

EEG corpus: what we hold, dataset by dataset

As of 2026-10-31. **Total received (unfiltered): 73,111 recordings, 28,820.3 h of unique EEG. Total usable (converted): 71,169 recordings, 28,778.5 h.** Subjects are summed over 43 datasets and Temple (19,992, not de-duplicated).

Dataset	Population	Recordings received	Recordings usable	Subjects	Hours (unique, usable)	EEG ch	Labels / metadata we hold	Assessments	Used in	Licence / access
TOTAL received (unfiltered)	Everything on disk, A-D; Temple identical copies counted once	73,111	n/a	19,992	28,820.3	n/a	Hours include 41.7 h of Temple not converted (short or flat)	43	n/a	see rows
TOTAL usable (converted)	Everything converted to our training formats, A-D	n/a	71,169	19,992	28,778.5	n/a	n/a	43	n/a	see rows
A. d10 training corpus: the 17 datasets the production encoder was pretrained on (evaluation cohorts held out)										
CAUEEG	Hospital EEG, Korea: dementia (AD, MCI, vascular), Parkinson's, normal	348	348	348	75.3	19	Diagnosis (AD, MCI, VaD, PD, normal), age, symptoms, event markers	10	Production pretraining + evaluation	Academic and research use only; access by approved request form
DS004504	Alzheimer's, frontotemporal dementia, healthy older adults; resting EEG	88	88	88	18.9	19	Diagnosis (AD / FTD / control), MMSE, age	5	Production pretraining + evaluation	CC0
PARKINSON_UCSD	Parkinson's disease and healthy controls, resting EEG (UC San Diego)	46	46	31	2.3	19	PD vs control, MMSE, NAAART, disease duration, age	1	Production pretraining + evaluation	CC0
PEARL	Polish cohort: EEG, Alzheimer's risk genes, lifestyle, cognition	233	233	79	42.3	19	Age, cognition (RPM, CVLT), BDI, APOE genotype, 27 blood markers	2	Production pretraining + evaluation	CC0
READTBI	TBI assessment, BrainScope handheld 6-electrode device (4 usable)	510	510	510	99.2	4	TBI vs control label; no age or demographics	1	Production pretraining + evaluation	Closed data (BrainScope READ-TBI study), not a public dataset; terms of use not verified
TD_BRAIN	Brainclinics: patients (MDD, ADHD, OCD, others) and healthy controls	2,692	2,692	1,274	171.3	19	Diagnosis, age, eyes open / closed; treatment outcome (V3.1 subset)	3	Production pretraining + evaluation	Data-use agreement signed before download; terms not reviewed
DORTMUND	Healthy adults aged 20-70, Dortmund Vital Study	3,263	3,263	608	163.4	64	Age, handedness (demographics); no clinical labels	1	Production pretraining + evaluation	CC0
LEMON	Healthy adults, MPI Leipzig Mind-Brain-Body study	214	214	214	60.6	61	Age (5-year bins), demographics; no clinical labels	0	Production pretraining	CC0
DEPRESSION	College-age adults, high vs low depression (BDI), resting EEG, U Arizona	243	243	122	21.2	60	BDI, HamD, SCID diagnosis, STAI, age	2	Production pretraining + evaluation	CC0
SRM	Healthy adults, resting EEG with neuropsychological test scores	152	152	110	10.1	62	Age; RAVLT, digit span, TMT, Stroop, verbal-fluency scores	0	Production pretraining	CC0
MUMTAZ2018	Major depressive disorder patients and healthy controls	180	180	64	19.6	19	MDD vs healthy; eyes-closed / eyes-open / task state	1	Production pretraining + evaluation	CC BY 4.0
HBN	Children and adolescents 5-21 y, Healthy Brain Network, six EEG tasks	24,514	24,514	3,015	1,790.4	19	Age, p-factor, attention, internalizing, externalizing scores	3	Production pretraining + evaluation	CC-BY-SA 4.0 (OpenNeuro, 11 releases); NEMAR mirror lists CC-BY-NC-SA 4.0; participants.tsv commercial_use = Yes for all 3,155
ERP_CORE	Healthy adults, six ERP paradigms (ERP CORE, 38 of 40 participants)	266	266	38	42.1	28	Paradigm / task labels (six ERP paradigms)	1	Production pretraining + evaluation	CC BY 4.0
SIENA	Adults with epilepsy, Siena Scalp EEG Database (14 patients)	40	40	14	20.0	27	Seizure start / end annotations; epilepsy diagnosis	0	Production pretraining	CC BY 4.0
CHB_MIT	Children with epilepsy, CHB-MIT Scalp EEG (23 patients)	665	665	23	658.1	23 (18-24)	Seizure onset / offset annotations; patient ID	1	Production pretraining + evaluation	ODC-By 1.0
CMU_MIGRAINE	Migraine patients and controls, high-density EEG (128-channel system)	116	116	39	20.2	70	Migraine vs control; sensory and rest paradigms	0	Production pretraining	CC BY 4.0
