RESEARCH STRATEGY EPIBIOS4RX – INFORMATICS AND ANALYTICS CORE

Significance: A fundamental challenge in discovering treatments and associated biomarkers of epileptogenesis is that this process is likely multifactorial and crosses multiple modalities. Investigators must have access to a large number of high quality, well-curated data points and study subjects in order for biomarker signals to be detectable above the noise inherent in complex phenomena like epileptogenesis, TBI, and conditions of data collection. Additionally, data generating and collecting sites are spread worldwide among different laboratories, clinical sites, heterogeneous data types, formats, and across multicenter preclinical trials. Before the data can even be analyzed, a central platform is needed to standardize these data and provide tools for searching, viewing, annotating, and analyzing them. **The Scientific Premise of EpiBioS4Rx is: Epileptogenesis after TBI can be prevented with specific treatments; the identification of relevant biomarkers and performance or rigorous preclinical trials will permit the future design and performance of economically feasible full-scale clinical trials of antiepileptogenic therapies.** The Informatics and Analytics Core (IAC) will centralize an enduring data archive and analytic tools that will allow the broader epilepsy research community to identify and validate biomarkers of epileptogenesis in images, electrophysiology, and molecular/serological/tissue studies. Beyond creating a centralized data repository, the IAC will pioneer innovative standardization/co-registration methods, fully supported by novel image and electrophysiology processing methods to extract candidate biomarkers from the diverse data. Not only will a well-curated and standardized multi-modal data set facilitate the development of models of epileptogenesis, it will also ensure that such models are statistically significant and can be validated. Based on our previous experience with similar multicenter projects, we are confident that our infrastructure will lead to success in this project. Just as the Alzheimer's Disease Neuroimaging Initiative (ADNI) has been a powerful catalyst for success in biomarker research in Alzheimer's disease (AD), the IAC has the power to foster a similar and potentially greater level of success for the study of epileptogenesis.

The sheer amount of data to be collected in these studies is unprecedented: video-EEG from scores of animals after TBI recorded continuously for 6 months, in addition to prolonged continuous ICU EEG recordings from humans and intermittent sampling of brain images, blood, and tissue data. The data, measured in tens to one hundred Terabytes, represents investigation on a scale that was not possible even 5 years ago. It leverages state of the art analysis tools to track candidate biomarkers and their statistical associations. To analyze these data properly, it requires a diverse, accomplished group of investigators spanning neurology, neuroscience, imaging, mathematics, engineering, and computer science, as well as collecting comprehensive data in parallel from humans and animal models after TBI. The IAC will be seamlessly integrated with *Projects 1-3*, assisting in collecting data and providing analytic tools that will lead to biomarkers of epileptogenesis. By combining new data capabilities and our powerful, best-in-class, interdisciplinary team, quantitative models of epileptogenesis may be possible. These types of models will enrich preclinical trial populations, expedite interventions to prevent epilepsy after brain insults, and document epilepsy before late seizures occur. Based on previous studies, it is likely that there are reproducible changes in biomarkers, such as occurrence of pHFOs in the intracranial EEG, which identify the presence of epilepsy before its overt clinical expression[64,74,75]. Our findings from this study could operationally change our definition of when epilepsy begins or is prevented.

The IAC will bring big data techniques and rigorous analysis to longitudinal data collected from humans and animal models of TBI, epilepsy, and their interaction. It will develop and implement new approaches, including novel graphical methods to visualize multivariable interactions, to quantify phenotype and molecular profiles in these disorders. A first-rate bioinformatics platform, LONI, will focus on TBI and epileptogenesis research. The tools, pipelines, and protocols developed for this proposal will be made available to the epilepsy research community, with the potential to change, long-term, the way that images, video, electrophysiology, proteomics, and metadata are analyzed in these fields. Quantitative and data mining methods will enable investigators to record and analyze gold-standard data and create a shared bioinformatics resource for epilepsy research that will live on long after the end of this project. Perhaps most importantly, the IAC will provide the technical sophistication to tease out the interaction between the complex processes studied in *Projects 1-3*, integrating multi-modal data in a way that has been beyond the capability of a single laboratory or center. The IAC will provide a lasting and open platform for standardized biomarker research in both TBI and epilepsy as well as engage with and guide the projects that, together, will lead to future clinical trial development.

Innovation: The central innovation in the IAC involves the development of techniques and standards to (1) import and interlink data and metadata from a variety of different modalities, (2) support highly flexible search and browsing of these data, and (3) enable analysis of the data and creation of models followed by sharing of results along with their provenance. This infrastructure and its related standards (derived from the Common

Data Elements) are vital to promoting broad cross-community data sharing and collaboration. Our effort builds upon robust existing platforms and joint expertise in computer science data integration, image analysis, electrophysiology analysis, and big data. Novel techniques will interlink and co-register data, search across link relationships, and pivot from data item to related data item. The IAC provides the platform and lower-level primitives upon which data mining and analysis for biomarkers will occur. These “coupled analytics” will be used in our efforts to establish data relationships and to identify variables, weighting, and other statistics necessary to formulate models of epileptogenesis.

There are eight major innovations in this project:

(**1) A single automated image processing platform for both human and animal imaging data:** The IAC, under the auspices of LONI, will provide state-of-the-art data processing and statistical techniques to analyze complex MRI data sets, generate adjudicated imaging data sets (similar to that shown as preliminary data in *Project 1*), and identify key temporal and spatial imaging features associated with increased risk of developing seizures. The study of epileptogenesis in vivo has never incorporated these tremendously powerful tools, and the CWOW’s skillset is uniquely positioned to deploy them effectively for the first time within the field. Table 1 outlines the main hypotheses in the projects, the statistical models to be used, and impact towards planning a clinical trial.

(2) More powerful mathematical techniques applied to human and animal EEG analysis: The CWOW’s prolonged continuous data streams of intracranial, cortical surface, and scalp EEG from humans and animal models of epilepsy spanning months represent a potentially overwhelming amount of data. However, the IAC team’s expertise in both neuroscience and mathematics makes it uniquely positioned to discover previously unidentified biomarkers of epileptogenesis. By applying cutting edge mathematical tools via supervised and unsupervised learning methods, the IAC will be able to subject a robust data set to recently pioneered data analysis tools. The IAC will adapt methods used to analyze prolonged continuous intracranial EEG data sets from humans spanning over 18 months but adapt them to the specific panel of electrophysiological biomarkers, outlined in the Approach section below[11].

(3) Validating multi-modal data using novel statistical tools: Comparing multi-modal data from humans and animals has been limited by concerns that such comparisons might not be appropriate. By applying novel, validated statistical methods to elucidate epileptogenesis, we can ensure the reliability of any such comparisons. These include robust methods for logistic regressions, composite endpoint analyses, mixed models analyses, and multimodel inference techniques. Associating and integrating electrophysiological and blood biomarker data and multi-modal neuroimaging data will greatly aid the identification of time courses of variables that change in concert, and consequently, potential stages of epileptogenesis (see *Project 3*). The process of validation will rely on modern approaches to appraising propagation of uncertainty wherein computational modeling is recursively benchmarked against the data, along with concomitant system and statistical analyses. This involves a variety of both deterministic and nondeterministic validation metrics, which themselves are subject to boundary conditions and uncertainty estimates[10,59,60]. Monte Carlo simulations, sensitivity analyses and related cutting-edge approaches will be employed.

