Received: 27 February 2025 | Revised: 16 May 2025 | Accepted: 19 May 2025

DOI: 10.1111/epi.18487

Epilepsia®

RESEARCH ARTICLE

Harvard Electroencephalography Database: A comprehensive clinical electroencephalographic resource from four Boston hospitals

Chenxi Sun[1,2] | **Jin Jing**[1,2] | **Niels Turley**[1,2] | **Callison Alcott**[2,3] | **Wan-Yee Kang**[1,2] | **Andrew J. Cole**[2,4] | **Daniel M. Goldenholz**[1,2] | **Alice Lam**[2,4] | **Edilberto Amorim**[5] | **Catherine Chu**[2,4] | **Sydney Cash**[2,4] | **Valdery Moura Junior**[2,4] | **Aditya Gupta**[1,2,4] | **Manohar Ghanta**[1,2,4] | **Bruce Nearing**[1,2] | **Fábio A. Nascimento**[4,6] | **Aaron Struck**[7] | **Jennifer Kim**[8] | **Shadi Sartipi**[2,4] | **Alexandra-Maria Tauton**[2,4] | **Marta Fernandes**[2,4] | **Haoqi Sun**[1,2] | **Grace Bayas**[1,2] | **Kaileigh Gallagher**[1,2] | **Joost B. Wagenaar**[9] | **Nishant Sinha**[10] | **Christopher Lee-Messer**[11] | **Christine Tsien Silvers**[12] | **Bharath Gunapati**[12] | **Jonathan Rosand**[2,4] | **Jurriaan Peters**[2,3] | **Tobias Loddenkemper**[2,3] | **Jong Woo Lee**[2,13] | **Sahar Zafar**[2,4] | **M. Brandon Westover**[1,2]

[1]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

[2]Harvard Medical School, Boston, Massachusetts, USA

[3]Boston Children's Hospital, Boston, Massachusetts, USA

[4]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA

[5]University of California, San Francisco, San Francisco, California, USA

[6]Department of Neurology, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA

[7]Department of Neurology, University of Wisconsin–Madison, Madison, Wisconsin, USA

[8]Department of Neurology, Yale School of Medicine, New Haven, Connecticut, USA

[9]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[10]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[11]Department of Neurology, Stanford University School of Medicine, Stanford, California, USA

[12]Amazon Web Services, Seattle, Washington, USA

[13]Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts, USA

Correspondence
Chenxi Sun, Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA.
Email: csun8@bidmc.harvard.edu

Abstract

Objective: This article presents the Harvard Electroencephalography Database (HEEDB), a large-scale, deidentified, and standardized electroencephalographic (EEG) resource supporting artificial intelligence-driven and reproducible research in epilepsy and broader clinical neuroscience.

Chenxi Sun and Jin Jing are considered co-first authors.

Jurriaan Peters, Tobias Loddenkemper, Jong Woo Lee, Sahar Zafar, and M. Brandon Westover are considered co-senior authors.

Funding information
American Heart Association, Grant/Award Number: 20CDA35310297 and 24DIVSUP1274116; U.S. Department of Defense, Grant/Award Number: ERP 220036, W81XWH-19-1-0861 and W81XWH-21-C-0075; National Science Foundation, Grant/Award Number: 2014431; National Institutes of Health, Grant/Award Number: 1OT2OD032701, K23NS119794, K23NS124656, K99NS138680, R01AG073410, R01AG073598, R01HL161253, R01NS126282, R01NS128342, R01NS130119, R01NS131347, RF1AG064312 and RF1NS120947; Cures Within Reach, Grant/Award Number: 2022CAL-Amorim; Zoll Foundation; Hellman Foundation; Beth Israel Deaconess Medical Center

Methods: HEEDB aggregates more than 280 000 EEG recordings from more than 108 000 patients across four Harvard-affiliated hospitals. Data are harmonized using the Brain Imaging Data Structure and hosted on the Brain Data Science Platform. EEG data are linked with clinical notes, International Classification of Diseases, 10th Revision codes, medications, and EEG reports. Deidentification follows Health Insurance Portability and Accountability Act Safe Harbor standards.
Results: The database includes routine, epilepsy monitoring unit, and intensive care unit EEGs across all age groups, with 73% linked to deidentified clinical reports and 96% of those matched to recordings. Findings are extracted using expert curation, regular expressions, and medical natural language processing models. Auxiliary data include diagnoses, medications, and hospital course, supporting multimodal analysis.
Significance: HEEDB fills a critical gap in EEG data availability for epilepsy research. By enabling large-scale, privacy-compliant, and clinically relevant analysis, it accelerates the development of diagnostic tools, improves training datasets for machine learning, and promotes data-sharing in alignment with FAIR (Findable, Accessible, Interoperable, Reusable) and National Institutes of Health data policies.

KEYWORDS
AI for neurology, Data-driven EEG analysis, Deidentified clinical data, EEG data platform, EEG large-scale database

1 | INTRODUCTION

1.1 | Growing importance of big data in biomedical research

Over the past quarter-century, biomedical research has experienced a transformative surge in data generation across diverse domains, paired with increasing demands for reproducibility and methodological rigor. The National Institutes of Health (NIH) and technology companies have invested in data-sharing frameworks, comprehensive research repositories, and cloud-based computational platforms to meet these challenges.

Major initiatives include the NIH's creation of more than 130 biomedical data repositories, the BD2K (Big Data to Knowledge) program,[1] and the launch of cloud platforms like BioData Catalyst.[2] In 2023, the NIH strengthened these efforts through its formal Data Management and Sharing (DMS) requirements to ensure publicly funded research data are systematically archived for broader reuse.[3] Meanwhile, industry-led programs such as Google Cloud Life Sciences[4] and the Amazon Web Services (AWS) Open Data Sponsorship Program[5] enable researchers to host and access large-scale datasets within high-performance and low-cost platforms. Such initiatives have reshaped the landscape of cancer research,[6] genomics,[7] and epidemiology.[8]

Key points

- A large-scale, deidentified EEG database with >100 000 unique patients is presented.
- The database includes diverse data structures and modalities.
- It is well structured and user-friendly and includes statistical insights.
- It serves as a robust foundation for scalable research and AI-driven analysis.

1.2 | Physiological big data: Gaps and opportunities

Alongside genomics and imaging, physiological measurements represent a promising frontier in "big data." Large volumes of continuous or near-continuous records—from wearable heart rate monitors to multiday polysomnography—are now readily collected. Efforts such as PhysioNet[9] and the National Sleep Research Resource[10] have expanded cardiac, critical care, and sleep-related data-sharing, advancing cardiovascular and sleep research. However, resources explicitly dedicated to large-scale electroencephalography (EEG) remain underrepresented despite EEG's central role in epilepsy diagnosis, consciousness disorders, and neurological research.

1.3 | Underrepresentation of brain-focused physiological data

EEG captures the brain's dynamic, time-varying electrical activity at high temporal resolution, making it indispensable in clinical and research settings, particularly in epilepsy, intensive care monitoring, and developmental neuroscience. Yet existing public EEG databases are limited in scope and size; sleep datasets often have few channels and focus on cardiopulmonary outcomes. Consequently, many neurological studies requiring large samples and long recordings remain unexplored.

1.4 | Clinical EEG as a big data resource

In the United States alone, more than 1 million EEG tests are performed yearly, including approximately 250 000 continuous EEG studies in critical care units.[11] Diagnostic errors can substantially affect patient outcomes,[12,13] yet up to 75% of EEGs are interpreted by clinicians not specially trained in neurophysiology.[14] Globally, 50 million people have epilepsy,[15] most lacking EEG access. The rise of artificial intelligence (AI) and machine learning (ML) in health care amplifies the need for diverse, large-scale EEG data to train and validate new algorithms and tools.[14]

1.5 | Bridging the gap and objectives of this paper

To address the shortage of large-scale, clinically representative EEG repositories, we introduce the Harvard Electroencephalography Database (HEEDB; https://bdsp.io/content/harvard-eeg-db/3.0/). This cloud-based resource standardizes EEG data from multiple Boston-area hospitals.

This article explores how linking EEG with electronic health records (EHRs) advances research while addressing data harmonization and privacy challenges. HEEDB mitigates these issues by adopting the Brain Imaging Data Structure (BIDS) and aligning with Findable, Accessible, Interoperable, Reusable (FAIR) principles. HEEDB's deidentification protocols ensure privacy while preserving clinical relevance. We highlight future directions, emphasizing multimodal data and collaboration to maximize EEG's potential in neurology, critical care, and beyond.