PHYSIONET_MMI	Healthy volunteers, motor movement / imagery (BCI2000), 109 subjects	215	215	109	46.4	62	Task / run labels (movement vs imagery)	1	Production pretraining + evaluation	ODC-By 1.0
Subtotal A	d10 corpus; the pretraining split (train + validation) is 2,675 h	33,785	33,785	6,686	3,261.4	n/a	n/a	32	n/a	n/a
B. Processed, not in a production training run: evaluation cohorts and research datasets (PARKINSON_TLV also exists as shards; counted once, here)										
ICARE	Comatose cardiac-arrest patients, 5 hospitals, ICU EEG (I-CARE)	1,619	1,619	579	899.4	19	Outcome (CPC), poor-outcome flag, age, hospital, arrest details	0	Not yet used	CC BY-NC-SA 4.0: non-commercial, share-alike
NEONATAL_HELSINKI	Newborns in intensive care, Helsinki: seizure-annotated EEG (79 babies)	79	79	79	110.9	19	Seizure annotations from 3 raters; clinical information	0	v11 research run only	CC BY 4.0
CHBMP	Healthy young / middle-aged adults, Cuban Human Brain Mapping Project	250	250	250	105.6	60 (56-60)	Demographics; cognitive scales (WAIS-III, MMSE) in the source files	0	v11 research run only	not verified: Synapse team request with stated use; no licence on page
POSTCOVID	COVID-19 survivors and close contacts, Cuba; 19-channel EEG	169	169	169	35.2	19	Group, COVID severity, 70+ neurological-symptom fields	0	v11 research run only; evaluated at chance	CC0
PHANTOM	Phantom-limb pain: amputees (with / without pain) and controls, HD-EEG	152	152	38	16.5	60 (60-63)	Amputation and pain status, pain questionnaires (35 fields)	0	v11 research run only	CC0
BRAINLAT	Latin America: AD, bvFTD, Parkinson's, MS, healthy controls; 128-ch	124	124	124	15.1	128	Diagnosis, age, education, cognitive and functional scales	8	Evaluation; v11 research run only	not verified at source (Synapse page states none); our docs say CC BY 4.0
PARKINSON_TLV	Parkinson's disease and controls, Tel Aviv; resting and walking EEG	144	144	144	9.6	54	PD vs control; UPDRS-III, MoCA, LEDD, duration, tremor / PIGD scores	1	Evaluation; v11 research run only	CC0
PSYCHOSIS	First-episode psychosis patients and healthy controls, resting EEG	82	82	77	6.8	58	FEP vs control; about 240 clinical fields (BPRS, GAF, ...)	0	v11 research run only; evaluated at chance	CC0
DS004504_PHOTIC	Same subjects as DS004504; eyes-open photic stimulation	68	68	68	6.7	19	Diagnosis group, MMSE, age	0	v11 research run only	CC0
AD_RISK	Alzheimer's-risk cohort for an EEG-biomarker study (44 subjects)	44	44	44	3.7	56	Age, education, cognitive screening scores (16 fields)	0	v11 research run only	CC0
PD_MORTALITY	Parkinson's disease patients with mortality follow-up	94	94	94	3.6	61 (61-63)	Survival status	1	Evaluation; v11 research run only	CC0
Subtotal B	n/a	2,825	2,825	1,666	1,213.1	n/a	n/a	10	n/a	n/a
C. Converted to training shards in October 2026 (none used in a production run yet)										
CINC2018	Patients at the MGH sleep lab: overnight polysomnography (CinC 2018)	1,983	1,983	1,983	15,261.2	6	Age; 30-s sleep stages and arousals on the 994 training records	0	Not yet used	ODC-By 1.0
DS008108	Obstructive sleep apnea (56) and controls (86): PSG + cognition, Dalian	142	142	142	1,119.6	8 (6-8)	OSA group and severity (AHI), age, questionnaires, hypnogram	0	Not yet used	CC0
CAPSLPDB	Sleep disorders and 16 healthy sleepers, Parma: 108 recordings	108	108	108	1,004.0	13 (2-18)	Diagnosis, age, sleep stages, CAP A-phase annotations	0	Not yet used	ODC-By 1.0
NMT	Routine clinical EEG, Pakistan: 2,002 normal / 415 abnormal (43% < 18 y)	2,417	2,409	2,409	488.9	21	Normal / abnormal (neurologist), age, official train / eval split	1	Evaluation (external); trainer pilot	CC BY-SA 4.0 (share-alike), per the dataset authors' README
BIGP3BCI	P300 brain-computer-interface users incl. 48 with ALS; 19 studies	6,980	6,979	327	267.1	32 (16-32)	ALS diagnosis; ALSFRS-R score for 29 participants; study, device	0	Not yet used	CC BY 4.0
CORK_HIE	Term newborns with hypoxic-ischaemic encephalopathy, Cork; 53 babies	169	169	53	169.0	9	Background-severity grade 1-4 per hour, seizure flag	0	Not yet used	CC BY 4.0
DS008768	Parkinson's (202) and controls (110), resting EEG, Iowa	358	358	312	20.3	61 (61-63)	PD vs control; age, MoCA, UPDRS-III, LEDD, ON / OFF state	0	Not yet used	CC0