(4) Machine learning using novel statistical methods that may lead to the development of models of epileptogenesis: Change Point Analysis, Censor Analysis, Diffusion Maps, and Hidden Markov Models will be adapted to unsupervised learning. To our knowledge, these techniques have not previously been applied to studies of epileptogenesis (See Previous Work and Preliminary Analysis).

(5) Innovative methods to interrogate and understand interrelationships between EEG and MRI: Our preliminary work in TBI has fostered the understanding of injuries and progressive loss of white matter in addition to how these affect epileptiform source localization[70,33,23]. Such integrative approaches can provide critical insights into the brain-wide effects of TBI, PTE, and their modulation over time, and then form the basis for improvements in outcome prediction accuracy.

(6) Integrating multimodal data with serological data: We will integrate serum biomarker results from Dr. Agoston, with imaging, EEG, and outcomes data from all projects and integrate between animals and humans. Table 2 outlines the multivariable biomarker data to be collected, integrated, and analyzed.

(7) Web-based data entry tools for each project, including animals and humans: Multiple clinical sites will input meta-data, EEG, MRI, and outcomes electronically using standardized format. The IAC will integrate animal data with common data tags incorporating timing after injury, anatomic location, experimental interventions, and preclinical drug trial results, thus facilitating analysis and translation of findings.

(8) Multi-modal preclinical trial: While a wide range of data have been collected from different epilepsy subjects, there is no consolidated clinical data set available to the research community that is multi-modal and comprises both humans and animal models. Beyond consolidating existing data, EpiBioS4Rx will conduct a preclinical trial to create a data set, which, by using a fixed group of subjects, will control for the effects of variation between individuals and allow for cross-comparison between human and animal data.

Project	Hypothesis	Statistical Methods	Clinical Trial Planning
1	Temporal lobe TBI in the rat will result in electrophysiological (seizures, pHFOs, rHFOSs) abnormalities in the perilesional cortex and hippocampus	A, B	1, 3, 5
1	Temporal lobe TBI in the rat will result in structural pathology in entorhinal-hippocampal, septo-hippocampal, and thalamo-cortical networks	A, B	1, 2, 3, 5
1	TBI induces changes in plasma proteins and/or microRNAs, which signal neuronal/glial degeneration, axonal and dendritic injury, neuroinflammation, and metabolic changes	C, D, E, F	1, 2,5
2	A multi-modality rapid screening platform for target relevance and engagement, modification of early stage post-TBI seizures, EEG and plasma biomarkers and persistence of effects beyond treatment exposure helps select optimal treatment protocols for candidate AEG treatments for PTE	A-H	2, 3, 5
2	Early stage treatments that have long-lasting modifying effects on relevant targets and MRI/EEG/plasma TBI biomarkers can also have AEG effects for PTE epileptogenesis	B, G, H, I	2, 3, 4, 5
3	Early post-traumatic epileptic EEG activity (seizures, pHFOs, rHFOSs) indicates the presence of an epileptogenic process in patients after moderate to severe TBI.	A, G	1, 4, 5
3	Acute structural/functional disconnections abnormalities within hippocampal or thalamo-cortical networks circuits are associated with epileptogenesis after severe TBI.	A, G	1, 3, 4, 5
3	Specific epileptogenic pathways amenable to therapeutic interventions will generate biomarkers that can be monitored in the post-traumatic patient.	C, D, E, F	1, 3, 4, 5
3	Prospective implementation of high resolution advanced EEG methods among our TBI-ICU-EEG-study sites will enable the selection of an enriched population of patients at high risk based of developing PTE that can be targeted in a future interventional trial.	G, I	3, 4, 5

Statistical Methods: (A) Conventional and robust regression and logistic regression analyses, (B) Composite endpoint analyses, (C) Change point analyses, (D) Censor analyses, (E) Hidden Markov modeling, (F) Hierarchical clustering analyses, (G) Multimodel inference techniques, (H) Mixed models analyses

Clinical Trial Planning Contribution: (1) Proof of mechanism, (2) Treatment effect in animals, (3) Validation of biomarkers, (4) Biomarkers can be monitored in humans, (5) Provides DSMB and *Public Engagement Core* with data for interventional trial plan

Table 1: Main hypotheses, the statistical models to be used, and impact towards planning a clinical trial.

Biomarker	Mechanism	Expected change	Treatment in Project 2	Effect of Treatment in Animals	Statistical Testing
EEG: pHFO, rHFOSs, seizures	Clustering of Synchronized depolarization	Present in 75% of those who go on to get PTE	Multiple drugs: Z994, Sodium Selenate; Deferiprone; Vineret +- VX765	50% reduction	A, I
Tau (total)	Protein Phosphatase 2	Increase in PTE 150-200%	Sodium Selenate	60% reduction	A, I

IL1β	CNS inflammation	Increase in PTE 100%	Anti-inflammatory, miRNAs	50% reduction	A, I
TBI biomarkers GFAP, S100B, MBP, GFAP	CNS injury	Increased in PTE 50% above TBI background	Multiple drugs	30% reduction in one or more outcomes	A, B, I
miRNA such as 106b-5p	Inflammation	Increase in PTE 50% above TBI background	Antiinflammatories	60% reduction to normal	A, I
Hippocampal /Thalamic structural functional changes	Network plasticity	Altered connectivity as compared with TBI background	Multiple	40% reduction in loss of hippocampal volume after drug given	A, I

Table 2: Multivariable biomarker data to be collected, integrated, and analyzed.

Power Analysis (*Project 3*): In this study, we assessed the number of human subjects necessary to answer several aims, and hence, the numbers vary across aims[9]. For Specific Aim 1, the number to determine if early seizures (recorded on depth or surface EEG) predict late PTE, based on a 50% incidence of early seizures and 50% incidence of PTE (in this highly selected group), was n=144, assuming a 30% attrition rate, initial n=187. For HFO, we have only animal data to consider for power estimates, with 60% of animals having HFOs early after TBI. We anticipate incomplete overlap between early seizures and HFOs, hence an additional 100 subjects are being added empirically for depth EEG pHFO and rHFOSs studies. For Specific Aim 2, we powered the MRI studies based on our existing work regarding temporal lobe injury being prevalent in the PTE group (74% in PTE vs 35% in non-PTE; effect size h=0.805, total n=120, or initial n before attrition=172). For Specific Aim 3, we used results from animal studies on the effect of PTE on tau[45] and the effects of sodium selenate to mitigate seizures and tau expression. We also used existing human studies of serum tau levels after TBI to model the expected increase in serum tau that occurs due to trauma complicated by seizures. We powered the study in order to demonstrate between group differences in those patients with severe TBI plus seizures vs those with severe TBI alone during the initial week after TBI. Based on published values for serum tau (ranging from 10 pg/ml (SD 15)[61] in mild TBI to 436 pg/ml (SD 472)[45], the number of subjects needed was n=108, and with 30% attrition, n=144. We plan to enroll 300 subjects, assuming some non-overlapping results in Specific Aims 1-3.