With more than 280 000 EEG recordings accompanied by metadata, clinical notes, and physician interpretations, HEEDB is an actively maintained and continuously expanding resource. In addition to EEG reports, we are integrating a broader range of patient data over time—including but not limited to medication records, International Classification of Diseases (ICD)-coded diagnoses, and neuroimaging—to support diverse research and clinical applications. HEEDB enables large-scale EEG data-sharing, accelerates AI-driven diagnostics, and supports training for neurologists, data scientists, and biomedical engineers. By combining comprehensive data with cloud-based analytic tools, we aim to spark a new era of EEG research that transcends institutional boundaries and drives brain-focused data science forward.

Therefore, this paper aspires to:

1. **Increase awareness** of EEG's value—and the value of other clinical brain-health data—for advancing precision medicine and neuroscience research.
2. **Explain the dataset's organization and content** so investigators can effectively navigate HEEDB.
3. **Inform future data collection and annotation** procedures, promoting robust data harmonization across institutions.
4. **Prepare EEG researchers** to meet NIH data-sharing requirements and leverage large-scale resources like HEEDB for reproducible science.
5. **Encourage other groups** to publish or contribute their own EEG databases, using the Brain Data Science Platform (BDSP)'s ability to host data at no cost and promote more transparent science.
6. **Invite new collaborators** to join the Brain Data Science Platform Contributing Collaborators Consortium (BDSPC3). This enables secure sharing of deidentified data even before formal publication, accelerating the collective understanding of clinical neuroscience.

2 | BRAIN DATA SCIENCE PLATFORM

HEEDB is hosted on BDSP,[16] a platform supported by multiple NIH grants (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119) to share deidentified brain-related clinical data. Its repositories include the Human Sleep Project and I-CARE (International Cardiac Arrest Research Consortium). On November 21, 2024, the NIH recognized BDSP as an approved data-sharing repository, aligning with the NIH's DMS policy and underscoring its commitment to open science and collaborative neuroscience research.

Data storage within BDSP is sponsored by the AWS Open Data Sponsorship Program, enabling secure, scalable, and free dataset access and sharing. BDSP fosters collaborative research through initiatives like the BDSPC3, encouraging contributions of deidentified data to enrich the repository and advance neurology, critical care, and cognitive neuroscience. For more information, visit bdsp.io (https://bdsp.io/) or explore BDSP resources (https://bdsp.io/content/).

3 | HARVARD ELECTROENCEPHALOGRAPHY DATABASE

3.1 | Contributing sites

HEEDB comprises more than 280 000 EEG recordings from more than 100 000 patients across four Harvard-affiliated hospitals: Massachusetts General Hospital (MGH), Brigham and Women's Hospital (BWH), Beth Israel Deaconess Medical Center (BIDMC), and Boston Children's Hospital (BCH). These hospitals are Level 4 Comprehensive Epilepsy Centers, designated by the National Association of Epilepsy Centers, providing top-tier epilepsy care, including advanced neurodiagnostic monitoring, epilepsy surgery, and multidisciplinary treatment. They operate more than 3000 inpatient beds, with MGH having 1059, BWH 826, BIDMC 743, and BCH 466. All four hospitals offer the full range of EEG services, including routine outpatient EEGs for diagnostic testing and epilepsy management, epilepsy monitoring units for inpatient evaluations, and intensive care unit (ICU) EEG services for critically ill patients. As a dedicated pediatric hospital, BCH leads in pediatric epilepsy care and neurodiagnostic procedures. At the same time, MGH supports a smaller and highly competent pediatric epilepsy service alongside its adult services.

3.2 | Demographics

HEEDB includes 108 341 unique patients from 2003 to 2024. The cohort comprises 51% males and 49% females. The median age at the time of procedures is 45 years (range = neonatal to 90+ years, with ages >90 years anonymized). Age distribution varies by site (Figure 1A), providing a diverse cohort with a broad age distribution. Patient counts and median ages are as follows: MGH, 37 394, age 53 years; BWH, 29 160, age 59 years; BIDMC, 17 578, age 60 years; and BCH, 25 045, age 7 years. Race (Figure 1B) follows US Census Bureau classifications[17] and are fully represented: White (63.51%), Black (7.92%), American Indian/Alaska Native (.14%), Asian (2.88%), Native Hawaiian/Pacific Islander (.04%), Multiracial (.91%), and Other (5.65%). Within this population, 7.56% are Hispanic, 58.24% are non-Hispanic, and 34.20% have unknown ethnicity. Detailed demographic statistics are provided in Table S1.

3.3 | EEG recordings

HEEDB contains 284 341 EEG recordings. All patients have at least one EEG recording. The median number of EEG recordings per patient is 2 (range = 1–151; 55% have one, 17% two, 10% three, and 19% more than three).

EEG files are structured into four BIDS-compliant levels (version 1.7.0), a standardized framework ensuring interoperability and clarity across datasets. As illustrated in Figure 2, the root level contains metadata (e.g., dataset_description.json, participants.tsv, README). The subject-level folders, named using unique identifiers (e.g., sub-SiteIdPatientId), contain participant-specific data. Session-level subfolders (e.g., ses-01) organize EEG studies chronologically. At the EEG data level, files include raw EEG data (.edf), annotations (_annotations.tsv), channel descriptions (_channels.tsv), and metadata (_eeg.json). This structure ensures consistency and supports efficient data retrieval and analysis.

3.4 | EEG annotations

EEG annotations are provided in three forms: EEG reports, structured data extracted from EEG reports, and informal annotations.

3.4.1 | EEG reports

Clinical EEG reports filed in the electronic medical record describe the overall content of the EEG recordings. These reports indicate whether the recording was normal or abnormal for age, whether it included sleep (and details about normal or abnormal sleep microstructure), and whether epileptiform discharges, seizures, or other findings were observed. They often document patient behavior during seizures or other episodes, aiding in differentiating seizures, encephalopathy, and artifacts. Authored collaboratively by neurologists-in-training (fellows) and board-certified attending clinical neurophysiologists, these reports vary in style from free-form narratives to semistructured or structured formats. HEEDB is a continuously updated, deidentified dataset, with EEG reports incorporated incrementally through a structured deidentification and harmonization process. As of early 2025, 73.27% of HEEDB patients have deidentified clinical EEG reports.

