A. OVERALL COVER PAGE

Project Title: Automatic discovery and processing of EEG cohorts from clinical records	
Grant Number: 5U01HG008468-03	**Project/Grant Period:** 06/01/2015 - 05/31/2018
Reporting Period: 06/01/2016 - 05/31/2017	**Requested Budget Period:** 06/01/2017 - 05/31/2018
Report Term Frequency: Annual	**Date Submitted:**
Program Director/Principal Investigator Information: JOSEPH PICONE , MS PHD **Phone number:** 2152044841 **Email:** joseph.picone@temple.edu	**Recipient Organization:** TEMPLE UNIV OF THE COMMONWEALTH TEMPLE UNIVERSITY 1801 N Broad Street, 401 Conwell Hall PHILADELPHIA, PA 191226003 **DUNS:** 057123192 **EIN:** 1231365971A1 **RECIPIENT ID:**
Change of Contact PD/PI: N/A	
Administrative Official: JOHN D PENNER Student Faculty Center 3340 North Broad Street, 4th Floor, Suite 427 Philadelphia, PA 191405104 **Phone number:** 215-707-3887 **Email:** john.penner@temple.edu	**Signing Official:** KAREN DENISE MITCHELL Student Faculty Center 3340 North Broad Street, 4th Fl, Ste 427 Philadelphia, PA 191405104 **Phone number:** 215-707-7547 **Email:** grantsmanagement@temple.edu
Human Subjects: No	**Vertebrate Animals:** No
hESC: No	**Inventions/Patents:** No

B. OVERALL ACCOMPLISHMENTS

B.1 WHAT ARE THE MAJOR GOALS OF THE PROJECT?

The specific aims have not changed since the start of the project. The percentages listed are relative to the overall goals of the multi-year project since many of the sub-tasks span multiple years.

Specific Aim 1: Automatically recognize and time-align EEG events that contribute to a diagnosis: We will develop automated techniques to discover and align the underlying EEG events that led to a diagnosis using data-driven approaches and semi-supervised learning. Five classes of events will be identified: spike and sharp wave; generalized periodic epileptiform discharges; periodic lateralized epileptiform discharges, eye blink and artifact. Everything else is considered background. This will make the data more useful to a wide range of clinical research, and support a new form of biomedical knowledge derived from BigData repositories.

YR1 Sub-Tasks: Annotation Development (50%), Iterative Training and Bootstrapping (50%)

Specific Aim 2: Automatically recognize critical concepts in the EEG reports: We will automatically recognize clinical events (e.g. “intermittent bursts of paroxysmal high amplitude activity”) and their types: EEG-specific ACTIVITY (e.g. “beta frequency activity”), EEG-specific PATTERN (e.g. “burst suppression pattern”), or CLINICAL DEPARTMENT (e.g. “coded for 30 minutes in the emergency room”). In addition we shall automatically distinguish the clinical events’ polarity (POSITIVE, NEGATIVE) and modality (e.g. CONDITIONAL, POSSIBLE). In EEG reports, mentions of clinical events also have dense spatial and temporal information associated with them that will be mined automatically. Spatial expressions (e.g. “bilateral”, “diffuse”) and their spatial roles to the clinical events shall be discovered. Similarly, temporal expressions (e.g. “every ten seconds”) and their temporal links to the clinical events shall be automatically mined. In addition, because EEG reports describe also the clinical picture of patients, we shall identify automatically several types of medical concepts in the form of medical problems (e.g. “epilepsy”), tests and treatments.

YR1: Clinical Events (80%), Medical Concepts (75%)

Specific Aim 3: Automatic patient cohort retrieval: We shall develop a patient cohort retrieval system that will identify patients having EEGs relevant to a query or similar to a given EEG. Central to the patient cohort retrieval system is a qualified medical knowledge graph, generated automatically by using a BigData solution based on MapReduce operating on the knowledge automatically extracted in aims 1 and 2. In this way, the patient cohort retrieval system will be designed to search both free-text chart notes and EEG signals. Searching both areas will enhance retrieval for those medical events or concepts recorded in only one place. In addition, a spatial and temporal characterization of the way in which events in an EEG are narrated by physicians and the validation of these across a BigData resource are important contributions to basic science.

YR1: Generation of QMKG with MapReduce (33%), Query Expansion (33%), Index Generation (95%), Learning to Rank Based on Feedback (100%)

Specific Aim 4: Evaluation and analysis of the results of the patient cohort retrieval: To evaluate the cohort identification system clinicians and medical students shall design sets of queries that model inclusion criteria that describe the kinds of patients desired for comparative studies on EEG data. In addition, the experts will select subsets of EEGs to retrieve similar EEG data automatically from the cohort identification system. Relevance judgements produced by clinical experts shall be used to qualify the degrees of relevance of the patients identified. For each query, medical experts shall examine the top-ranked cohorts for common precision errors (false positives), and the bottom five ranked common recall errors (false negatives). User validation testing will be performed using live clinical data and the feedback will enable an analysis of the errors that will be used to better rank EEG reports. This will enhance the quality of the cohort identification system. User acceptance studies shall also be conducted and information about the perceived value of the system shall be collected.

An annotated big data archive of EEGs will greatly increase accessibility for non-experts in neuroscience, bioengineering and medical informatics who would like to study EEG data and demonstrate that a much wider range of big data bioengineering applications are now tractable. The cohort retrieval system and annotated EEG signals will greatly reduce training times for medical students pursing careers in neuroscience.

YR 1: Generation of Queries (33%), Evaluation of Patient Cohort System (33%), Analysis of Results (33%), Component Evaluation (33%), Demonstration / Feedback (25%)

B.1.a Have the major goals changed since the initial competing award or previous report?

No

B.2 WHAT WAS ACCOMPLISHED UNDER THESE GOALS?

File uploaded: acc_main_v01.pdf

B.3 COMPETITIVE REVISIONS/ADMINISTRATIVE SUPPLEMENTS

For this reporting period, is there one or more Revision/Supplement associated with this award for which reporting is required?

Yes

Revision/ Supplements #	Revision/ Supplements Title	Specific Aims	Accomplishments
3U01HG008468-02S1	Scalable EEG interpretation using Deep Learning and Schema Descriptors	Aim 1: Automatic labeling of the TUH EEG Corpus for seizure events. Aim 2: Application of deep learning sequential modeling techniques for EEGs to predict seizures. Characterize performance as a function of latency. Aim 3: Defining and generating Hierarchical epileptiform Activity Descriptors (HAD) for EEGs using deep learning. Aim 4: Automated Tagging of HADs in medical texts using deep learning.	Please see the end of the Accomplishments section above for a complete description of our accomplishments on the supplemental. Given the size of the award, we felt it best to write a detailed account of our accomplishments.

B.4 WHAT OPPORTUNITIES FOR TRAINING AND PROFESSIONAL DEVELOPMENT HAS THE PROJECT PROVIDED?

File uploaded: training_v00.pdf

B.5 HOW HAVE THE RESULTS BEEN DISSEMINATED TO COMMUNITIES OF INTEREST?

We continue to disseminate our resources through traditional methods such as publication. We actively maintain three web sites related to this project:

(1) The Cohort Retrieval Project Web Site: https://www.isip.piconepress.com/projects/nih_cohort/

We upgraded this web site to use Drupal so that it is more user-friendly and contains more frequent announcements about resources. We post periodic updates about the project as needed.

(2) The NEDC Web Site: https://www.nedcdata.org/

We make major announcements about data and resources to this site.

(3) The TUH EEG Database Web Site: https://www.isip.piconepress.com/projects/tuh_eeg/

This is the location of the TUH EEG Corpus and its various subsets. We make all databases available from this site.

We also will be hosting the IEEE Signal Processing in Medicine and Biology Symposium in December 2017 for the 4th year. We use this conference as a way to advertise our research and data resources.

Further, we now have our data resources cross-listed on a number of sites popular within the neuroscience community. We continue to reach out to a wide range of neurologists, and have a presence at major neurology conferences, in an effort to promote our programs.

B.6 WHAT DO YOU PLAN TO DO DURING THE NEXT REPORTING PERIOD TO ACCOMPLISH THE GOALS?

Primary Award:

Aim 1: Automatically recognize and time-align events in EEG Signals: We will continue our efforts to improve the performance of our deep learning classifiers by exploring transfer learning approaches and pre-training strategies. We will also investigate ways we can incorporate confidence measures to reduce the high false alarm rate.

Aim 2: Automatically recognize critical clinical concepts in the EEG Reports: We will continue developing an integration of the EEG signal and EEG report event detection systems so that we can query both data resources simultaneously. EEG reports have dense spatial and temporal information associated with clinical events and medical concepts, which we propose to automatically discover as well. Prof. Harabagiu's teams have participated in the 2012 Semeval Task on automatic Spatial Role labeling as well as in the 2012 i2b2 Challenge that focused on the automatic processing of temporal expressions and the temporal ordering of clinical events. We propose to use annotations focusing on the spatio-temporal specifics of EEG reports in order to (a) detect automatically spatial expressions and (b) temporal expressions, facilitating (c) the recognition of spatial relations to clinical events or medical concepts and (d) identification of spatial relations and spatial containers. Specifically, while considering the annotation scheme for spatial information reported in (Kordjamshidi et al., 2012), from which the Trajector, Landmark and Spatial Indicator spatial roles shall be derived, we shall also add new roles that have been considered in the spatial features of EEG reports (Bebiczky et al., 2013), e.g. Laterality (e.g.

left/right/midline/bilateral/diffuse) and Regions (e.g. frontal, temporal, central, parietal, occipital). The spatial role Trajector is defined as the clinical event or medical concept whose location is being described, whereas the Landmark spatial role is defined as reference entity in relation to which the location is defined (e.g. “over the occipital region”); the Spatial Indicator spatial role defines the constraints of spatial properties of the Landmark. The spatial role of Laterality is a special case of Spatial Indicator, whereas Regions is a special case of Landmark. To detect these spatial expressions and to identify the special roles and their relations to clinical events and medical concepts, we propose to use a joint classification framework that was very successful when we built the UTDSpRL spatial role labeler for the Semeval 2012 evaluation (Roberts & Harabagiu, 2012b). The spatial information that shall be identified in the EEG reports will also enable new indexing schemes for searching the EEG big data and identifying patient cohorts.

Aim 3: Automatic Patient Cohort Retrieval: To achieve the goals pertaining to Aim 3 of the project, in the next period we shall organize the acquired clinical knowledge in a graph, which will benefit from our recently developed medical knowledge embeddings (MKE). We shall incorporate in the MKE, which are superior to the qualified medical knowledge graph (QMKG), temporal and spatial relations in addition to the temporal and spatial attributes defined in the Hierarchical EEG Activity Tags. We also plan to use the MKE to enhance a novel learning to rank methodology for patient cohort identification. Lastly, we plan to continue enhancing our multi-modal indexing and retrieval scheme for clinical data, to include clinical events and spatial and temporal information as well as EEG events derived from the EEG signals.

Aim 4: Evaluation and Analysis of the Results of the Patient Cohort Retrieval: To achieve the goals pertaining to Aim 4 of the project, in the next period we shall continue the collaboration between Temple University and UTD to validate the results of the patient cohort identification system and to generate the final set of clinically-relevant queries. The patient cohort retrieval system shall be evaluated by considering that it shall return a list of ranked EEG reports and signals that satisfy the inclusion criteria set by an expended set of queries. Like in in the first year of the project, the development of the queries for evaluation will also consider exclusion criteria determined by the TUH EEG Corpus. Specifically, the main reasons for excluding a candidate query shall be either that the topic of the query will not be a good fit for the TUH EEG Corpus or there will be too few or too many EEG reports returned by a Boolean retrieval system that the UTD team shall build on top of the Lucene-based Index of the EEG corpus.

Another important aspect of the evaluation of the patient cohort retrieval system will focus on relevance judgments. The team from Temple University shall recruit judges who are physicians in residence and medical students at Temple University. We propose to recruit five judges. Judges will be instructed to rate each EEG report deemed relevant by the patient cohort retrieval system to determine whether such a patient would be a candidate for a clinical study on the topic of the query. A definitely relevant judgment will mean that the patient is unequivocally a candidate for the study. A possibly relevant judgment meant that the patient might be a candidate for the study but insufficient information was available for a definitive decision. A not relevant judgment will mean that the patient was not a candidate for the clinical study. The judgments will also enable an analysis of the results of the patient cohort retrieval system that can lead to learning to rank from the feedback provided by expert clinicians. We propose to categorize the reasons for incorrectly retrieving EEG reports and to derive their characterizing features. This will allow us to learn optimal ranking functions on the EEG corpus.

We will also start conducting user acceptance testing using three focus groups: expert annotators, clinicians and medical students.

Supplemental Award:

Aim 1: Automatic labeling of the TUH EEG Corpus for seizure events: We will continue developing the seizure detection corpus and refining the detail in which each event is annotated so that we can build event-specific models. We will also validate our results on several other databases we have recently acquired.

Aim 2: Application of deep learning sequential modeling techniques for EEGs: We will improve performance of our baseline prediction technology using more accurate deep learning models. We will explore in more detail what cues in the signal correlate with patients’ ability to predict seizures. Our goal is to be able to predict seizures with some degree of accuracy 30 minutes or more in advance with a low false alarm rate.

Aim 3: Defining Hierarchical epileptiform Activity Descriptors (HAD) for EEGs: We will generate a schema of Hierarchical epileptiform Activity Descriptors (HAD) and a hierarchical structure which will be rooted into the HAD tag, while organizing hierarchies for (1) the epileptiform activity waveform; (2) the epileptiform activity frequency band; (3) the epileptiform activity anatomical location; (4) the epileptiform activity position; (5) the epileptiform activity distribution; (6) the epileptiform activity frequency; (7) the epileptiform activity magnitude.

Aim 4: Automated Tagging of HADs in medical texts: We will produce the HAD tags automatically using several deep learning frameworks. We will automatically recognize the relations between the HAD tags. We will automatically annotate EEG signal recordings with the same HAD tags and the relations between them that we discovered from the EEG reports that interpret the EEG signal recordings. Clinical decision support (CDS) will be provided by retrieving scientific articles that document the medical care of similar patients before, during and after their seizures.

Accomplishments

Goals Specific to Aim 1: *Automatically recognize and time-align events in EEG Signals:* Identification of the type and temporal location of EEG signal events such as spikes or generalized epileptiform discharges in the EEG signal are critical to the interpretation of an EEG.

Our main goals concerning Aim 1 for this year included (1) using domain knowledge to improve the previously reported performance, (2) developing a self-training method that would allow to increase the amount of annotated data, and (3) explore and implement more efficient training algorithms.

The probability for the occurrence of a specific EEG activity can vary largely across channels. We can say, for example, that the observation of alpha rhythm is more likely in the occipital region than in other channels, or that it is much more likely to observe ocular artifacts in the frontal channels rather than in the temporal or posterior channels. Taking this spatial information into account, we decided to consider the possibility of training channel dependent models with our currently established HMM+SdA system.

In our standard HMM+SdA baseline system, which is channel independent, we train one model for each event that we wish to decode in the EEG signal. In this case, we have six models: Spike and Sharp Wave (SPSW), Generalized Periodic Epileptiform Discharge (GPED), Periodic Lateralized Epileptiform Discharge (PLED), Eye Movement (EYEM), Artifact (ARTF) and Background (BCKG). In the channel dependent variation of the experiment, we train one model per event for each channel. Basically, we have six models for each channel.