Approach: Previous and Preliminary Work: Below we briefly present some examples of pertinent preliminary work by the LONI team.

Regional Brain Shape and Volume Analysis: We have applied local shape analysis modeling for comparing morphometric group differences from serial brain images in TBI. Using non-parametric permutation-based tests, we have computed probability values (p-values) quantifying the differences between TBI patients with or without post-traumatic seizures. **Figure 1** (item 2) shows the end-to-end pipeline workflow demonstrating the pre-processing, local shape analysis, and visualization of the imaging and clinical data. **Figure 1** (items 4&5) shows the p-values obtained by using all subjects' left and right thalamic shapes and generating average surfaces, volumes, and connections. Using this approach, we found significant changes in hippocampal volume, thalamic volume, and combined models of anatomic volume, in addition to EEG spike counts for patients who developed epilepsy after TBI (**Figure 7**). See preliminary data in Project 1 (Table 3 in *Project 1*).

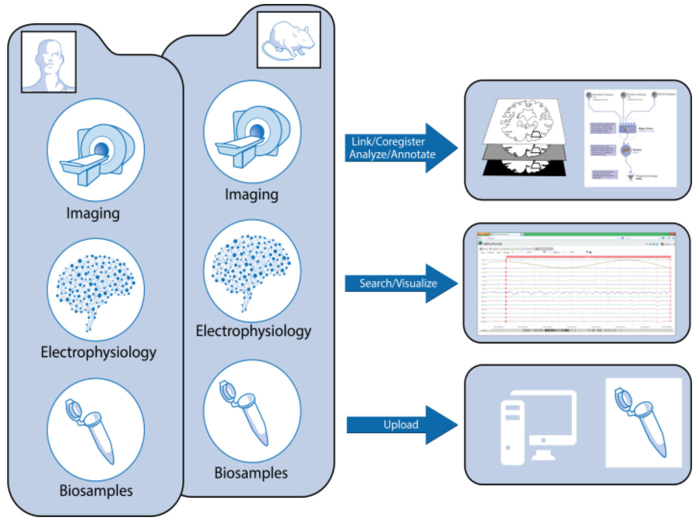


Figure 1. The EpiBioS4Rx portal supports hetero¬geneous data and provides essential features including upload/ingest, search/visualize, link/co-register, and analyze/annotate, enabling the specific aims of the IAC.

Image Data Analysis and Modeling from Non-Human Species: Data acquisition protocols (anatomical, DTI, and resting state fMRI) for animal studies will parallel human studies and leverage the analysis pathways described above for human data. This includes analyzing anatomical data and connectivity as well as quantifying lesions and their progression (Project 2). Changes and rate of progression/correlation with electrophysiology and other biomarkers will be performed separate from, but in parallel with, human analysis. Animal pipeline design and testing have been initiated.

A broad range of data will be measured in hospitals, animal laboratories, and multicenter trials conducted worldwide. Such a multidisciplinary translational research agenda requires handling large, highly heterogeneous data, including electrophysiology, imaging, molecular/serological, clinical measurements, and demographic metadata. Data integration is required at a semantic level with harmonization facilitated by data converters and mappings, common ontologies, and common data dictionary elements. Moreover, the data are likely to double in size every two years, requiring an adaptive, distributed infrastructure. The EpiBioS4Rx Portal will provide a web-based platform with an array of capabilities illustrated in **Figure 1**, to handle data integration, storage, annotation, provenance, and servicing.

Streamlined Data Consolidation: Users will be able to upload their raw data files directly to an extended version of the LONI infrastructure where they will automatically be classified, converted, and annotated. By automating much of this process, researchers both uploading and downloading data will be spared the time and effort currently involved in accessing and sharing epilepsy data. This streamlined data consolidation will increase the financial efficiency and scientific productivity of the CWOW and the broader epilepsy research community. In addition, physical samples will be brought together in a single biobank, further reducing coordination challenges. Much of these data have already been collected by the PIs presenting this proposal; however, the huge size of these data combined with the many different file formats makes effective navigation currently frustratingly labor-intensive and error-prone.

User-friendly data search and navigation: By converting data to consistent file formats and tagging that data with metadata, the CWOW will enable Google-style search of all available epilepsy data. However, because data will be interlinked and co-registered across data sets and modalities, the search functionality will not simply match data against individual items like Google—rather, it will find interlinked combinations of data (even across modalities and data sources) that match the desired criteria. This enables sophisticated custom searches that match the functionality of predefined query forms. Users will be able to browse data in its most appropriate visual representation and pivot from one data view or modality to another. Through our experiences with LONI, we have learned the access control and sharing mechanisms required by the community and how effectively to enable inter-project as well as community-scale data sharing. Key components include giving users explicit access control for their data and results as well as providing project groups for larger-scale permissions management. Furthermore, comparable tools that can be repurposed were developed for LONI as part of our planning grant and projects like ADNI.

Automated analysis: The LONI Pipeline contains a common framework for visual and programmatic construction of data-driven workflows for electrophysiology, imaging, and biosample data. We will customize these tools to the study of epilepsy data. With the aid of LONI's workflow builder, complex analyses are represented visually, further supporting researchers' investigations. Examples of Pipeline applications include developing a unified coordinate space for seizure locations across organisms (Specific Aim 2), using string similarity and value overlap to predict that different contributor metadata fields are the same, and providing graphical interfaces for linking data. Co-registration algorithms will typically be invoked at upload-time but may be triggered later manually for further refinement. The IAC will provide MRI supervision and integration from different scanners and centers by supervising phantom studies, assessing quality, and fixing problems with heterogeneity. While LONI has primarily used Pipeline for human data, Dr. Dong et al. have studied neural networks of the mouse neocortex using these tools[78]. The CWOW would further expand these capabilities, providing robust workflow pipelines for both humans and animal models.

Iterative improvement using novel analytical tools: The sheer quantity of data and the noise inherent in the data necessitates the development of novel analytical tools, incorporating the most recently developed mathematical and statistical tools to discover previously undetected biomarkers. Dr. Bragin et al. recently discovered a novel biomarker, repetitive high frequency oscillations and spikes (rHFOSs)[4]. Dr. Gotman, a consultant on this project, has established tools to study the relationship between spikes and HFOs and showed their links to epileptogenesis[40]. Dr. Duncan has developed sophisticated mathematical methods[18] to analyze both animal and human data separately and also to process for trans-species comparisons.