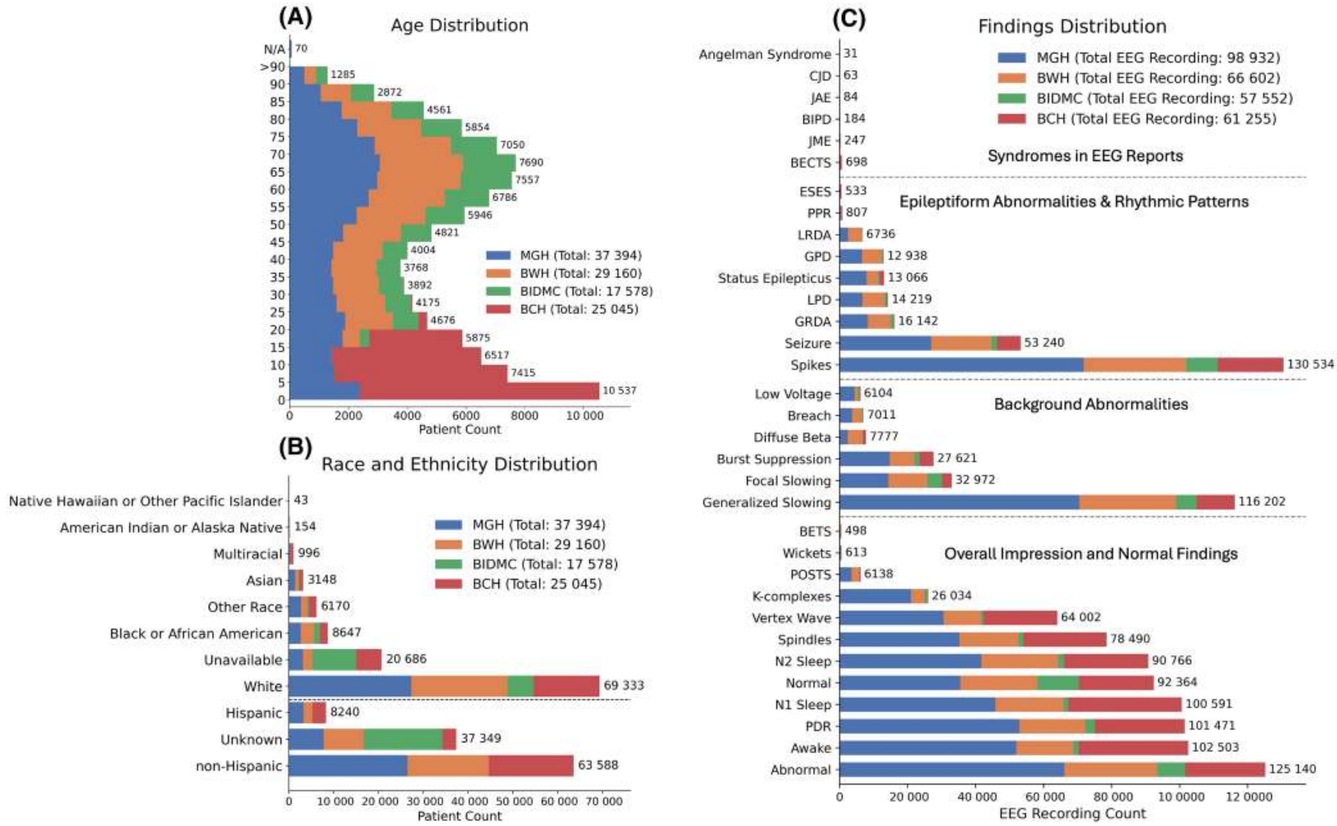

FIGURE 1 Demographics. (A) Age distribution of patients. (B) Race and ethnicity distribution of patients. (C) Findings distribution of electroencephalographic (EEG) recordings. In each plot, the stacked bars show contributions to the overall counts from each hospital. In panels A and B, statistics are presented at the patient level, whereas in panel C, they are based on individual EEG recordings. BCH, Boston Children's Hospital; BECTS, Benign Epilepsy with Centrotemporal Spikes; BETS, Benign Epileptiform Transients of Sleep; BIDMC, Beth Israel Deaconess Medical Center; BIPD, Bilateral Independent Periodic Discharges; BWH, Brigham and Women's Hospital; CJD, Creutzfeldt-Jakob Disease; ESES, Electrical Status Epilepticus in Sleep; GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; JEA, Juvenile Absence Epilepsy; JME; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; MGH, Massachusetts General Hospital; PDR, Posterior Dominant Rhythm; POSTS, Positive Occipital Sharp Transients of Sleep; PPR, Photoparoxysmal Response.

Among these, 95.6% of reports are matched to their corresponding EEGs, whereas 4.4% are unmatched due to time-recording inconsistencies in the reports.

3.4.2 | Informal annotations

Informal annotations by EEG reviewers, including technologists, nurses, and physicians, offer additional insights into EEG findings. Although often tentative, informal, and nonstandardized, these labels nevertheless come with time stamps and can be helpful in training ML algorithms for EEG interpretation.

3.4.3 | EEG findings in HEEDB

EEG findings are extracted from annotations and reports, categorized into four major categories and 33 subcategories. Figure 1C presents the distribution of findings across different hospitals and categories. Each EEG with an available report has at least a normal or abnormal label. Statistical details of the findings are provided in Table S2.

Findings are identified using customized regular expressions that match full-term abbreviations commonly used as shorthand by annotator. These findings provide event-level weak labels, localized to specific EEG segments. However, due to the expertise-dependent nature of annotations, they serve as references rather than definitive ground truth.

EEG report findings are extracted using a combination of customized regular expressions and a fine-tuned Llama 3 7B model trained on a medical corpus of reports.[18] We designed highly constrained prompts for the medical Llama model, applying strict criteria to minimize ambiguity and reduce false results. These findings serve as EEG-level labels, describing entire or partial recordings without precise localization. Compared to annotations, they provide stronger

```
Bids-root-folder/
        └──── dataset_description.json
        └──── participants.json
        └──── participants.tsv
        └──── README
        └──── sub-Id/
                 └──── ses-01/
                 └──── sub-SiteIdPatientId_ses-1_scans.tsv
                 └──── eeg/
                          └──── sub-Id_ses-1_task-eeg_annotations.tsv
                          └──── sub-Id_ses-1_task-eeg_channels.tsv
                          └──── sub-Id_ses-1_task-eeg_eeg.edf
                          └──── sub-Id_ses-1_task-eeg_eeg.json
                          └──── sub-Id_ses-1_task-eeg_pre.csv
```

FIGURE 2 Hierarchical organization of electroencephalographic (EEG) data in the Brain Imaging Data Structure (BIDS) format. The BIDS format (version 1.7.0) organizes EEG data into four hierarchical levels to ensure interoperability and clarity across datasets. At the root level, general metadata files (e.g., dataset_description.json, participants.json, participants.tsv, and a README) provide high-level information about the dataset and participants. The subject level groups data by unique participant identifiers (e.g., sub-SiteIdPatientId), containing all data related to a specific participant. Within each subject folder, the session level arranges EEG recordings chronologically into subfolders (e.g., ses-01), representing individual recording sessions. The EEG data level includes essential files such as the raw EEG data files (.edf), annotations (_annotations.tsv), channel descriptions (_channels.tsv), and additional metadata (_eeg.json). This structure ensures consistency and facilitates efficient data analysis and retrieval.

labels, as they are reviewed and confirmed by at least two clinicians, including a senior clinician. Method details and validation are provided in Supplementary A in Data S1.

3.5 | EHR data

Beyond EEG signals, HEEDB includes auxiliary data situating recordings in a clinical context, enabling research on health conditions and EEG findings. These data provide a comprehensive framework for understanding the medical, behavioral, and contextual factors surrounding EEG studies, enhancing its neurophysiological and clinical research value.

Structured data, including ICD codes, medication records, and other relevant elements, are stored in Parquet format for efficient querying. Unstructured data, such as EEG reports and clinical notes, are available as plain text for natural language processing (NLP) applications. The combination of structured and unstructured data forms a robust resource for examining neurological conditions and their clinical implications.

3.5.1 | Medical diagnostic codes

HEEDB includes diagnostic ICD-9 and ICD-10 codes from clinical visits. These codes enable researchers to link EEG findings with diagnostic categories and health conditions. To ensure consistency, ICD-9 codes are mapped to ICD-10, with all diagnoses standardized to ICD-10. Figure 3A,B shows the total ICD-10 codes per patient (median = 191) and unique diagnostic categories per patient (median = 44). To enhance the understanding of diagnostic codes, we provide an ICD-10 neurology classification, grouping diagnoses into 18 major categories.[19] Figure 3C illustrates their distribution across different hospitals. Notably, seizure, headache, and sleep disorders are the most common neurology-related diagnoses in HEEDB. Table S3 details classifications and statistics. See Supplementary B in Data S1 for more details.

3.5.2 | Medications

Medication data include outpatient prescriptions and inpatient administration records, detailing drug names, dosages, routes (e.g., oral, intravenous, subcutaneous), and timing. The timing and dosage information for inpatient medications enables pharmacokinetic and pharmacodynamic studies that correlate specific drug regimens with EEG findings, offering insights into the neurophysiological impacts of pharmacological interventions. Figure 4A,B shows the total medications per patient (median = 46) and the unique medication types per patient (median = 15). To summarize drug information across the cohort, we categorized drugs using the Anatomical Therapeutic Chemical (ATC) system.[20] Figure 4C shows 14 major categories, with nervous system, alimentary

