically comprise thousands of participants (controls and/or cases; *see* Table 1), and data are inherently heterogeneous, each site having specific recruitment and protocols.

Other neuroimaging MRI datasets have focused on twins and siblings (*see* Subheading 5.1) and include the Queensland Twin Imaging (QTIM) study, the Queensland Twin Adolescent Brain Project, the Vietnam Era Twin Study of Aging (VETSA) [28], and the Human Connectome Project (HCP) [29] (*see* Subheading 5.1). In addition, there are many more datasets available on neurological disorders, which may be explored via XNAT, LONI, or the Dementias Platform UK (DPUK), to name a few, PPMI (Parkinson' Progression Markers Initiative) [30], MEMENTO (deterMinants and Evolution of AlzheiMer's disEase aNd relaTed disOrder) [31], EPAD (European Prevention of Alzheimer's Dementia) [32], and ABIDE (Autism Brain Imaging Data Exchange) [33, 34].

3.2 Positron Emission Tomography (PET–MRI)

Positron emission tomography (PET) images are 3D images that highlight the concentration of a radioactive tracer administered to the patient. Here, we will focus on brain PET images, although other parts of the body may also be imaged. The different tracers allow to measure several aspects of brain metabolism (e.g., glucose) or spatial distribution of a molecule of interest (e.g., amyloid).

PET relies on the nuclear properties of radioactive materials that are injected in the patient intravenously. When the radioactive isotope disintegrates, it emits a photon that will be detected by the scanner. This signal is used to find the position of the emitted positrons which allow us to reconstruct the concentration map of the molecule we are tracing [35].

As for MRI, PET images are available as a series of DICOM files or a single NIfTI format. They are composed of voxels (3D pixel), and their size (e.g., 1 × 1 × 1 mm) corresponds to the image resolution. A BIDS extension has also been developed for positron emission tomography, in order to standardize data organization for research purposes.

PET is considered invasive due to the injection of the tracer, which results in very small risk of potential tissue damage. Overall, the quantity of radioactive isotope remains small enough to make it safe for most people, but this limits its widespread acquisition in research, especially on healthy subjects or in children. Moreover, PET requires to have a high-cost cyclotron to produce the radiotracers nearby because the half-life of the radioisotopes is typically short (between a few minutes to few hours).

Several tracers are used for brain PET imaging, one of the most common ones being the ^{18}F-fluorodeoxyglucose (^{18}F-FDG). ^{18}F-FDG concentrates in areas that consume a lot of glucose and will thus highlight brain metabolism. In practice, ^{18}F-FDG PET

images are often used to study neurodegenerative disease by revealing hypometabolism that characterizes some dementia [36, 37]. Other diseases such as epilepsy and multiple sclerosis can be studied through this modality, but since it is not part of clinical routine, data are rare, and we are not aware of publicly available datasets.

In whole-body PET scans, ^{18}F-FDG is used to detect tumors, which consumes a lot of glucose. However, the brain consumes a lot of glucose as part of its normal functioning, and brain tumors are not noticeable using this tracer. Instead, clinicians would use ^{11}C-choline that will also accumulate in the tumor area but is not specifically used by the brain otherwise. In addition to glycemic radiotracers, oxygen-15 is also used to measure blood flow in the brain, which is thought to be correlated with brain activity. In practice, this tracer is less used than ^{18}F-FDG because of its very short half-life. Other tracers are used to show the spatial concentration of specific biomarkers: for instance, ^{18}F-florbetapir (AV45), ^{18}F-flutemetamol (Flute), Pittsburgh compound B (PiB), and ^{18}F-florbetaben (FBB) are amyloid tracers used to highlight ß-amyloid aggregation in the brain, which is a maker of Alzheimer's disease. Finally, ^{11}C-5-hydroxytryptamine (5-HT) neurotransmitter is used to expose the serotonergic transmitter system.

We have made a non-exhaustive list of publicly available datasets containing PET scans with different tracers. Most datasets focused on neurodegenerative disorders and also collected brain MRI (see previous section). The Alzheimer's Disease Neuroimaging Initiative (ADNI) is one of the largest datasets with PET images for Alzheimer's disease [11]. ADNI used F-FDG-PET as well as PET amyloid tracers: FBB, AV45, and PiB. The Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) only collected amyloid tracers of PET images: PiB, AV45, and Flute [12]. The Open Access Series of Imaging Studies v3 (OASIS3) includes PET imaging from three different tracers, PIB, AV45, and ^{18}F-FDG [13].

In addition to those neurodegenerative datasets, PET is available in the Lundbeck Foundation Centre for Integrated Molecular Brain Imaging (CIMBI) database and biobank established in 2008 in Copenhagen, Denmark [38]. CIMBI shares structural MRI, PET, genetic, biochemical, and clinical data from 2000 persons (around 1600 healthy subjects and almost 400 patients with various pathologies). Tracer used for PET is the ^{11}C-5-HT which is relevant to study the serotonergic transmitter system. Applications to access the data can be made on their website by completing a form (*see* Table 2).

The ChiNese brain PET Template (CNPET) dataset has been developed by the Medical Imaging Research Group (https://biomedimg-dlut-edu.cn/), from Dalian University of Technology (China) [39]. The database contains 116 records of ^{18}F-FDG-PET

from healthy patients, which has been used to make a Chinese population-specific statistical parametric mapping (SPM, i.e., average template used for PET processing). The data used to build the PET brain template have been released and are available on NeuroImaging Tools and Resources Collaboratory (NITRC, https://www.nitrc.org/) platform.

4 EEG/MEG

Electroencephalography (EEG) measures the electrical activity of the brain [40–42]. Signals are captured through sensors distributed over the scalp (noninvasive) or by directly placing the electrodes on the brain surface, a procedure that requires a surgical intervention [43]. This technique is characterized by its high temporal resolution, enabling the study of dynamic processes such as cognition or the diagnosis of conditions such as epilepsy. Yet, EEG signals are nonstationary and have a non-linear nature, which makes it difficult to get useful information directly in the time domain. Nonetheless, specific patterns can be extracted using advanced signal processing techniques.

Another technique that captures brain activity is magnetoencephalography (MEG). This technology maps the magnetic fields induced above the scalp surface. Similar to EEG, MEG provides high time resolution, but it is preferentially sensitive to tangential fields from superficial sources [44, 45]. This could be considered as an advantage, since magnetic fields are less sensitive to tissue conductivities, facilitating source localization. However, MEG instrumentation is more expensive and not portable [46, 47].

During signal recording, undesirable potential coming from sources other than the brain may alter the quality of the signals. These artifacts should be detected and removed in order to improve pattern recognition. Multiple methods could be applied depending on the artifact to be eliminated: re-referencing with common average reference (CAR), ICA decomposition to remove other physiological sources as eye movements or cardiac components, notch filter to get rid of power line noise, and pass-band filtering to keep the physiological rhythms of interest, among others [48–51]. Other spatial filters such as common spatial pattern (CSP) for channel selection or filter bank CSP (FBCSP) for band elimination are largely used in motor decoding [52, 53].

Other signal processing tools allow the user to extract features describing relevant information contained in the signals. Subsequently, those patterns may be used as input for a classification pipeline. The target features vary according to the condition under study. Generally, the domain of clinical diagnostics focuses either on event-related potentials (ERP) or on spectral content of the signal [54, 55]. The first refers to voltage fluctuations associated

with specific sensory stimuli (e.g., P300 wave) or task, like motor preparation and execution, covert mental states, or other cognitive processes. The amplitude, latency, and spatial location of the resulting waveform activity reveal the underlying mental state [56]. On the other side, spectral analysis refers to the computation of the energy distribution of the signals in the frequency domain. Most spectral estimates are based on Fourier transform; this is the case of non-parametric methods, such as Welch periodogram estimation, which based their computation on data windowing [57].

Another approach is to study the interactions across sources (inferring connection between two electrodes by means of temporal dependency between the registered signals), which is known as functional connectivity. Multiple connectivity estimators have been developed to quantify this interaction [58]. Through these functional interactions, complex network analysis can also be implemented, where sensors are modeled as nodes and connectivity interactions as links [59–61].

EEG and MEG are essential to evaluate several types of brain disorders. One of the most documented is epilepsy, based on seizure detection and prediction [62–64]. Other neurological conditions can be characterized like Alzheimer's disease, associated with changes in signal synchrony [65, 66]. Furthermore, motor task decoding in brain–computer interfaces (BCI) offers a promising tool in rehabilitation [67]. This type of data, from healthy to clinical cases, can be found on multiple open-access repositories, such as Zenodo (https://zenodo.org) or PhysioNet ((1)), as well as via collaborative projects such as the BNCI Horizon 2020 (http://bnci-horizon-2020.eu), which gathers a collection of BCI datasets (*see* Table 2). These repositories are also valuable in that they contribute to establishing harmonization procedures in processing and recording. All dataset-collected informed consent and data were anonymized to protect the participants' privacy. Moreover, regulations may vary from one country to another, which require, for example, studies to be approved by ethics committees. Additionally, licensing (that define copyrights of the dataset) must be considered depending on the intended use of the open-access datasets.

Data come in different formats according to the acquisition system or the preprocessing software. The most common formats for EEG are .edf, .gdf, .eeg, .csv, or .mat files. For MEG, it is very often .fif and .bin (*see* Table 1). The different formats can create challenges when working with multiple datasets. Luckily, some tools have been developed to handle this problem, for example, FieldTrip [68] or Brainstorm [69] implemented in MATLAB, or the Python modules mne [70] and moabb [71]. Of note, these tools also contain sets of algorithms and utility functions for analysis and visualization.

5 Genetics

5.1 Twin Samples

Twins provide a powerful method to estimate the importance of genetic and environmental influences on variation in complex traits. Monozygotic (MZ, aka identical) twins develop from a single zygote and are (nearly) genetically identical. In contrast, dizygotic (DZ, aka fraternal) twins develop from two zygotes and are, on average, no more genetically related than non-twin siblings. In the classical twin design, the degree of similarity between MZ and DZ twin pairs on a measured trait reveals the importance of genetic or environmental influences on variation in the trait. Twin studies often collect several different data types, including brain MRI scans, assessments of cognition and behavior, self-reported measures of mental health and wellbeing, as well as biological samples (e.g., saliva, blood, hair, urine). Datasets derived from twin studies are text-based and include phenotypic data and background variables (e.g., individual and family IDs, sex, zygosity, age). Notably, the correlated nature of twin data (i.e., the non-independence of participants) should be considered during analysis as it may violate statistical test assumptions [72, 73].

Raw data is typically stored locally by the data owner, with de-identified data available upon request. In larger studies, data is stored and distributed through online repositories. Recently, the sharing of publicly available de-identified data with accompanying publications has become commonplace.

Several extensive twin studies combine imaging, behavioral, or biological data (*see* Table 2). These studies cover the whole life span (STR) as well as specific age periods, for example, children/adolescents (QTAB), young (QTIM, HCP-YA), middle-aged (VETSA), or older (OATS) adults.