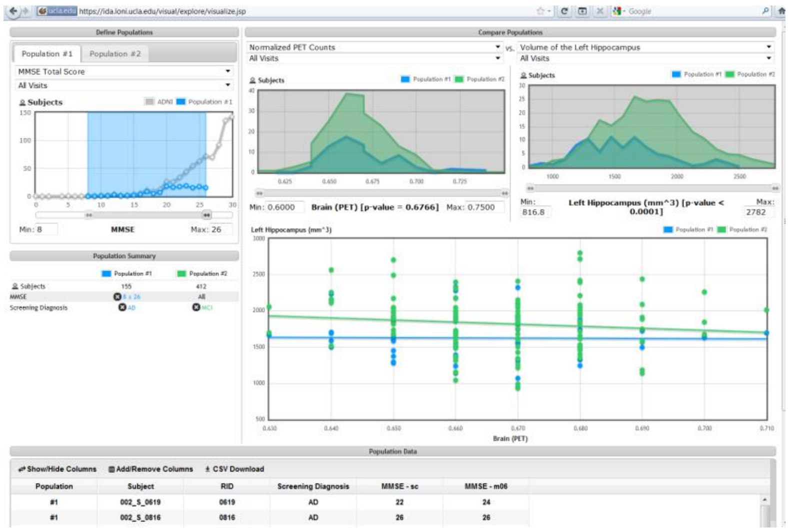
Both conventional and robust regression as well as logistic regression techniques will be used to address the principal questions of reduction of incidence of PTE (e.g., R packages “Robust” and “rlm”) since varying degrees of heteroscedasticity are fully expected in these data, which in continuous and binary outcome models affect both the betas and their standard errors. Composite endpoint analyses will provide a tool to build causal links between markers and related outcomes[52]. Hierarchical clustering will provide mechanisms for defining meaningful data clusters based on several choices of distance metrics (R packages hclust and rpud). Multimodel inference is a powerful tool for exploring the relative statistical standing of competing models and performing model selection and averaging[8] (R package “MuMIn”). Mixed models analyses are widely recognized as essential for studies in which repeated measurements are made on clusters of related statistical units (see R packages “nlme” and “lmer”). Change point analyses will be used to address when and if one or more changes occur and the confidence bounds around such changes (R package “changepoint”). Censor analyses will be used to assess relationships across variables with either left- or right-censoring due to values exceeding threshold; such analyses are an important component of what are known as vector generalized models (R package “VGAM”). Diffusion mapping is an advanced method of data parametrization within which statistical visualization joins with clustering using diffusion K-means and adaptive regression modeling (see R Package “DiffusionMap”). Hidden Markov modeling is an advanced method for assessing time series or stochastic processes in the context of unobserved Markov processes, using specialized algorithms (see R packages “HiddenMarkov”, “HMM,” and “depmixS4”). All analyses will be controlled for family-wise error rates.

We are keenly aware of multiple conceptual and technical issues regarding data analyses with biomarkers, including within subject correlation, multiplicity, multiple clinical endpoints, and selection bias21,59. Utilization of multiple statistical approaches will allow us to address these concerns in full. Additionally these methods will allow us to meet the full range of challenges posed by Barker-Haliski et al2. as to the ways in which clinical development should best link with translational science.

Standardized sample collection, shipping, and biobank storage protocols: The project will define methods for harvesting, freezing, and storing tissue and other biosamples (i.e. serum). Data and tissue stored for collaborating preclinical trials, such as TRACK TBI, ALLO, PPMI, TRACKHD, ICBM, AIBL, ACE, ABIDE, 4RTNI, Mapp, and the Human Connectome Project already have these protocols in place with informatics provided by LONI. Animal protocols for storing parallel samples to humans will be stored and treated in a similar fashion to compare findings from parallel human and animal studies.

A data safety monitoring board (DSMB) will advise us as we perform a rigorous multicenter preclinical antiepileptogenesis trial using a blinded, vehicle-controlled randomized study design to determine the antiepileptogenic effect of the lead compound. The results of the three projects integrated with the IAC, following the close guidance of the DSMB, will assist in planning the optimal design of a future clinical antiepileptogenesis trial for successful drugs.

A major challenge of the CWOW lies in managing the heterogeneity and scale of the potentially relevant data. This occurs (1) while integrating and interlinking the data such that it can be stored, accessed, searched, and analyzed and (2) while browsing and algorithmically analyzing the data in search of biomarkers, where the space of possibly relevant features is extremely high. Our team has world-class expertise in supporting both of these tasks and to ensure rigorous experimental design for robust and unbiased results. Two example analyses using our existing platforms demonstrate these capabilities.



https://ida.loni.ucla.edu/visual/explore/visualize.jsp

Define Populations

Population #1 | Population #2

MMSE Total Score

All Visits

Min: 8 MMSE Max: 26

Compare Populations

Normalized PET Counts vs. Volume of the Left Hippocampus

All Visits | All Visits

Min: 0.6000 Brain (PET) [p-value = 0.6766] Max: 0.7500

Min: 816.8 Left Hippocampus (mm^3) [p-value < 0.0001] Max: 2782

Population Summary

	Population #1	Population #2
Subjects	155	412
MMSE	8 ≤ 26	All
Screening Diagnosis	AD	MCI

Population Data

Show/Hide Columns | Add/Remove Columns | CSV Download

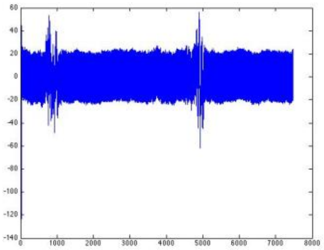
Population	Subject	RID	Screening Diagnosis	MMSE - sc	MMSE - m06
#1	002_S_0619	0619	AD	22	24
#1	002_S_0816	0816	AD	26	26

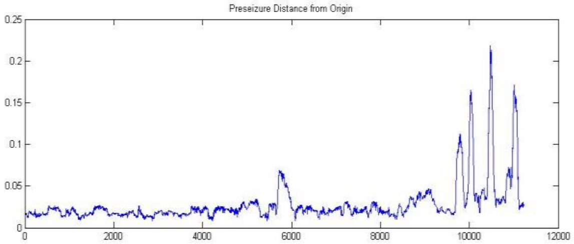
Figure 2. ADNI analysis of patients relating positron emission tomography tracers and hippocampal volume measured in MRI.

Example 1, LONI and ADNI: ADNI (http://ADNI.loni.usc.edu) is supported by the LONI platform and comparable in scope to the proposed CWOW. It contains neuroimaging data from 58 centers and links molecular, serum, and other data. LONI provides access to these distributed data through a data visualization and analysis platform that can be queried to locate data elements. Data can be visualized, queried, and analyzed using an extensive, evolving library of analysis tools to look for associations of features across imaging, molecular biomarkers, and other data[68]. **Figure 2** shows an example of such an analysis query in LONI. In this case, the analysis tests 3 characteristics of 2 patient groups from the ADNI database (from top left clockwise): (A) mini-mental state examination (MMSE) in study patients (blue) compared to ADNI controls (grey), (B) Patient (dark green) PET metabolism (1st plot) and hippocampal volume (2nd plot) compared to controls (light green), and (C) comparison of normalized PET to hippocampal volume (graph bottom right). These workflows and capabilities are directly adaptable to the proposed CWOW project.