DS004579	Parkinson's (94) and controls (45), interval-timing task, Iowa	139	139	139	55.7	61 (61-63)	PD vs control; age, MoCA, UPDRS	0	Not yet used	CC0
DS003509	Parkinson's (28), controls (28): Simon-conflict task, ON / OFF, UNM	84	84	56	48.5	63	PD vs control; medication state; age	0	Not yet used	CC0
DS004584	Parkinson's (100) and controls (49), rest eyes-open, Iowa	149	149	149	6.6	61 (61-63)	PD vs control; age, MoCA, UPDRS	0	Not yet used	CC0
DS003490	Parkinson's (25), controls (25): auditory oddball + rest, ON / OFF, UNM	75	75	50	12.8	63	PD vs control; medication state; age	0	Not yet used	CC0
DS004902	Resting EEG after normal sleep vs sleep deprivation (71 participants)	218	213	71	17.7	61	Sleep-deprivation condition; PSQI, PANAS and other questionnaires	0	Not yet used	CC0
DS007615	Healthy women: LDAEP and resting EEG (69)	192	192	69	18.5	62	Age, education, BDI, UPPS, hormonal and lifestyle questionnaires	0	Not yet used	CC0
DS004000	Ultimatum game in schizophrenia, Fribourg (42 participants with EEG)	86	86	43	12.3	64	Group and role only	0	Not yet used	CC0
DS007609	Resting EEG and trait anxiety, 51 participants, 256-channel	51	51	51	4.1	64	Age, STAI state / trait, BIS, TIPI	0	Not yet used	CC0
Subtotal C	n/a	13,151	13,137	5,962	18,506.3	n/a	n/a	1	n/a	n/a
D. Temple University Hospital EEG (TUH), whole recordings, converted Oct 28 2026; a recording is assigned to one corpus, so subsets do not double count										
TUEG (general clinical)	Temple Univ. Hospital clinical EEG: routine, ICU, epilepsy unit; 2000-2019	11,098	10,001	3,987	3,131.1	19	Age on 9,911 of 10,001; no diagnoses, annotations or physician reports	0	Screening runs (not production)	NEDC page: "can be used for both research and commercialization"; signed NEDC form; agreement not reviewed
TUSZ	TUEG subset annotated for seizure events (start, stop, channel, type)	8,104	7,505	675	1,460.5	19	Seizure annotations on all 7,505; age on 7,425; normal / abnormal on 192	0	Screening runs (not production)	Same NEDC terms as TUEG
TUAB	TUEG subset labelled normal / abnormal by clinicians	1,727	1,720	1,316	654.5	19	Normal / abnormal on 1,720 of 1,720; age on 1,712	0	Screening runs (not production)	Same NEDC terms as TUEG
TUEP	TUEG subset: 100 epilepsy + 100 non-epilepsy patients, clinical metadata	2,052	1,831	155	451.3	19	Epilepsy yes / no on 1,831 of 1,831; seizure annotations; age on 1,831	0	Screening runs (not production)	Same NEDC terms as TUEG
TUEV	TUEG subset with six annotated EEG-event classes	322	318	28	86.8	19	Event annotations on 318 of 318; no patient IDs; age on 312	0	Screening runs (not production)	Same NEDC terms as TUEG
TUAR	TUEG subset with five artifact types annotated	47	47	36	13.7	19	Artifact annotations on 47 of 47; age on 43	0	Screening runs (not production)	Same NEDC terms as TUEG
TUSL	TUEG subset annotated for slowing, seizure and complex background	0	0	38	0.0	19	Slowing annotations; all 112 files duplicate recordings counted elsewhere	0	Screening runs (not production)	Same NEDC terms as TUEG
Subtotal D	26,977 raw files = 23,350 unique recordings; 21,422 converted	23,350	21,422	5,678	5,797.7	19	5,669 of 5,678 patients converted; unique hours on disk 5,839.5	0	n/a	n/a

Footnote. Received = files as downloaded (Temple: raw EDF files, identical copies counted once; shard datasets: files listed at conversion; older datasets: files in our training format, raw counts not tracked). Usable = converted and passed QC. Hours = unique EEG time from file headers; no overlapping windows, no augmentation. Subjects are summed over datasets and not de-duplicated (some cohorts overlap, e.g. DS004504 / DS004504_PHOTIC and the Iowa PD sets). PARKINSON_TLV also exists as shards and is counted once, in B. Assessments = comparisons on the 45-assessment accuracy slide: the rows sum to 43, plus 1 pooled age assessment (CAUEEG + DS004504 + PARKINSON_UCSD) and 1 sleep-staging assessment on the external Sleep-EDF cohort, which is not in the corpus. v11 = a research retrain with ten extra datasets, not the production encoder; screening runs = short (3,000-step) Temple pretraining experiments. Licence entries quote the source named in the workbook, read on 2026-10-31; "not verified" means the source could not be read.