We evaluated the performance of the first pass of decoding (P1), which is an HMM-GMM system, for each channel individually. Table 1 shows the individual performance for each channel. The error rate for the channel independent system is 16.49%, which is outperformed by several of the channel dependent models and their average, which is 14.14%.

The output for P1 was then used as an input for a second pass of processing (P2) that consists of a deep learning system known as Stacked denoising Autoencoders (SdA). We selected the channel that showed the best performance for P1, which was F4-C4, to run the experiment for P2. The error rate for P2 for the channel dependent model was evaluated to be 29.59%, while the analogous channel independent model has an error rate of 8.71%. Figure 1 shows the Detection Error Tradeoff (DET) curve for P1 and P2 of both systems.

The sudden drop in performance between the first and the second passes of processing for the channel independent modes can be attributed to the fact that there is not enough data for accurate training of the model. It is widely accepted that deep learning

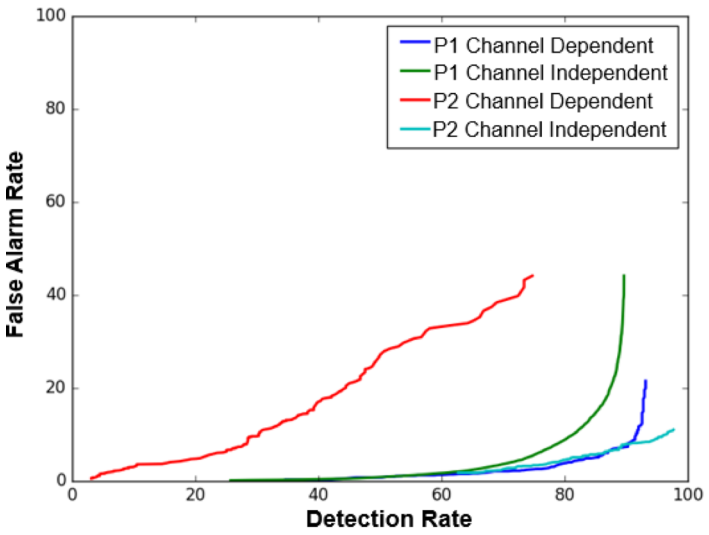

Figure 1. DET Curve for the first (P1) and second (P2) passes of processing for the channel dependent and independent models respectively.

Channel #	Channel Label	Error (%)
ch000	Fp1-F7	17.74
ch001	F7-T3	16.38
ch002	T3-T5	13.68
ch003	T5-O1	11.55
ch004	Fp2-F8	14.39
ch005	F8-T4	17.34
ch006	T4-T6	16.85
ch007	T6-O2	16.13
ch008	A1-T3	10.64
ch009	T3-C3	14.23
ch010	C3-CZ	15.31
ch011	CZ-C4	12.17
ch012	C4-T4	13.43
ch013	T4-A2	12.95
ch014	FP1-F3	15.62
ch015	F3-C3	14.64
ch016	C3-P3	15.73
ch017	P3-O1	12.65
ch018	FP2-F4	12.51
ch019	**F4-C4**	**8.87**
ch020	C4-P4	15.95
ch021	P4-O2	12.27

Table 1. HMM performance of channel dependent models reported individually for each channel.

systems need vast amounts of data to achieve good performance levels, and dividing the available data among the 22 channels does not benefit the deep learning system. We can see that, contrary to the previous case, the channel independent model is improved with the second pass of processing.

In general, we observe that the channel dependent models do improve the performance of the system up to the first pass of processing. To obtain a performance boost after P2, however, it would be necessary to expand the amount of labeled data for each channel. By the results presented so far, we can project that channel dependent models trained with large amounts of data could significantly improve the detection rate and decrease the false alarm rate after P2.

To address the lack of annotated data, we developed a self-training approach to iteratively annotate a large clinical EEG corpus. The main motivation for this task was to create and explore the impact of an algorithm that would work on large clinical EEG data resources. The principal outcome of this work is the ability to automatically annotate the entire TUH EEG database with a level of accuracy comparable to that of manual annotations made by human experts, allowing the implementation of more sophisticated deep learning systems.

The self-training algorithm that we developed was tested and implemented with the six EEG events that we mentioned above (SPSW, GPED, PLED, EYEM, ARTF, BCKG). A model for each class was first trained with a small pool of high confidence annotations made by experts. This trained model was then used to decode non-labeled data. The decoded events (1 second epochs) with the highest posterior probabilities (highest confidence labels) were then selected and added to the training pool to retrain the models. The retrained models were then evaluated in an open evaluation set and used to decode more unlabeled data in the corpus. These steps were repeated until the entire database was annotated with high confidence labels. Figure 2 Depicts an overview of the entire process.

To obtain an early assessment of the effectiveness of this newly developed technique, we evaluated the sensitivity for each class after one iteration of the algorithm. Table 2 summarizes the observations of this experiments. It is possible to see that the sensitivity improved for the GPED, PLED and SPSW models. This improvement was less evident for the EYEM model, which only exhibited an improvement of ~0.3%. The BCKG and ARTF models, on the other hand showed a degradation in performance. Since the last three classes are all background events and are largely available across the corpus, the implementation of the self-training algorithm for them is not as critical as it is for the first three classes. As a matter of fact, the event with the least occurrences in the original annotated data is SPSW, and is therefore the class that we focus on expanding for the rest of the analysis.

Class	Sensitivity (%)	
	Before	After
GPED	52.8	56.5
PLED	54.2	60.4
SPSW	41.6	49.6
EYEM	81.8	82.1
BCKG	72.1	71.2
ARTF	41.2	39.1

Table 2. Sensitivity of the six EEG event models before and after the first iteration of the self-training algorithm.

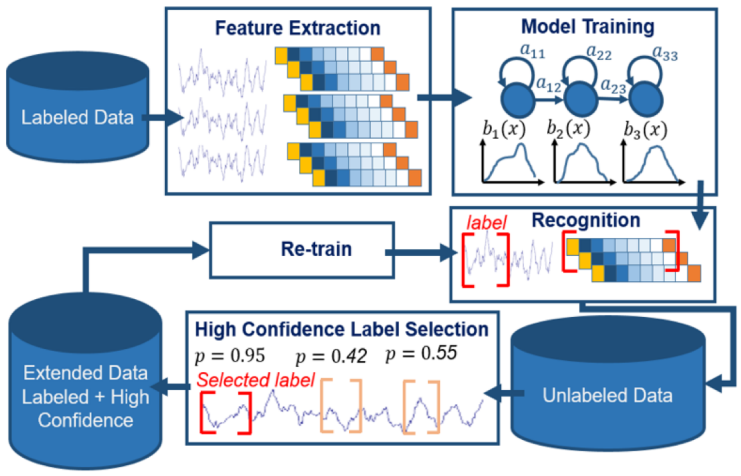

Figure 2. A generic approach to self-training.

This algorithm is innovative because, besides operating with high levels of confidence in a large clinical data resource, it works without any human supervision. For this reason, however, it is necessary to ensure that the algorithm's parameters are properly adjusted to select the correct number of high confidence events and maintain the performance of the system (or improve it) for future iterations. To optimize the algorithm and find a proper experimental scheme, we conducted two different types of tests: (1) analysis of the number of high confidence events to select for retraining, and (2) an analysis of the threshold for the posterior probabilities to be included for retraining.

First, we investigated the impact of different rankings of epochs. We focused on the SPSW class for this parameter analysis. Figure 3 shows the trend in recognition performance when we reduce the number of included SPSW events during re-training. In this analysis, we controlled the amount of highly ranked epochs: by reducing the preserved features for re-training, an increasing of the recognition performance was observed. As shown in Figure 3 we began by augmenting the training set with the top 10% of the decoded features. As we tightened the inclusion thresholds, recognition performance increased.

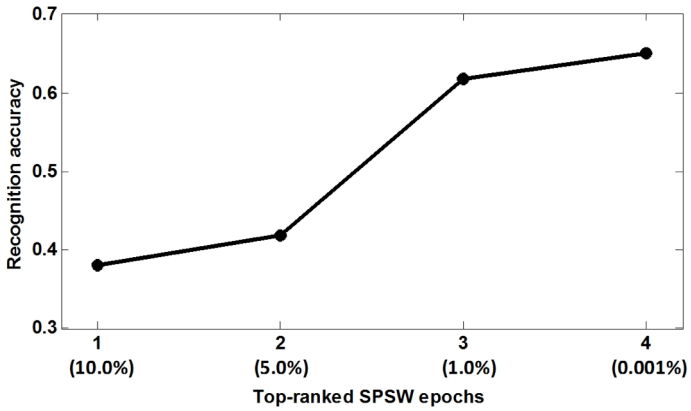

Figure 3. Number of decoded SPSW epochs used for retraining.

We conducted a series of experiments with different posterior probability (log likelihood) thresholds for the selected epochs. The results for the first part of the analysis helped to select an initial threshold, which was then varied to find is optimal value. Figure 4 depicts the recognition accuracy for SPSW events as a function of the threshold. Even though the figure shows that the optimal log-likelihood threshold for event selection is 355, the performance of the baseline is not compromised by threshold values as high as 375.

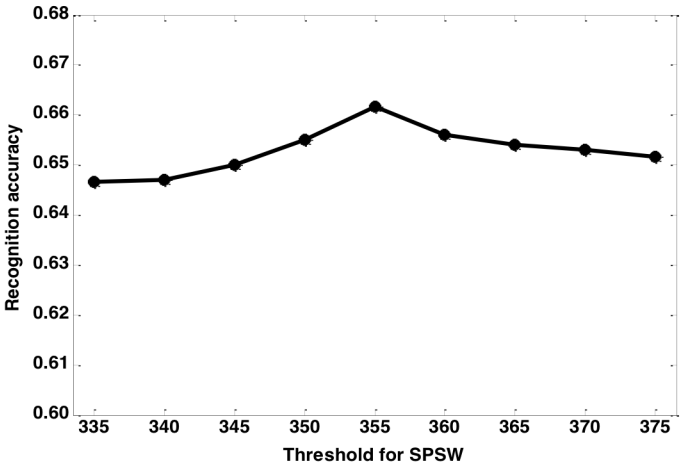

Figure 4. Effect of probability threshold variation over performance.

The experiments that we present above allowed to properly tune the algorithm for the labeling of new SPSW epochs. To further test the variability of the system's performance for a large amount of automatically labeled data, we ran the algorithm for 5 iterations. As is shown in Figure 5, a large number of SPSW epochs (almost 30,000 new labels) were automatically labeled by our algorithm, while maintaining a performance comparable to that of the baseline system for SPSW.

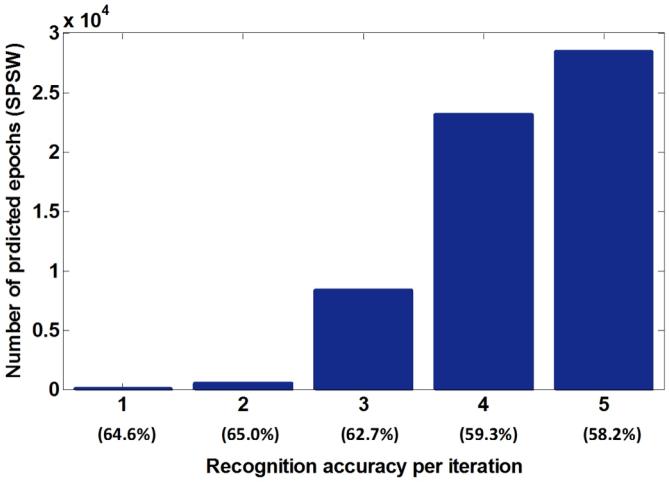

Figure 5. Effect of the added epochs for SPSW over performance for 5 iterations of the self-training algorithm.

The experiments that we have described to this point show that the approach that we have designed works for a large clinical EEG Corpus, TUH EEG, and provides high confidence annotations without human supervision. In other words, the implementation of this algorithm and future more sophisticated variations of it will not only allow the implementation of deep learning models for the decoding of EEG signals, but will also become a crucial tool in the annotation of the entire TUH EEG Corpus.

As an activity to support the increasingly complex training processes that we must use to train models with larger datasets, we invested time in the complete parallelization of our HMM training algorithms. More specifically, we substituted our isolated unit training system with an embedded training system that resembles algorithms that have been widely used in the speech recognition field to build sub-word systems. In essence, our new training procedure simultaneously updates all of the HMMs in a system using all of the input training data. Figure 6 depicts the training method that we were implementing before the modifications. This approach, although effective, is not suitable for operation in larger databases, since it estimates the HMM parameters in a

serial way. The parallel solution that we have adapted to our system, shown in Figure 7, solves the problem of long training times by parallelizing the training operations. It can be seen that the input data is divided, and each partition is used to estimate parameters that are later combined by an accumulator, and used to update the HMMs.

We investigated the effectiveness of the new training paradigm through the implementation of the same experiment with both sequential and parallel training. The experiment that we decided to use for comparison of the two training techniques was a seizure detection problem, which typically requires many hours of training, due to the long annotated seizure segments. The system was based on a left-to-right 3-state HMM for each class. The probabilities yielded by this model were later post-processed with a Stacked denoising Autoencoder for temporal and spatial context integration. For simplicity, we trained only two models: seizure and background (non-seizure). The training set contained 172 EEG files recorded with 22 channels. We confirmed the efficiency of the parallel training method when we observed that this technique, which used 150 cores, took about 2 hours for training, whereas the sequential training approach exceeded 22 hours.

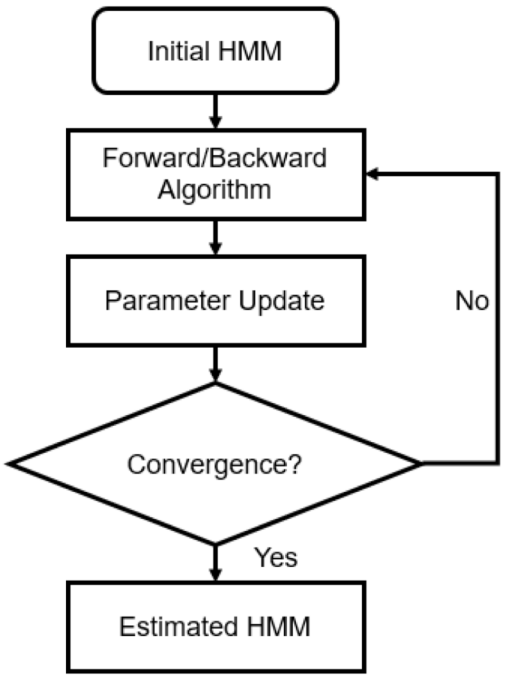

Figure 6. Process followed for the isolated unit training approach.

The performance of the systems for the two different training implementations was comparable. This behavior was expected, since the system modeled was essentially the same. The sensitivity of the trained background model for parallel training (29.07%) was higher than that of the same model trained sequentially (27.38%). On the contrary, the sensitivity of the seizure model trained in parallel was higher (18.67%) than the sequentially trained model (16.60%). The sensitivity for the second pass of processing (P2) for the system trained in parallel was 96.16% (SEIZ) and 93.85% (BCKG). For the sequentially trained models, the sensitivity was 95.84% (SEIZ) and 93.13% (BCKG) respectively.