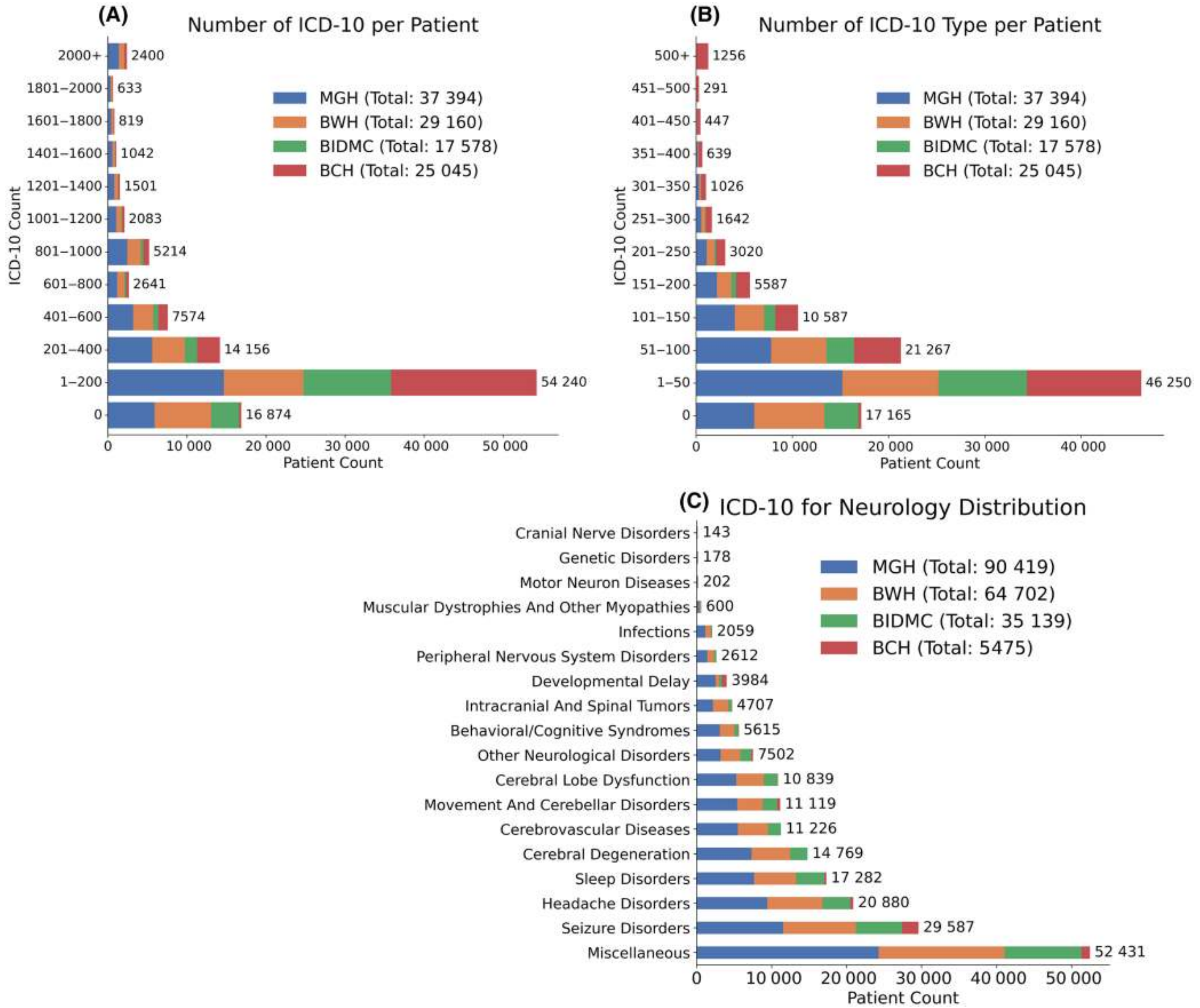

FIGURE 3 Medical diagnostic codes. (A) The number of ICD-10 codes per patient in intervals of 200. (B) The number of unique International Classification of Diseases, 10th Revision (ICD-10) types per patient in intervals of 50. (C) Counts of how many patients have ICD-10 codes within key categories (ICD-10, Clinical Modification for Neurology). In each plot, the stacked bars show contributions to the overall counts from each hospital. ATC, Anatomical Therapeutic Chemical; BCH, Boston Children's Hospital; BIDMC, Beth Israel Deaconess Medical Center; BWH, Brigham and Women's Hospital; MGH, Massachusetts General Hospital.

tract and metabolism, and cardiovascular drugs most prescribed. A more detailed breakdown of classifications and statistics is provided in Table S4. See Supplementary C in Data S1 for more details.

3.5.3 | Admission–discharge–transfer information

Admission–discharge–transfer records track patient movements within the hospital system, distinguishing between outpatient visits, hospital admissions, and ICU stays, providing context for understanding the progression and severity of neurological conditions.

3.5.4 | Vital signs

Vital sign measurements, including heart rate, blood pressure, and oxygen saturation, add a physiological dimension to the dataset, complementing the EEG data and offering additional insights into each patient's overall health status during the EEG recordings.

3.5.5 | Clinical notes

The database incorporates clinical notes, such as discharge summaries, progress notes, and therapy evaluations, which capture nuanced clinical details. These

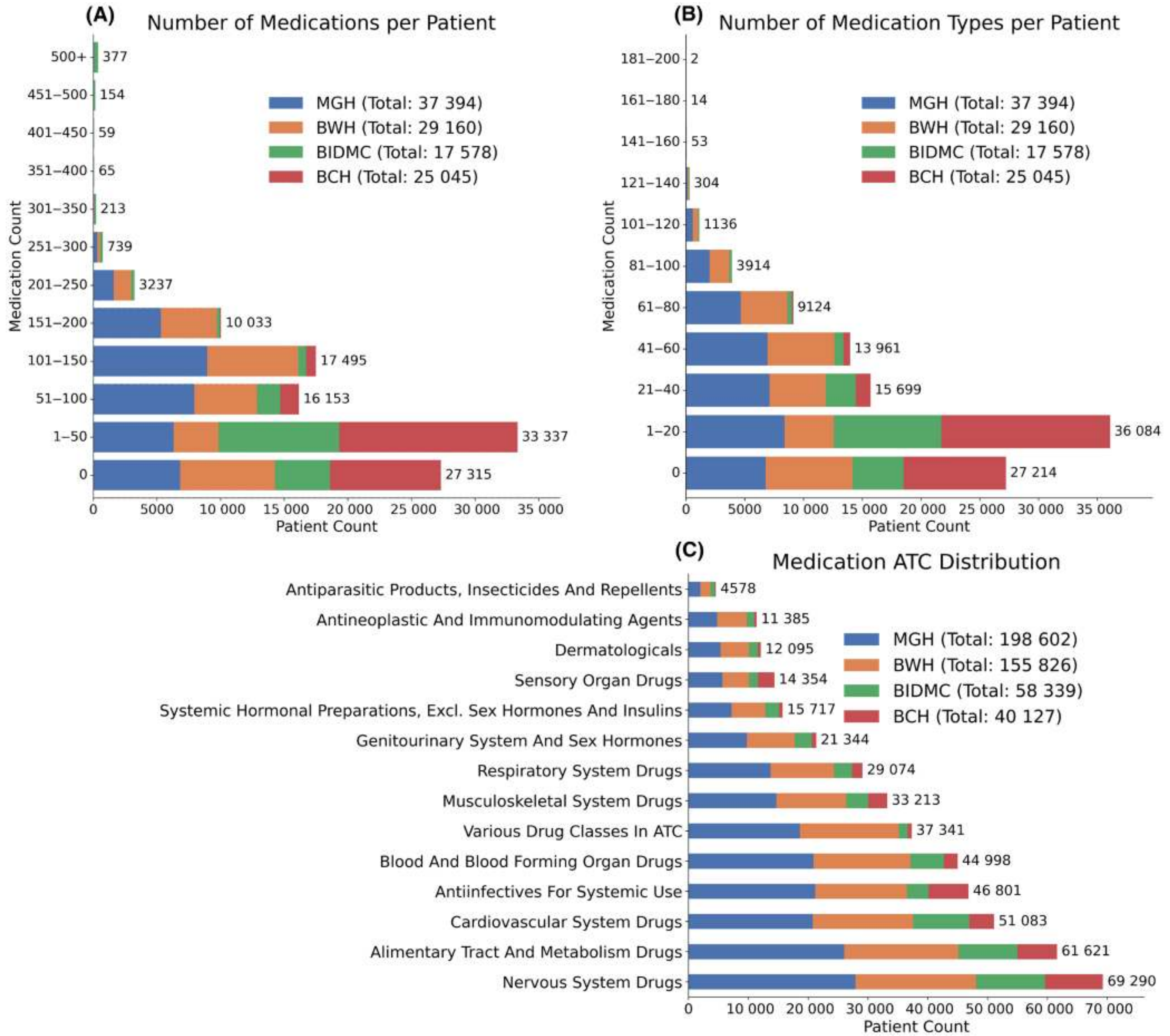

FIGURE 4 Medications. (A) The number of medications per patient in intervals of 50. (B) The number of unique medication types per patient in intervals of 20. (C) Counts of how many patients have been prescribed medications within key categories (Anatomical Therapeutic Chemical classification system). In each plot, the stacked bars show contributions to the overall counts from the four contributing hospitals. BCH, Boston Children's Hospital; BIDMC, Beth Israel Deaconess Medical Center; BWH, Brigham and Women's Hospital; ICD, International Classification of Diseases; MGH, Massachusetts General Hospital.

unstructured data sources are particularly valuable for phenotype extraction using NLP algorithms. Several advanced NLP tools developed by the authors allow identifying clinical phenotypes with greater accuracy than traditional ICD code-based methods. These algorithms and the data used for their development will soon be available on bdsp.io, further enhancing the dataset's accessibility and research potential.