Example 2, EEG: Figure 3 (b) shows an example of one of the innovative mathematical tools used to analyze the EEG data. A new algorithm, called Diffusion Component Analysis, is used for the purpose of dimensionality reduction to deal with the massive amount of data as well as for pattern recognition to search for various biomarkers in the data. **Figure 3 (b)** shows how the method can produce an automatic, unsupervised embedding to visually depict spikes in the data more accurately than spike detection methods that have been used in the past. Furthermore, **Figure 4** shows an example of how the resulting embeddings from the algorithm can be quantified by calculating the Euclidean distance of each point in the 3D embedding from the origin and plotting those distances over time. This particular graph shows how the method finds more variability in this example in the EEG data closer to the time of the onset of a seizure.

These examples demonstrate LONI's ability to handle large, diverse, multicenter, and heterogeneous data sets. We now describe how these capabilities will be combined to support the CWOW.





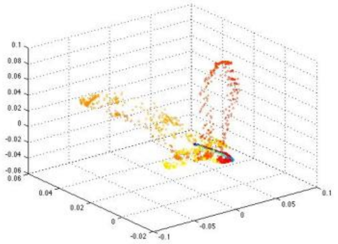


Figure 3 (a) and (b). Example of raw EEG data (a) on the left and the corresponding embedding (b) on the right using our algorithm that is based on diffusion maps for the purpose of spike detection.

Figure 4. The Euclidean distance from the origin, 30 minutes before a seizure (the x-axis represents time windows). The largest 3 eigenvectors in the pre-seizure segment are used for the diffusion maps embedding.

Specific Aim 1: Centralized data repository and innovative standardization

Data Upload and Quality Control: The IAC will be responsible for study-wide data quality control. The LONI Neuroimaging Quality Control System (LONI QC) will be used for all multi-modal data, checked for quality, and reviewed by participating collaborators from *Projects 1-3*. We have considerable experience dealing with inter-site variability in data, annotation, and models. We have developed tools to normalize and harmonize signal, image, and other data. Images uploaded from participating centers will be processed using LONI's multicenter data review and assessment system (https://qc.loni.usc.edu/). This system allows automated pre-processing that generates vector statistics and derived images to assess data quality. Data will be run through automated artifact detection algorithms in preparation for initial biomarker processing[5,19]. This system is web-accessible, user friendly, simple to navigate, and will provide a long-term resource for this field. Data will be sent back to *Projects 1-3* to inform the results in an unbiased way.

In support of the CWOW, we will develop an extensible, robust infrastructure for integrating, searching, and analyzing multi-modal data. Our team has deep expertise in data integration[17,25,26,67,66], existing infrastructure for workflow tasks to convert data, and a suite of import and conversion tools developed for LONI.

Data upload: Building on our expertise in data integration[17,25,26,67,66] and the import tools developed for LONI, we will develop automated data import pipelines whereby data can be directly contributed to our platform. Data transformation software[2] will automatically detect new data, validate it, and map the data to a common data model (where applicable), and also pre-index the data by features and values to aid in search and co-registration. Our common data model will align with the NINDS common data elements[27]. Through a federated architecture, key components of the data may be distributed across the LONI platform. We will deploy additional capabilities for handling molecular and cellular data, leveraging infrastructure used for other data types. Quality control and provenance information will be maintained with all data.

Data search: Data will be made searchable and accessible through web clients as well as programmatic (MATLAB, C, Python, Java, and R) interfaces. The web interface will provide query forms for standard search criteria. However, in searching for biomarkers, investigators may need more flexible ways of finding their data. In order to enable novel queries, we will build upon keyword search techniques for text, also looking across linked data items[66] in novel ways. Terms in the search will be matched against different data objects (e.g., annotations, PDF documents, and metadata), and the system will traverse links among the data matches to return joint answers. Under this model, the user can ask, for example:

Seizure onset regions in MRI of patients with left frontal lobe seizures and at least 3 seizure events: the system would separately match the search terms against data (MRI, patients, EEG annotations); then find subsets of these that match the constraints and are "connected," i.e., a subset of the patients have left frontal lobe seizures in their case history. These are linked to sets of electrophysiology and image data, thus, the electrophysiology and image data are (manually or automatically) annotated with information about seizure occurrences and onset regions.

Access control and sharing: Access control and encryption will be provided to support a wide variety of data sharing policies. The database will ensure that IRB and HIPAA regulations are followed using our integrated data de-identification components[14,17,54]. There is a quick and intuitive mechanism for tracking the status of all data sets and providing an audit trail (data provenance) so that investigators may understand who processed the data[60] when and in what way.

Data processing pipeline: For images, we will use the LONI Pipeline Workflow Environment[14,15,16,48,54], used to analyze neuroimaging data for hundreds of research publications. An extensible framework of protocols contains validated tools for spatial normalization, intensity inhomogeneity correction, template matching, atlas construction, statistical mapping (e.g. variation or asymmetry), and format conversion. The pipeline has been recently extended to genomic data processing[31] and will easily be adapted to human and animal data streams generated by CWOW investigators. Once the data are identified, through Application Programming Interfaces, a variety of technologies can be used to access the data and process it in parallel, e.g., MapReduce, parallel MATLAB, or MPI. Specific analysis operations can be composed graphically into "pipelines" or "workflows" that can be applied to multiple data sets. We have a robust and established computational infrastructure, powerful validated software tools, distributed web service architectures, and significant informatics experience. We will follow project management and informatics approaches successfully validated in other projects[31,61,77,22].

We briefly describe how different data modalities will be cataloged before discussing co-registration.

Molecular & cellular: We will optimize protocols for collecting animal and human tissue, providing analysis and integration methods for genomic, epigenetic, proteomic, and metabolomics studies in *Projects 1-3*.

Pathology: Participating sites in *Projects 1-3* will standardize the acquisition and storage of pathological data and develop algorithms for analysis, necessary for integration of pathological data elements into our bioinformatics platform. These include methods for harvesting, freezing, and storing tissue and other biosamples (i.e. CSF). Data and tissue stored for collaborating preclinical trials, (e.g. TRACK TBI) already have these protocols in place. We will ensure that animal protocols for storing parallel samples to humans are stored and treated in a similar fashion so that findings from parallel human and animal studies can be compared.

A vital part of the proposed work will be registering all of the data collected within the CWOW. These will be mapped into a central temporal and spatial coordinate framework for analysis. Specimens will be identified by their subject number, generating lab, temporal and spatial coordinates, and information to be stored in the IAC. We will describe the amount of TBI, the stage and severity of epilepsy, and alterations in behavior over time. Here we will adopt established scales, benchmarks, and Common Data Elements. Temporal measures will include age of animal at time/date of injury and every biological sample or measurement made will be logged by absolute time, able to be referred back to these temporal landmarks. Spatial descriptions and co-registrations of lesions will be made according to detailed coordinate/imaging maps of the brain, co-registered to sensors, such as implanted or scalp electrodes, when possible. This coordinate system will be implemented for humans and for animal models of epilepsy, for comparison across species. With the aid of the LONI Pipeline, much of this work can be automated.

Traumatic lesions will be quantified according to this co-registration. Tissue samples taken will then be referenced to imaging coordinates so that cellular and molecular markers can be localized to the same coordinate reference frame as electrophysiological and imaging measures. In this way, straight-forward comparisons can be made across the heterogeneity of patients and animals.