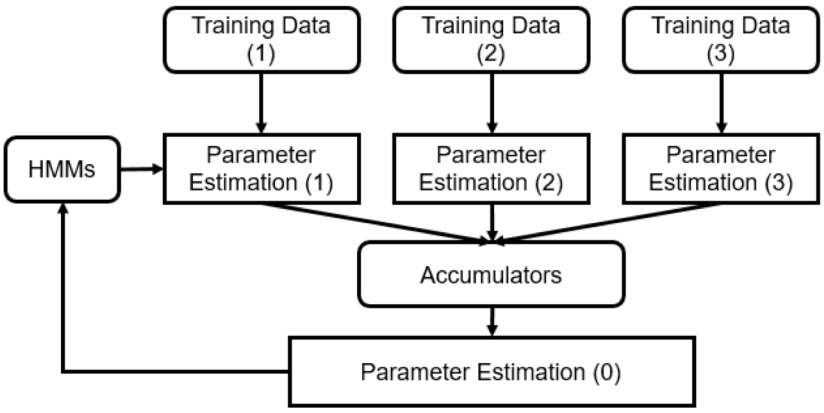

Figure 7. Parallelized training approach.

One thing that we observed in the seizure detection experiment mentioned above, however, was the high False Alarm rate (FA), which reached 1754 false detections per 24-hour period for the first pass of processing and 14146 for the more sensitive second pass. Decreasing the FA is a crucial aspect for the development of a clinical event detection system. In this sense, we explored the possibility of making some changes in our system that would allow us to reach a good compromise between the FA and the sensitivity.

The sensitivity is a one-to-one comparison between the ground truth label and the prediction, i.e. there is no imposter case in this process. This could lead to an extreme case: when the window duration is small enough, theoretically we may achieve rather high sensitivity since the tiny window can lose most of the class-specific information in feature extraction. This conjecture is verified by our results in P2: the sensitivity is high but the FA is also very large. In order to better deal with the imposter case (related to FA), each feature needs to better represent the characteristics of specific class. Studies in the EEG signal processing field, have shown that, depending on the classification scenarios, the optimal window size for EEG analysis ranges from 3s to 30s.

With the dataset and goal from the seizure detection system mentioned above, we conducted a series of experiments using different window sizes for decoding, we found that increasing the window duration can reduce the FA significantly. For P1 stage (HMM), increasing the window duration from 1 second (default) to 25 seconds the total FA rate reduced from 1754 per 24 hour to 404 per 24 hours. A similar FA reduction trend was observed from the P2 SdA processing stage: the FA rate reduced from 14146 per 24 hours to 141 per 24 hours. It is worth mentioning that since the individual EEG recordings may contain multiple classes (labels), the performance will reach an optimal point and then begin to drop as window duration continues to increase. The current experimental design clearly shows performance improvements, but since the same window length was applied globally, the big window may have covered multiple short events which belong to different classes during the evaluation. This will certainly reduce the sensitivity, which was noticed in our experiments too. Therefore, it is worth to investigate the topic of automatically adjusting window size locally using deep learning technology.

Goals Specific to Aim 2: *Automatically recognize critical concepts in an EMR:* (1) EEG activities and their attributes, (2) EEG events, (3) medical problems, (4) medical treatments and (5) medical tests mentioned in the narratives of the EEG reports, along with their inferred forms of *modality* and *polarity*. When we considered the recognition of the modality, we took advantage of the definitions used in the 2012 i2b2 challenge on evaluating temporal relations in medical text. In that challenge, modality was used to capture whether a medical event discerned from a medical record actually happened, is merely proposed, mentioned as conditional, or described as possible. We extended this definition such that the possible modality values of “factual”, “possible”, and “proposed” indicate that medical concepts mentioned in the EEGs are actual findings, possible findings and findings that may be true at some point in the future, respectively. For identifying polarity of medical concepts in EEG reports, we relied on the same definition used in the 2012 i2b2 challenge, considering that each concept can have either a “positive” or a “negative” polarity, depending on any absent or present negation of its finding. Through the identification of modality and polarity of the medical concepts, we aimed to capture the neurologist’s beliefs about the medical concepts mentioned in the EEG report. Some of the medical concepts mentioned in the EEG reports that describe the clinical picture of a patient are similar to those evaluated in the 2010 i2b2 challenge, as they represent medical problems, tests and treatments, thus we could take advantage of our participation in that challenge and use many of the features we have developed for automatically recognizing such medical concepts. However, EEG reports also contain a substantial number of mentions of EEG activities and EEG events, as they discuss the EEG test.

In the second year of the project, we developed the ability to automatically annotate all medical concepts from the EEGs, creating an annotation schema after consulting numerous neurology textbooks and inspecting a large number of EEG reports from TUH-EEG. The annotation schema also represents the first step in our Multi-Task Active Deep Learning (MTADL) paradigm developed in the second year of the project, which required the following 5 steps:

STEP 1: *The development of an annotation schema;*
STEP 2: *Annotation of initial training data;*
STEP 3: *Design of deep learning methods that are capable to be trained on the data;*
STEP 4: *Development of sampling methods for Multi-task Active Deep Learning system*
STEP 5: *Usage of the Active Learning system which involves:*
  Step 5.a.: *Accepting/Editing annotations of sampled examples*
  Step 5.b.: *Re-training the deep learning methods and evaluation the new system.*

After developing the annotation schema, we performed manual annotations of an initial training set, which enable the design and development of two deep learning methods that were trained on that data. In addition, we developed sampling mechanisms that enabled the “active” component of the deep leaning, which then was used by (a) accepting or (b) editing the samples examples, followed by the re-training of the deep learning methods. The entire architecture of MTADL is illustrated in training the two deep learning architectures illustrated in Figure 8.

We also evaluated the impact of Multi-task Active Deep Learning (MTADL) on the performance of our model. Specifically, we measured the change in performance after each additional round of annotations. Figure 9 presents these results. Clearly the impact of MTADL on the performance of our model across all tasks was significant allowing it to achieve high performance after as few as 100 additional EEG Reports have been annotated. We plan to continue using the MTDADL annotation tool to vastly improve the accuracy of identifying

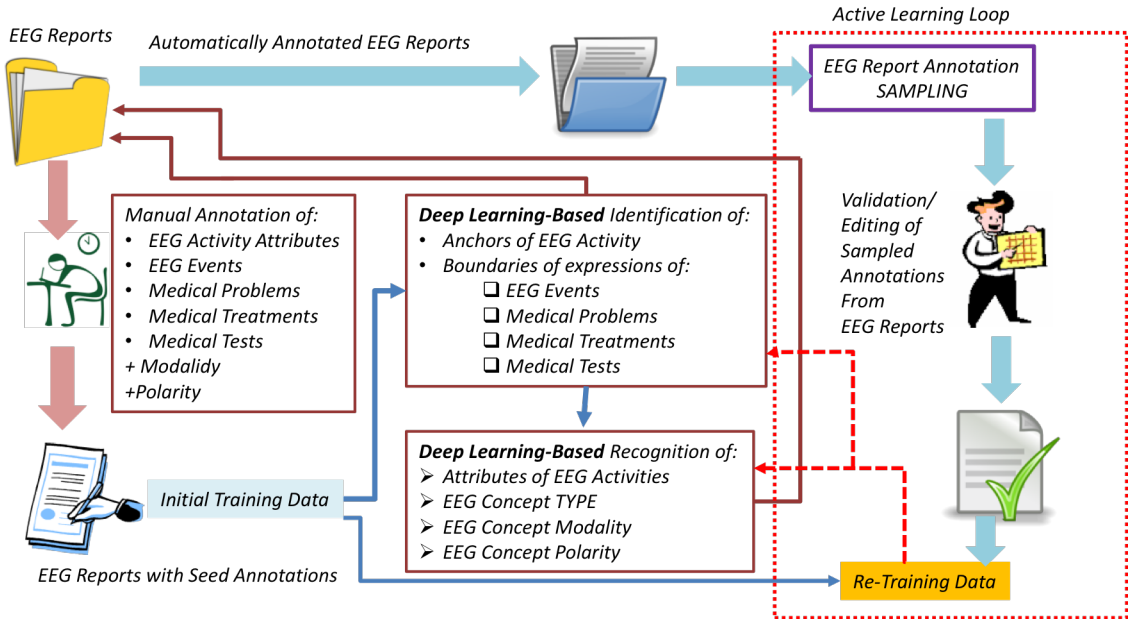

Figure 8. Architecture of the Multi-Task Active Deep Learning for annotating EEG reports.

automatically (1) the attributes of EEG activities as well as (b) the attributes of the other medical concepts mentioned in the EEG reports. As shown in Figure 9, we already have achieved impressive performance on identifying the anchors of EEG activities as well as the boundaries of all medical concepts.

Goals Specific to Aim 3: *Retrieve patient cohorts from the EMRs that document their hospital visits.* The Multi-Modal EEG Patient Cohort Retrieval system called *MERCuRY* (an acronym for Multi-modal EncephalogRam patient Cohort discoverY), was developed at University of Texas at Dallas by relying on the UTD team's prior experience with patient cohort identification based on principles of Information retrieval (IR). In the second year of the project, The UTD team aimed to improve this patient cohort retrieval system by making use of the medical knowledge that can be automatically acquired from the EEG reports. We generated qualified medical knowledge graphs (QMKGs) automatically discerned from the EEG reports by using big-data techniques, similar to our prior work on retrieving patient cohorts in the TREC Medical Records track (TRECMed), a task developed in 2011 and 2012 as an Information Retrieval challenge pertinent to real-world clinical medicine and evaluated in the annual TExt Retrial Conference (TREC) hosted by the National Institute for Standards and Technology (NIST).

We also developed another knowledge representation for yet another TREC special track on Clinical Decision Support (TREC-CDS), which a Clinical Picture and Therapy Graph (CPTG) as a factorized Markov network. Finally we developed a novel representation, namely medical knowledge embeddings (MKE) which is a new probabilistic knowledge representation which is superior to the QMKG or the CPTG because (1) the relationships are not informed only by cohesive properties of texts, but by patterns of interactions between medical concepts, as captured by deep learning methods; and (2) similar medical concepts and relations share the same neighborhoods in the multi-dimensional space enabled by the knowledge embeddings. The latter property resolves semantic heterogeneity which arises when disparate terminology is used to refer to the same concepts or relations while identical terms may refer to distinct concepts.

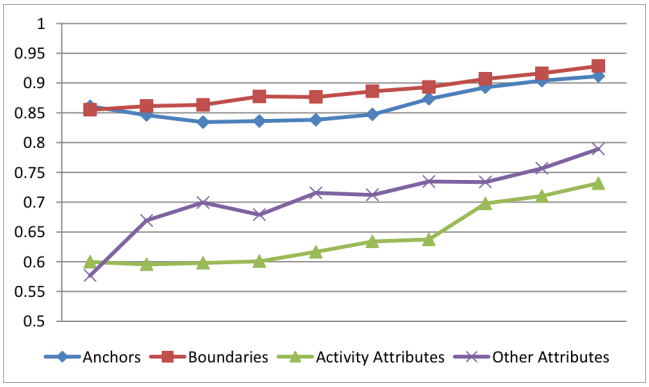

Figure 9. Learning curves for all annotations, shown over the first 100 EEG Reports annotated and evaluated with F_1 measure.

As noted in Sahoo et al. (2014) a seizure with alteration of consciousness may be referred to as complex partial seizure, dialeptic seizure or focal dyscognitive seizure by different epilepsy experts. An MKE representation places all these expressions in the same location of the multi-dimensional space, as it learns that they are involved in the same relations

with other epilepsy-relevant concepts. Thus, unlike the Epilepsy and Seizure Ontology (EpSO), the MKE representation does not require reconciliation of semantic heterogeneity, while being used for retrieving patient cohorts from medical records.

Bottom-up knowledge acquisition methods rely on the automatic identification of concepts and relations from data to enable (i) the population of the knowledge representation and (ii) linking the acquired knowledge to existing ontologies. In learning medical knowledge embeddings (MKE) from EEG reports we did not only perform bottom-up acquisition of medical knowledge from EEG reports, but we also represented the knowledge probabilistically in a multi-dimensional space and performed inference on it. To do so, we followed a methodology which involved the following four steps:

STEP 1: Decide which medical concepts and which relations between them are expressed in the EEG reports;
STEP 2: Automatically extract medical concepts and relations from the EEG reports;
STEP 3: Learn the MKE and generate the associated MKE graph;
STEP 4: Perform inference with MKE.

It is to be noted that the MKE represent only knowledge available from the EEG reports, which do not discuss the taxonomic organization of medical concepts or their partonymy relations. These forms of relations are encoded in medical ontologies, thus the MKE provide complementary knowledge to medical ontologies. However, many of the concepts represented in the MKE are also encoded in existing medical ontologies, providing a simple mechanism of linking the MKE to various ontologies available in BioPortal. For example, the clinical history and the medication list of EEG reports mention multiple medical concepts already encoded in the Unified Medical Language System (UMLS) ontology. Medical problems such as seizures, and treatments such as “Keppra”, “Lamictal” are encoded in UMLS while concepts such as idiopathic generalized epilepsy will be linked both to UMLS and the ESSO ontology. However, these ontologies do not capture relations between such concepts that are implied in the EEG reports, e.g. which brain activities evidence some epilepsy specific medical problems.

Our four-step methodology aims to capture and represent such relationships, while also providing their probabilistic likelihood, learned automatically from the medical practice evidenced in the large corpus of EEG reports. In addition to medical problems and treatments that describe the clinical picture and therapy of a patient, EEG reports mention EEG events, which represent stimuli that activates the EEG (e.g. hyperventilation) and EEG activities, representing brain waves or sequences of waves. The description section of the EEG reports describing the EEG record mention a multitude of EEG activities and events recognized by the neurologist from the analysis of the EEG signal. EEG activities are also mentioned in the impression section and in the clinical correlation section. Thus we decided to encode in the MKE four types of medical concepts: (1) EEG events; (2) EEG activities; (3) medical problems and (4) treatments. Whenever these concepts are also encoded in other ontologies, we linked to them. For example, medical problems such as idiopathic generalized epilepsy, when identified in an EEG report, with methods developed in the STEP 2 of our methodology, shall be linked to UMLS through its concept unique identifier (CUI). In addition to these four types of concepts, we decided to discern four types of binary relations that are implicit in the EEG reports. Each of these relations operates between a source argument and a destination argument. The relations along with examples of the four types of medical concepts are illustrated in Figure 10.

The EVIDENCES relation (from Figure 3) operates between: (a) EEG events, EEG activities, treatments, and (b) medical problems as providing evidence for the medical problem from the clinical correlation section of the EEG report. The EVOKES binary relation always has an EEG activity as a destination concept, as it attempts to capture the medical concepts that evoke the respective EEG activity. Those medical concepts can be either EEG events, or other EEG activities, medical problems or treatments followed by the patient. The third relation, namely OCCURS-WITH constraints both its arguments to be of the same type, e.g. either EEG activities, medical problems or treatments. The TREATMENT-FOR relation capture the treatments followed to prescribed for certain medical problems.