The Swedish Twin Registry (STR) was established in the late 1950s with the primary aim to explore the effect of environmental factors (e.g., smoking and alcohol) on disorders [74]. Data were first collected through questionnaires and interviews with the twins and their parents. Later, the STR incorporated data from biobanks, clinical blood chemistry assessments, genotyping, health checkups, and linkages to various Swedish national population and health registers [74]. The STR is now one of the largest twin registers in the world [75] with information on more than 87,000 twin pairs (https://ki.se/en/research/the-swedish-twin-registry). It has been used extensively for the research of health and illness, including various neurological disorders, including dementia [76], Parkinson's disease [77], and motor neuron disease [78].

The Queensland Twin Adolescent Brain (QTAB, 2015–present) was enabled through funding from the Australian National Health and Medical Research Council (NHMRC). It focuses on the period of late childhood/early adolescence, with brain imaging,

cognition, mental health, and social behavior data collected over two waves (age 9–14 years at baseline, $N = 427$). A primary objective is to chart brain changes and the emergence of depressive symptoms throughout adolescence. Biological samples (blood, saliva), sleep (self-report), and motor activity measures (see Section 8) were also collected. Data is available from the project owners upon request.

The Queensland Twin IMaging (QTIM, 2007–2012) study, funded through the National Institutes of Health (NIH) and NHMRC, was a collaborative project between researchers from QIMR Berghofer Medical Research Institute, the University of Queensland, and the University of Southern California, Los Angeles. Brain imaging was collected in a large genetically informative population sample of young adults (18–30 years, $N > 1200$) for whom a range of behavioral traits, including cognitive function, were already characterized (as a component of the Brisbane Adolescent Twin Study, QIMR Berghofer Medical Research Institute [79]). Notably, the dataset includes a test–retest neuroimaging subsample ($n = 75$) to estimate measurement reliability. Data is available from the project owners upon request.

The Human Connectome Project Young Adult (HCP-YA, 2010–2015) study, funded by the NIH, is based at Washington University, University of Minnesota, and Oxford University. Investigators spent 2 years developing state-of-the-art imaging methods [29] before collecting high-quality neuroimaging, behavioral, and genotype data in ~1200 healthy young adult twins and non-twin siblings (22–35 years). HCP-YA data has been used widely in twin-based analyses, examining genetic influences on network connectivity [80], white matter integrity [81], and cortical surface area/thickness [82]. Open-access HCP-YA data is available from the Connectome Coordination Facility following registration (https://db.humanconnectome.org), with additional data use terms applicable for restricted data (e.g., family structure, age by year, handedness).

The Vietnam Era Twin Study of Aging (VETSA, 2003–present), funded by the NIH, started as a study of cognitive and brain aging but has since pivoted to the early identification of risk factors for mild cognitive impairment and Alzheimer's disease [28]. In addition to neuroimaging and cognitive data, the VETSA study includes health, psychosocial, and neuroendocrine data collected across three waves (baseline mean age 56 years, follow-up waves every 5–6 years) [83]. VETSA data is available following registration (https://medschool.ucsd.edu/som/psychiatry/research/VETSA/Researchers/Pages/default.aspx).

The Older Australian Twins Study (OATS, 2007–present) [84], funded by the NHMRC and Australian Research Council, is a longitudinal study of genetic and environmental contributions to brain aging and dementia. The project includes neuroimaging and

cognitive data collected across four waves (baseline mean age 71 years, follow-up waves every 2 years). OATS was expanded in wave 2 to include positron emission tomography (PET) scans to investigate the deposition of amyloid plaques in the brain. Data is available from the project owners upon request.

There is a wealth of twin studies worldwide in addition to those mentioned here (see [85] for an overview). Foremost is the Netherlands Twin Registry [86], a substantial data resource with dedicated projects investigating neuropsychological, biomarker, and behavioral traits. In addition, several extensive family/pedigree imaging studies exist, including the Genetics of Brain Structure and Function study [87] and the Diabetes Heart Study-Mind Cohort [88]. Further, the previously mentioned ABCD study [89] includes embedded twin subsamples.

Twin datasets have been used to estimate the heritability (the proportion of observed variance in a phenotype attributed to genetic variance) of phenotypes derived through machine learning, such as brain aging [90–92] and brain network connectivity [93]. Further, machine learning models have been trained to discriminate between MZ and DZ twins based on dynamic functional connectivity [94] and psychological measures [95]. In addition, machine learning has been used to predict co-twin pairs based on functional connectivity data [96].

5.2 Molecular Genetics

5.2.1 The UK Biobank

The UK Biobank (UKB) is one of the largest population-based cohorts, comprising nearly half a million adult participants (aged over 40 years at the time of recruitment), recruited across over 20 assessment centers in the UK. The UKB resource is accessible to the research community through application (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access) and, as of the end of 2021, counted more than 28,000 registered approved researchers worldwide. In 2021, UKB launched a cloud-based Research Analysis Platform (RAP), which provides computational tools for data visualization and analysis, thereby aiming to democratize access for researchers lacking such infrastructure. The associated fees for using the UKB resource include the yearly tier-based access fees, depending on the type of data accessed, as well as the cost of running the analyses and storing the generated data, while the storage of the UKB dataset itself is provided free of charge. Certain emerging datasets (e.g., whole exome and genome sequences) will be only available for analysis through the platform, both due the enormous size and tighter regulation around those datasets. Upon publication, researchers are required to return their results, including the methodology and any essential derived data fields, back to the UKB, which are subsequently incorporated into the resource in order to promote reproducible research.

The cohort is deeply phenotyped with thousands of traits measured across multiple assessments. The initial assessment visit took place from 2006 to 2010, where ~502,000 participants consented to participate (each keeping the right to withdraw their consent and be removed from the study at any time), completed the interview, filled questionnaires, underwent multiple measurements, and donated blood urine and saliva samples (see https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Reception.pdf).
The first repeat assessment was conducted in 2012–2013 and included approximately 120,000 participants. Next, the participants were invited to attend the imaging visits: the initial (2014+) and the first repeat imaging visit (2019+). So far, 50,000 initial imaging visits have been conducted, with a target to image 100,000 participants (10,000 repeat). The imaging data includes brain [14, 97], heart [98], and abdominal MRI scans [99], with both bulk images and image-derived measures available for analysis, as well as retinal OCT images, whole body MRI, and carotid ultrasound [100]. Finally, follow-up information from the linked health and medical records is regularly collected and updated in the resource, including data for COVID-19 research. The showcase of the available anonymous summary information is available at https://biobank.ndph.ox.ac.uk/showcase/.

The interim release of the genotyping data comprised ~150,000 samples and was released in 2015, followed by the full release of 488,000 genotypes in the middle of 2017. The available genotype data included variant calls from UK BiLEVE and UK Biobank Axiom arrays (autosomes, sex chromosomes, and mitochondrial DNA) as well as phased haplotype values and imputation to a combined panel of Haplotype Reference Consortium (HRC) and the merged UK10K and 1000 Genomes phase 3 reference panels [101], also known as v2 release. Subsequently, the v2 imputation was replaced by imputation to HRC and UK10K haplotype resource only (v3), after a problem was discovered for the set imputed to UK10K + 1000 Genomes panel (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100319). The genotypes of approximately 3% of the participants remained not assayed due to DNA processing issues. To note, ~50,000 individuals included in the interim genotype release were involved in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) project, and their genotypes were assayed on a different but very closely related array than the rest of the participants (https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/genotyping_qc.pdf). The UK BiLEVE focused on genetics of respiratory health, and the participants were selected based on lung function and smoking behavior [102].

Whole exome sequencing (WES) and whole genome sequencing (WGS) have been funded through the collaboration between the UK Biobank and biotechnology companies Regeneron and GlaxoSmithKline (GSK). The first UKB release of WES data

included 50,000 participants, prioritized based on the availability of MRI data, baseline measurements, and linked hospital and primary care records and enriched in patients diagnosed with asthma [103]. Recently (November 2021), the new data release included $N = 200{,}000$ WGS and $N = 450{,}000$ WES [104]. WGS for the remaining participants is currently underway. For all the past and future timelines, see https://biobank.ctsu.ox.ac.uk/showcase/exinfo.cgi?src=timelines_all.

Most of the UKB participants reported their ethnic background as White British/Irish or any other white background (~94%), which was coherent with the observed genetic ancestries [101]. For example, the ancestries identified from genetic markers showed a predominant European ancestry ($N \sim 464{,}000$), followed by South Asian (~12,000), African (~9000), and East Asian ancestry (~2500) [105]. As a population-based cohort, the UKB mostly comprises unrelated participants. While the pedigree information was not collected as a part of assessment, the genetic analysis has identified approximately 100,000 pairs of close relatives (third degree or closer, including 22,000 sibling pairs and 6000 parent–offspring pairs) [101]. This amount of relatedness is, however, larger than expected for a random sample from a population and reflects a participation bias toward the relatives of the participants. Moreover, the UKB sample is, on average, healthier, more educated, and less deprived than the general UK population [2].

5.3 Genetic Consortia

5.3.1 ENIGMA Consortium

The Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium was formed in 2009 with the goal of conducting large-scale neuroimaging genetic studies of human brain structure, function, and disease [27]. Currently, more than 2000 scientists from 400 institutions around the world with neuroimaging (including structural and functional MRI) and electroencephalography (EEG) data have joined the consortium and formed 50 working groups that focus on different psychiatric and neurological disorders as well as healthy variation, method development, and genomics [27].

To date, the ENIGMA Genetics Working Group (for an overview, see [106]) have conducted genome-wide association meta-analyses for hippocampal and intracranial volume [107–109], subcortical volume [110, 111], and cortical surface area and thickness [112]. The ENIGMA Genetics Working Group provides researchers imaging and genetic protocols to enable each group to conduct their own association analyses before contributing summary statistics to the meta-analysis. While these genome-wide association studies have focused on structural phenotypes and the analysis of common single nucleotide polymorphisms (SNPs), the ENIGMA EEG Working Group have recently conducted a genome-wide association meta-analysis for oscillatory brain activity [113], and the ENIGMA Copy Number Variant (CNV) Working Group,

which formed in 2015, is currently investigating the impact of rare CNVs beyond the 22q11.2 locus on cognitive, neurodevelopmental, and neuropsychiatric traits [114].

The sample sizes of the ENIGMA Genetics and CNV Working Groups continuously increase as new cohorts with MRI and genetic data join the consortium. As of 2020, the CNV Working Group sample comprises of 38 ENIGMA cohorts [114], while the latest Genetics Working Group genome-wide association meta-analysis [112] consisted of a discovery sample of 49 ENIGMA cohorts and the UK Biobank ($N = 33{,}992$ individuals of European ancestry), a replication sample of 2 European ancestry cohorts ($N = 14{,}729$ participants), and 8 ENIGMA cohorts of non-European ancestry ($N = 2994$ participants). This meta-analysis identified 199 genome-wide significant variants that were associated with either the surface area or thickness of the whole human cortex and 34 cortical regions with known functional specializations. They also found evidence that the genetic variants that influence brain structure also influence brain function, such as general cognitive function, Parkinson's disease, depression, neuroticism, ADHD, and insomnia [112].