3.6 | Data deidentification and privacy

HEEDB implements comprehensive deidentification procedures in compliance with Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor requirements.[21] For EEG data, custom software removes patient identifiers and shifts temporal information (month and year) per HIPAA guidelines. This date-shifting process is consistently applied across all associated data types to maintain temporal relationships while protecting patient privacy. Clinical data undergo multistep deidentification; structured data (e.g., lab results, demographics) have Protected Health Information (PHI) removed, and unstructured text (e.g., clinical notes, EEG reports, and imaging reports) is processed using validated automated deidentification software (described below) that has been successfully employed in similar large-scale clinical databases. Complete anonymization (rather than deidentification) is performed for data subject to General Data Protection Regulation requirements. Security is further reinforced through mandatory data use agreements (DUAs) that users must sign when requesting access to the data, which prohibit reidentification attempts and require reporting any inadvertently discovered PHI.

Clinical narrative text (e.g., EEG reports, radiology reports, and other clinical notes) undergoes deidentification using PHILter,[22] a certified deidentification system that employs rule-based pattern matching and statistical methods to identify and remove PHI. With high precision and recall, its output is externally audited and certified as fully deidentified. See Supplementary D in Data S1 for more details.

3.7 | Tools for EEG analysis

Tools for analyzing clinical EEG data are provided to accompany the EEGs in HEEDB, including:

1. An automated detector for epileptiform discharges (SpikeNet 1.0[23]): https://github.com/bdsp-core/SpikeNet1.
2. Two automated detectors for electrographic seizures, lateralized and generalized rhythmic delta activity, and lateralized and generalized periodic discharges:
   a. SPaRCNet[24]: https://github.com/bdsp-core/IIIC-SPaRCNet/tree/main/SPaRCNet.
   b. TEEGLLTEEG1: https://github.com/chengstark/ProtoPMed-EEG/tree/main.
3. A tool for efficient manual annotation of continuous EEG recordings: https://github.com/bdsp-core/Rapid_IIIC_Labeling_GUI_MultipleEEGs?tab=readme-ov-file.

Figure 5A shows spectrograms from an 8-h-long continuous EEG recording and automated detections of recurrent seizures and generalized rhythmic delta activity. Figure 5B shows a close-up of one of the seizures. Figure 6 shows a 15-s sample of EEG and a spike–wave discharge detected by the SpikeNet 1.0 model. Links to additional tools will be provided at bdsp.io as they become available.

3.8 | Adherence to FAIR principles

Adherence to FAIR principles is a cornerstone of HEEDB for all BDSP datasets. These principles align with recent federal data-sharing requirements and ensure that HEEDB serves as a sustainable, scalable resource for the neuroscience research community. From its inception, the design and implementation of HEEDB have been guided by the collaborative efforts of neurologists, clinical researchers, computer scientists, and data scientists, fostering a platform optimized for usability and accessibility.

Key elements of HEEDB's adherence to FAIR principles include:

1. A **streamlined registration and access process**, allowing researchers to request access to EEG and auxiliary data under a unified DUA.
2. A **secure and scalable infrastructure**, leveraging AWS Open Data Sponsorship, to facilitate long-term availability and the reliable transfer and analysis of EEG files and accompanying EHR data.
3. **Standardized metadata and documentation**, ensuring precise and consistent descriptions of datasets, variables, and study designs.
4. **Harmonization of EEG data**, including adherence to the BIDS format and standardized metadata for EEG recordings and associated variables.
5. **Signal processing protocols** that ensure uniformity in file formats, annotations, and sampling rates across contributing institutions.

These processes were developed iteratively, reflecting the requirements of the NIH's Data Management and Sharing Plan (2023), and have been disseminated through collaborations with professional societies and workshops. By adhering to FAIR principles, HEEDB facilitates rigorous, reproducible research and promotes cross-disciplinary collaboration, driving innovation in clinical neuroscience.

3.9 | Data use agreement

Access to HEEDB is facilitated through the BDSP (bdsp.io), ensuring secure and controlled use of the data. Although the data have been deidentified per HIPAA, HEEDB is classified as “restricted access,” allowing only credentialed researchers who meet specific criteria. Users must sign a DUA committing to data security, prohibiting reidentification, and restricting use to lawful scientific research. Data-sharing is forbidden, commercial use requires a separate agreement, and any