The extraction of medical knowledge from EEG reports consists of (1) automatic identification of medical concepts and (2) binary relation detection. Medical Concept Identification was performed by taking advantage of our existing MTADL active deep learning methodology, which was developed for Aim 2 of the main project. Detecting Relations between Medical Concepts was possible when pairs of medical concepts identified in the

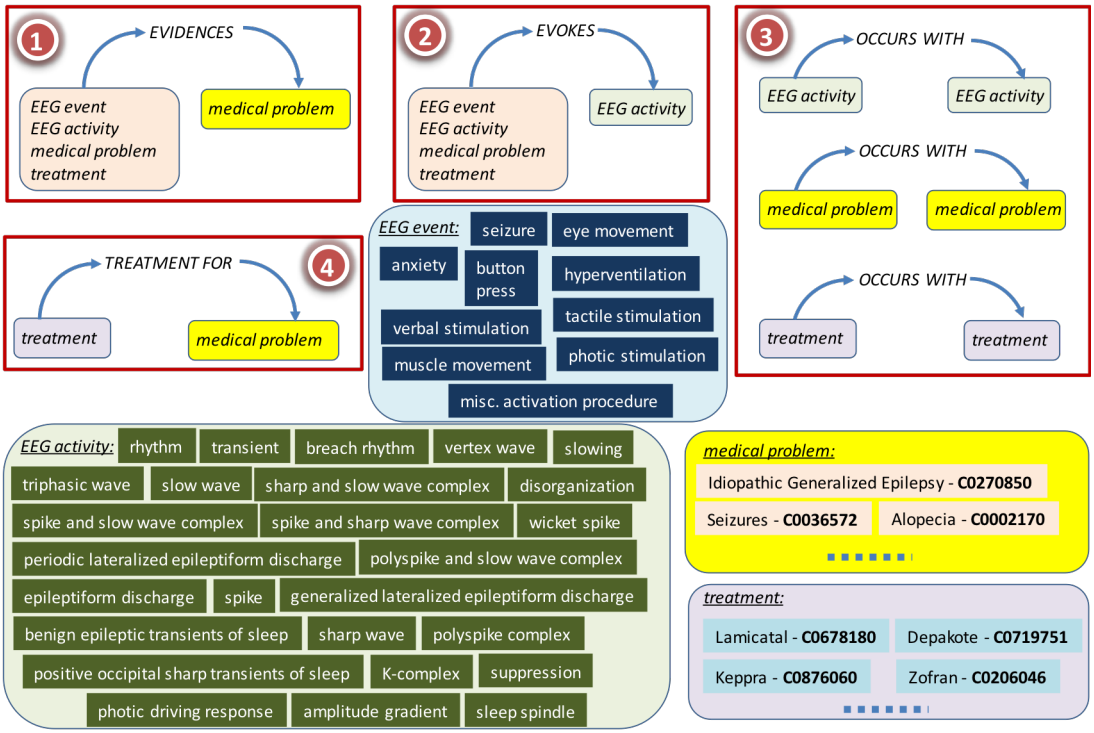

Figure 10. Medical concepts and relations represented in the Medical Knowledge Embeddings (MKE).

same EEG report were considered. Specifically, we established the four types of relations illustrated in Figure 10 by considering: (1) a *potential* EVIDENCES relation between any medical concepts from an EEG report and a medical problem identified in its clinical correlation section; (2) a *potential* EVOKES relation between any medical concept and an EEG activity, provided that the treatments were not identified in the clinical correlation section, as they may indicate possible or recommended treatments; (3) a *potential* OCCURS WITH relation between pairs of EEG activities, medical problems and treatments that are identified in the same section of the EEG report; and (4) a *potential* TREATMENT FOR relation between any treatment and a medical problem identified in the history section of the EEG report. All these potential relations are indicative of implied relations, that are not always directly stated in the text of the EEG report.

The likelihood of the relations represented in the MKE are learned by relying on TransE, a framework that represents relations between medical concepts as *translation vectors,* connecting its arguments, i.e. the two medical concepts in the embedding space. TransE learns an embedding, $\vec{c_i}$, for each medical concept c_i and an embedding, and an embedding $\vec{r}$, for each relation type *r* such that the relation embedding is a translation vector between the two concept embeddings representing its arguments. This means that for any medical concept c_i, the concept most likely to be related to c_i by the relation *r* should be the medical concept whose embedding is closest to ($\vec{c_i}$+$\vec{r}$) in the embedding space. By modeling the medical concepts as points in the embedded space and the relations between them as translation vectors, we were able to measure the plausibility of any potential relation between any pair of concepts using the geometric structure of the embedding space. The plausibility of a relation between a source medical concept and a destination medical concept, represented as a triple, $<c_s;r;c_d>$, is inversely proportional to the distance in the embedding space between the point predicted by our model ($\vec{c_s}$+$\vec{r}$) and the point in the embedding space representing the destination argument of the relation, i.e. $\vec{c_d}$. For this purpose, we used the *Manhattan Distance* as our distance function. Details of the formal models and the optimization functions were described in a paper that we submitted for the Annual Symposium of the American Medical Informatics Association (AMIA), entitled "*Deep Learning Meets Biomedical Ontologies: Knowledge Embeddings for Epilepsy*".

The relations represented in the MKE were evaluated in terms of (a) their plausibility; and (b) their completeness. The plausibility of relations encoded in MKE was assessed in three ways, measuring how well MKE ranks triples from a test set *T*, of 1,000 relation triples held out from the data used to train the MKE. For each triple *t* in the test set, we randomly removed either the source or destination argument and produced a set of candidate triples by replacing the removed argument with every medical concept automatically identified

with the methodology developed for Aim 2 of the main project. We ranked the candidate triples in ascending order according to the distance function. This allowed us to calculate the following metrics using the rankings produced from every triple in the test set: (1) Mean Reciprocal Rank (MRR); (2) Precision at 10 (P@10) and (3) Hits at K (i.e. H@10 and H@100). The micro-averaged Mean Reciprocal Rank of 83.33% indicates that for the majority of triples in the test set, the top ranked candidate triple is correct. The P@10 metric showed that 66.73% of the top 10 ranked triples were correct, in general. It is interesting to note that the results for the Hits@10 metric have the most variability between relation types. In general, the Hits@100 results showed that the MKE correctly ranks test relations *t* in the top 5% of candidate triples 81.3% of the time. Future work will consider techniques for learning plausibility thresholds that will allow MKE to be considered for curation and acceptance in existing, expert and community-validated biomedical ontologies.

By applying knowledge graph embedding techniques, we were able to discover data-driven knowledge which can be linked to other ontologies from the BioPortal. Experimental results demonstrated the promise of this approach and highlight the potential of the MKE for bridging the knowledge gaps of existing neurological ontologies. The MKE developed for this aim showcased the way in which deep learning techniques applied to large collections of medical records can supply medical knowledge derived from clinical practice and meet the ontological commitments encoded in existing biomedical ontologies. By representing medical knowledge probabilistically, the MKE will also enable probabilistic reasoning on its knowledge.

Finally, the team at The University of Texas at Dallas has developed a web interface that allows users to (1) enter and search arbitrary patient cohort queries, (2) browse the retrieved (and ranked) EEG reports produced by their MERCuRY system for the given query, and (3) view the content of the of each retrieved EEG report. The team at Temple University has developed a system that performs automatic time-aligned EEG event recognition and displays the recognized signals through a local application. To efficiently integrate the knowledge learned from the two applications, we designed an Application Programming Interface (API) that allows requests and responses from the MERCuRY system. Figure 11 shows a screenshot of the user interface for the patient cohort system, which allows to visualize the annotated EEG signals and medical reports.

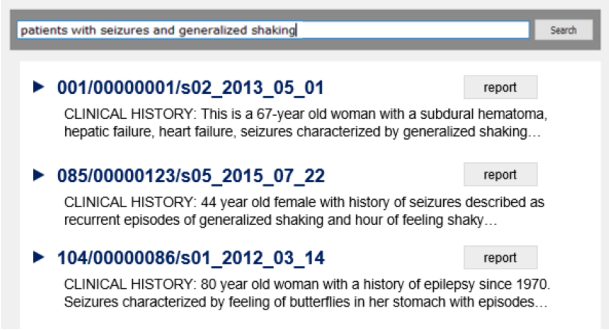

Figure 11. The graphical user interface for the patient cohort retrieval system.

Goals Specific to Aim 4: *Validate the usefulness of the patient cohort identification system by collecting feedback from clinicians and medical students.* For each query, medical experts shall examine the top ranked cohorts for common precision errors (false positives), and the bottom five ranked common recall errors (false negatives). In a very fruitful collaboration, both the Temple University team and the UTD team have participated in the evaluation and validation of the patient cohort identification system implemented in the MERCuRY system. Our immediate goal this second year was to assemble *120 clinically relevant queries* that are used by neurologists to evaluate the quality of the EEG reports/records considered relevant by the patient cohort retrieval system in its current form. To assembled to targeted number of queries, we have performed several rounds of query generation. In each round, the neurologists have provided clinical rationales for the queries, suggested some, invalidated some and are continued to be interviewed. We retained only patient cohort queries that have been validated independently by two neurologists and compute the inter-neurologist agreement rate. In addition, we are collecting examples of queries which were not deemed clinically relevant by at least two neurologists, along with their rationales. In the process, we also assigned evaluation tasks to the neurologists that either proposed a query or validated a query. Examples of the clinically relevant queries provided by neurologists and analyzed by us are:

No	QUERY	Agreed By	Rationale
GQ1	*Patients experiencing **seizures** and generalized shaking*	*Drs. Cheng, Ellis & Tobochnick*	*Dr. Cheng's rationale*: This query defines a specific clinical population for which a few specific EEG patterns are typically associated with. May not exclude psychogenic non-epileptogenic seizures. *Dr. Ellis's rationale*: I think "seizures" is certainly a relevant query. In theory it would also be relevant to query the clinical semiology ("generalized shaking").
GQ2	*Multiple sclerosis and seizure*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: Seizures are more prevalent in patients with multiple sclerosis compared to the general population. This search would isolate those patients with MS who had an EEG ordered for seizure evaluation. *Dr. Tobochnick's rationale*: Same
GQ3	*Patients with **anoxic brain injury** and EEGs demonstrating **sharp waves, spikes, or spike/polyspike and wave activity** or seizures.*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: This query helps to identify the prevalence of anoxic brain injury patients with a predisposition towards, or captured evidence of, seizures. While not often a consequence of anoxic brain injury, seizures may occur. *Dr. Tobochnick's rationale*: Same
GQ4	*Patients under 18 years old with absence seizures*	*Drs. Barnes, Ellis, Cheng & Tobochnick*	*Dr. Barnes's rationale*: This is a nice, well-defined query. *Dr. Cheng's rationale*: Same as Dr. Tobochnick. *Dr. Ellis' rationale*: I assume we mean electrographic absences, that is, 3Hz generalized spike-and-wave discharges with associated loss of awareness, captured during EEG recording. *Dr. Tobochnick's rationale*: Specifies patient population with type of epilepsy consistent with age range in query and type of seizures with characteristic EEG pattern.
GQ5	*Patients over age 18 with history of developmental delay and EEG with electrographic seizures*	*Drs. Barnes, Ellis, Cheng & Tobochnick*	*Dr. Barnes's rationale*: The morphology of the ictal patterns would be of interest to epileptologists. *Dr. Cheng's rationale*: Same as Dr. Tobochnick. *Dr. Ellis' rationale*: A relevant and interesting cohort of patients (not all developmental delay is due to brain injury, but clearly the brain is functionally abnormal by definition). *Dr. Tobochnick's rationale*: Returns records from specific clinical patient population with seizures captured on EEG. Many patients with developmental delay have higher risk of epilepsy due to prior brain injury.
GQ6	*History of seizures and EEG with TIRDA without sharps, spikes, or electrographic seizures*	*Drs. Barnes, Ellis, Cheng & Tobochnick*	*Dr. Barnes's rationale*: TIRDA is a very relevant finding to EEG physicians. *Dr. Cheng's rationale*: Note that TIRDA can occur with or without associated epileptiform activity. This wording may exclude patients with both TIRDA and sharp waves. *Dr. Ellis' rationale*: TIRDA is definitely a clinically relevant EEG finding to query. I also like the option of querying for the absence of specific findings, as a way to limit the query results. *Dr. Tobochnick's rationale*: Queries patients with clinical history concerning for epilepsy and EEG with specific feature that has epileptic potential.
GQ7	*History of psychogenic non-epileptic seizures and EEG with sharp waves, spike/polyspike and wave or spikes*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: Epilepsy and psychogenic non-epileptic seizures are often co-morbid. This query would help to identify those with both epileptic and non-epileptic seizures. *Dr. Tobochnick's rationale*: Same
GQ8	*Patients with history of seizure and normal EEG*	*Drs. Barnes & Tobochnick*	*Dr. Barnes' rationale:* Could help to guide management of patients after first time seizure *Dr. Tobochnick's rationale*: Same
GQ9	*Patients evaluated for seizures vs **stroke***	*Drs. Barnes & Tobochnick*	*Dr. Barnes's rationale*: Although this will retrieve a large number of records, the question of TIA vs seizure is a clinically relevant one, and EEG may be helpful. We see teams consulting for this all the time. *Dr. Tobochnick's rationale*: Although this will retrieve a large number of records, the question of TIA vs seizure is a clinically relevant one, and EEG may be helpful.
GQ10	*Patients with stroke and an EEG demonstrating sharp waves, spikes, or spike/polyspike and wave activity*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: While individuals with stroke are at increased risk of having seizures, the majority do not. This query helps to identify those with EEGs demonstrating a predisposition to seizures, i.e. with epileptiform activity. *Dr. Tobochnick's rationale*: Same
GQ11	*Brain tumor and sharp waves,*	*Drs. Cheng*	*Dr. Cheng's rationale*: Identifies patients with a brain tumor which