Importantly, all imaging, EEG, and genetic (imputation and association analysis) protocols are freely available from the ENIGMA website (http://enigma.ini.usc.edu/). However, to access the summary statistics for each published genome-wide association meta-analysis, researchers need to complete an online Data Access Request Form (http://enigma.ini.usc.edu/research/download-enigma-gwas-results/). If a researcher wants to propose new genetic analyses that cannot be conducted with these publicly available summary statistics, they need to become a member of ENIGMA. Researchers can join the consortium by (a) contributing a cohort with MRI and genetic data, (b) collaborating with another research group that does have MRI and genetic data, or (c) contributing their expertise in genomic or methodological areas that are inadequately addressed by other consortium members. Of note, since storage of the MRI and genetic data is not centralized, each ENIGMA cohort can choose to contribute or not to new proposed analyses.

5.3.2 The Psychiatric Genomics Consortium (PGC)

The Psychiatric Genomics Consortium (PGC) began in 2007. The central idea of the PGC is to use a global cooperative network to advance genetic discovery in psychiatric disorders in order to identify biologically, clinically, and therapeutically meaningful insights. To date, the PGC is one of the largest, most innovative, and productive consortia in the history of psychiatry. The Consortium now consists of workgroups on 11 major psychiatric disorders, a Cross-Disorder Workgroup, and a Copy-Number Variant Workgroup. In addition, the PGC provides centralized support to the PGC researchers with a Statistical Analysis Group, Data Access

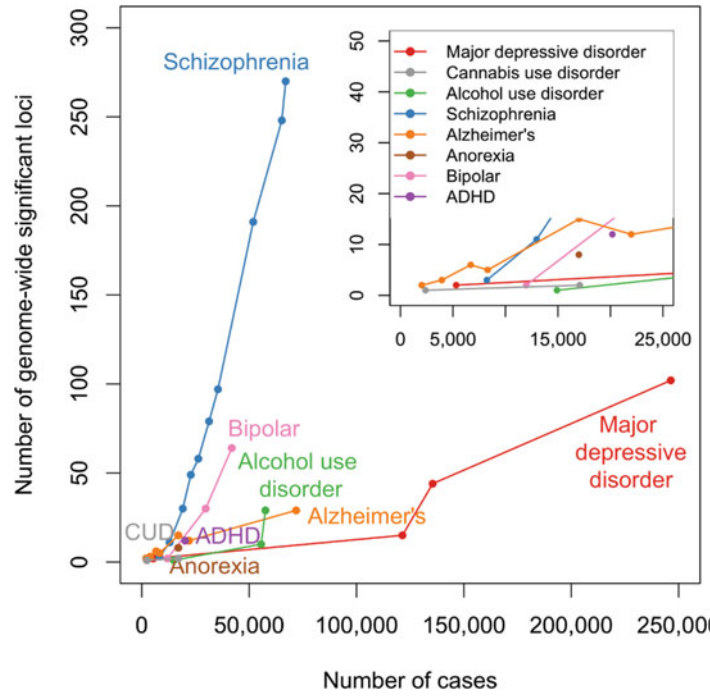

Fig. 1 PGC discoveries over time

Committee, and Dissemination and Outreach Committee. To increase ancestral diversity, the Consortium established the Cross-Population Workgroup in 2017 for outreach and developing/deploying trans-ancestry analysis methods [115]. The Consortium outreach expands ancestry diversity by adding non-European cases and controls. The PGC continues to unify the field and attract outstanding scientists to its central mission (800+ investigators from 150+ institutions in 40+ countries). PGC work has led to *320 papers*, many in high-profile journals (*Nature* 3, *Cell* 5, *Science* 2, *Nat Genet* 27, *Nat Neurosci* 9, *Mol Psych* 37, *Biol Psych* 25, *JAMA Psych* 12). The full results from all PGC papers are freely available, and the findings have fueled analyses by non-PGC investigators (sample sizes and findings for eight major psychiatric disorders are summarized in Fig. 1)

Computation and data warehousing for the PGC are non-trivial. The PGC uses the Netherlands "LISA" computing cluster. LISA compute cluster in Amsterdam which is used for most analyses (occasional analyses are done on other clusters, but 90% of PGC computation is done on LISA). The core software is the RICOPILI data analytic pipeline [116]. This pipeline has explicit written protocols for uploading data to the cluster in the Netherlands that one uses for quality control, imputation, analysis, meta-analysis, and bioinformatics. The actual mega-analyses are conducted by PGC analysts under the direction of a senior statistical geneticist, geneticist, or highly experienced analyst.

The PGC has a proven commitment to open-source, rapid progress science. All PGC results are made freely available as soon as a primary paper is accepted (GWAS summary statistics available at https://www.med.unc.edu/pgc/download-results/). The researchers can obtain access to the individual-level data either through controlled-access repositories (e.g., the Database of Genotypes and Phenotypes, dbGaP, or the European Genome-phenome Archive) or via the PGC streamlined process for secondary data analyses (https://www.med.unc.edu/pgc/shared-methods/data-access-portal/) [117].

PGC analyses have always been characterized by exceptional rigor and transparency. PGC analysts will enhance this by publishing markdown notebooks for all papers on the PGC GitHub site (https://github.com/psychiatric-genomics-consortium) to enable precise reproduction of all analyses (containing code, documentation of QC decisions, analyses, etc.).

5.4 Exome and Whole Genome Sequencing: Trans-Omics for Precision Medicine (TOPMed)

The Trans-Omics for Precision Medicine (TOPMed) program, sponsored by the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute (https://topmed.nhlbi.nih.gov), is part of a broader Precision Medicine Initiative, which aims to provide disease treatments tailored to an individual's unique genes and environment. TOPMed contributes to this Initiative through the integration of whole genome sequencing (WGS) and other omics data. The initial phases of the program focused on whole genome sequencing of individuals with rich phenotypic data and diverse backgrounds. The WGS of the TOPMed samples was performed over multiple studies, years, and sequencing centers [118, 119]. Available data are processed periodically to produce genotype data "freezes." Individual-level data is accessible to researchers with an approved dbGaP data access request (https://topmed.nhlbi.nih.gov/data-sets), via Google and Amazon cloud services. More information about data availability and how to access it can be found on the dataset page (https://topmed.nhlbi.nih.gov/data-sets).

As of September 2021, TOPMed consists of ~180 K participants from >85 different studies with varying designs. Prospective cohorts provide large numbers of disease risk factors, subclinical disease measures, and incident disease cases; case-control studies provide improved power to detect rare variant effects. Most of the TOPMed studies focus on HLBS (heart, lung, blood, and sleep) phenotypes, which leads to 62 K (~35%) participants with heart phenotype, 50 K (~28%) with lung data, 19 K (~11%) with blood, 4 K (~2%) with sleep, and 43 K (~24%) for multi-phenotype cohort studies. TOPMed participants' diversity is assessed using a combination of self-identified or ascriptive race/ethnicity categories and observed genetics. Currently, 60% of the 180 K sequenced

participants are of non-European ancestry (i.e., 29% African ancestry, 19% Hispanic/Latino, 8% Asian ancestry, 4% other/multiple/unknown).

Whole genome sequencing is performed by several sequencing centers to a median depth of 30× using DNA from blood, PCR-free library construction, and Illumina HiSeq X technology (https://topmed.nhlbi.nih.gov/group/sequencing-centers). Randomly selected samples from freeze 8 were used for whole exome sequence using Illumina v4 HiSeq 2500 at an average 36.4× depth. A trained machine learning algorithm with known variants and Mendelian inconsistent variants is applied by the Informatics Research Centre for joint genotype calling across all samples to produce genotype data "freezes" (https://topmed.nhlbi.nih.gov/group/irc). In TOPMed data freeze 8 ($N \sim 180$ K) (https://topmed.nhlbi.nih.gov/data-sets), variant discovery identified 811 million single nucleotide variants and 66 million short insertion/deletion variants. In the latest data freeze 9 (https://topmed.nhlbi.nih.gov/data-sets), variant discovery was initially made on ~206 K samples including data from Centers for Common Disease Genomics (CCDG). This data was re-subset to ~158,470 TOPMed samples plus 2504 from 1000 Genomes samples were used for variant re-discovery. Then, a total of 781 million single nucleotide variants and 62 million short insertion/deletion variants were identified and passed variant quality controls. These variant counts in freeze 9 are slightly smaller than that of freeze 8 due to monomorphic sites in TOPMed samples. A series of data freezes is being made available to the scientific community as genotypes and phenotypes via dbGaP (https://www.ncbi.nlm.nih.gov/gap/); read alignments are available via the Sequence Read Archive (SRA) and variant summary information via the Bravo variant server (https://bravo.sph.umich.edu/freeze8/hg38/) and dbSNP (https://www.ncbi.nlm.nih.gov/snp/).

TOPMed studies provide unique opportunities for exploring the contributions of rare and noncoding sequence variants to phenotypic variation. For instance, [119] used 53,831 samples from freeze 5 (https://topmed.nhlbi.nih.gov/data-sets) to investigate the role of rare variants into mutational processes and recent human evolutionary history. The recent TOPMed freeze 8 were used (together with WGS from the UK Biobank) to assess effect size of casual variants for gene expression using 72 K African American and ~298 K European American [120]. Similarly, a large set of multi-ethnic samples from freeze 5, 8, and 9 were used to develop comprehensive tools such as the STAAR and SCANG pipelines, which are used to identify noncoding rare variants [121] and to build predictive models for protein abundances [122] and discovery of causal genetic variants for different phenotypes [123, 124]. Overall, the Trans-Omics for Precision Medicine (TOPMed) program has the potential to help in improving

diagnosis, treatment, and prevention of major diseases by adding WGS and other "omics" data to existing studies with deep phenotyping.

6 Genomics

6.1 Methylation

DNA methylation (DNAm) is a covalent molecular modification by which methyl groups (CH_3) are added to the DNA. In vertebrates—and eukaryotes in general—the most common methylation modification occurs at the fifth carbon of the pyrimidine ring (5mC) at cytosine–guanine dinucleotides (CpG). Most bulk genomic methylation patterns are stable across cell types and throughout life, changing only in localized contexts, for example, due to disease-associated processes.

There are numerous ways of measuring DNAm at a genome-wide level, with bisulfite conversion-based methods being the most popular in the field of epidemiological epigenetics. These methods consist of bisulfite-induced modifications of genomic DNA, which results in unmodified cytosine nucleotides being converted to uracil, while 5mC remain unaffected. Of all these bisulfite conversion-based technologies—including sequencing-based methods—hybridization arrays are the most widely used, primarily due to their low cost and high-throughput nature.

The current Illumina Infinium® HumanMethylation450 (or 450 K) and Illumina Infinium® HumanMethylation850 (or EPIC) arrays assess around 450,000 and 850,000 methylation sites across the genome, respectively, covering 96% of the CpG islands (i.e., genomic regions with high CpG frequency), 92% of the CpG islands' shores [125, 126] (<2 kb flanking CpG Islands), and 86% of the CpG islands' shelves (<2 kb flanking outward from a CpG shore), which have been shown to be more dynamic than CpG islands [127]. Although most current studies have used the 450 K array [128], the EPIC array covers >90% of the 450 K sites plus additional CpG sites in the enhancer regions identified by the ENCODE and FANTOM5 projects [129].