Specific Aim 2: Novel processing methods to extract candidate biomarkers from multi-modal data

The LONI Pipeline consists of an already robust suite of analytic tools, which can be extended to cover the full range of data types for both humans and animal models. A workflow is included below to show a sample of the platform's functionality:

Below are details on how different types of data will be analyzed by the LONI Pipeline Workflow Environment:

Gross Pathology Classification and Longitudinal Change: We will follow state-of-the-art approaches to extract features from study images, some of which we have pioneered[33]. Raw clinical images, with a variety of injury characteristics, will be processed to yield subject and group level statistics (**Figure 5**). Imaging sequences will highlight tissue types present after TBI, including hemorrhage, edema, necrotic, atrophic, and enlarged CSF spaces (see *Projects 1 and 2*). We will extract the extent and distribution of these lesions using innovative mathematical techniques to assess longitudinal changes in anatomy, including thalamocortical and hippocampal structures and their connections[34,72,47,35]. Dr. Duncan will lead this effort and integrate with *Projects 1 and 2*. A nonlinear and local network approach will be used to determine if the early occurrence of specific electrophysiological features of epileptogenesis (i.e., interictal epileptiform activity or morphologic changes in spikes and seizures) during the initial week after TBI predicts the development of PTE.

Imaging Analysis: Patterns of cortico-thalamic and hippocampal structure and functional connectivity will be analyzed. Studies will be performed on both humans and animals; segmentation using personalized atlas construction and topological change estimation will be used[71]. Statistical parametric mapping (SPM) of voxel based morphometry (VBM), which has demonstrated differing patterns of hippocampal atrophy, as well as extensive areas of neocortical atrophy and thalamic involvement in MTLE will be applied[42,43,51]. Furthermore, automated shape analysis will be used as a more sensitive measure to distinguish changes in the MRI[13,55,61,62]. This imaging approach can now be applied to animals at UCLA, Kuopio, Einstein, and Melbourne. Other neuroimaging features that have been implicated as biomarkers and will be examined with our bioinformatics platform include: 1. hippocampal T2 signal changes[24,53], 2. EEG-fMRI changes[44], 3. structural connectivity using diffusion tensor imaging (DTI)[41,50], and 4. resting state functional connectivity MRI (fcMRI)[29,30]. Animal model surveys have the advantage that they can be obtained at intervals following the epileptogenic insult and can be correlated with concurrent long-term video-EEG monitoring results.

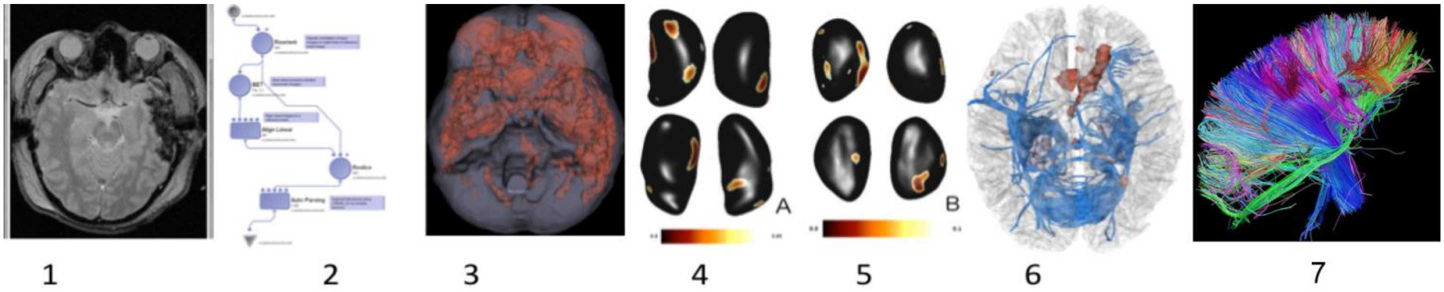


Figure 5. Example of transformation of raw clinical imaging data set (GRE, DTI, FLAIR, MPRAGE) into group data showing lesion location across the entire cohort of patients (n=46), and advanced diffusion tractography in individual patients. This makes use of modular and automated LONI pipeline techniques and other processing methods. 1) raw GRE image, 2) exampled LONI pipeline, 3) group hemorrhage load across 46 patients, 4&5) LONI pipeline output demonstrating thalamic (A-acute) and (B-chronic) in PTE, 6) DTI tractography denoting changes in white matter in chronic TBI, and 7) Connectomics.

Analyses of White Matter Connectivity: We have introduced methods for TBI connectomics[70,36,37,69], which we will use. We will use DTI to extract connectivity between all pairs of gyral and sulcal structures in the presence of brain trauma. Diffusion tractography will be used to determine connectivity properties (WM bundle length, connectivity density, fractional anisotropy (FA)) and each subject's weighted connectivity matrix. WM fiber tracking of inter-regional connectivity will be conducted using TrackVis (trackvis.org) and/or other tractography tools. Connectivity between regions (i.e., thalamo-coritical connections and hippocampal connections) will be assessed systematically within each patient using purpose-built workflows for multi-modal co-registration of MRI. This will be followed by calculation of (i) inter-regional connectivity matrices and (ii) longitudinal changes in connectivity topology using network-theoretic descriptors of nodal and network-wide segregation (clustering coefficient, modularity, etc.) and integration (characteristic path length, global efficiency, etc.). Additional network-theoretic measures (scale freedom, small worldness, robustness, centrality, degree distribution, and communication efficiency)[20,65,73,58] will be computed. Results and changes over time in each patient will be visualized and analyzed using connectograms[38]. Workflows will be integrated with the LONI Pipeline.

Serum Biomarker Data Analysis and Comparison: Serum Biomarker data will be analyzed using ANOVA in combination with multiple comparisons post hoc tests in consultation with statisticians at the Center of Biostatistics, Uniformed Services University, and additional miRNA analysis by *Project 1* investigators. The independent variables are injury (injury or sham) and treatment (drug or vehicle). The dependent variables will be the concentrations of the selected protein marker per mg protein (RPPM and ELISA). For each of our numerical measures, we will determine whether there are statistically significant differences between or among the various treatment groups. The dependent variables will be tested in this manner. Statistical tests will be applied according to established ANOVA procedures ($p<0.05$)[28]. Following statistical analysis, values will be directly compared among the various experimental groups.

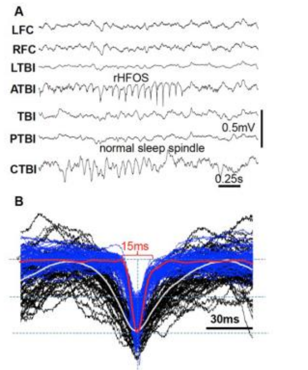


Figure 6. A – An example of rHFOS recorded 10 days after TBI in the area anterior to TBI (ATBI) and normal spindle that coincidently occurred in the area of contralateral to the TBI core (CTBI). B – Fifty five superimposed single waves form normal (black) and pathological (blue) spindle-like activity. White and red thick lines are normalized averages of correspondingly normal pathological events.