	spike/polyspike and wave or spikes.	*& Tobochnick*	demonstrates epileptogenicity, i.e. a predisposition to having seizures. *Dr. Tobochnick's rationale*: Same
GQ12	*Autism and sharp waves, spike/polyspike and wave or spikes*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: Identifies patients with autism that demonstrate epileptogenicity, i.e. a predisposition to having seizures. *Dr. Tobochnick's rationale*: Same
GQ13	*EEGs without sharp waves, spikes, or spike/polyspike and wave activity in patient's diagnosed with* ***epilepsy***	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: While individuals with stroke are at increased risk of having seizures, the majority do not. This query helps to identify those with EEGs demonstrating a predisposition to seizures, i.e. with epileptiform activity. *Dr. Tobochnick's rationale*: Same
GQ14	*Patients taking topiramate (Topamax) with a diagnosis of headache and EEGs demonstrating sharp waves, spikes, or spike/polyspike and wave activity.*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: Topiramate (Topamax) is indicated for primary use in management of seizures and migraine headaches. This query identifies how common it is for patients with both conditions to be prescribed topiramate. *Dr. Tobochnick's rationale*: Same
GQ15	*Patients with anoxic brain injury and EEG reports denoting brain death.*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale*: Brain death remains a clinical diagnosis. However, there is no universal consensus on brain death criteria, and ancillary tests such as EEG are often used to assist with this. However, electroencephalographers are wary of declaring brain death from an EEG, as even electrodes on gelatin can demonstrate a signal. Nonethelss, this practice persists. This query would help to identify how often EEG contributes to brain death determination in clinical practice. *Dr. Tobochnick's rationale*: Same
GQ16	*Patients with a history of anoxic brain injury*	*Drs. Barnes & Tobochnick*	*Dr. Barnes' rationale*: Often ordered consult, can help researchers propose a manner in which to decide on who requires an EEG as a part of their care. *Dr. Tobochnick's rationale*: EEG can be helpful in evaluating for epileptic myoclonus after anoxic brain injury, although again the utility of this query would be greater with information regarding current mental status (awake, obtunded, coma) and timing of brain injury (2 years ago vs 2 days ago).
GQ17	***EEG showing*** *electrocerebral silence, aka electrocerebral inactivity (ECI)*	*Drs. Ellis & Tobochnick*	*Dr. Ellis' rationale*: This is the EEG finding in brain death. *Dr. Tobochnick's rationale*: Same.
GQ18	*EEG showing triphasic waves.*	*Drs. Ellis & Tobochnick*	*Dr. Ellis' rationale*: This EEG pattern is often seen in patients with hepatic encephalopathy due to liver failure, or other metabolic encephalopathies. Generally this pattern is not considered epileptiform, though that claim is somewhat controversial. *Dr. Tobochnick's rationale*: Same
GQ19	*EEG showing periodic lateralized epileptiform discharges (PLEDs)*	*Drs. Ellis & Tobochnick*	*Dr. Ellis' rationale*: This EEG pattern is common in patients with highly epileptogenic brain lesions, such as HSV encephalitis, brain tumors, etc. It portends very high risk for seizures. Management of this EEG pattern itself is controversial. *Dr. Tobochnick's rationale*: Same
GQ20	*EEG showing generalized periodic epileptiform discharges (GPEDs)*	*Drs. Ellis & Tobochnick*	*Dr. Ellis' rationale*: This EEG pattern is often seen after global brain injury such as post-cardiac arrest, and is thought to portend a poor prognosis, but the clinical management is controversial. *Dr. Tobochnick's rationale*: Same
GQ21	*EEG showing burst suppression*	*Drs. Ellis & Tobochnick*	*Dr. Ellis' rationale*: burst-suppression is seen after severe global brain injury (for example anoxic injury after cardiac arrest) or due to anesthetic medications (the so-called "pharmacologic coma"). Whether or not the burst-suppression pattern is medication induced would also be clinically relevant, but searching the EEG reports themselves unlikely to reliably provide this data. *Dr. Tobochnick's rationale*: burst suppression pattern can sometimes add prognostic value.
GQ22	*Patients under 50 with FIRDA*	*Drs. Barnes & Tobochnick*	*Dr. Barnes' rationale*: FIRDA is a normal variant above 50, is a nonspecific finding sometimes associated with encephalopathy. May be beneficial to allow for the search so that the data can be compared to the clinical course of the patient *Dr. Tobochnick's rationale*: Same
GQ23	*Pediatric patients with posterior dominant rhythms slower than anticipated for age*	*Drs. Barnes, Cheng & Tobochnick*	*Dr. Barnes's rationale*: same as Dr. Tobochnick. *Dr. Cheng's rationale*: same as Dr. Tobochnick. *Dr. Tobochnick's rationale*: Query provides specific EEG abnormality in an appropriately specific patient population, the

			causes of which are many.
GQ24	*ICU patients with altered mental status and EEG showing non-convulsive status epilepticus (NCSE)*	*Drs. Barnes, Ellis, Cheng & Tobochnick*	*Dr. Barnes's rationale*: Up to 20% of unconscious patients in the ICU setting have status epilepticus, and we have scant data on them, this would be a query that would prove very useful. *Dr. Cheng's rationale*: same as Dr. Tobochnick. *Dr. Tobochnick's rationale*: Specific population of patients with indication for EEG and finding of seizure activity on EEG without clinical convulsions. "ICU" is optional but provides a much more homogenous patient population. NCSE is common in ICU patients with typical presentation only of altered mental status.
GQ25	*Patients with a history of head trauma and **abnormal EEG***	*Drs. Barnes & Tobochnick*	*Dr. Barnes' rationale*: Specific patient population that may have focal changes on EEG of interest to a corpus user *Dr. Tobochnick's rationale*: Same
GQ26	*Patients with a history of migraine and abnormal EEG*	*Drs. Barnes & Tobochnick*	*Dr. Barnes' rationale*: Could lend to helping influence decisions on which headache patients we should send for an EEG, pattern that we don't know may arise *Dr. Tobochnick's rationale*: Complex migraines may occasionally have focal features that could potentially be confused for seizures
GQ27	*Patients with numbness or paresthesias as an indication for EEG, with EEG both with and without epileptiform activity*	*Drs. Cheng & Tobochnick*	*Dr. Cheng's rationale:* Surface EEG is not high yield for seizure onset zones with a small surface area, and for auras such as sensory symptoms, though it is often ordered. This query would find patients who presented with such symptoms, and help to estimate predictive value of such an investigation. *Dr. Tobochnick's rationale*: Same
GQ28	*Patients younger than 30 years old with wicket noted on EEG*	*Drs. Barnes & Tobochnick*	*Dr. Barnes' rationale*: Abnormal finding for a patient of that age, but nonspecific, could yield a previously unrecognized pattern *Dr. Tobochnick's rationale*: Same
GQ29	*Patients with excess theta in drowsiness*	*Drs. Barnes & Tobochnick*	*Dr. Barnes's rationale*: Who are we to determine what someone will want to look at? May be useful in the future. *Dr. Tobochnick's rationale*: This query provides a specific EEG finding and a specific clinical background. The ultimate clinical significance of this scenario is pretty unremarkable, but the query itself is appropriate.
GQ30	*Epileptiform discharges with (or without) a clinical correlate*	*Drs. Ellis & Tobochnick*	*Dr. Ellis' rationale*: relevant to distinguish whether epileptiform discharges cause clinical manifestations. *Dr. Tobochnick's rationale*: suggestive of diagnosis of epilepsy

We developed an evaluation protocol which was followed by both teams through a secure-interface generated at UTD. We primarily evaluated the MERCuRY system according to its ability to retrieve patient cohorts. To this end, we generated a set of 115 evaluation queries. For each query, we retrieved the ten most relevant patients as well as a random sample of ten additional patients retrieved between ranks eleven and one hundred. We asked six relevance assessors to judge whether each of these patients belonged or did not belong to the given cohort. Moreover, the order of the documents (and queries) were randomized and judges were not told the ranked position of each patient. Each query and patient pair was judged by at least two relevance assessors, obtaining an inter-annotator agreement of 80.1% (measured by Cohen's kappa).

This experimental design allowed us to evaluate not only the set of patients retrieved for each cohort, but also the individual rank assigned to them. Specifically, we adopted standard measures for information retrieval effectiveness, where patients labeled as belonging to the cohort were considered *relevant* to the cohort query, and patients labelled as not belonging to the cohort were considered as *non-relevant* the cohort query. Note that because our relevance assessments consider only a sample of the patients retrieved for each query, we adopted two measures of ranked retrieval quality: the Mean Average Precision[28] (MAP) and the Normalized Discounted Cumulative Gain (NDCG). The MAP provides a single measurement of the quality of patients retrieved at each rank for a particular topic. Likewise, the NDCG measures the *gain* in overall cohort quality obtained by including the patients retrieved at each rank. This gain is accumulated from the top-retrieved patient to the bottom-retrieved patient, with the gain of each patient discounted at lower ranks. Lastly, we computed the "Precision at 10" metric (P@10), which measures the ratio of patients retrieved in the first ranks which belong to the patient cohort.

A number of clinicians have evaluated the clinical queries provided by UTD. The evaluation has been conducted for two rounds: in the first round, 15 queries were composed by students who have no clinical background, these queries were then sent to 4 neurologists to evaluate their clinical relevance: each

neurologist made his/her judgment for the queries. The neurologists were also asked to provide another 5 clinical-related queries. For the original 15 queries, depending on the judgements from the neurologists the confidence of clinical-relevance was ranked, i.e. analyze the inter-rater agreement amongst the four neurologists. This information was provided to our UTD colleagues to tune their query retrieval system. In the second round, 60 different queries were presented to 2 clinical consultants for new the round of judgments, each of them was also asked to provide 15 additional queries. These query evaluations were then sent to UTD team for the analysis and model training.

Accomplishments – Supplemental

Aim 1: *Automatic labeling of the TUH EEG Corpus for seizure events.*

We have labeled the data using AutoEEG and Persyst.

Aim 2: *Application of deep learning sequential modeling techniques for EEGs* to predict seizures. Characterize performance as a function of latency.

We have characterized performance using oour HMM-SdA system. See plot.

Aim 3: *Defining and generating Hierarchical epileptiform Activity Descriptors (HAD) for EEGs using deep learning.*

Identification of epileptiform activities, seizures and the specific EEG patterns that accompany epilepsy syndromes remains an electroencephalographer's most critical task. EEG signals record both epileptiform activities and EEG events. While the Hierarchical Event Descriptors (HED) (available from http://www.hedtags.org) have defined many types of EEG experimental events, no existing components of schema.org standardize the epileptiform activities and their attributes. We filled this gap by generating a schema of Hierarchical epileptiform Activity Descriptors (HAD). Similarly to the Hierarchical Event Descriptors (HED), we generated a hierarchical structure for the Hierarchical epileptiform Activity Descriptors (HAD), which will be rooted into the HAD tag, while organizing hierarchies as attributes for (1) the epileptiform activity waveform; (2) the epileptiform activity frequency band; (3) the epileptiform activity anatomical location; (4) the epileptiform activity position; (5) the epileptiform activity distribution; (6) the epileptiform activity frequency; (7) the epileptiform activity magnitude. We generated 1890 HAD attributes, organized in seven hierarchies: (a) a hierarchy of *morphology* HAD tags; (b) a hierarchy of *frequency bands*; (c) a hierarchy of *magnitude* tags; (d) a hierarchy of *recurrence* HAD tags; (e) a hierarchy of *dispersal* HAD tags; (f) a hierarchy of brain *hemisphere* HAD tags; (g) a hierarchy of *brain location* HAD tags; and (f) a small hierarchy of HAD tags to indicate whether the EEG activity occurs in the *background* or not. It is to be noted that out of the seven attribute hierarchies, three correspond to spatial properties and one to temporal properties of the EEG activities.
We designed a fine-grained ontology of annotations to account for the variation in the medical language used by neurologists when generating the EEG reports. For example, for the "Morphology" attribute we have defined 2 sub-types, namely: (1) "Rhythm" and (2) "Transient". "Transient", in turn, contains 3 sub-types: "Pattern", "Complex", and "Single Wave". We also noted the importance of the attribute denoted as "Frequency Band". The frequency band associated with an EEG activity is important for diagnosis. A polyspike, for example, represents a transient waveform whose frequency can lie within the alpha, beta, delta or theta brands. Knowing the frequency of the activity can change the clinical significance. For example, in http://www.ncbi.nlm.nih.gov/books/NBK2608/figure/ch10.f5/ , the authors report that polyspikes with a 14 Hz (Alpha or Beta band) frequency are associated with Doose Syndrome, while in http://www.ncbi.nlm.nih.gov/books/NBK98213/ the authors report that absence seizures are associated with polyspikes with 2 to 4 Hz frequencies (Delta band). Together, the "Frequency Band" and "Morphology" attributes allows us to indicate the frequency of not only rhythmic waves, but of transients (like polyspikes) as well. For example, "alpha waves" would correspond to an EEG annotation of "waves" with "Morphology=rhythm" and "Frequency Band=alpha".
The full HAD tag hierarchies are:

Morphology ::= represents the type or "form" of EEG activity waves.

- Rhythm
- Transient
  - Pattern
    - Burst suppression
    - Slowing
    - Benign epileptic transients of sleep (BETS)
    - Photic driving (response)
    - Periodic Laterilized Epilepitiform Discharges (**PLEDs**) were defined as repetitive periodic, focal, or hemispheric epileptiform discharges (spikes, spike and waves, polyspikes, sharp waves) usually recurring every 1 to 2 seconds.
    - Generalized periodic epileptiform discharges (**GPEDs**) are very rare patterns and are classified as periodic short-interval diffuse discharges (PSIDDs), periodic long-interval diffuse discharges (PLIDDs) and suppression-burst patterns according to the interval between the discharges.
    - Epileptiform discharge (unspecified)
  - Complex: A sequence of two or more waves having a characteristic form or recurring with a fairly consistent form, distinguished from background activity.
    - K-complex
    - Sleep spindles
    - Spike-and-sharp-wave complex
    - Spike-and-slow-wave complex
    - Sharp-and-slow-wave complex
    - Triphasic wave: High-amplitude (over 70 mV) positive sharp transients, which are preceded and followed by relatively low-amplitude negative waves. The first negative wave generally has a lower amplitude than the negative afterwave. The distribution is generalized, and frequently the largest deflections in a bipolar fronto-occipital derivation occur at the frontal electrodes. Triphasic waves tend to have a repetition rate of ca. 1±2 Hz.
    - Polyspike complex
    - Polyspike-and-slow-wave complex
  - (Single) Wave:
    - V wave
    - Wicket spikes
    - Spike
    - Sharp wave
    - Slow wave
    - Positive occipital sharp transients of sleep (POSTS) / Lambda Wave

Frequency Band ::= Clinically relevant frequency bands (Details in http://emedicine.medscape.com/article/1139332-overview#a2)

- Alpha (8 – 13 Hz)
- Beta (13 – 32 Hz)
- Delta (< 4 Hz)
- Theta (4 - 8 Hz)
- Gamma (> 32 Hz)
- N/A

Magnitude ::= describes the amplitude of the EEG activity if it is emphasized in the EEG report

The MAGNITUDE attribute of the EEG Activity may have the following values:

- Low: e.g.*: subtle (spike), small (polyspike discharge)*
- High: e.g.*: high (voltage burst); high amplitude (spike); excess (theta)*
- Normal

Background ::= this is a binary attribute to denote if an EEG activity occurs in the background or not.

The BACKGROUND attribute of the EEG Activity may have the following values:

- Yes
- No

Recurrence (TEMPORAL) ::= describes how often the EEG activity occurs. The RECURRENCE attribute of the EEG Activity may have the following values: • Continuous (e.g. “rhythmic”) • Repeated (e.g. “intermittent”, “regular”) • None (e.g. “burst”)
Dispersal (SPATIAL) ::= describes the spread of the activity over regions of the brain The DISPERSAL attribute of the EEG Activity may have the following values: • Localized (e.g. focal): limited to a small area of the brain • Generalized (e.g. diffuse): occurring over a large area of the brain or both sides of the head • N/A (if not specified)
Hemisphere (SPATIAL) ::= describes which hemisphere of the brain does the activity occur in. The HEMISPHERE attribute of the EEG Activity may have the following values: • Right • Left • Both • N/A (if no hemispheric information is provided)
Brain Location (SPATIAL) ::= describes the region of the brain in which the EEG activity occurs The BRAIN LOCATION attribute of the EEG Activity indicates the location/area of the activity (corresponding to electrode placement). It may have the following values: • Frontal (i.e. Anterior): Corresponds to the frontal region of the brain including all F*, Fp* and AF* electrodes • Occipital (i.e. Posterior): Corresponds to the occipital region of the brain including all O* electrodes • Temporal: Corresponds to the temporal region of the brain including all T* electrodes • Central: Corresponds to the central region of the brain including all C* electrodes • Parietal: Corresponds to the parietal region of the brain including all P* electrodes • Frontocentral: Corresponds to the area between the frontal and central regions of the brain including all FC* electrodes • Frontotemporal: Corresponds to the region between the frontal and temporal regions of the brain including all FT* electrodes • Centroparietal: Corresponds to the region between the central and parietal regions of the brain including all CP* electrodes • Parieto-occiptal: Corresponds to the region between the parietal and occipital regions of the brain including all PO* electrodes • N/A (if no location information is provided)

The “Recurrence” HAD tag captures whether an EEG activity re-occurs, and, if so, whether the recurrence is continuous or not. It is a special form of temporal attribute, which, to our knowledge, has not been studied in previous annotations of temporal information in EHRs. In contrast, the "Dispersal" attribute does not capture any temporal information, but instead indicates whether the EEG activity was limited to a small "localized" region of the brain, or that it occurred throughout a "generalized" or large region of the brain. Thus it is one of the *spatial attributes* that we discern from the EEG reports. The second spatial attribute capture the hemisphere of the brain where the EEG activity is noticed. The third spatial attribute is the "BRAIN LOCATION", defined as a multi-valued attribute which organize the values into a hierarchy, such that each electrode is a leaf under the corresponding brain region. E.g. "Brain Location=Temporal>T1" if we know the location of the EEG activity is T1, or just "Brain Location=Temporal" if we don't know the location, but only the area. This works with multiple values, for example "polyspike in T1, T2 spreading into parietal lobe" would produce the attribute "Brain Location=(Temporal>T1, Temporal>T3, Parietal)".