After probe hybridization and extension steps, the array is scanned, and the intensities of the unmethylated and methylated bead types are measured. DNAm values are then represented by the ratio of the intensity of the methylated bead type to the combined locus intensity. These are known as *beta (β)* values and are continuous variables between 0 and 1 (Equation 1), although a value of 1 is impossible to achieve in practice, due to the addition of a stabilizing α offset (to handle low-intensity signals):

Equation 1 DNA methylation β values as measured by the Illumina Infinium® methylation arrays M = methylated intensity, U = unmethylated intensity, α = arbitrary offset to handle signals with low readings (usually 100)

$$\beta = \frac{M}{M + U + \alpha} \quad (1)$$

These raw intensities are then stored in binary IDAT files (one for each of the red and green channels). The bulk of each file consists of four fields: the ID of each bead type on the array, the mean and standard deviation of their intensities, and the number of beads of each type, generated per sample. This raw data format allows for flexible use, including differing preprocessing strategies [130]. However, these files are usually not readily available in public repositories (e.g., Gene Expression Omnibus [131] or GEO), due to their large size. For example, a compressed .tar file of IDATs for a sample size of around 700 individuals, measured with EPIC arrays, is about 10 Gb. Instead, researchers usually upload the processed DNAm β values (following normalization) as compressed .txt or .csv files with columns representing samples and rows the measured *loci*. This can be a problem for reproducibility, as different research groups tend to prefer their own preprocessing or normalization methods—and there are many [132]! On this note, there has been a recent push in the field, for standardization of DNAm array preprocessing pipelines, including the user-friendly *Meffil* pipeline [133].

Reproducibility and interpretation of DNAm studies are subject to additional factors outside of data processing methods. For comparison, genetic data is (mostly) germline determined and can be assumed to be randomly assigned with respect to characteristics of individuals. Thus, a case-control (or cross-sectional) design has an inference of association through causality and can convey information of liability to disease. This contrasts with DNAm data which is a reversible process influenced by a large range of biological, technical, and environmental factors (e.g., medication and complications of the disease itself) and is thus more susceptible to spurious cryptic association or reverse causation [134, 135]. DNAm studies will therefore benefit from longitudinal designs, both for biomarker discovery and mechanistic insights [134, 136].

Reed et al. [137] provide one good example of this. Briefly, the authors generated a DNAm score for body mass index (BMI) within the ARIES subsample of the Avon Longitudinal Study of Parents and Children birth cohort (ALSPAC), using effect sizes of 135 CpG sites from a published meta-analysis of DNAm and BMI [138]. Using multiple time points for matched mothers and children using linear and cross-lagged models to explore the causal relationship between phenotypic BMI and the DNAm scores, they

found a strong linear association within time points [137]. However, when testing for temporal associations, DNAm scores at earlier time points showed no association with future BMI, indicating that a DNAm score generated from a reference cross-sectional study performs better as a biomarker of extant BMI, but poorly as a predictor for future BMI.

In Table 2, we have compiled a list of the largest and/or most used DNAm array datasets—including the Genetics of DNA Methylation Consortium (goDMC), an international collaboration of human epidemiological studies that comprises >30,000 study participants with genetic and DNAm array data [139]. These samples are usually integrated in larger genetic/epidemiological studies, except for perhaps the NIH Roadmap Epigenomics Mapping Consortium [140], which was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research, and the BLUEPRINT project [141, 142], which aims to generate at least 100 reference epigenomes of distinct types of hematopoietic cells from healthy individuals and of their malignant leukemic counterparts. Lastly, in contrast to genetic data, the de-identified DNAm data—either raw or preprocessed—is typically open access in public repositories such as GEO [131], or dbGAP [143], or the web portals provided by the respective projects. However, access to accompanying phenotypic data may require additional approval by the managing committees of each individual project.

6.2 Gene Expression Data: GTEx

Launched in 2010, the Genotype-Tissue Expression (GTEx) project is an ongoing effort that aims to characterize the genetic determinants of tissue-specific gene expression [144]. It is a resource database available to the scientific community, which is comprised of multi-tissue RNA sequencing (RNA-seq: gene expression) and whole genome sequence (WGS) data collected in 17,382 samples across 54 tissue types from 948 postmortem donors (version 8 release). Sample size per tissue ranges from $n = 4$ in kidney (medulla) to $n = 803$ in skeletal muscle. The majority of donors are of European ancestry (84.6%) and male (67.1%) with ages ranging from 20–70 years old. The primary cause of death for donors 20–39 years old was traumatic injury (46.4%) and heart disease for donors 60–70 years (40.9%).

Data is constantly being added to the database using sample data from the GTEx Biobank. For example, recent efforts have focused on gene expression profiling at the single-cell level to achieve a higher resolution understanding of tissue-specific gene expression and within tissue heterogeneity. As a result, single-cell RNA-seq (scRNA-seq) data was generated in 8 tissues from 25 archived, frozen tissue samples collected on 16 donors. Further, the Developmental Genotype-Tissue Expression (dGTEx) project (https://dgtex.org/) is a relatively new extension of GTEx that was

launched in 2021 that aims to understand the role of gene expression at four developmental time points: postnatal (0–2 years of age), early childhood (2–8 years of age), pre-pubertal (8–12.5 years of age), and post-pubertal (12.5–18 years of age). It is expected that molecular profiling (including WGS, bulk RNA-seq, and, for a subset of samples, scRNA-seq) will be performed on 120 relatively healthy donors (approximately 30 donors per age group) in 30 tissues. Data from this study would provide, for example, a baseline for gene expression patterns in normal development for comparison against individuals with disease.

GTEx provides extensive documentation on sample collection, laboratory protocols, quality control and standardization, and analytical methods on their website (https://gtexportal.org/home/). This allows for replication of their protocols and procedures in other cohorts to aid in study design and for researchers to further interrogate the GTEx data to answer more specific scientific questions. Processed individual-level gene expression data is made freely available on the GTEx website for download, while controlled access to individual-level raw genotype and RNA sequencing data are available on the AnVIL repository following approval via the National Center for Biotechnology Information's database of Genotypes and Phenotypes (dbGAP, dbGaP accession phs000424), a data archive website that stores and distributes data and results investigating the relationship between genotype and phenotype (https://www.ncbi.nlm.nih.gov/gap/). Clinical data collected for each donor is categorized into donor-level (demographics, mediation use, medical history, laboratory test results, death circumstances, etc.) and sample-level (tissue type, ischemic time, batch ID, etc.) data and is also available through dbGAP.

Over the many years, data from the GTEx project has provided unprecedented insight into the role genetic variation plays in regulating gene expression and its contribution to complex trait and disease variation in the population. The latest version 8 release from GTEx comes with a comprehensive catalogue of variants associated with gene expression, or eQTLs (expression quantitative trait loci), across 49 tissues or cell lines (derived from 15,201 samples and 838 donors) (GTEx Consortium, 2020). This analysis has demonstrated that gene expression is a highly heritable trait, with millions of genetic variants affecting the expression of thousands of genes across the genome. These pairwise gene variant associations can be classified as either *cis*- or *trans*-eQTLs, which describes proximal (i.e., within a predefined window of the target gene) or distal (i.e., beyond the predefined window or on a different chromosome from the target gene) genetic control, respectively. Indeed, it has been shown that 94.7% of all protein-coding genes have at least one *cis*-eQTL. In addition, 43% of genetic variants (minor allele frequency $> 1\%$) have been found to affect gene expression in at least one tissue, and the majority of *cis*-eQTLs appear to be shared

across the sexes and ancestries (GTEx Consortium, 2020). Relatively few *trans*-eQTLs have been identified due to limitations in sample sizes; however, these typically affect gene expression in one or very few tissues, with about a third of *trans*-eQTLs mediated by *cis*-eQTLs [144]. Importantly, GTEx provides full eQTL summary statistics for download and an interactive portal (https://gtexportal.org/home/) for quick searches. As most trait-associated loci identified in genome-wide association studies (GWAS) are in noncoding regions of the genome, the eQTL data generated by GTEx has been leveraged to provide insight into the genetic and molecular mechanisms that underlie complex traits and diseases. Indeed, GWAS trait-associated variants are enriched for *cis*-eQTLs, and genetic variants that affect multiple genes in multiple tissues are found to also affect many complex traits (GTEx Consortium, 2020). This indicates that *cis*-eQTLs have a high degree of pleiotropy and exert their effect on complex traits and diseases by regulating proximal gene expression.

In addition to the comprehensive catalogue of multi-tissue eQTLs to understand gene regulation, additional flagship GTEx studies include understanding sex-biased gene expression across tissues [145], functional rare genetic variation [146], cell type-specific gene regulation [147], and predictors of telomere length across tissues [148].

The extensive publicly available data generated by the GTEx project is a valuable resource to the scientific community and will allow for further data interrogation for many years to come.

7 Electronic Health Records

7.1 Clinical Data Warehouse: Example from the Parisian Hospitals (APHP)

Clinical data warehouses (CDW) gather electronic health records (EHR), which can gather demographic data, results from biological tests, prescribed medications, and images acquired in clinical routine, sometimes for millions of patients from multiple sites. CDW can allow for large-scale epidemiological studies, but they may also be used to train and/or validate machine learning (ML) and deep learning (DL) algorithms in a clinical context. For example, several computer-aided diagnosis tools have been developed for the classification of neurodegenerative diseases. One of their main limitations is that they are typically trained and validated using research data or on a limited number of clinical images [149–154]. It is still unclear how these algorithms would perform on large clinical dataset, which would include participants with multiple diagnoses and more generally heterogeneous data (e.g., multiple scanners, hospitals, populations).

One of the first CDW in France was launched in 2017 by the AP-HP (Assistance Publique – Hôpitaux de Paris), which gathers most of the Parisian hospitals [155]. They obtained the

authorization of the CNIL (Commission Nationale de l'informatique et des Libertés, the French regulatory body for data collection and management) to share data for research purposes. The aim is to develop decision support algorithms, to support clinical trials, and to promote multicenter studies. The AP-HP CDW keeps patients updated about the different research projects through a portal (as authorized by CNIL), but, according to French regulation, active consent was not required as these data were acquired as part of the routine clinical care of the patients.

Accessing the data is possible with the following procedure. A detailed project must be submitted to the Scientific and Ethics Board of the AP-HP. If the project holders are external to the AP-HP, they have to sign a contract with the Clinical Research and Innovation Board (Direction de la Recherche Clinique et de l'Innovation). Once the project is approved, data are extracted and pseudo-anonymized by the research team of the AP-HP. Data are then made available in a specific workstation via the Big Data Platform, which is internal to the AP-HP. The Big Data Platform supports several research environments (e.g., JupyterLab Environment, R, MATLAB) and provides computational power (CPUs and GPUs) to analyze the data.

An example of the research possible using such CDW is the APPRIMAGE project, led by the ARAMIS team at the Paris Brain Institute. The project was approved by the Scientific and Ethics Board of the AP-HP in 2018. It aims to develop or validate algorithms that predict neurodegenerative diseases from structural brain MRI, using a very large-scale clinical dataset. The dataset provided by the AP-HP gathers all T1w brain MRI of patients aged more than 18 years old, collected since 1980. It therefore consists of around 130,000 patients and 200,000 MRI which were made available via the Big Data Platform of the AP-HP. Of note, clinical data was available for only 30% of the imaged participants (>30,000 patients) as it relies on the ORBIS Clinical Information System (Agfa HealthCare), installed more recently in the hospitals. The sheer size of the data poses obvious computational challenges, but other difficulties include harmonizing clinical reports collected in the different hospitals or handling the general heterogeneity of the data (e.g., hospitals, acquisition software, populations). To tackle this issue, we have developed a pipeline for the quality control of the MR images [156].