Analyses of Electrophysiology: EEG data from *Projects 1, 2, and 3* will be analyzed in the projects and by the IAC. We plan to study the variability of the statistics of the EEG using various measures, such as covariance matrices, local power spectra, and the Mahalanobis distance of local covariance matrices. We will show how these various statistics change over time and determine if we can use these methods to distinguish between patients who develop PTE and those who do not. Whereas most EEG studies use Principal Components Analysis (PCA) as a dimensionality reduction tool, we will use a more recently developed method called diffusion maps. Unlike PCA that is linear and global, diffusion mapping is nonlinear and local, so it is more applicable to our complex EEG data. Not only does diffusion mapping allow for dimensionality reduction of the data, but this method also provides pattern recognition so that specific parts of the data may be analyzed more closely to reveal potential biomarkers for epileptogenesis. We have recorded high dimensional data that measure brain activity, and we assume that for the problem of detecting seizures and predicting epileptogenesis, there are a few brain activity factors that govern the occurrence of seizures. The goal is to find those controlling low dimensional factors based on the observable high dimensional data. Our algorithm allows us to separate the signal from the noisy EEG data by the combination of the Mahalanobis distance measure and inverse covariance matrices. Furthermore, we have studied rat EEG to detect and analyze paroxysmal events like pathological high frequency oscillations (pHFOs) and repetitive high frequency oscillations and spikes (rHFOSs). rHFOSs, which can be distinguished from sleep spindles, could be noninvasive biomarkers of epileptogenesis (**Figure 6**). HFOs (ripples and fast ripples) have been studied as biomarkers for epileptogenesis as well (**Figures 8 and 9**) and found in seizure onset zones.

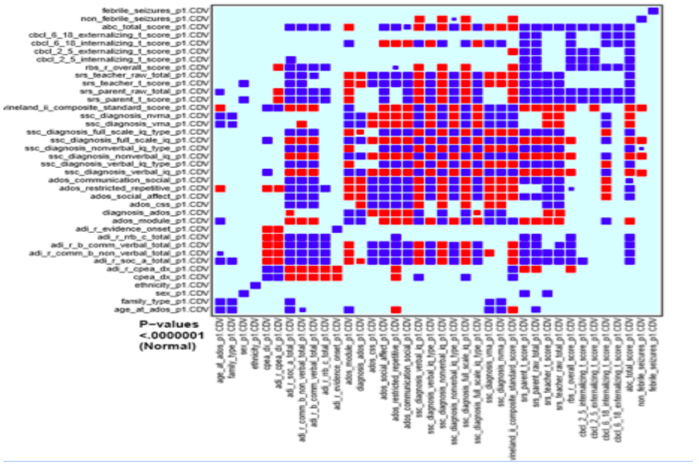


Figure 7. P-value block plot for variable associations of P<.0000001. Variables are listed on x and y axes. Colors denote different classes of significance in the model. Movies of this tool and their analysis will reveal correlations between biomarkers, which will be extracted and plotted. Such multivariate visualization methods will be used to observe graphically how multiple, seemingly unrelated parameters vary over time.

Data collected in *Projects 1-3* will be organized both temporally and spatially to integrate measures in images, electrophysiology, and tissues (including blood and CSF) along with clinical data. The data will be initially analyzed at the center where they are collected and then sent to LONI for further, more thorough analysis where all types of data collected will be compared and analyzed. This provides a huge variable space to navigate for associations. Two main methodologies will be used to look for correlations in these measures. For temporal data, we will use multivariate visualization methods to observe graphically how multiple, seemingly unrelated parameters vary over time (**Figure 7**)[12,72].

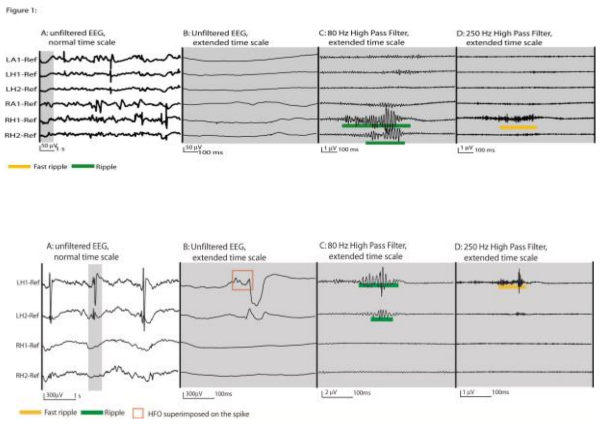


Figure 1:
A: unfiltered EEG, normal time scale
B: Unfiltered EEG, extended time scale
C: 80 Hz High Pass Filter, extended time scale
D: 250 Hz High Pass Filter, extended time scale
LA1-Ref
LH1-Ref
LH2-Ref
RA1-Ref
RH1-Ref
RH2-Ref
50 µV 1 s
50 µV 100 ms
1 µV 100 ms
1 µV 100 ms
Fast ripple
Ripple
A: unfiltered EEG, normal time scale
B: Unfiltered EEG, extended time scale
C: 80 Hz High Pass Filter, extended time scale
D: 250 Hz High Pass Filter, extended time scale
LH1-Ref
LH2-Ref
RH1-Ref
RH2-Ref
300µV 1 s
300µV 100ms
2 µV 100ms
1 µV 100ms
Fast ripple
Ripple
HFO superimposed on the spike

Figures 8 and 9. In (8), the unfiltered EEG shows frequent spikes in both mesial TL structures. The gray section in A is expanded in time and amplitude in B, C, and D to demonstrate the co-occurrence of ripples and fast ripples (pHFOs) outside spikes in channel RH1, and a ripple is seen in channel RH2. In (9), the same unfiltered EEG is shown as in (8); this time the selection includes 2 spikes over LA1 and LH1. Both spikes were outside of the seizure onset zone and in the healthier mesial temporal lobe. There were no HFOs during this time period (B, C, D).

We will use dynamic visualization methods to observe movies of these multivariate trajectories, adjusted for multiplicity, and extract quantitative measures from them. Variations of these techniques are called the Lineup Method and Association Navigator, and they have been used successfully to assess similar large pools of multi-modal data from multicenter studies of human autism[6,32]. Similar methods will be used for spatial graphical rendering of changes in images, adjusted for statistical significance, in order to isolate changes in specific brain regions statistically associated with developing epilepsy. In addition to these methods, we will use techniques of multi-model inference[7] to select the best models from possible combinations of features, including a newly available software application in R[3].

Specific Aim 3: Models of Epileptogenesis: Build models of epileptogenesis, based on heterogeneous biomarkers, that track the probability of developing epilepsy over time

As data flow into the IAC and are analyzed, the analysis may lead to models of epileptogenesis that can be used to forecast the probability of seizure onset. Input to the models would be primary biomarkers found by single variable analyses in *Projects 1-3* and then combinations of variables revealed in Specific Aim 2. These 3 specific aims will proceed simultaneously, and the goal is to construct and adapt models as adequate primary data are recorded, iterating over time, to improve precision. Predictive models will consist of hierarchical Bayesian nonparametric Markov switching processes[76] and probabilistic forecasting methods used successfully to convert changes in EEG features into probabilistic distributions of seizure likelihood that have been implemented in real time in seizure prediction devices[11]. These methods will be adapted to the multivariate distributions determined above by clustering on multiple features, rather than electrophysiology alone. These methods are outlined in detail in a previous publication from our group[63].