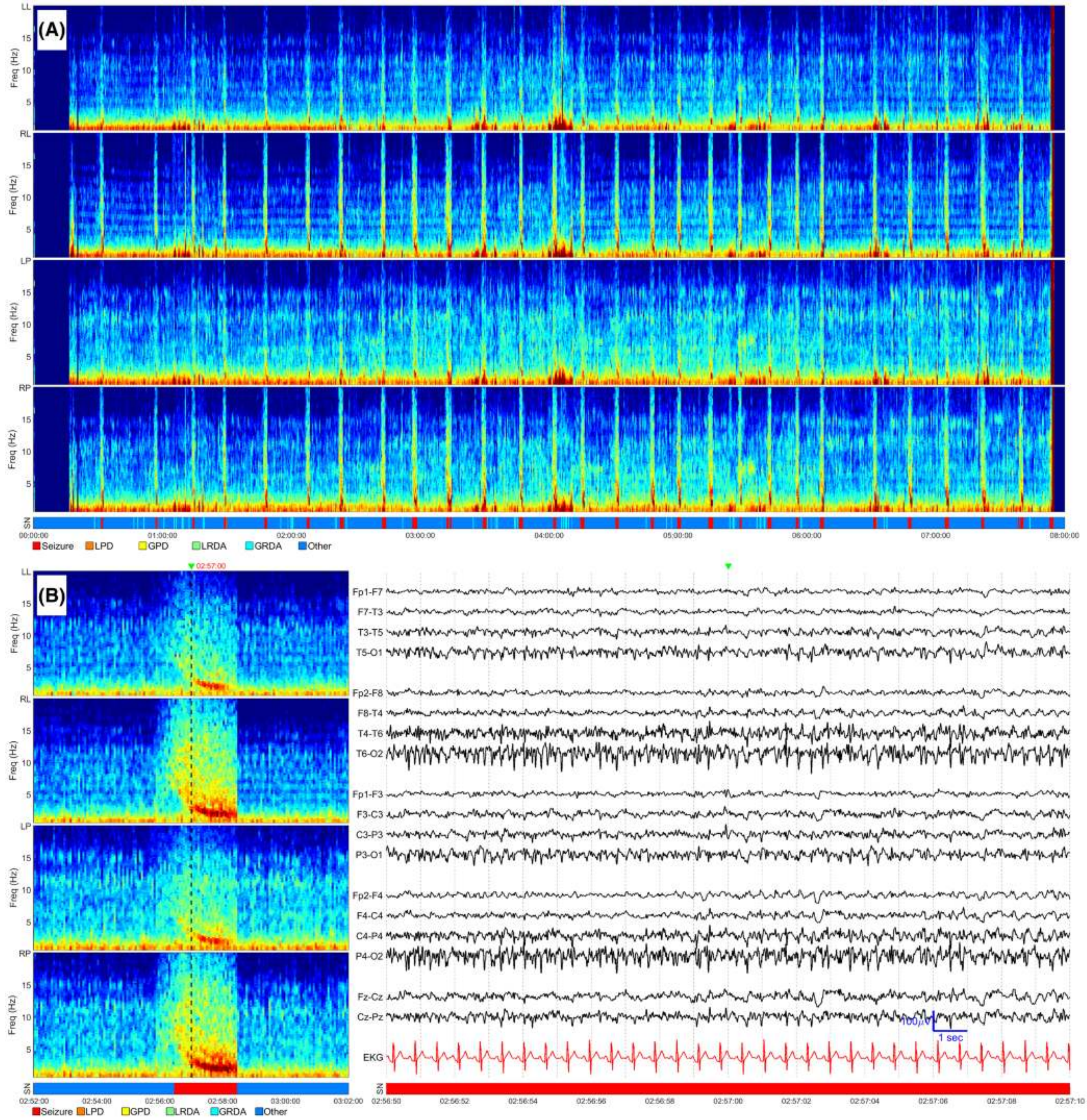

FIGURE 5 Seizure detection tool. (A) Spectrogram of an 8-h continuous electroencephalographic (EEG) recording. The four panels depict the average spectral power over the left lateral (LL), right lateral (RL), left paracentral (LP), and right paracentral (RP) regions. The spectrogram shows a narrowband monotonous pattern of delta activity, interrupted at semiregular intervals by 28 distinct flame-shaped electrographic seizures. Automated detections of seizures by a neural network model (SPaRCNet) are indicated in red in the color bar below. (B) A close-up of one of the seizures shown in panel A. The spectrogram is demonstrated over 10 min. A 20-s sample of the raw EEG is shown in bipolar montage, centered at the location indicated in the spectrogram by the vertical dashed line and green triangle. GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; SN, SPaRCNet, Seizure, Periodic and Rhythmic Continuum pattens Deep Neural Network.

inadvertent PHI discovery must be reported. Data are accessed through a secure, cloud-based platform hosted on AWS as part of the AWS Open Data Sponsorship Program rather than via direct downloads, ensuring privacy and scalability. Full DUA and access details are on the bdsp.io's "Harvard Electroencephalography Database" project page (https://bdsp.io/content/harvard-eeg-db/3.0/).

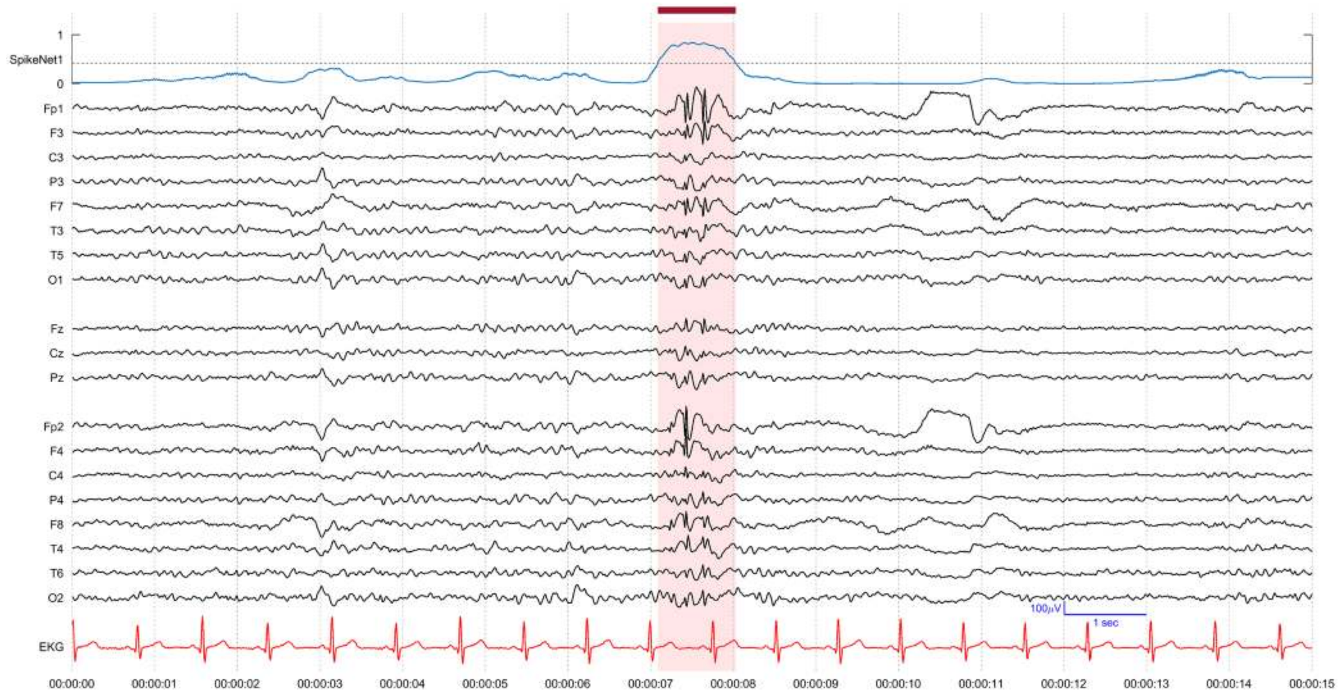

FIGURE 6 Epileptiform discharge detection tool. An automatically detected bifrontal spike-and-wave epileptiform discharge is shown. The common average referential montage displays the electroencephalogram (EEG) over 15 s. The output of the SpikeNet 1.0 algorithm is shown above the EEG. The red overbar and transparent pink vertical bar indicate the period when the probability exceeds the detection threshold (theta = .42).

3.10 | How to access the data

To access and interact with HEEDB data via BDSP, users must complete setup steps to ensure secure and efficient access (see Supplementary E in Data S1). These steps empower researchers to fully leverage the rich dataset while maintaining strict data security. Instructions are provided on bdsp.io (https://bdsp.io//about/howto_accessdata/).

3.11 | Data use cases

The database has already been pivotal in developing and validating ML tools to improve patient care in neurology.

3.11.1 | Epilepsy

HEEDB has contributed to multiple epilepsy-related advances, including rapid annotation methods,[25,26] interictal epileptiform discharge (a.k.a. spike and sharp wave) detection,[27,28] children seizure risk analysis,[29] epilepsy diagnosis in resource-limited settings,[30] and posttraumatic epilepsy predictions.[31] See Supplementary F.1 in Data S1 for additional research.

3.11.2 | Sleep

Investigations leveraging automated and semiautomated neural network approaches have explored sleep architecture and the relationship between sleep patterns and brain health[32] and with neurodegenerative disease, including Alzheimer disease.[33]

3.11.3 | Critical care

Harmful brain activity

HEEDB data supports research on ICU EEG monitoring,[34,35] seizure risk,[36] delayed cerebral ischemia,[37] drug effect tracking,[38] and EEG-based neurologic outcome analysis.[39] These studies have advanced automated seizure detection,[40,41] seizure impact assessment,[42] other harmful brain activity detection,[43,44] and NLP-based clinical data extraction.[45] See Supplementary F.3 in Data S1 for additional research.

Coma

ML models derived from HEEDB have contributed to improved coma prognostication capabilities following cardiac arrest.[46,47]

Encephalopathy

Studies on EEG patterns in patients with altered mental status are refining physiologically based methods for detecting and grading delirium, encephalopathy,[48,49] and chimeric antigen receptor T-cell neurotoxicity.[50]

3.11.4 | Future opportunities

The ML studies (Supplementary F in Data S1) have deepened our understanding of how distinct EEG patterns associate with cognitive function, disease progression, and treatment response. We hope future investigators will capitalize on HEEDB's wealth of multimodal information—from continuous EEG recordings to extensive clinical annotations—to build predictive models that integrate physiological, imaging, and laboratory data. New frontiers include advancing personalized interventions for neurocritical care, refining sedation protocols, and developing precise diagnostic tools for emerging neurological disorders. As suggested by the diverse work cited above, continued research using HEEDB data promises to extend the impact of HEEDB across epilepsy, neurocritical care, and the broader neurosciences.

Future researchers can explore additional avenues. The database's extensive EEG and clinical records present rich opportunities for pioneering brain health and neurological disease management discoveries.

4 | CHALLENGES AND OPPORTUNITIES

4.1 | Limitations

Although the HEEDB and bdsp.io initiatives lay a strong foundation for collaborative neuroscience research, several challenges remain. Current EEG annotations are provided by single experts, lacking multirater consensus and serving as weak supervision rather than ground truth. To address this, we are implementing a multiexpert annotation workflow to improve label consistency and quality. We also assign different reliability levels to findings and labels using a tiered metadata system—"annotation," "report," and "verified" (annotation < report < verified)—to help users better assess data trustworthiness and make informed use of the dataset.

The dataset's demographic distribution is skewed, with White patients overrepresented and limited inclusion of Asian or Indigenous groups. Institutional differences in documentation formats further complicate data integration. To mitigate this, we are developing NLP tools to standardize records across sites.