7.2 Swedish National Registries

In Sweden, a unique 10-digit personal identification number has been assigned to each individual at birth or migration since 1947, which allows linkages across different Swedish population and health registers with almost 100% coverage [157]. The Swedish Total Population Register (TPR) was established in 1968 and is maintained by Statistics Sweden to obtain data on major life events, such as birth, vital status, migration, and civil status [158]. TPR is a

key source to provide basic information in medical and social research in Sweden. The Swedish Population and Housing Censuses (1960–1990) and the Swedish Longitudinal Integrated Database for Health Insurance and Labour Market Studies (Swedish acronym LISA) (since 1990) provide information on demographic and socioeconomic status for the Swedish population, including the highest attained educational level and household income [159]. The Swedish Multi-Generation Register (MGR) provides information on familial links for individuals born since 1932 onward in Sweden [160], which makes it possible to perform family studies to investigate familial risk of different health outcomes and control for familial confounding when needed.

The Swedish National Patient Register (NPR) is a valuable source for medical research, which has since 1964 collected data on inpatient care (nationwide coverage since 1987) and outpatient care (more than 85% of the entire country since 2001) [161]. Diagnoses are according to the Swedish revisions of the International Classification of Disease codes (ICD codes). The positive predictive value of the diagnoses is high, ranging from 85% to 95%, in NPR [161]. NPR has been used in studies of different diseases including many neurological disorders such as Alzheimer's disease [162], Parkinson's disease [163], and amyotrophic lateral sclerosis [164]. The Swedish Cancer Register (SCR) has been used extensively in Swedish cancer research, especially cancer epidemiology. SCR was established in 1958 and includes data on all newly diagnosed malignant and benign tumors, including different kinds of brain tumors [165, 166]. The Swedish Medical Birth Register (MBR) was established in 1973 and contains information on almost all deliveries (from prenatal to postnatal) in Sweden [167]. MBR has contributed mainly to the reproductive epidemiologic research in Sweden and has also been used in epidemiological studies of diseases later in life including different neurological disorders [168, 169]. The Swedish Causes of Death Register (CDR) includes information on virtually all deaths in Sweden since 1952 [170] and has been used to identify various causes of death in medical research, including deaths due to neurological disorders [171]. The Swedish Prescribed Drug Register (PDR) was founded in July 2005 and provides information on all prescription drugs dispensed from pharmacies in Sweden [172, 173]. PDR has been used to study patterns of use as well as consequences of medication use, including memantine [174] and dopaminergic anti-Parkinson drug [175].

In addition to these general health registers, there are also hundreds of disease quality registers that are used for patient care and research in Sweden. For instance, the Swedish Dementia Registry (SDR) was established in 2007 to achieve high quality of diagnostics and care for patients with dementia [176]. The Swedish Neuro-Register (SNR) was founded in 2001 (web-based since

2004, originally named as the Swedish Multiple Sclerosis Quality Registry) with the primary aim to improve care of patients with different neurological disorders including multiple sclerosis, Parkinson's disease, severe neurovascular headache, myasthenia gravis, narcolepsy, epilepsy, inflammatory polyneuropathy, as well as amyotrophic lateral sclerosis in Sweden [177, 178]. The Swedish Stroke Register is one of the world's largest stroke registers, which was established in 1994 and has included data from almost all hospitals that admit acute stroke patients in Sweden [179].

In Sweden, individual-level data in public registers are strictly protected by several laws, including the Ethics Review Act, the General Data Protection Regulation (GDPR), and the Public Access to Information and Secrecy Act (OSL). The Swedish Ethical Review Authority (Etikprövningsmyndigheten in Swedish) assesses projects according to the Ethics Review Act and requires a Swedish responsible person (Forskningshuvudman in Swedish) for the research. In addition to ethical approval, the Statistics Sweden (SCB) and the National Board of Health and Welfare (Socialstyrelsen in Swedish) also need to make an assessment according to GDPR and OSL, to determine whether individual-level data can be made available for potential research purposes. It generally takes around 1–6 months from contact person assignment to delivery of microdata in the SCB (www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/) and around 3–6 months to process applications for individual-level data in the Socialstyrelsen (www.socialstyrelsen.se/en/statistics-and-data/statistics/).
According to standard legal provisions and procedures, the SCB and Socialstyrelsen only provide data to researchers working in Sweden, and researchers in other countries need to cooperate with Swedish colleagues to apply for the data.

According to the General Data Protection Regulation (GDPR), online access (e.g., through virtual machines) or transfer of individual-level data is allowed in countries of the European Union (EU) or European Economic Area (EEA), after proper legal agreements. Online access or transfer of individual-level data to an external partner in a third country outside EU/EEA is also permitted, if the third country has been approved by the European Commission and the external partner signs and complies with legal agreements that include requirements for how data must be protected, including Data Transfer Agreement (DTA), Data Processing Agreement (DPA), Material Transfer Agreement (MTA), as well as Research Collaboration Agreements.

8 Smartphone and Sensors

Smartphones and sensors allow for the unobtrusive collection of behavioral and physiological data. For instance, smartphones are commonly used in ecological momentary assessment (EMA) studies [180], resulting in continuous, real-time assessment of participant behavior, symptoms, and experiences. In addition, the built-in microphone and touchscreen of smartphones/tablets can record speech and motor movement. Recent advances in smartwatch technology has enabled many commercial devices (e.g., Fitbit, Garmin, Apple) to track physiological metrics (e.g., heart rate variability, pulse oximetry, temperature) in addition to traditional physical activity data (e.g., step count, Global Positioning System, exercise tracking). Sensors are also commonly used to collect data without requiring participant interaction. Wearable sensor devices (e.g., wrist-worn accelerometers) can collect data on sleep, activity, and physiology without burdening participants or influencing their behavior. Datasets derived from smartphone and sensor studies are typically text-based, though raw data may be proprietary. The analysis of smartphone and sensor data typically requires complex algorithms/machine learning approaches due to the complexity of data collected (in the frequency of hundreds of observations per second, from many different sensors collecting data simultaneously). Raw data is typically stored locally by the data owner, with de-identified data available upon request. In more extensive studies, data is stored and distributed through online repositories.

Several studies have collected real-world behavioral and physiological data using smartphone and sensor devices (*see* Table 2), including community twin studies (BATS, QTAB), large-scale biomedical databases (UK Biobank), and studies focusing on specific disorders (mPower).

The Brisbane Adolescent Twin Study (BATS) and the Queensland Twin Adolescent Brain (QTAB) projects are twin studies sourced from the Queensland Twin Registry (QTwin). The BATS project, enabled through funding from the NHMRC, was a longitudinal study of adolescent twins, which collected accelerometry data over three waves between 2014 and 2018 (ages 12, 14, and 16 years). The Queensland Twin Adolescent Brain study (QTAB, 2015–present), previously discussed in Subheading 5.1, collected accelerometry data over two waves (age 9–14 years at baseline). In both studies, participants wore a wrist-mounted accelerometry recording device for 2 weeks (day and night, removed only for bathing) and completed a daily sleep diary. Raw accelerometry data were processed and consolidated with sleep diary data to produce sleep onset, wake, and sleep duration estimates. The BATS and QTAB datasets include behavioral and psychological measures (e.g., assessments of cognition and behavior, self-

reported mental health and well-being) for further investigation of accelerometry measures. BATS and QTAB data is available from the project owners upon request.

The UK Biobank, previously discussed in Subheading 5.2, collected accelerometry data in 100,000 participants between 2013 and 2016. Participants wore a wrist-mounted activity monitor to capture physical activity and sleep patterns for 7 days. Since 2018, repeat measures have been collected for a subset of participants every quarter to examine seasonal influences on measurements. Data is available in raw (measured every 5 s) and average (by day and hour) acceleration formats. The deep phenotyping of the UK Biobank has allowed for accelerometry-based measures to be examined alongside several other measures, including brain structure [181], mood disorders [182], and Alzheimer's disease [183]. UK Biobank data is available online following registration (https://bbams.ndph.ox.ac.uk/ams/).

The mPower study (2015–present), sponsored by Sage Bionetworks with funding from the Robert Wood Johnson Foundation, aims to establish the baseline variability of real-world activity measurements of individuals with Parkinson's disease. Data is collected through an iPhone application, with minimal interruption to the daily life of participants. The initial data release (collected over 6 months) included health survey and sensor-based activity (e.g., gait and balance) data for ~8000 participants (with ~1000 self-identified as having a professional diagnosis of Parkinson's disease). In addition, approximately 900 participants contributed at least five separate days' worth of data. mPower data is accessible through the data sharing service Synapse (https://www.synapse.org/mpower).

A recent review [184] provides an overview of studies using smartphones to monitor symptoms of Parkinson's disease and in-depth descriptions of the methodology involved in these types of studies. Additionally, studies have used smartphone-based EMA to detect or treat mood disorders (see [185] for a review). Further, the Mobile Motor Activity Research Consortium for Health (MMARCH; http://mmarch.org/) is a collaborative international network working to standardize the analysis of actigraphy data in studies investigating motor activity, mood, and related disorders.

Machine learning approaches have been widely applied to data collected from smartphone and sensor devices, most notably in studies of Parkinson's disease. For example [186], used machine learning classifiers applied to accelerometry data from the UK Biobank to classify individuals with Parkinson's disease with an area under the curve of 0.85 (based on gait and low movement data). Another study [187] used data from the mPower study to detect dopaminergic medication response by applying machine learning techniques to the tapping task performance (measured via the mPower smartphone application) of Parkinson's disease patients before and after medication. Further, classifiers have been

used to detect states of deep brain stimulation (i.e., distinguishing between "On" and "Off" settings) in Parkinson's disease patients using accelerometer and gyroscope signals from smartphones [188]. Machine learning approaches have also shown promise for other disorders. For instance, machine learning algorithms within a smartphone application have helped identify individuals with obstructive sleep apnea, using actigraphy, body position assessment, and audio recordings [189]. Lastly, some developed a pipeline for personalized modeling of depressed mood (based on EMA) and smartwatch-derived sleep and physical activity measures [190].

Acknowledgments and Fundings

This research was supported by the Australian National Health and Medical Research Council (1,078,037, 1,078,901, 1,113,400, 1,161,356, and 1,107,258), the Australian Research Council (FT180100186 and FL180100072), the Sylvia and Charles Viertel Charitable Foundation, the program "Investissements d'avenir" ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) and reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the European Union H2020 program (project EuroPOND, grant number 666992), the joint NSF/NIH/ANR program "Collaborative Research in Computational Neuroscience" (project HIPLAY7, grant number ANR-16-NEUC-0001-01), the ICM Big Brain Theory Program (project DYNAMO, project PredictICD), and the Abeona Foundation (project Brain@Scale). BCD is supported by a CJ Martin Fellowship (APP1161356).