Aim 4: *Automated Tagging of HADs in medical texts using deep learning.*

The definition of the HAD tags allowed us to develop deep neural learning architectures capable of annotating the tags in the EEG reports, and in many other biomedical texts. For example, a mention of a PLED in the EEG report would correspond to the attribute annotations "*Morphology=Transient>Pattern>PLED*" with other attributes depending on the context. However, our fine-grained attributes allows us to detect circumlocutious or implied PLEDs. For example, "*bursts of frontally predominant high amplitude spike or sharp activity*" -- a type of PLED -- would correspond to the EEG Activity "*spike or sharp activity*" with the following attributes:

- *Morphology=Transient>Complex>Spike-and-sharp-wave complex*
- *Frequency Band: N/A*
- *Magnitude: High*
- *Background: No*
- *Recurrence: Repeated*
- *Dispersal: Localized*
- *Hemisphere: N/A*
- *Brain Location: Frontal*

Thus, because a PLED is defined as repetitive periodic, focal, or hemispheric epileptiform discharges (spikes, spike and waves, polyspikes, sharp waves), we can infer implied PLEDs as EEG Activities whose Morphology attribute has the value "spike", "spike and wave", "polyspikes", or "sharp waves" and its Recurrence has the value "Repeated". Automatically producing HAD tag annotations was made possible by two deep learning architectures informed by two feature vector representations, that considered the features illustrated in Table 1.

*Features used for **Deep Learning-Based** Identification of Anchors of EEG Activity Attributes and*	*Features used for **Deep Learning-Based Recognition** of Attributes EEG Activities*
1. The lemma of the token and the previous/next tokens *2. The PoS of the token and the previous/next tokens* *3. The phrase chunk of the token and the previous/next tokens* *4. The lemmas of the previous, current, and next tokens* *5. The Brown cluster of the token* *6. The UMLS Concept Unique Identifier (cui) of UMLS concepts containing the token* *7. The title of the section containing the token*	*1. The medical concept mention itself* *2. The lemmas of each token in the medical concept mention* *3. The PoS of each token in the medical concept mention* *4. The lemmas of 3 tokens before/after the medical concept mention* *5. The title of the section containing the token* *Context Features: For each token, **t**, in the sentence:* *6. The syntactic dependency path to **t**.* *7. The number of words between the medical concept mention and **t*** *8. The number of "hops" in the syntactic dependency path from the head of the medical concept mention to **t*** *9. The number of medical concepts between the medical concept mention and **t***

Table 1: Features vectors used for automatic annotation of HAD tags

We used the GENIA tagger for tokenization, lemmatization, Part of Speech (PoS) recognition, and phrase chunking. Stanford CoreNLP was used for syntactic dependency parsing. Brown Cluster features generated from the entire TUH EEG corpus were used in both feature vector representations listed in Table 1. Brown clustering is an unsupervised learning method that discovers hierarchical clusters of words based on their contexts. We also used in the feature vector representation medical knowledge available from the Unified Medical Language System (UMLS).

EEG reports mention multiple medical concepts in the narratives used in each report section. To find the spans of text that correspond to EEG activities and should receive HAD tags, we trained a stacked Long Short-Term Memory (LSTM) network for detecting EEG Activity anchors. The stacked LSTM network processes each document at the sentence level. To do this, we represented each sentence as a sequence of tokens [w_1, w_2,...,

w_N], and train both LSTMs to assign a label $b_i \in$ { *"I", "O", "B"*} to each token w_i such that it will receive a label b_i=*"B"* if the token w_i is at the beginning of a mention of a medical concept, a label b_i=*"I"* if the token w_i is inside any mention of a medical concept and a label b_i=*"O"* if the token w_i is outside any mention of a medical concept.

For example, the token sequence *left anterior temporal [sharp and complexes]*$_{ACT}$" would correspond sequence [*O,O, O,O,B,I,I,I*], tokens {*occasional, left, anterior,* are all assigned labels of *O*, as part of the anchor of an EEG although they describe its token {*sharp*} is assigned a label tokens {*and, slow, wave,* are all assigned labels of *I*. This allows medical concept mentions identified by continuous tokens starting with a token optionally followed tokens labeled *I*.

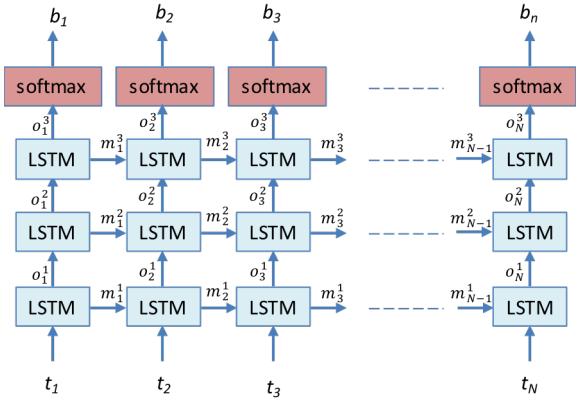

Figure 4: Deep Learning architecture for the identification of the EEG activity anchors .

"*occasional slow wave* to the label where *temporal*} they are not activity, attributes, of *B*, and the *complexes*} *IOB* notation to be sequences of labeled *B*

To be able to use a deep learning architecture for automatically identifying the anchors of EEG activities we first tokenized all reports, and represented each token w_i as a feature vector, t_i obtained by considering the features illustrated in Table 1. As illustrated in Figure 2, the features vectors t_1, t_2, …, t_N are provided as input to the stacked LSTMs to predict a sequence of output labels, b_1, b_2, …, b_N. To predict each label b_i, the deep learning architecture considers (1) the vector representation of each token, t_i; as well as (2) the vector representation of all previous tokens from the sentence by updating a *memory* state that is shared throughout the network. LSTM cells also have the property that they can be "stacked" such that the outputs of cells on level l are used as the inputs to the cells on level on level $l+1$. We used a stacked LSTM with 3 levels where the input to the first level is a sequence of token vectors and the output from the top level is used to determine the IOB labels for each token. The output from the top level, o_i^3, is a vector representing token w_i and every previous token in the sentence. To determine the IOB label for token w_i, the output o_i^3 is passed through a softmax layer. The softmax layer produces a probability distribution over all IOB labels. This is accomplished by computing a vector of probabilities, q_i such that $q_{i,1}$ is the probability of label "I", $q_{i,2}$ is the probability of label "O", and $q_{i,3}$ is the probability of label "B". The predicted IOB label is then chosen as the label with highest probability, $y_i = \underset{j}{\operatorname{argmax}}\, q_{ij}$. We used the architecture illustrated in Figure 4 to decide where HAD tags can be placed. In order to decide which tags should be selected, we considered 16 possible attributes for EEG Activities as well as polarity and modality, and type, modality, and polarity. Traditionally, attribute classification is performed by training a classifier, such as an SVM, to determine the value for each attribute. This approach would require training 18 separate attribute classifiers for EEG Activities. However, by leveraging the power of deep learning, we could simplify this task by creating one multi-purpose, high-dimensional vector representation of an EEG activity, or *embedding*, and use this representation to determine each attribute simultaneously with the same deep learning network. Using a shared embedding allows important information to be shared between individual tasks. To accomplish this, we use the Deep Rectified Linear Network (DRLN) for multi-task attribute detection, illustrated in Figure 5.

Training and Professional Development

Temple University Postdoc: Our postdoc arrived in late March 2016. Over the first few months of the project, he struggled with computer programming and computational issues. We provided remedial training on C++ and Python programming, and provided instruction on Linux clusters. In Fall 2016, he sat in on PI Picone's Software Tools for Engineers course (*https://www.isip.piconepress.com/courses/temple/ece_3822/*), in which he was introduced to many aspects of our computing environment.

We also emphasize communication skills. We supported him in presenting his work at the 2016 IEEE Signal Processing in Medicine and Biology Symposium, which we host at Temple University. He learned how to write a standard conference paper and how to present in this type of technical forum.

He was trained how to annotate EEG data for seizures to increase his understanding of the problem space. He worked closely with our annotation team and received instruction from our senior student working on this research. He is now able to annotate seizures in signals with a reasonable degree of accuracy.

Temple University Graduate Students: Graduate student training consisted mainly of programming skill development. We are making heavy use of Python and C++ in this project. We have defined a methodology for developing, documenting and releasing C++, Python and MATLAB code, and our graduate students have been trained on this. They were mentored by the PIs.

We have also adopted the Keras toolkit as a high-level interface to Theano. This has greatly reduced the time it takes to prototype new deep learning systems.

Temple University Undergraduate Students: We have hired several undergraduate students to provide IT support for our team members. These students administer the software and hardware production computing environment. We have trained several of these students since they had no prior system administration experience for Linux clusters. We have also trained a student to actively maintain our project web sites using tools such as Drupal. These types of state of the art computational skills make these students very attractive to future employers. We expect several of these students to continue in our MS program.

University of Texas at Dallas Graduate Students: Three PhD students have been advised for their research conducted for this project at University of Texas at Dallas.

Travis Goodwin is a 5th year PhD student in Computer Science at UTD who has developed novel research in the area of multi-modal indexing, inference of underspecified information in the EEG reports and interaction of various factors in the EEG reports. Travis has also been working on defining the HAD tags under the supplement project. He has also worked on using deep learning methods for the automatic annotation of HAD tags as well for generating data-driven neural knowledge representations of the knowledge discerned from the EEG reports. Travis has successfully submitted 10 conference papers and has received this year the best student paper award at the IEEE CIKM Conference, a major conference on knowledge managements and information retrieval. In addition, he has submitted successfully 5 papers to at *the American Medical Informatics Association Joint Summits on Clinical Research Informatics (AMIA-CRI)* or *the American Medical Informatics Association Annual Symposium (AMIA)* as well as 2 journal papers. These accomplishments meet Travis's Individual Development Plans

(IDPs).

Ramon Maldonado is a 2nd year PhD student in Computer Science at UTD who has performed research on automatically identifying all medical concepts from the Temple University Hospital EEG data, in the forms of EEG reports documenting 25,000 sessions and 15,000 patients collected over 12 years at Temple University Hospital. During the past year, Ramon became a qualified PhD student, by passing a set of qualifying exams, while developing new techniques for his research, which is remarkable. He is has submitted a paper which was accepted and is presented at *the American Medical Informatics Association Joint Summits on Clinical Research Informatics (AMIA-CRI)*, San Francisco, CA. Ramon has also recently submitted a paper to *the American Medical Informatics Association Annual Symposium (AMIA)* as well as a journal paper to the *Journal of Biomedical Informatics.* Ramon plans to submit a second paper to the *Journal of Biomedical Informatics* soon, based on his work in the SEMEVAL 2016 evaluation on clinical narratives describing events, event containers as well as temporal relations. These accomplishments meet Ramon's Individual Development Plans (IDPs).

Stuart Taylor is a 1st year PhD student in Computer Science at UTD who has worked on the initial development of the active deep learning for annotating EEG reports. He also worked on the generation of queries for the evaluation of patient cohorts. Stuart has submitted a poster to the to the *American Medical Informatics Association Annual Symposium (AMIA)* in 2017 and has plan on working on the recognition of HAD tags in biomedical texts. These accomplishments meet Stuart's Individual Development Plans (IDPs).

General Training: In Spring 2017, PI Picone offered a three credit hour independent study course titled "Information Theory" (ECE 8526: *https://www.isip.piconepress.com/courses/temple/ece_8526/*). This course was attended by several of students who are either currently contributing to the project or will be joining the team in the coming year.

C. OVERALL PRODUCTS

C.1 PUBLICATIONS

Are there publications or manuscripts accepted for publication in a journal or other publication (e.g., book, one-time publication, monograph) during the reporting period resulting directly from this award?

Yes

Publications Reported for this Reporting Period

Public Access Compliance	Citation
Complete	Goodwin T, Harabagiu SM. A Probabilistic Reasoning Method for Predicting the Progression of Clinical Findings from Electronic Medical Records. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science. 2015;2015:61-5. PubMed PMID: 26306238; PubMed Central PMCID: PMC4525214.
Complete	Harati A, Golmohammadi M, Lopez S, Obeid I, Picone J. Improved EEG Event Classification Using Differential Energy. ... IEEE Signal Processing in Medicine and Biology Symposium. IEEE Signal Processing in Medicine and Biology Symposium. 2015 December;2015. PubMed PMID: 27213180; PubMed Central PMCID: PMC4874511.
Complete	López S, Suarez G, Jungreis D, Obeid I, Picone J. Automated Identification of Abnormal Adult EEGs. ... IEEE Signal Processing in Medicine and Biology Symposium. IEEE Signal Processing in Medicine and Biology Symposium. 2015 December;2015. PubMed PMID: 27195311; PubMed Central PMCID: PMC4868184.
Complete	Goodwin T, Harabagiu SM. Inferring the Interactions of Risk Factors from EHRs. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science. 2016;2016:78-87. PubMed PMID: 27595044; PubMed Central PMCID: PMC5001781.
Complete	Goodwin TR, Harabagiu SM. Multi-modal Patient Cohort Identification from EEG Report and Signal Data. AMIA ... Annual Symposium proceedings. AMIA Symposium. 2016;2016:1794-1803. PubMed PMID: 28269938; PubMed Central PMCID: PMC5333290.
In Process at NIHMS	Goodwin T, Harabagiu S. Inferring the Interactions of Risk Factors from EHRs. Proceedings of the 2016 American Medical Informatics Association (AMIA) Summit on Clinical Research Informatics (CRI). 2016 March;:78-87.
In Process at NIHMS	Goodwin T, Harabagiu S. Embedding Open-domain Common-Sense Knowledge from Text. Proceedings of the Language Resources and Evaluation Conference (LREC-2016). 2016 May.
In Process at NIHMS	Goodwin T, Harabagiu S. Medical Question Answering for Clinical Decision Support. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016 October 01.
In Process at NIHMS	Goodwin T, Harabagiu S. Multi-Modal Patient Cohort Identification from EEG Report and Signal Data. Proceedings of the American Medical Informatics Association Annual Symposium (AMIA). 2016 November.
N/A: Not Journal	Picone J, Obeid Iyad. Fundamentals in Data Science: Data Wrangling, Normalization, Preprocessing of Physiological Signals. The BD2K Guide Lecture Series; 2016 November 18; Bethesda, Maryland, USA.
N/A: Not Journal	Harabagiu S. Fundamentals in Data Science: Active Deep Learning-Based Annotation of Electroencephalography Reports for Patient Cohort Identification. The BD2K Guide Lecture Series; 2016 December 02; Bethesda, Maryland, USA.
Non-Compliant	Lopez S, Gross A, Yang S, Golmohammadi M, Obeid Iyad, Picone J. An Analysis of Two Common Reference Points for EEGs. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium. 2016 December 03;N/A(N/A):N/A.
In Process at NIHMS	Yang S, Lopez S, Golmohammadi M, Obeid Iyad, Picone J. Semi-automated Annotation of Signal Events In Clinical EEG Data. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium. 2016 December 03;N/A(N/A):N/A.