A second model we will use is called Change Point Analysis. In this method, one or multiple parameters are monitored for abrupt or gradual deviations that predict regime change. The method will be used to predict when biomarkers in epilepsy-prone animals or humans deviate statistically from controls at “change points,” and their probability of getting epilepsy at each time will be computed. This is used successfully in industry to predict device failures or new operating regimes[46].

Why this team for the IAC? LONI has provided the informatics and analytics for major projects in the study of a range of neurological diseases. PPMI (http://www.ppmi-info.org/) is an international study to identify biomarkers of Parkinson’s disease progression. Sites in the U.S., Europe, and Australia collect imaging, biologic sampling, and clinical assessments. Since 2005, the longitudinal ADNI has been validating the use of biomarkers, including blood tests, tests of cerebrospinal fluid, and MRI/PET imaging for AD clinical trials and diagnosis. It is one of the largest and most successful big data exchanges (now with whole genome sequencing of subjects), with heterogeneous data collected from almost 60 sites and distributed to thousands. We are responsible for the informatics of ADNI. The ENIGMA project (http://enigma.loni.usc.edu) is a network of imaging and informatics researchers studying brain structure and function, based on MRI, DTI, fMRI, and genome-wide association scan (GWAS) data. The Big Data for Discovery Science Center (BDDS) is comprised of leading experts in biomedical imaging, proteomics, and computer science. Our BDDS center streamlines big data management, aggregation, manipulation, integration, and the modeling of biological systems across spatial and temporal scales (http://bd2k.ini.usc.edu). By combining LONI’s Pipeline Workflow Environment and big data experience with some of the world’s foremost experts in epilepsy research, we are confident that EpiBioS4Rx can extend the successes of PPMI, ADNI, and ENIGMA to the study of epileptogenesis.

We have a team of LONI Pipeline support staff and programmers who will be available to help with any part of the data upload, management, and analysis process. This team includes Karen Crawford, Database Manager;

Rita Esquivel, Project Assistant; Dr. Scott Neu, Software Developer; Petros Petrosyan, Pipeline Manager; Samuel Hobel and Alexander Nizni, Programmers. Co-Investigator Dr. Rema Raman will serve a critical role throughout the duration of this project on approaches for data harmonization across datasets, implementation of mixed effects models across cohorts, and cross-validation between cohorts. Co-Investigator Dr. John Van Horn will design the software architecture for acquiring, archiving, managing, and displaying bioinformatics data as well as aid with consortium efforts. Furthermore, our consultants, Dr. Brian Litt and Dr. Jean Gotman, will provide their EEG analysis expertise and guidance for this project.

Clearly, a project of this scale requires leadership with a demonstrable track record of program administration and scientific excellence. The USC group has demonstrated leadership in large scientific efforts with experience in neuroscience imaging, epilepsy, data integration[39,25,26,67,63], scalable cloud technologies[49], electrophysiology, imaging, and multifactorial data analysis. We have an unmatched resource of knowledge to integrate and analyze multi-modal data towards the discovery of biomarkers of epilepsy. We have either led or participated in several national and international big-data related initiatives, various NIH P- and U-class projects, and NSF projects related to big data computing.

In short, our team represents a group of leading investigators particularly well suited toward addressing the aforementioned issues and providing novel solutions for meeting the challenges presented by big clinical data sets. Our team members have direct and immediate experience in managing large, multi-site investigations for brain research, are well known for our leading work in computer science, engineering, workflow development technologies, and informatics, as well as our world-leading efforts to develop systems-level computational approaches toward modeling biological processes.

Why this organizational model for the IAC? A central IAC Statistics Committee group will guide experimental design, power and experimental analyses, provide service to the investigators, and work closely with the Project Steering Committee. The IAC will train members of the individual projects to use its resources and central, web-based tools that will link our collaborative CWOW community. The IAC statistics committee, composed of LONI statisticians, computer scientists, and IAC PIs, possess expertise and experience in analyzing neuroscience imaging, epilepsy, and data integration, as well as electrophysiology, genetics, imaging, and multifactorial data analysis. We also have considerable experience with large-scale biomedical data analytics and take part in several national and international big-data related initiatives, including the ADNI, PPMI, ENIGMA, BD2K, coordinating data for the DARPA RAM centers, and other NIH and NSF projects. Our group will house an ILAE Web Atlas for animal models of epilepsy and work with a host of industry research partners. We are confident that this experience will ensure similar success and productivity of our EpiBioS4Rx CWOW consortium.

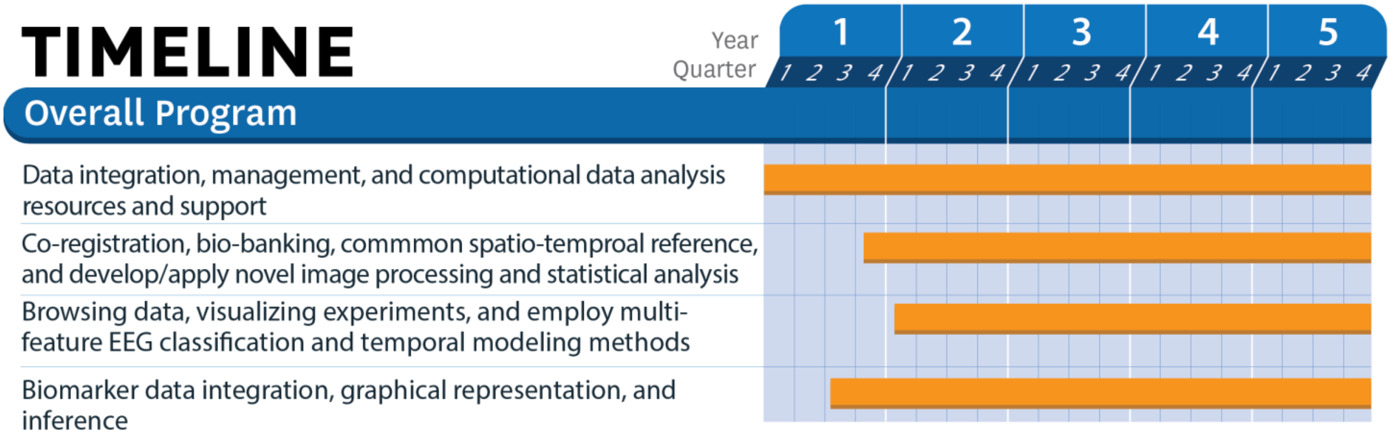


Figure 10. Timeline. Data integration, management, and providing data analysis workflows using the LONI Pipeline will take place throughout the entirety of the project. Co-registration, bio-banking, common spatio-temporal references, and the development of novel processing and modeling algorithms will commence in Q4 of Year 1, once data have begun to accumulate in earnest. Browsing data, visualizing experiments, electrophysiological modeling, and source localization will begin in Year 2, while statistical comparisons between brain morphometry, connectomics, and their potential integration with electrophysiological models will commence in Q2 of Year 1 and take place throughout the funding period.