References

1. UNESCO Recommendation on Open Science - UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
2. Fry A, Littlejohns TJ, Sudlow C et al (2017) Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol 186:1026–1034
3. Ben-Eghan C, Sun R, Hleap JS et al (2020) Don't ignore genetic data from minority populations. Nature 585:184–186
4. Yang H-C, Chen C-W, Lin Y-T et al (2021) Genetic ancestry plays a central role in population pharmacogenomics. Commun Biol 4: 1–14
5. Barton NTL Paul Resnick, and Genie (2019) Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/
6. Courtland R (2018) Bias detectives: the researchers striving to make algorithms fair. Nature 558:357–360
7. Bustamante CD, De La Vega FM, Burchard EG (2011) Genomics for the world. Nature 475:163–165
8. Sirugo G, Williams SM, Tishkoff SA (2019) The missing diversity in human genetic studies. Cell 177:26–31

9. Herrick R, Horton W, Olsen T et al (2016) XNAT central: open sourcing imaging research data. NeuroImage 124:1093–1096
10. Routier A, Burgos N, Díaz M et al (2021) Clinica: an open source software platform for reproducible clinical neuroscience studies. Front Neuroinform 15:689675
11. Petersen RC, Aisen PS, Beckett LA et al (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 74:201–209
12. Ellis KA, Bush AI, Darby D et al (2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 21:672–687
13. LaMontagne PJ, Keefe S, Lauren W et al (2018) OASIS-3: longitudinal neuroimaging, clinical and cognitive dataset for normal aging and Alzheimer's disease. Alzheimers Dement J Alzheimers Assoc 14:P1097
14. Miller KL, Alfaro-Almagro F, Bangerter NK et al (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19:1523–1536
15. Alfaro-Almagro F, Jenkinson M, Bangerter NK et al (2018) Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage 166:400–424
16. Casey BJ, Cannonier T, Conley MI et al (2018) The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev Cogn Neurosci 32:43–54
17. Barch DM, Albaugh MD, Avenevoli S et al (2018) Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: rationale and description. Dev Cogn Neurosci 32:55–66
18. Schmaal L, Hibar DP, Sämann PG et al (2017) Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. Mol Psychiatry 22:900–909
19. Hoogman M, Muetzel R, Guimaraes JP et al (2019) Brain imaging of the cortex in ADHD: a coordinated analysis of large-scale clinical and population-based samples. Am J Psychiatry 176:531–542
20. van Rooij D, Anagnostou E, Arango C et al (2018) Cortical and subcortical brain morphometry differences between patients with autism Spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD Working Group. Am J Psychiatry 175:359–369
21. Logue MW, van Rooij SJH, Dennis EL et al (2018) Smaller hippocampal volume in posttraumatic stress disorder: a multisite ENIGMA-PGC study: subcortical volumetry results from posttraumatic stress disorder consortia. Biol Psychiatry 83:244–253
22. Boedhoe PSW, Schmaal L, Abe Y et al (2018) Cortical abnormalities associated with pediatric and adult obsessive-compulsive disorder: findings from the ENIGMA Obsessive-Compulsive Disorder Working Group. Am J Psychiatry 175:453–462
23. Mackey S, Allgaier N, Chaarani B et al (2019) Mega-analysis of gray matter volume in substance dependence: general and substance-specific regional effects. Am J Psychiatry 176:119–128
24. van Erp TGM, Walton E, Hibar DP et al (2018) Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the Enhancing Neuroimaging Genetics Through Meta-analysis (ENIGMA) Consortium. Biol Psychiatry 84:644–654
25. Hibar DP, Westlye LT, Doan NT et al (2018) Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Mol Psychiatry 23:932–942
26. Dima D, Modabbernia A, Papachristou E, et al (2021) Subcortical volumes across the lifespan: data from 18,605 healthy individuals aged 3–90 years. Hum Brain Mapp
27. Thompson PM, Jahanshad N, Ching CRK et al (2020) ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. Transl Psychiatry 10:1–28
28. Kremen WS, Franz CE, Lyons MJ (2013) VETSA: the Vietnam era twin study of aging. Twin Res Hum Genet 16:399–402
29. Van Essen DC, Smith SM, Barch DM et al (2013) The WU-Minn Human Connectome Project: an overview. NeuroImage 80:62–79
30. Marek K, Chowdhury S, Siderowf A et al (2018) The Parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. Ann Clin Transl Neurol 5: 1460–1477
31. Dufouil C, Dubois B, Vellas B et al (2017) Cognitive and imaging markers in non-demented subjects attending a memory clinic: study design and baseline findings of the MEMENTO cohort. Alzheimers Res Ther 9:67

32. Ritchie CW, Muniz-Terrera G, Kivipelto M et al (2020) The European Prevention of Alzheimer's Dementia (EPAD) Longitudinal Cohort Study: Baseline Data Release V500.0. J Prev Alzheimers Dis 7:8–13
33. Di Martino A, Yan C-G, Li Q et al (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19: 659–667
34. Di Martino A, O'Connor D, Chen B et al (2017) Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data 4:170010
35. Sharp, PF, Welch, A (2005) Positron emission tomography. In: Sharp PF, Gemmell HG, Murray AD (eds) Practical Nuclear Medicine. Springer, London. https://doi.org/10.1007/1-84628-018-4_3
36. Herholz K (1995) FDG PET and differential diagnosis of dementia. Alzheimer Dis Assoc Disord 9:6–16
37. Sudre CH, Cardoso MJ, Modat M et al (2020) Chapter 15 - Imaging biomarkers in Alzheimer's disease. In: Zhou SK, Rueckert D, Fichtinger G (eds) Handbook of medical image computing and computer assisted intervention. Academic Press, pp 343–378
38. Knudsen GM, Jensen PS, Erritzoe D et al (2016) The Center for Integrated Molecular Brain Imaging (Cimbi) database. NeuroImage 124:1213–1219
39. Wang H, Tian Y, Liu Y et al (2021) Population-specific brain [18F]-FDG PET templates of Chinese subjects for statistical parametric mapping. Sci Data 8:305
40. Jackson AF, Bolger DJ (2014) The neurophysiological bases of EEG and EEG measurement: a review for the rest of us. Psychophysiology 51:1061–1071
41. Niedermeyer E and da da Silva FHL (2005) Electroencephalography: basic principles, clinical applications, and related fields, Lippincott Williams & Wilkins
42. Nunez PL, Srinivasan R (2006) Electric fields of the brain: the neurophysics of EEG. Oxford University Press, New York
43. Engel AK, Moll CKE, Fried I et al (2005) Invasive recordings from the human brain: clinical insights and beyond. Nat Rev Neurosci 6:35–47
44. Cohen D (1972) Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer. Science 175:664–666
45. Hämäläinen M, Hari R, Ilmoniemi RJ et al (1993) Magnetoencephalography---theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 65:413–497
46. Lopes da Silva F (2013) EEG and MEG: relevance to neuroscience. Neuron 80:1112–1128
47. Malmivuo J (2012) Comparison of the properties of EEG and MEG in detecting the electric activity of the brain. Brain Topogr 25:1–19
48. Bertrand O, Perrin F, Pernier J (1985) A theoretical justification of the average reference in topographic evoked potential studies. Electroencephalogr Clin Neurophysiol 62: 462–464
49. de Cheveigné A, Nelken I (2019) Filters: when, why, and how (not) to use them. Neuron 102:280–293
50. Michel CM, Brunet D (2019) EEG source imaging: a practical review of the analysis steps. Front Neurol 10
51. Subha DP, Joseph PK, Acharya UR et al (2010) EEG signal analysis: a survey. J Med Syst 34:195–212
52. Ang KK, Chin ZY, Zhang H, et al (2008) Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 2390–2397
53. Müller-Gerking J, Pfurtscheller G, Flyvbjerg H (1999) Designing optimal spatial filters for single-trial EEG classification in a movement task. Clin Neurophysiol 110:787–798
54. Pfurtscheller G, Lopes da Silva FH (1999) Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 110:1842–1857
55. Sur S, Sinha VK (2009) Event-related potential: an overview. Ind Psychiatry J 18:70–73
56. Neuper C, Wörtz M, Pfurtscheller G (2006) ERD/ERS patterns reflecting sensorimotor activation and deactivation. Prog Brain Res 159:211–222
57. Crone NE, Miglioretti DL, Gordon B et al (1998) Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain J Neurol 121(Pt 12):2301–2315
58. Bastos AM, Schoffelen J-M (2016) A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Front Syst Neurosci 9:175

59. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10:186–198
60. De Vico Fallani F, Richiardi J, Chavez M et al (2014) Graph analysis of functional brain networks: practical issues in translational neuroscience. Philos Trans R Soc Lond Ser B Biol Sci 369:20130521
61. Gonzalez-Astudillo J, Cattai T, Bassignana G et al (2021) Network-based brain–computer interfaces: principles and applications. J Neural Eng 18:011001
62. Mirowski P, Madhavan D, LeCun Y et al (2009) Classification of patterns of EEG synchronization for seizure prediction. Clin Neurophysiol 120:1927–1940
63. Siddiqui MK, Morales-Menendez R, Huang X et al (2020) A review of epileptic seizure detection using machine learning classifiers. Brain Inform 7:5
64. Smith SJM (2005) EEG in the diagnosis, classification, and management of patients with epilepsy. J Neurol Neurosurg Psychiatry 76: ii2–ii7
65. Dauwels J, Vialatte F, Cichocki A (2010) Diagnosis of Alzheimer's disease from EEG signals: where are we standing? Curr Alzheimer Res 7:487–505
66. Vecchio F, Babiloni C, Lizio R et al (2013) Resting state cortical EEG rhythms in Alzheimer's disease: toward EEG markers for clinical applications: a review. Suppl Clin Neurophysiol 62:223–236
67. Lotte F, Bougrain L, Cichocki A et al (2018) A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. J Neural Eng 15:031005
68. Oostenveld R, Fries P, Maris E et al (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:156869
69. Tadel F, Baillet S, Mosher JC et al (2011) Brainstorm: a user-friendly application for MEG/EEG analysis. Comput Intell Neurosci 2011:879716
70. Gramfort A, Luessi M, Larson E et al (2013) MEG and EEG data analysis with MNE-python. Front Neurosci 7:267
71. Jayaram V, Barachant A (2018) MOABB: trustworthy algorithm benchmarking for BCIs. J Neural Eng 15:066011
72. Carlin JB, Gurrin LC, Sterne JA et al (2005) Regression models for twin studies: a critical review. Int J Epidemiol 34:1089–1099
73. Sainani K (2010) The importance of accounting for correlated observations. PM&R 2: 858–861
74. Zagai U, Lichtenstein P, Pedersen NL et al (2019) The Swedish twin registry: content and management as a research infrastructure. Twin Res Hum Genet 22:672–680
75. Magnusson PKE, Almqvist C, Rahman I et al (2013) The Swedish Twin Registry: establishment of a biobank and other recent developments. Twin Res Hum Genet 16:317–329
76. Tomata Y, Li X, Karlsson IK et al (2020) Joint impact of common risk factors on incident dementia: a cohort study of the Swedish Twin Registry. J Intern Med 288:234–247
77. Wirdefeldt K, Gatz M, Pawitan Y et al (2005) Risk and protective factors for Parkinson's disease: a study in Swedish twins. Ann Neurol 57:27–33
78. Fang F, Kamel F, Lichtenstein P et al (2009) Familial aggregation of amyotrophic lateral sclerosis. Ann Neurol 66:94–99
79. Wright MJ, Martin NG (2004) Brisbane Adolescent Twin Study: outline of study methods and research projects. Aust J Psychol 56:65–78
80. Miranda-Dominguez O, Feczko E, Grayson DS et al (2018) Heritability of the human connectome: a connectotyping study. Netw Neurosci 2:175–199
81. Kochunov P, Jahanshad N, Marcus D et al (2015) Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA-DTI data. NeuroImage 111:300–311
82. Schmitt JE, Raznahan A, Liu S et al (2020) The genetics of cortical myelination in young adults and its relationships to cerebral surface area, cortical thickness, and intelligence: a magnetic resonance imaging study of twins and families. NeuroImage 206:116319
83. Kremen WS, Franz CE, Lyons MJ (2019) Current status of the Vietnam Era Twin Study of Aging (VETSA). Twin Res Hum Genet 22:783–787
84. Sachdev PS, Lammel A, Trollor JN et al (2009) A comprehensive neuropsychiatric study of elderly twins: the Older Australian Twins Study. Twin Res Hum Genet 12:573–582
85. Hur Y-M, Bogl LH, Ordoñana JR et al (2019) Twin family registries worldwide: an important resource for scientific research. Twin Res Hum Genet 22:427–437
86. Ligthart L, van Beijsterveldt CEM, Kevenaar ST et al (2019) The Netherlands twin register: longitudinal research based on twin and twin-