N/A: Not Journal	Obeid Iyad, Picone J. Biomedical Signal Processing in Big Data. 1 ed. Sejdik E, Falk T, editors. Boca Raton, Florida, USA. CRC Press; 2017. Chapter N/A, Machine Learning Approaches to Automatic Interpretation of EEGs; N/Ap. N/Ap.
PMC Journal - In process	Goodwin T, Harabagiu S. Read, Decide and Explain! Recollective (Explanatory) Question Answering. Annual Conference of the Association of Computational Linguistics (ACL). 2017 February.
PMC Journal - In process	Goodwin T, Harabagiu S. Deep Learning from EEG Reports for Inferring Underspecified Information. Proceedings of the American Medical Informatics Association Joint Summits on Clinical Research Informatics (AMIA-CRI). 2017 March.
PMC Journal - In process	Goodwin T, Harabgiu S. Inferring Clinical Correlations from EEG Reports with Deep Neural Learning. American Medical Informatics Association Annual Symposium (AMIA). 2017 March.
PMC Journal - In process	Goodwin T, Maldonado R, Harabagiu S. Automatic Recognition of Symptom Severity from Psychiatric Evaluation Reports. Journal of Biomedical Informatics. 2017 March.
PMC Journal - In process	Goodwin T, Harabagiu S. Knowledge Representations and Inference Techniques for Medical Question Answering. ACM Transactions on Intelligent Systems and Technology. 2017 March.
PMC Journal - In process	Maldano R, Goodwin T, Harabagiu S. Active Deep Learning-Based Annotation of Electroencephalography Reports for Cohort Identification. Proceedings of the American Medical Informatics Association Joint Summits on Clinical Research Informatics (AMIA-CRI). 2017 March 01.
PMC Journal - In process	Maldonado R, Goodwin T, Skinner M, Harabagiu S. Deep Learning Meets Biomedical Ontologies: Knowledge Embeddings for Epilepsy. American Medical Informatics Association Annual Symposium (AMIA). 2017 March.
PMC Journal - In process	Taylor S, Goodwin T, Harabagiu S. An Evaluation of Syntactic Dependency Parsers on Clinical Data. American Medical Informatics Association Annual Symposium (AMIA). 2017 March.
N/A: Not Journal	Golmohammadi M, Shah V, Lopez S, Ziyabari S, Camaratta J, Obeid Iyad, Picone J. The TUH EEG Seizure Corpus. Annual Meeting of the American Clinical Neurophysiology Society; 2017 February 08; Phoenix, Arizona, USA.
N/A: Not Journal	Golmohammadi M, Ziyabari S, Lopez S, Krome E, Thiess M, Obeid Iyad, Picone J, Yang S. EEG Event Detection Using Deep Learning. Big Data to Knowledge All Hands Grantee Meeting; 2016 November 29; Bethesda, Maryland, USA.
N/A: Not Journal	Harabagiu S, Goodwin T, Maldonado R, Taylor S. Active Deep Learning-Based Annotation of Electroencephalography Reports for Cohort Identification. Big Data to Knowledge All Hands Grantee Meeting; 2016 November 29; Bethesda, Maryland, USA.
N/A: Not Journal	Harabagiu S, Goodwin T. Deep Learning-Based Multi-Modal Indexing of Heterogeneous Clinical Data for Patient Cohort Retrieval. Big Data to Knowledge All Hands Grantee Meeting; 2016 November 29; Bethesda, Maryland, USA.
N/A: Not Journal	Picone J, Obeid Iyad, Harabagiu S. Automatic Discovery and Processing of EEG Cohorts from Clinical Records. Big Data to Knowledge All Hands Grantee Meeting; 2016 November 29; Bethesda, Maryland, USA.
N/A: Not Journal	Picone J, Obeid Iyad, Harabagiu S. Scalable EEG interpretation using Deep Learning and Schema Descriptors. Big Data to Knowledge All Hands Grantee Meeting; 2016 November 29; Bethesda, Maryland, USA.
N/A: Not Journal	Somaru Pat, Obeid Iyad, Picone J. Low-Cost High-Performance Computing Via Consumer GPUs. Picone J, editor. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium. IEEE Signal Processing in Medicine and Biology Symposium; 2016 December 03; Philadelphia, Pennsylvania, United States.
N/A: Not Journal	Thiess M, Krome E, Golmohammadi M, Obeid Iyad, Picone J. Enhanced Visualizations for Improved Real-Time EEG Monitoring. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium. IEEE Signal Processing in Medicine and Biology Symposium; 2016 December 03; Philadelphia, Pennsylvania, USA.

In Process at NIHMS	Active Deep Learning-Based Annotation of Electroencephalography Reports for Cohort Identification. Proceedings of the American Medical Informatics Association Joint Summits on Clinical Research Informatics (AMIA-CRI).

C.2 WEBSITE(S) OR OTHER INTERNET SITE(S)

Category	Explanation
Research Material	https://www.isip.piconepress.com/projects/nih_cohort: This is the URL from which we distribute project information and related resources.
Data or Databases	https://www.isip.piconepress.com/projects/tuh_eeg: This is the URL from which we distribute data and resources related to the TUH EEG Corpus.

C.3 TECHNOLOGIES OR TECHNIQUES

Category	Explanation
Software	AutoEEG Demonstration System: A visualization tool that includes annotation of EEG signals. This tool has been used to generate the data described in the report.
Software	MERCuRY (Multi-modal EncephalogRam patient Cohort discoveRY): A demonstration of our cohort retrieval system.

C.4 INVENTIONS, PATENT APPLICATIONS, AND/OR LICENSES

Have inventions, patent applications and/or licenses resulted from the award during the reporting period?

No

C.5 OTHER PRODUCTS AND RESOURCE SHARING

NOTHING TO REPORT

D. OVERALL PARTICIPANTS

D.1 WHAT INDIVIDUALS HAVE WORKED ON THE PROJECT?

Commons ID	S/K	Name	Degree(s)	Role	Cal	Aca	Sum	Foreign Org	Country	SS
JOSCONE	Y	Picone, Joseph	MS,PHD	PD/PI	0	1	2			NA
SHARABAGIU	Y	Harabagiu, Sanda Maria	PHD	PD/PI	0	3	3			NA
OBEID07	Y	Obeid, Iyad	BS,MS,PHD	PD/PI	0	0	1			NA
RAMONMALDONADO	N	Maldonado, Ramon	BS	Graduate Student (research assistant)	12	0	0			NA
GOLMOHAMMADI	N	Golmohammadi, Meysam	MS	Graduate Student (research assistant)	12	0	0			NA
DAWSHED	N	Jamshed, Dawer	PhD	Graduate Student (research assistant)	5	0	0			NA
AILOPEZ	N	Lopez, Silvia	BS	Graduate Student (research assistant)	12	0	0			NA
CHRBELL	N	Campbell, Chris	HS	Undergraduate Student	1	0	0			NA
STEWONG	N	Wong, Steven	HS	Undergraduate Student	1	0	0			NA
PATSOMARU	N	Somaru, Pat	HS	Undergraduate Student	1	0	0			NA
NICECCA	N	Mecca, Nicholas	HS	Undergraduate Student	1	0	0			NA
MATIESS	N	Thiess, Matthew	HS	Undergraduate Student	1	0	0			NA
JASRGEY	N	Bergey, Jason	HS	Undergraduate Student	1	0	0			NA
JAMHUGH	N	McHugh, James	HS	Undergraduate Student	1	0	0			NA
EVALTIN	N	von Weltin, Eva	HS	Undergraduate Student	1	0	0			NA
ELLROME	N	Krome, Elliott	HS	Undergraduate Student	1	0	0			NA
DTREJO	N	Trejo, Devin	HS	Undergraduate Student	1	0	0			NA
STUARTTAYLOR	N	Taylor, Stuart	BS	Graduate Student (research assistant)	12	0	0			NA

TRAVISGOODWIN	N	Goodwin, Travis	MS	Graduate Student (research assistant)	12	0	0			NA
SCOTTYANG	N	Yang, Su	PHD	Postdoctoral Scholar, Fellow, or Other Postdoctoral Position	12	0	0			NA
TAMHSAN	N	Ahsan, Tameem	HS	Undergraduate Student	1	0	0			NA

Glossary of acronyms:
S/K - Senior/Key
DOB - Date of Birth
Cal - Person Months (Calendar)
Aca - Person Months (Academic)
Sum - Person Months (Summer)
Foreign Org - Foreign Organization Affiliation
SS - Supplement Support
RE - Reentry Supplement
DI - Diversity Supplement
OT - Other
NA - Not Applicable

D.2 PERSONNEL UPDATES

D.2.a Level of Effort

Will there be, in the next budget period, either (1) a reduction of 25% or more in the level of effort from what was approved by the agency for the PD/PI(s) or other senior/key personnel designated in the Notice of Award, or (2) a reduction in the level of effort below the minimum amount of effort required by the Notice of Award?

Yes

Unexpected staffing changes and a late start with postdoc staffing, along with additional external funding awards, have resulted in a proposal to adjust staffing in the third year of the project. We have also employed more undergraduates for data preparation since we were not able to utilize as much neurologist time on preparing data. This is described in more detail in the report.

D.2.b New Senior/Key Personnel

Are there, or will there be, new senior/key personnel?

No

D.2.c Changes in Other Support

Has there been a change in the active other support of senior/key personnel since the last reporting period?

No

D.2.d New Other Significant Contributors

Are there, or will there be, new other significant contributors?

No

D.2.e Multi-PI (MPI) Leadership Plan

Will there be a change in the MPI Leadership Plan for the next budget period?

No

E. OVERALL IMPACT

E.1 WHAT IS THE IMPACT ON THE DEVELOPMENT OF HUMAN RESOURCES?

Not Applicable

E.2 WHAT IS THE IMPACT ON PHYSICAL, INSTITUTIONAL, OR INFORMATION RESOURCES THAT FORM INFRASTRUCTURE?

There are two types of infrastructure impact this project has: computational and archival. The computing cluster we have developed allows our students to run significant computer simulations at an overall system cost much lower than available cloud-based services. Its cost-effective approaches to disk space (an order of magnitude cheaper than commercial storage area network solutions) and compute capacity (using commodity GPUs in bulk) have enabled us to run very large-scale simulations on big data. We have been able to demonstrate this cluster to our engineering students in an undergraduate class that PI Picone teaches (ECE 3822: Software Tools for Engineers), so the cluster is also have educational impact. For many of our students this is their first exposure to clustered computing.

In terms of archival, the database of clinical EEGs we are building has become the primary archive for hospital clinicians. They routinely consult us when they want to do an historical search on their patients, or need to locate an older EEG that has been purged from their local cache. The database itself continues to grow and evolve, and now includes annotations of subsets of the data for seizure detection and normal/abnormal classification. This continues to make TUH EEG the premier resource for machine learning research on clinical EEG data.

E.3 WHAT IS THE IMPACT ON TECHNOLOGY TRANSFER?

Not Applicable

E.4 WHAT DOLLAR AMOUNT OF THE AWARD'S BUDGET IS BEING SPENT IN FOREIGN COUNTRY(IES)?

NOTHING TO REPORT

F. OVERALL CHANGES

F.1 CHANGES IN APPROACH AND REASONS FOR CHANGE

Not Applicable

F.2 ACTUAL OR ANTICIPATED CHALLENGES OR DELAYS AND ACTIONS OR PLANS TO RESOLVE THEM

Over the past year we have solicited about 100 neurologists to find consultants willing to mark up data. Approximately 20 of these agreed to annotate data. Unfortunately, they have not provided results in a timely fashion. They have been responsive on analyzing queries and usability engineering issues, but have not been responsive on annotating EEG signal data. Therefore, we shifted our strategy to do two things. First, we are implementing a pseudo-crowd sourcing strategy using an FAQ-type approach. We will use this to solicit interest from an international pool of neurologists and to stimulate discussion on particular issues (e.g., short seizures). Second, we have trained a team of undergraduates to annotate signals. Their performance is very competitive with our neurologists. In fact they have demonstrated an ability to annotate signals in a much more detailed manner than neurologists typically do. This has allowed us to generate data required for our machine learning systems. We continue to attempt to use neurologists for query reviews and annotation in an effort to certify our data and conduct inter-rater agreement studies.

F.3 SIGNIFICANT CHANGES TO HUMAN SUBJECTS, VERTEBRATE ANIMALS, BIOHAZARDS, AND/OR SELECT AGENTS

F.3.a Human Subjects

No Change

F.3.b Vertebrate Animals

No Change

F.3.c Biohazards

No Change

F.3.d Select Agents

No Change

G. OVERALL SPECIAL REPORTING REQUIREMENTS

G.1 SPECIAL NOTICE OF AWARD TERMS AND FUNDING OPPORTUNITIES ANNOUNCEMENT REPORTING REQUIREMENTS

NOTHING TO REPORT

G.2 RESPONSIBLE CONDUCT OF RESEARCH

Not Applicable

G.3 MENTOR'S REPORT OR SPONSOR COMMENTS

Not Applicable

G.4 HUMAN SUBJECTS

G.4.a Does the project involve human subjects?

No

G.4.b Inclusion Enrollment Data

Not Applicable

G.4.c ClinicalTrials.gov

Does this project include one or more applicable clinical trials that must be registered in ClinicalTrials.gov under FDAAA?

G.5 HUMAN SUBJECTS EDUCATION REQUIREMENT

Are there personnel on this project who are newly involved in the design or conduct of human subjects research?

G.6 HUMAN EMBRYONIC STEM CELLS (HESCS)

Does this project involve human embryonic stem cells (only hESC lines listed as approved in the NIH Registry may be used in NIH funded research)?

No

G.7 VERTEBRATE ANIMALS

Does this project involve vertebrate animals?

No

G.8 PROJECT/PERFORMANCE SITES

Organization Name:	DUNS	Congressional District	Address
Primary: Temple University - Of The Commonwealth System of	057123192	PA-001	TEMPLE UNIV OF THE COMMONWEALTH 1947 N. 12th Street Philadelphia PA 191226099
The University of Texas at Dallas	800188161	TX-032	Office of Sponsored Projects, AD15 800 West Campbell Road Richardson TX 75080
TEMPLE UNIVERSITY	057123192		TEMPLE UNIVERSITY 1801 N Broad Street, 401 Conwell Hall PHILADELPHIA PA 191226003

Temple University - Of The Commonwealth System of	057123192	PA-001	TEMPLE UNIV OF THE COMMONWEALTH 1947 N. 12th Street Philadelphia PA 191226099
The University of Texas at Dallas	800188161	TX-032	Office of Sponsored Projects, AD15 800 West Campbell Road Richardson TX 75080
TEMPLE UNIVERSITY	057123192		TEMPLE UNIVERSITY 1801 N Broad Street, 401 Conwell Hall PHILADELPHIA PA 191226003
Temple University - Of The Commonwealth System of	057123192	PA-001	TEMPLE UNIV OF THE COMMONWEALTH 1947 N. 12th Street Philadelphia PA 191226099
The University of Texas at Dallas	800188161	TX-032	Office of Sponsored Projects, AD15 800 West Campbell Road Richardson TX 75080
TEMPLE UNIVERSITY	057123192		TEMPLE UNIVERSITY 1801 N Broad Street, 401 Conwell Hall PHILADELPHIA PA 191226003
Temple University - Of The Commonwealth System of	057123192	PA-001	TEMPLE UNIV OF THE COMMONWEALTH 1947 N. 12th Street Philadelphia PA 191226099
The University of Texas at Dallas	800188161	TX-032	Office of Sponsored Projects, AD15 800 West Campbell Road Richardson TX 75080

G.9 FOREIGN COMPONENT

No foreign component

G.10 ESTIMATED UNOBLIGATED BALANCE

G.10.a Is it anticipated that an estimated unobligated balance (including prior year carryover) will be greater than 25% of the current year's total approved budget?