Together, these efforts aim to enhance data harmonization and support more scalable, generalizable research. Despite the challenges, the growing bdsp.io ecosystem offers unprecedented opportunities to drive innovation in brain health.

4.2 | Further expanding the frontier of brain data science

The HEEDB initiative represents a significant step toward democratizing EEG research and advancing brain data science. Through bdsp.io, researchers can contribute new datasets, establish complementary repositories, and leverage our open-access deidentification tools, creating new opportunities to accelerate discovery. This approach supports compliance with NIH data-sharing requirements, facilitates exact reproducibility of results, and promotes transparency and collaboration within the scientific community. Researchers validate their work and provide a foundation for others to build upon by publishing data alongside their findings.

Beyond EEG data, bdsp.io supports diverse brain-related datasets, including brain magnetic resonance images, intracranial pressure waveforms, polysomnography, and other physiological monitoring data. These multimodal datasets enable researchers to examine neurological conditions comprehensively, paving the way for integrative and innovative approaches to brain health.

A key example of collaborative growth is the Critical Care EEG Monitoring Research Consortium (CCEMRC), comprising more than 50 sites that have conducted EEG research for more than a decade. CCEMRC members are actively joining the BDSPC3 consortium, with 22 sites already participating and committing to contribute at least 100 EEG cases each. The BDSPC3 framework simplifies research data-sharing; sites sign a standardized DUA, secure local institutional review board approval, and collaborate on approved projects. Participating institutions share data under embargo, conduct multicenter research, and publish their findings and associated datasets on bdsp.io with citable DOIs.

This model extends beyond individual datasets, encouraging the creation of tailored repositories that address specific research questions. By integrating multimodal data and fostering a culture of open collaboration, bdsp.io expands the boundaries of precision brain health. This vision empowers the research community to collectively transform neurological conditions' study, diagnosis, and treatment, driving a new era of brain data science innovation.

5 | CONCLUSIONS

HEEDB represents a transformative leap forward in clinical neuroscience by addressing critical gaps in large-scale EEG data accessibility. With more than 280 000 EEG recordings integrated with comprehensive electronic health records, HEEDB democratizes access to brain-focused physiological data and enables cutting-edge research in neurology, critical care, sleep medicine, and beyond.

By adhering to standardized frameworks such as BIDS and aligning with FAIR principles, HEEDB ensures that this valuable resource is interoperable, accessible, and ready for use in advanced ML- and AI-driven diagnostic tools. Furthermore, its robust deidentification processes maintain patient privacy while preserving clinical relevance, establishing it as a biomedical research model for ethical data-sharing.

Through its diverse applications—from epilepsy diagnostics to critical care interventions and sleep research—HEEDB has already facilitated significant advances in understanding neurological conditions. Looking ahead, it holds immense potential to drive innovations in personalized medicine, enhance diagnostic accuracy, and enable the development of novel biomarkers and therapeutic strategies.

HEEDB's collaborative model, supported by the Brain Data Science Platform and the AWS Open Data Sponsorship Program, invites researchers worldwide to contribute to and benefit from this unique repository. By fostering a global community of contributors and users, HEEDB is poised to accelerate discoveries and improve clinical neuroscience outcomes, helping to shape the future of brain health research and improve patient care.

AUTHOR CONTRIBUTIONS

Chenxi Sun, Jin Jing, and M. Brandon Westover have full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. *Concept and design:* Jurriaan Peters, Tobias Loddenkemper, Jong Woo Lee, Sahar Zafar, and M. Brandon Westover. *Acquisition, analysis, or interpretation of data:* All authors. *Drafting of the manuscript:* Chenxi Sun, Jin Jing, and M. Brandon Westover. *Critical revision of the manuscript for important intellectual content:* All authors. *Statistical analysis:* Chenxi Sun, Jin Jing, and M. Brandon Westover. *Administrative, technical, or material support:* Jurriaan Peters, Tobias Loddenkemper, Jong Woo Lee, Sahar Zafar, and M. Brandon Westover. *Supervision:* M. Brandon Westover.

ACKNOWLEDGMENTS

This work was funded by grants from the National Institutes of Health (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119, K23NS124656, K23NS119794, 1OT2OD032701, R01NS128342, K99NS138680), the National Science Foundation (2014431), the Department of Defense (ERP 220036, W81XWH-19-1-0861, W81XWH-21-C-0075), the American Heart Association (AMFDP Faculty Development Program [843457], 20CDA35310297, 24DIVSUP1274116), the Epilepsy Research Fund, the Regents of the University of California, Cures Within Reach (2022CAL-Amorim), the Zoll Foundation, the Hellman Foundation, and Beth Israel Deaconess Medical Center.

CONFLICT OF INTEREST STATEMENT

M.B.W. is a cofounder of and consultant to Beacon Biosignals, with personal equity, and receives royalties from Wolters Kluwer and Demos Medical. D.M.G. is an unpaid advisor for Epilepsy AI and Eysz, and a paid advisor for Magic Leap. He has received speaker fees from AAN, AES, ACNS, NNS, and AI in Epilepsy and Neurology, and served as a consultant for Neuro Event Labs, IDR, LivaNova, and Health Advances. T.L. is an inventor on patents and patent applications related to the detection, prediction, management, and treatment of epilepsy and seizures; has received device donations from Epitel and Empatica; has received travel support from academic and scientific organizations; and hosts international fellows. C.T.S. and B.G. are employed by and hold equity in Amazon Web Services. J.R. is the founder of the Global Brain Care Coalition and cofounder of McCance for Brain Health, has consulted for the NFL and Eli Lilly, and holds leadership roles at Columbia University, *The European Stroke Journal*, and *The Lancet Neurology*. The remaining authors report no conflicts of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

DATA AVAILABILITY STATEMENT

All data needed to reproduce the results are available at bdsp.io at https://bdsp.io/content/harvard-eeg-db/4.1/.

ORCID

Chenxi Sun https://orcid.org/0000-0002-1762-0877
Niels Turley https://orcid.org/0009-0009-4806-578X
Andrew J. Cole https://orcid.org/0000-0002-0828-826X
Daniel M. Goldenholz https://orcid.org/0000-0002-8370-2758
Alice Lam https://orcid.org/0000-0001-7754-4637
Catherine Chu https://orcid.org/0000-0001-7670-9313
Aditya Gupta https://orcid.org/0000-0002-5243-368X
Marta Fernandes https://orcid.org/0000-0002-7203-2832
Kaileigh Gallagher https://orcid.org/0009-0005-2802-9790

Jong Woo Lee https://orcid.org/0000-0001-5283-7476
Sahar Zafar https://orcid.org/0000-0001-5252-5376
M. Brandon Westover https://orcid.org/0000-0003-4803-312X