family designs. Twin Res Hum Genet 22: 623–636

87. Glahn DC, Winkler AM, Kochunov P et al (2010) Genetic control over the resting brain. Proc Natl Acad Sci 107:1223–1228
88. Raffield LM, Cox AJ, Hugenschmidt CE et al (2015) Heritability and genetic association analysis of neuroimaging measures in the Diabetes Heart Study. Neurobiol Aging 36:1602.e7–1602.15
89. Iacono WG, Heath AC, Hewitt JK et al (2018) The utility of twins in developmental cognitive neuroscience research: how twins strengthen the ABCD research design. Dev Cogn Neurosci 32:30–42
90. Brouwer RM, Schutte J, Janssen R et al (2021) The speed of development of adolescent brain age depends on sex and is genetically determined. Cereb Cortex 31:1296–1306
91. Cole JH, Poudel RPK, Tsagkrasoulis D et al (2017) Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163:115–124
92. Vandenbosch MMLJZ, van't Ent D, Boomsma DI et al (2019) EEG-based age--prediction models as stable and heritable indicators of brain maturational level in children and adolescents. Hum Brain Mapp 40:1919–1926
93. Chung MK, Lee H, DiChristofano A et al (2019) Exact topological inference of the resting-state brain networks in twins. Netw Neurosci 3:674–694
94. Yamin MA, Dayan M, Squarcina L, et al (2019) Investigating the impact of genetic background on brain dynamic functional connectivity through machine learning: a Twins Study. In: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), pp. 1–4
95. Han Y, Adolphs R (2020) Estimating the heritability of psychological measures in the Human Connectome Project dataset. PLoS One 15:e0235860
96. Demeter DV, Engelhardt LE, Mallett R et al (2020) Functional connectivity fingerprints at rest are similar across youths and adults and vary with genetic similarity. iScience 23: 100801
97. Elliott LT, Sharp K, Alfaro-Almagro F et al (2018) Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562:210–216
98. Pirruccello JP, Bick A, Wang M et al (2020) Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat Commun 11:2254
99. Liu Y, Basty N, Whitcher B et al (2021) Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. elife 10:e65554
100. Chua SYL, Dhillon B, Aslam T et al (2019) Associations with photoreceptor thickness measures in the UK Biobank. Sci Rep 9: 19440
101. Bycroft C, Freeman C, Petkova D et al (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562:203–209
102. Wain LV, Shrine N, Miller S et al (2015) Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med 3:769–781
103. Van Hout CV, Tachmazidou I, Backman JD et al (2020) Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 586:749–756
104. Backman JD, Li AH, Marcketta A et al (2021) Exome sequencing and analysis of 454,787 UK Biobank participants. Nature 599:628–634
105. Wang Y, Guo J, Ni G et al (2020) Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun 11:3865
106. Medland SE, Grasby KL, Jahanshad N, et al (2020) Ten years of enhancing neuroimaging genetics through meta-analysis: an overview from the ENIGMA Genetics Working Group. Hum Brain Mapp
107. Adams HHH, Hibar DP, Chouraki V et al (2016) Novel genetic loci underlying human intracranial volume identified through genome-wide association. Nat Neurosci 19: 1569–1582
108. Hibar DP, Adams HHH, Jahanshad N et al (2017) Novel genetic loci associated with hippocampal volume. Nat Commun 8:13624
109. Stein JL, Medland SE, Vasquez AA et al (2012) Identification of common variants associated with human hippocampal and intracranial volumes. Nat Genet 44:552–561
110. Hibar DP, Stein JL, Renteria ME et al (2015) Common genetic variants influence human subcortical brain structures. Nature 520: 224–229
111. Satizabal CL, Adams HHH, Hibar DP et al (2019) Genetic architecture of subcortical brain structures in 38,851 individuals. Nat Genet 51:1624–1636

112. Grasby KL, Jahanshad N, Painter JN et al (2020) The genetic architecture of the human cerebral cortex. Science 367:eaay6690
113. Smit DJA, Wright MJ, Meyers JL et al (2018) Genome-wide association analysis links multiple psychiatric liability genes to oscillatory brain activity. Hum Brain Mapp 39:4183–4195
114. Sønderby IE, Ching CRK, Thomopoulos SI, et al (2021) Effects of copy number variations on brain structure and risk for psychiatric illness: large-scale studies from the ENIGMA working groups on CNVs. Hum Brain Mapp
115. Peterson RE, Kuchenbaecker K, Walters RK et al (2019) Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. Cell 179:589–603
116. Lam M, Awasthi S, Watson HJ et al (2020) RICOPILI: rapid imputation for COnsortias PIpeLIne. Bioinforma 36:930–933
117. Sullivan PF, Kendler KS (2021) The state of the science in psychiatric genomics. Psychol Med 51:2145–2147
118. Stilp AM, Emery LS, Broome JG et al (2021) A system for phenotype harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) program. Am J Epidemiol 190: 1977–1992
119. Taliun D, Harris DN, Kessler MD et al (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. Nature 590:290–299
120. Patel RA, Musharoff SA, Spence JP, et al (2021) Effect sizes of causal variants for gene expression and complex traits differ between populations
121. Li Z, Li X, Zhou H, et al (2021) A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies
122. Schubert R, Geoffroy E, Gregga I, et al (2021) Protein prediction for trait mapping in diverse populations
123. Hindy G, Dornbos P, Chaffin MD, et al (2021) Rare coding variants in 35 genes associate with circulating lipid levels – a multi-ancestry analysis of 170,000 exomes
124. Selvaraj MS, Li X, Li Z, et al (2021) Whole genome sequence analysis of blood lipid levels in >66,000 individuals
125. Bibikova M, Barnes B, Tsan C et al (2011) High density DNA methylation array with single CpG site resolution. Genomics 98: 288–295
126. Yong W-S, Hsu F-M, Chen P-Y (2016) Profiling genome-wide DNA methylation. Epigenetics Chromatin 9:26
127. Ziller MJ, Gu H, Müller F et al (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500:477–481
128. Maden SK, Thompson RF, Hansen KD et al (2021) Human methylome variation across Infinium 450K data on the gene expression omnibus. NAR Genom Bioinf 3:lqab025
129. Moran S, Arribas C, Esteller M (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics 8:389–399
130. Sala C, Di Lena P, Fernandes Durso D et al (2020) Evaluation of pre-processing on the meta-analysis of DNA methylation data from the Illumina HumanMethylation450 BeadChip platform. PLoS One 15:e0229763
131. Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 41: D991–D995
132. Wang T, Guan W, Lin J et al (2015) A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. Epigenetics 10:662–669
133. Min JL, Hemani G, Davey Smith G et al (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinforma 34:3983–3989
134. Birney E, Smith GD, Greally JM (2016) Epigenome-wide association studies and the Interpretation of Disease -Omics. PLoS Genet 12:e1006105
135. Michels KB, Binder AM (2018) Considerations for design and analysis of DNA methylation studies. Methods Mol Biol 1708:31–46
136. Mill J, Heijmans BT (2013) From promises to practical strategies in epigenetic epidemiology. Nat Rev Genet 14:585–594
137. Reed ZE, Suderman MJ, Relton CL et al (2020) The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers. Clin Epigenetics 12:50
138. Mendelson MM, Marioni RE, Joehanes R et al (2017) Association of Body Mass Index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian Randomization Approach. PLoS Med 14:e1002215
139. Min JL, Hemani G, Hannon E et al (2021) Genomic and phenotypic insights from an

atlas of genetic effects on DNA methylation. Nat Genet 53:1311–1321

140. Chadwick LH (2012) The NIH Roadmap Epigenomics Program data resource. Epigenomics 4:317–324
141. Fernández JM, de la Torre V, Richardson D et al (2016) The BLUEPRINT data analysis portal. Cell Syst 3:491–495.e5
142. Martens JHA, Stunnenberg HG (2013) BLUEPRINT: mapping human blood cell epigenomes. Haematologica 98:1487–1489
143. Tryka KA, Hao L, Sturcke A et al (2014) NCBI's database of genotypes and phenotypes: dbGaP. Nucleic Acids Res 42:D975–D979
144. Lonsdale J, Thomas J, Salvatore M et al (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585
145. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S et al (2020) The impact of sex on gene expression across human tissues. Science 369:eaba3066
146. Ferraro NM, Strober BJ, Einson J et al (2020) Transcriptomic signatures across human tissues identify functional rare genetic variation. Science 369:eaaz5900
147. Kim-Hellmuth S, Aguet F, Oliva M et al (2020) Cell type-specific genetic regulation of gene expression across human tissues. Science 369:eaaz8528
148. Demanelis K, Jasmine F, Chen LS et al (2020) Determinants of telomere length across human tissues. Science 369:eaaz6876
149. Burgos N, Colliot O (2020) Machine learning for classification and prediction of brain diseases: recent advances and upcoming challenges. Curr Opin Neurol 33:439–450
150. Koikkalainen J, Rhodius-Meester H, Tolonen A et al (2016) Differential diagnosis of neurodegenerative diseases using structural MRI data. NeuroImage Clin 11:435–449
151. Morin A, Samper-Gonzalez J, Bertrand A et al (2020) Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort. J Alzheimers Dis 74:1157–1166
152. Rathore S, Habes M, Iftikhar MA et al (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage 155:530–548
153. Samper-González J, Burgos N, Bottani S et al (2018) Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. NeuroImage 183:504–521
154. Wen J, Thibeau-Sutre E, Diaz-Melo M et al (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63:101694
155. Daniel C, Salamanca E (2020) Hospital Databases. In: Nordlinger B, Villani C, Rus D (eds) Healthcare and artificial intelligence. Springer International Publishing, Cham, pp 57–67
156. Bottani S, Burgos N, Maire A et al (2022) Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. Med Image Anal 75:102219
157. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU et al (2009) The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. Eur J Epidemiol 24:659–667
158. Ludvigsson JF, Almqvist C, Bonamy A-KE et al (2016) Registers of the Swedish total population and their use in medical research. Eur J Epidemiol 31:125–136
159. Ludvigsson JF, Svedberg P, Olén O et al (2019) The longitudinal integrated database for health insurance and labour market studies (LISA) and its use in medical research. Eur J Epidemiol 34:423–437
160. Ekbom A (2011) The Swedish multi-generation register. Methods Mol Biol 675:215–220
161. Ludvigsson JF, Andersson E, Ekbom A et al (2011) External review and validation of the Swedish national inpatient register. BMC Public Health 11:450
162. Song H, Sieurin J, Wirdefeldt K et al (2020) Association of stress-related disorders with subsequent neurodegenerative diseases. JAMA Neurol 77:700–709
163. Fang F, Zhan Y, Hammar N et al (2019) Lipids, apolipoproteins, and the risk of Parkinson Disease. Circ Res 125:643–652
164. Longinetti E, Mariosa D, Larsson H et al (2017) Neurodegenerative and psychiatric diseases among families with amyotrophic lateral sclerosis. Neurology 89:578–585
165. Barlow L, Westergren K, Holmberg L et al (2009) The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol 48:27–33
166. Tettamanti G, Ljung R, Ahlbom A et al (2019) Central nervous system tumor registration in the Swedish Cancer Register and Inpatient Register between 1990 and 2014. Clin Epidemiol 11:81–92