Yes

Estimated unobligated balance: 601373

G.10.b Provide an explanation for unobligated balance:
We began the project approximately half a year behind schedule due to unexpected staffing issues that were described in the previous annual report. To offset lower than expected productivity from neurologists, we have employed undergraduate annotators, as described in the accomplishments section.

The supplemental award was awarded in October 2016 - the middle of the semester. This made it difficult to staff the project until January 1, 2017. Also, delays in invoicing contribute to the increased amount of unspent funds. However, approximately 50% of the supplemental funds have been obligated and will be spent by 5/31/2017. The remaining funds should be spent by December 31, 2017.

G.10.c If authorized to carryover the balance, provide a general description of how it is anticipated that the funds will be spent
Regarding the supplement, we are on track to expend the funds over roughly a one-year period starting January 1, 2017 and ending December 31, 2017. Because the award arrived in the middle of the Fall 2016 semester, we had to delay staffing the project until January 1.

Regarding the main award, we have made two major changes to the budget for FY3 to accelerate progress on the project. First, starting in FY2, we realized that it was going to be difficult to engage neurologists for data annotation. Therefore, starting in FY2, we have trained a group of undergraduates to do data annotation. They have been highly productive and are much less expensive. This has allowed us to increase the amount of data generated. We benchmarked these students against data that was annotated by a small number of neurologists and demonstrated that their accuracy was comparable or better than the experts. Therefore, we have relied on this mechanism to generate data.

We also plan to staff the project in FY3 with additional grad students who are fully trained on the technology and available because a concurrent related funded project (NSF STTR Phase 1) is coming to a close. These students will bring additional research on deep

learning techniques to our project to improve event detection performance.

G.11 PROGRAM INCOME

Is program income anticipated during the next budget period?

No

G.12 F&A COSTS

Not Applicable

RESEARCH & RELATED BUDGET - SECTION A & B

ORGANIZATIONAL DUNS*: 057123192

Budget Type*: ● Project ❍ Subaward/Consortium

Enter name of Organization: TEMPLE UNIV OF THE COMMONWEALTH

Start Date*: 06-01-2017 **End Date*:** 05-31-2018

A. Senior/Key Person

	Prefix	First Name*	Middle Name	Last Name*	Suffix	Project Role*	Base Salary ($)	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits ($)*	Funds Requested ($)*
1.	Dr	Iyad		Obeid		Project Co-PI	112,716.00		0.75	1.0	21,917.00	3,685.00	25,602.00
2.	Dr	Joseph		Picone		Project Lead	136,125.00		0.75	1.0	26,468.00	4,451.00	30,919.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name:

Total Senior/Key Person **56,521.00**

B. Other Personnel

Number of Personnel*	Project Role*	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits*	Funds Requested ($)*
1	Post Doctoral Associates	9.65			39,669.00	11,226.00	50,895.00
1	Graduate Students	12.0			22,472.00	3,798.00	26,270.00
1	Undergraduate Students	6.0			10,084.00	257.00	10,341.00
	Secretarial/Clerical						
3	**Total Number Other Personnel**					**Total Other Personnel**	**87,506.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**144,027.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E

ORGANIZATIONAL DUNS*: 057123192

Budget Type*: ● Project ❍ Subaward/Consortium

Enter name of Organization: TEMPLE UNIV OF THE COMMONWEALTH

Start Date*: 06-01-2017 **End Date*:** 05-31-2018

C. Equipment Description	
List items and dollar amount for each item exceeding $5,000	
Equipment Item	**Funds Requested ($)***
Total funds requested for all equipment listed in the attached file	**0.00**
Total Equipment	**0.00**

Additional Equipment: File Name:

D. Travel	**Funds Requested ($)***
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	16,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**16,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)***
1. Tuition/Fees/Health Insurance		0.00
2. Stipends		0.00
3. Travel		0.00
4. Subsistence		0.00
5. Other:		
0 Number of Participants/Trainees	**Total Participant Trainee Support Costs**	**0.00**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K

ORGANIZATIONAL DUNS*: 057123192

Budget Type*: ● Project ❍ Subaward/Consortium

Enter name of Organization: TEMPLE UNIV OF THE COMMONWEALTH

Start Date*: 06-01-2017 **End Date*:** 05-31-2018

F. Other Direct Costs	Funds Requested ($)*
1. Materials and Supplies	1,750.00
2. Publication Costs	0.00
3. Consultant Services	25,000.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	93,905.00
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
Total Other Direct Costs	**120,655.00**

G. Direct Costs	Funds Requested ($)*
Total Direct Costs (A thru F)	**280,682.00**

H. Indirect Costs

Indirect Cost Type	Indirect Cost Rate (%)	Indirect Cost Base ($)	Funds Requested ($)*
1. F&A	56.0	186,777.00	104,595.00
2. F&A	53.0	93,905.00	49,770.00
		Total Indirect Costs	**154,365.00**

Cognizant Federal Agency

(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	Funds Requested ($)*
Total Direct and Indirect Institutional Costs (G + H)	**435,047.00**

J. Fee	Funds Requested ($)*
	0.00

K. Budget Justification*	File Name: budget_justification_v03.pdf
	(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

Budget Justification

Our budget request for the third year of this project includes the amount originally awarded for the third year ($435,047).

Budget Justification (Temple University)

Dr. Iyad Obeid, Ph.D. (Temple University, co-PI): Dr. Obeid will spend one summer month and 0.75 months in the academic year on the primary award. His duties will include supervising the signal processing components of the EEG event detection project, and annotation of the EEG event data. He will also supervise the development and dissemination of the annotated version of the TUH EEG Corpus and the assessment phases of the project. Dr. Obeid is an Associate Professor of Electrical and Computer Engineering in the College of Engineering, and is jointly appointed to the Department of Bioengineering. He is also Director of the Neural Engineering Data Consortium. His expertise is in brain machine interfaces, including both hardware and software. He provides extensive knowledge on EEG technology as well.

Joseph Picone, Ph.D. (Temple University, Principal Investigator): Dr. Picone will spend one summer month and 0.75 months in the academic year on the primary award. His duties will include serving as the contact P.I. and providing overall coordination of the project. He will also direct the machine learning aspects of both projects. Dr. Picone is a Professor of Electrical and Computer Engineering in the College of Engineering. He is also co-Director of the Neural Engineering Data Consortium. His expertise is in machine learning specifically in signal and image processing applications such as speech recognition.

Postdoctoral Researcher: We have hired one postdoctoral researcher for the primary project. This person is responsible for managing annotations of the EEG signal data, the EEG reports, and all related information developed by the annotators. We expect this postdoc to transition to a new project in February 2018.

Graduate Students: We have allocated one graduate student to the project in the final year. This student will continue conducting research into improved deep learning methods.

Undergraduate Students: We have allocated funds for one undergraduate on this project. In the third year of this project this student will share time between publishing project-related information to the web and generating annotated data.

Travel: Participation in professional conferences, particularly for students, is an important part of the development process. A typical conference trip for a full-time staff person (PI or postdoc) averages about $2K, and approximately $1.5K for a graduate student or undergraduate student. In the third year, we have allocated $16K for travel for the primary project and $5K for the supplemental grant. These trips will include major neuroengineering conferences such as the IEEE EMBS Conference on Neural Engineering and first-tier machine learning conferences such as the Neural Information Processing Systems (NIPS) conference. We also plan to encourage our students to present at smaller regional conferences such as the IEEE Signal Processing in Medicine and Biology Symposium (which we host). We plan to present papers at these conferences as part of our participation. We have also allocated funds to attend the All-Hands meeting in Fall 2017. We will send the two PIs and four graduate students to this conference, and will be demonstrating a functional cohort retrieval system.

Commodities: The commodities budget includes $1750 for publications-related expenses, conference-related promotional materials, and incidental computer maintenance costs.

Consultants: We have reserved $25K for costs associated with neurologists serving as consultants. These neurologists will review queries and provide feedback on the relevance of the returned results, and overall user interface issues. We compensate them at $85/hr.

Additional Notes:

The on-campus research rate effective for the duration of this project is 56%. The fringe benefit rates used are as follows, also set at the beginning of the three-year project:

Category	Academic Year	Summer
Faculty	28.30%	28.30%
Postdoc	28.30%	28.30%
Graduate Students	16.90%	16.90%
Undergraduate Students (Summer)	8.20%	8.20%

RESEARCH & RELATED BUDGET - SECTION A & B

ORGANIZATIONAL DUNS*: 8001881610000

Budget Type*: ❍ Project ● Subaward/Consortium

Enter name of Organization: University of Texas at Dallas

Start Date*: 06-01-2017 **End Date*:** 05-31-2018

A. Senior/Key Person

	Prefix	First Name*	Middle Name	Last Name*	Suffix	Project Role*	Base Salary ($)	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits ($)*	Funds Requested ($)*
1.	Dr	Sanda		Harabagiu		Subcontractor: Natural Language Processing Expert	180,828.00		0.0	3.0	45,207.00	9,041.00	54,248.00

Total Funds Requested for all Senior Key Persons in the attached file

Additional Senior Key Persons: File Name:

Total Senior/Key Person **54,248.00**

B. Other Personnel

Number of Personnel*	Project Role*	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits*	Funds Requested ($)*
	Post Doctoral Associates						
2	Graduate Students	16.8			31,466.00	4,720.00	36,186.00
	Undergraduate Students						
	Secretarial/Clerical						
2	**Total Number Other Personnel**					**Total Other Personnel**	**36,186.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**90,434.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTION C, D, & E

ORGANIZATIONAL DUNS*: 8001881610000

Budget Type*: ❍ Project ● Subaward/Consortium

Enter name of Organization: University of Texas at Dallas

Start Date*: 06-01-2017 **End Date*:** 05-31-2018

C. Equipment Description	
List items and dollar amount for each item exceeding $5,000	
Equipment Item	**Funds Requested ($)***
Total funds requested for all equipment listed in the attached file	**0.00**
Total Equipment	**0.00**

Additional Equipment: File Name:

D. Travel	**Funds Requested ($)***
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	2,750.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**2,750.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)***
1. Tuition/Fees/Health Insurance		0.00
2. Stipends		0.00
3. Travel		0.00
4. Subsistence		0.00
5. Other:		
0 Number of Participants/Trainees	**Total Participant Trainee Support Costs**	**0.00**

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K

ORGANIZATIONAL DUNS*: 8001881610000

Budget Type*: ❍ Project ● Subaward/Consortium

Enter name of Organization: University of Texas at Dallas

Start Date*: 06-01-2017 **End Date*:** 05-31-2018

F. Other Direct Costs	Funds Requested ($)*
1. Materials and Supplies	721.00
2. Publication Costs	0.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	0.00
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
Total Other Direct Costs	**721.00**

G. Direct Costs	Funds Requested ($)*
Total Direct Costs (A thru F)	**93,905.00**

H. Indirect Costs

Indirect Cost Type	Indirect Cost Rate (%)	Indirect Cost Base ($)	Funds Requested ($)*
1. F&A	53.0	93,905.00	49,770.00
		Total Indirect Costs	**49,770.00**

Cognizant Federal Agency
(Agency Name, POC Name, and POC Phone Number)

I. Total Direct and Indirect Costs	Funds Requested ($)*
Total Direct and Indirect Institutional Costs (G + H)	**143,675.00**

J. Fee	Funds Requested ($)*
	0.00

K. Budget Justification*	File Name:
	11_budget_justification_v00_utd.pdf
	(Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

Salaries and Wages: Salaries for all personnel are based upon current University of Texas at Dallas academic and staff salary scales. Prof. Sanda Harabagiu, Principal Investigator, will provide the overall direction and management of the Aim 3: Defining Hierarchical epileptiform Activity Descriptors (HAD) for EEGs, and Aim 4: Automated Tagging of HADs in medical texts of the project entitled "Scalable EEG Interpretation Using Deep Learning and Schema Descriptors" as well as participating in the technical research and publication activities. Prof. Sanda Harabagiu, will collaborate closely with Professors Joseph Picone and Iyad Obeid from Temple University as well as the physicians from Temple Hospital involved in the project. She will work closely with Prof. Picone and Obeid to design the Hierarchical epileptiform Activity Descriptors developed for Aim 3 as well as on the automatic tagging of the TAG descriptors on the EEG reports from Temple University in her work. In addition, she will collaborate on the evaluation of the patient cohort retrieval system that uses the TAG meta-data as well as the evaluation of the results of the automatic retrieval of relevant medical articles that will inform the scalable interpretation of EEGs. Prof. Harabagiu will also supervise the research of two graduate PhD students. This research will form the core of two PhD dissertations. We request summer salary support for Prof. Sanda Harabagiu at a level of 3 months for the duration of the project. However, Prof. Harabagiu shall direct the research of this project throughout the year. We request salary support for 2 graduate research assistants. The graduate research assistants will have a support each year at 50% time during the academic year, and at 50% time during the summer months. The role of one graduate research assistant will be develop deep learning methods that will populate the schemas associated Hierarchical epileptiform Activity Descriptors by learning the nodes and the relations in the schema hierarchy developed for Aim 4. The second PhD student will focus on developing the deep learning methods that will automatically produce HAD tags both in the EEG reports and in medical articles describing the interpretation of EEGs and use the tags both in the patient cohort retrieval system and in a system capable to retrieve relevant medical articles in support of the interpretation of EEGs, developed in Aim 4. The two PhD students that will be assigned to work on this project will be paid at $2100/month for their activity.

Benefits: Employee benefits were estimated using the published University of Texas at Dallas rates. Benefit rates used in this proposal are 20% of salary of the PI and 15% for graduate students during the academic year and summer months.

Travel: A total of 3 domestic trips are requested yearly to attend technical meetings and workshops that are relevant to the project's overall research, and for scientific exchange. Domestic trips are estimated at $1500 each roundtrip. Expenses include estimates for airfare, ground transportation, hotel accommodations, registration for conferences and workshops (if applicable), and per diem. Relevant conferences where we anticipate publishing the results are the annual AMIA symposium and the annual AMIA Summit on Clinical Research..

Materials and Supplies: We have not budgeted any materials and supplies for FY3.

Publications Costs/page charges: A total of $721 is requested for the cost of preparing and publishing the results of the work conducted under the award.

Other: Tuition reimbursement is requested for the graduate research assistants at a level of $0 per year. Under agreement by The University of Texas at Dallas, tuition for "doctorally qualified" PhD students, who are supported by an RA on an externally funded grant, will have their tuition paid by the Provost. The students proposed to be used on this project qualify for this benefit, and therefore, no tuition is being requested.

Indirect Costs: Indirect costs were estimated in accordance with UTD's rate agreement, which was approved by DHHS, the Federal Cognizant Audit Agency for UTD on 9/01/08. The organized research F&A cost rate of 53% MTDC was used based on the nature of the proposed work.