REFERENCES

1. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21:957–8.
2. Ahalt S, Avillach P, Boyles R, Bradford K, Cox S, Davis-Dusenbery B, et al. Building a collaborative cloud platform to accelerate heart, lung, blood, and sleep research. J Am Med Inform Assoc. 2023;30:1293–300.
3. NOT-OD-21-013: Final NIH Policy for Data Management and Sharing. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html
4. Cloud Life Sciences public datasets | Cloud Life Sciences Documentation. Google Cloud. Available from: https://cloud.google.com/life-sciences/docs/resources/public-datasets
5. Open Data Sponsorship Program | AWS. Amazon Web Services, Inc. Available from: https://aws.amazon.com/opendata/open-data-sponsorship-program/.
6. Wu X, Li W, Tu H. Big data and artificial intelligence in cancer research. Trends Cancer. 2024;10:147–60.
7. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023;24:125–37.
8. Mooney SJ, Westreich DJ, El-Sayed AM. Epidemiology in the era of big data. Epidemiology. 2015;26(3):390–4.
9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101:E215–E220.
10. Zhang Y, Kim M, Prerau M, Mobley D, Rueschman M, Sparks K, et al. The National Sleep Research Resource: making data findable, accessible, interoperable, reusable and promoting sleep science. Sleep. 2024;47:zsae088.
11. Ostendorf AP, Ahrens SM, Lado FA, Arnold ST, Bai S, Bensalem Owen MK, et al. United States epilepsy center characteristics. Neurology. 2022;98:e449–e458.
12. Benbadis SR. Errors in EEGs and the misdiagnosis of epilepsy: importance, causes, consequences, and proposed remedies. Epilepsy Behav. 2007;11:257–62.
13. Benbadis SR. Misdiagnosis of epilepsy due to errors in EEG interpretation. Pract Neurol. 2007;7:323–5.
14. Struck AF, Rodriguez-Ruiz AA, Osman G, Gilmore EJ, Haider HA, Dhakar MB, et al. Comparison of machine learning models for seizure prediction in hospitalized patients. Ann Clin Transl Neurol. 2019;6:1239–47.
15. GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol. 2019;18:459–80.
16. Brain Data Science Platform increases EEG accessibility with open data and research enabled by AWS | AWS Public Sector Blog. Available from: https://aws.amazon.com/blogs/publicsector/brain-data-science-platform-increases-eeg-accessibility-with-open-data-and-research-enabled-by-aws/.
17. U.S. Census Bureau. Measuring Racial and Ethnic Diversity for the 2020 Census. [Accessed August 12, 2021]. Available from: https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html.
18. Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, Lee T, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. CoRR abs/2403.18421 2024.
19. Quest Diagnostics. ICD-10 Codes for Neurology (MI4957). 2015. Available from: https://www.questdiagnostics.com/content/dam/corporate/restricted/documents/icd-10-codes/cpt-2015/icd_10_codes_for_neurology-mi4957/ICD_10_Codes_for_Neurology-MI4957.pdf
20. World Health Organization. Anatomical Therapeutic Chemical (ATC) Classification. [Accessed January 2, 2025]. Available from: https://www.who.int/tools/atc-ddd-toolkit/atc-classification
21. U.S. Department of Health and Human Services. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Public Law 104–191. 1996. Available from: https://www.hhs.gov/hipaa/
22. Radhakrishnan L, Schenk G, Muenzen K, Oskotsky B, Ashouri Choshali H, Plunkett T, et al. A certified de-identification system for all clinical text documents for information extraction at scale. JAMIA Open. 2023;6(3):ooad045.
23. Jing J, Dauwels J, Cash S, Westover MB. SpikeGUI: software for rapid interictal discharge annotation via template matching and online machine learning. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc Annu Conf 2014, 4435–4438.
24. Jing J, Sun H, Kim JA, Herlopian A, Karakis I, Ng M, et al. Development of expert-level automated detection of Epileptiform discharges during electroencephalogram interpretation. JAMA Neurol. 2019;77:103–8. https://doi.org/10.1001/jamaneurol.2019.3485
25. Jing J, d'Angremont E, Ebrahim S, Tabaeizadeh M, Ng M, Herlopian A, et al. Rapid annotation of seizures and interictal-ictal-injury continuum EEG patterns. J Neurosci Methods. 2021;347:108956.
26. Zafar SF, Rosenthal ES, Jing J, Ge W, Tabaeizadeh M, Aboul Nour H, et al. Automated annotation of Epileptiform burden and its association with outcomes. Ann Neurol. 2021;90:300–11.
27. Jing J, Herlopian A, Karakis I, Ng M, Halford JJ, Lam A, et al. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. JAMA Neurol. 2020;77:49–57.
28. Jing J, Dauwels J, Rakthanmanon T, Keogh E, Cash SS, Westover MB. Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. J Neurosci Methods. 2016;274:179–90.
29. Chinappen DM, Xiao G, Jing J, Spencer ER, Eden UT, Kramer MA, et al. Spike height improves prediction of future seizure risk. Clin Neurophysiol. 2023;150:49–55.
30. McInnis RP, Ayub MA, Jing J, Halford JJ, Mateen FJ, Brandon Westover M. Epilepsy diagnosis using a clinical decision tool and artificially intelligent electroencephalography. Epilepsy Behav. 2023;141:109135.
31. Kim JA, Boyle EJ, Wu AC, Cole AJ, Staley KJ, Zafar S, et al. Epileptiform activity in traumatic brain injury predicts post-traumatic epilepsy. Ann Neurol. 2018;83:858–62.

32. Hogan J, Sun H, Paixao L, Westmeijer M, Sikka P, Jin J, et al. Night-to-night variability of sleep electroencephalography-based brain age measurements. Clin Neurophysiol. 2021;132:1–12.
33. Lam AD, Sarkis RA, Pellerin KR, Jing J, Dworetzky BA, Hoch DB, et al. Association of epileptiform abnormalities and seizures in Alzheimer disease. Neurology. 2020;95:e2259–e2270.
34. Moura LMVR, Shafi MM, Ng M, Pati S, Cash SS, Cole AJ, et al. Spectrogram screening of adult EEGs is sensitive and efficient. Neurology. 2014;83:56–64.
35. Sun H, Nagaraj SB, Akeju O, Purdon PL, Westover MB. Brain monitoring of sedation in the intensive care unit using a recurrent neural network. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf 2018.
36. Struck AF, Osman G, Rampal N, Biswal S, Legros B, Hirsch LJ, et al. Time-dependent risk of seizures in critically ill patients on continuous electroencephalogram. Ann Neurol. 2017;82:177–85.
37. Kim JA, Zheng WL, Elmer J, Jing J, Zafar SF, Ghanta M, et al. High epileptiform discharge burden predicts delayed cerebral ischemia after subarachnoid hemorrhage. Clin Neurophysiol. 2022;141:139–46.
38. Zafar SF, Rosenthal ES, Postma EN, Sanches P, Ayub MA, Rajan S, et al. Antiseizure medication treatment and outcomes in patients with subarachnoid hemorrhage undergoing continuous EEG monitoring. Neurocrit Care. 2022;36:857–67.
39. Fernandes MB, Valizadeh N, Alabsi HS, Quadri SA, Tesh RA, Bucklin AA, et al. Classification of neurologic outcomes from medical notes using natural language processing. Expert Syst Appl. 2023;214:119171.
40. Jing J, Ge W, Hong S, Fernandes MB, Lin Z, Yang C, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. Neurology. 2023;100:e1750–e1762.
41. Jing J, Ge W, Struck AF, Fernandes MB, Hong S, An S, et al. Interrater reliability of expert Electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. Neurology. 2023;100:e1737–e1749.
42. Westover MB, Shafi MM, Bianchi MT, Moura LMVR, O'Rourke D, Rosenthal ES, et al. The probability of seizures during EEG monitoring in critically ill adults. Clin Neurophysiol. 2015;126:463–71.
43. Westover MB, Ching S, Kumaraswamy VM, Akeju SO, Pierce E, Cash SS, et al. The human burst suppression electroencephalogram of deep hypothermia. Clin Neurophysiol. 2015;126:1901–14.
44. Zafar SF, Amorim E, Williamsom CA, Jing J, Gilmore EJ, Haider HA, et al. A standardized nomenclature for spectrogram EEG patterns: inter-rater agreement and correspondence with common intensive care unit EEG patterns. Clin Neurophysiol. 2020;131:2298–306.
45. Fernandes M, Westover MB, Singhal AB, Zafar SF. Automated extraction of stroke severity from unstructured electronic health records using natural language processing. J Am Heart Assoc. 2024;13:e036386.
46. Amorim E, Zheng W-L, Jing J, Ghassemi MM, Lee JW, Wu O, et al. Neurophysiology state dynamics underlying acute neurologic recovery after cardiac arrest. Neurology. 2023;101:e940–e952.
47. Pati S, McClain L, Moura L, Fan Y, Westover MB. Accuracy of limited-montage electroencephalography in monitoring Postanoxic comatose patients. Clin EEG Neurosci. 2017;48:422–7.
48. Sun H, Kimchi E, Akeju O, Nagaraj SB, McClain LM, Zhou DW, et al. Automated tracking of level of consciousness and delirium in critical illness using deep learning. NPJ Digit Med. 2019;2:89.
49. Luccarelli J, Smith JR, Fricchione G, Westover MB. Generalized periodic discharges associated with catatonia and delirium: a case series. J Neuropsychiatry Clin Neurosci. 2024;36:340–3.
50. Herlopian A, Dietrich J, Abramson JS, Cole AJ, Westover MB. EEG findings in CAR T-cell therapy-related encephalopathy. Neurology. 2018;91:227–9.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sun C, Jing J, Turley N, Alcott C, Kang W-Y, Cole AJ, et al. Harvard Electroencephalography Database: A comprehensive clinical electroencephalographic resource from four Boston hospitals. Epilepsia. 2025;66:3411–3425. https://doi.org/10.1111/epi.18487