167. Källén B, Källén K (2003) The Swedish Medical Birth Register - a summary of content and quality. 2003-112-3
168. Persson M, Razaz N, Tedroff K et al (2018) Five and 10 minute Apgar scores and risks of cerebral palsy and epilepsy: population based cohort study in Sweden. BMJ 360:k207
169. Tettamanti G, Ljung R, Mathiesen T et al (2016) Maternal smoking during pregnancy and the risk of childhood brain tumors: results from a Swedish cohort study. Cancer Epidemiol 40:67–72
170. Brooke HL, Talbäck M, Hörnblad J et al (2017) The Swedish cause of death register. Eur J Epidemiol 32:765–773
171. Subic A, Zupanic E, von Euler M et al (2018) Stroke as a cause of death in death certificates of patients with dementia: a Cohort Study from the Swedish Dementia Registry. Curr Alzheimer Res 15:1322–1330
172. Wallerstedt SM, Wettermark B, Hoffmann M (2016) The first decade with the Swedish prescribed drug register - a systematic review of the output in the scientific literature. Basic Clin Pharmacol Toxicol 119:464–469
173. Wettermark B, Hammar N, Fored CM et al (2007) The new Swedish Prescribed Drug Register--opportunities for pharmacoepidemiological research and experience from the first six months. Pharmacoepidemiol Drug Saf 16:726–735
174. Cermakova P, Nelson M, Secnik J et al (2017) Living alone with Alzheimer's disease: data from SveDem, the Swedish Dementia Registry. J Alzheimers Dis 58:1265–1272
175. Haasum Y, Fastbom J, Johnell K (2016) Use of fall-risk inducing drugs in patients using Anti-Parkinson Drugs (APD): a Swedish Register-Based Study. PLoS One 11: e0161246
176. Religa D, Fereshtehnejad S-M, Cermakova P et al (2015) SveDem, the Swedish Dementia Registry - a tool for improving the quality of diagnostics, treatment and care of dementia patients in clinical practice. PLoS One 10: e0116538
177. Hillert J, Stawiarz L (2015) The Swedish MS registry – clinical support tool and scientific resource. Acta Neurol Scand 132:11–19
178. Longinetti E, Regodón Wallin A, Samuelsson K et al (2018) The Swedish motor neuron disease quality registry. Amyotroph Lateral Scler Front Degener 19:528–537
179. Asplund K, Hulter Åsberg K, Appelros P et al (2011) The Riks-stroke story: building a sustainable national register for quality assessment of stroke care. Int J Stroke 6:99–108
180. Shiffman S, Stone AA, Hufford MR (2008) Ecological momentary assessment. Annu Rev Clin Psychol 4:1–32
181. Hamer M, Sharma N, Batty GD (2018) Association of objectively measured physical activity with brain structure: UK Biobank study. J Intern Med 284:439–443
182. Lyall LM, Wyse CA, Graham N et al (2018) Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK Biobank. Lancet Psychiatry 5:507–514
183. Huang J, Zuber V, Matthews PM et al (2020) Sleep, major depressive disorder, and Alzheimer disease: a Mendelian randomization study. Neurology 95:e1963–e1970
184. Little MA (2021) Smartphones for remote symptom monitoring of Parkinson's disease. J Parkinsons Dis 11:S49–S53
185. Yim SJ, Lui LMW, Lee Y et al (2020) The utility of smartphone-based, ecological momentary assessment for depressive symptoms. J Affect Disord 274:602–609
186. Williamson JR, Telfer B, Mullany R et al (2021) Detecting Parkinson's disease from Wrist-Worn Accelerometry in the U.-K. Biokank. Sensors 21:2047
187. Chaibub Neto E, Bot BM, Perumal T et al (2016) Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iPhone sensor data. Pac Symp Biocomput Pac Symp Biocomput 21:273–284
188. LeMoyne R, Mastroianni T, Whiting D et al (2019) Assessment of machine learning classification strategies for the differentiation of deep brain stimulation "on" and "off" status for Parkinson's disease using a smartphone as a wearable and wireless inertial sensor for quantified feedback. In: LeMoyne R, Mastroianni T, Whiting D et al (eds) Wearable and Wireless Systems for Healthcare II: movement disorder evaluation and deep brain stimulation systems. Springer, Singapore, pp 113–126
189. Behar J, Roebuck A, Shahid M et al (2015) SleepAp: an automated obstructive sleep apnoea screening application for smartphones. IEEE J Biomed Health Inform 19: 325–331
190. Shah RV, Grennan G, Zafar-Khan M et al (2021) Personalized machine learning of depressed mood using wearables. Transl Psychiatry 11:1–18
191. Vasanthakumar A, Davis JW, Idler K et al (2020) Harnessing peripheral DNA methylation differences in the Alzheimer's Disease

Neuroimaging Initiative (ADNI) to reveal novel biomarkers of disease. Clin Epigenetics 12:84

192. Cho H, Ahn M, Ahn S et al (2017) EEG datasets for motor imagery brain-computer interface. GigaScience 6:1–8
193. Cattan G, Rodrigues PLC, Congedo M (2018) EEG Alpha Waves Dataset. https://hal.archives-ouvertes.fr/hal-02086581
194. Abel JH, Badgeley MA, Meschede-Krasa B et al (2021) Machine learning of EEG spectra classifies unconsciousness during GABAergic anesthesia. PLoS One 16:e0246165
195. Wakeman DG, Henson RN (2015) A multi-subject, multi-modal human neuroimaging dataset. Sci Data 2:150001
196. Blankertz B, Dornhege G, Krauledat M et al (2007) The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. NeuroImage 37:539–550
197. Tangermann M, Müller K-R, Aertsen A et al (2012) Review of the BCI Competition IV. Front Neurosci 6:55
198. Mumtaz W, Xia L, Yasin MAM et al (2017) A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. PLoS One 12:e0171409
199. Mohammadi MR, Khaleghi A, Nasrabadi AM et al (2016) EEG classification of ADHD and normal children using non-linear features and neural network. Biomed Eng Lett 6:66–73
200. Jaramillo-Gonzalez A, Wu S, Tonin A et al (2021) A dataset of EEG and EOG from an auditory EOG-based communication system for patients in locked-in state. Sci Data 8:8
201. Shah V, von Weltin E, Lopez S et al (2018) The temple university hospital seizure detection corpus. Front Neuroinformatics 12:83
202. Andrzejak RG, Schindler K, Rummel C (2012) Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. Phys Rev E Stat Nonlinear Soft Matter Phys 86:046206
203. Goldberger AL, Amaral LA, Glass L et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101:E215–E220
204. Brunner C, Birbaumer N, Blankertz B et al (2015) BNCI Horizon 2020: towards a roadmap for the BCI community. Brain-Comput Interfaces 2:1–10
205. O'Callaghan VS, Hansell NK, Guo W et al (2021) Genetic and environmental influences on sleep-wake behaviours in adolescence, vol 2. SLEEP Adv, p zpab018
206. de Zubicaray GI, Chiang M-C, McMahon KL et al (2008) Meeting the challenges of neuroimaging genetics. Brain Imaging Behav 2:258–263
207. Relton CL, Gaunt T, McArdle W et al (2015) Data resource profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). Int J Epidemiol 44:1181–1190
208. Bonder MJ, Luijk R, Zhernakova DV et al (2017) Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet 49:131–138
209. Huan T, Joehanes R, Song C et al (2019) Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat Commun 10:4267
210. Sarnowski C, Satizabal CL, DeCarli C et al (2018) Whole genome sequence analyses of brain imaging measures in the Framingham study. Neurology 90:e188–e196
211. Westerman K, Sebastiani P, Jacques P et al (2019) DNA methylation modules associate with incident cardiovascular disease and cumulative risk factor exposure. Clin Epigenetics 11:142
212. Nabais MF, Lin T, Benyamin B et al (2020) Significant out-of-sample classification from methylation profile scoring for amyotrophic lateral sclerosis. Npj Genomic Med 5:1–9
213. Vallerga CL, Zhang F, Fowdar J et al (2020) Analysis of DNA methylation associates the cystine–glutamate antiporter SLC7A11 with risk of Parkinson's disease. Nat Commun 11:1238
214. Sullivan PF (2010) The psychiatric GWAS consortium: big science comes to psychiatry. Neuron 68:182–186
215. Sullivan PF, Agrawal A, Bulik CM et al (2018) Psychiatric genomics: an update and an agenda. Am J Psychiatry 175:15–27
216. Mak ACY, White MJ, Eckalbar WL et al (2018) Whole-genome sequencing of pharmacogenetic drug response in racially diverse children with asthma. Am J Respir Crit Care Med 197:1552–1564
217. Kurniansyah N, Goodman MO, Kelly T, et al (2021) A multi-ethnic polygenic risk score is associated with hypertension prevalence and progression throughout adulthood, medRxiv
218. Hu Y, Stilp AM, McHugh CP et al (2021) Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: the NHLBI TOPMed program. Am J Hum Genet 108:874–893

219. Wu K-D, Hsiao C-F, Ho L-T et al (2002) Clustering and heritability of insulin resistance in Chinese and Japanese hypertensive families: a Stanford-Asian Pacific Program in Hypertension and Insulin Resistance sibling study. Hypertens Res 25:529–536
220. Johnsen JM, Fletcher SN, Huston H et al (2017) Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the my life, our future initiative. Blood Adv 1:824–834
221. Zhao X, Qiao D, Yang C et al (2020) Whole genome sequence analysis of pulmonary function and COPD in 19,996 multi-ethnic participants. Nat Commun 11:5182
222. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369:1318–1330
223. Sletten TL, Rajaratnam SMW, Wright MJ et al (2013) Genetic and environmental contributions to sleep-wake behavior in 12-year-old twins. Sleep 36:1715–1722
224. Mitchell BL, Campos AI, Rentería ME et al (2019) Twenty-five and up (25Up) study: a new wave of the Brisbane Longitudinal Twin Study. Twin Res Hum Genet 22:154–163
225. Zietsch BP, Hansen JL, Hansell NK et al (2007) Common and specific genetic influences on EEG power bands delta, theta, alpha, and beta. Biol Psychol 75:154–164
226. Bot BM, Suver C, Neto EC et al (2016) The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 3:160011