Inferring Clinical Correlations from EEG Reports with Deep Neural Learning

Methods for Identification, Classification, and Association using EHR Data

Travis R. Goodwin (Presenter) & Sanda M. Harabagiu

University of Texas at Dallas
Disclosure

We have no relevant relationships with commercial interests to disclose.
Learning Objectives

After participating in this session, the learner should be better able to:

• design, train, or apply a deep neural network to automatically infer (missing) sections from electronic medical records
• design a neural network architecture for automatically extracting word- and report-level features
• design a recurrent neural language model for generating medical language
1. Introduction
2. Data
3. Methods
4. Evaluation
5. Conclusions
Diagnosing and managing neurological dysfunction often hinges on successful communication between the neurologist performing a diagnostic test (e.g., an EEG), and the primary physician or other specialists.

In Glick et al. (2005) studied malpractice claims against neurologists:

- 71% involved “an evident failure of communication by the neurologist”
- Majority resulted from deficient communication between the neurologist and the primary physician or other specialists.
- 62.5% of claims included diagnostic errors
- 41.7% involved errors in “ordering, interpreting, and reporting of diagnostic imaging, follow-through and reporting mechanisms.”

These types of errors could be reduced, and communication could be improved by developing tools capable of automatically analyzing medical reports.
Motivating Goal: Automatically extract and analyze the clinical correlations between any findings documented in a neurological report and the over-all clinical picture of the patient

- Enable future automatic systems to identify patients requiring additional follow-up by the primary physician, neurologist, or specialist.
- Ultimately provide a foundation for automatically identifying reports with incorrect, unusual, or poorly-communicated clinical correlations mitigating misdiagnoses and improving patient care.

Critical requirements:

1. inferring what the expected clinical correlations would be for the patient; and
2. quantifying the degree of disagreement or contradiction between the clinical correlations documented in a report and the expected clinical correlations for the patient.

In this initial study, we focus on the first requirement and study the clinical correlation section of EEG reports.
The role of the clinical correlation section of an EEG report:

Describe, explain, and justify the relationships between findings in the EEG report and the patient’s clinical picture so as to convince any interested health care professionals.

- Not sufficient to extract individual findings or medical concepts from the clinical correlation section!
- Clinical correlations are qualified and contextualized through all the subtlety and nuance enabled by natural language expression
  - Coherent natural language
  - Relies on medical knowledge and accumulated experience
Introduction: The Solution

Overview of our Approach:

- Leverage “big data” of EEG reports
- Identify and remove the clinical correlation section written by the neurologist and
- Train a model to infer the entire clinical correlation section from the remainder of each report

Deep Section Recovery Model (DSRM):

Step 1: word- and report- level features are automatically extracted from each EEG report to capture contextual, semantic, and background knowledge; and

Step 2: the most likely clinical correlation section is jointly (a) inferred and (b) expressed through automatically generated natural language
1. Introduction

2. Data

3. Methods

4. Evaluation

5. Conclusions
The Data: TUH EEG Corpus

Temple University Hospital (TUH) EEG Corpus

Largest publicly available dataset of EEG data

- 25,000 EEG sessions
- 15,000 patients
- Collected over 12 years

Contains both EEG Reports and EEG signal data

- (we only considered the textual reports)
# The Data: Examples of Sections

## Table 1: Examples of EEG Report sections from the TUH EEG Corpus (each section was taken from a different EEG report).

<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CLINICAL HISTORY:</strong></td>
<td>An elderly woman with change in mental status, waxing and waning mental status, COPD, morbid obesity, and markedly abnormal EEG. Digital EEG was done on XXXX XX, XXXX.</td>
</tr>
<tr>
<td><strong>INTRODUCTION:</strong></td>
<td>The EEG was performed using the standard 10/20 electrode placement system with an EKG electrode and anterior temporal electrodes. The EEG was recorded during wakefulness and photic stimulation, as well as hyperventilation, activation procedures were performed.</td>
</tr>
<tr>
<td><strong>MEDICATIONS:</strong></td>
<td>Keppra, Aricept, Senna, Aricept, ASA, famotidine</td>
</tr>
<tr>
<td><strong>DESCRIPTION:</strong></td>
<td>In wakefulness, the background EEG is very low voltage, relatively featureless with some 10 Hz activity in the background and a posterior dominant rhythm, which may be estimated at 7 Hz. The patient seems to have very brief lapses into sleep with diffuse 10 to 13 Hz activity and then spontaneous arousals. This pattern is a beta spindle and then an arousal can be identified throughout the record. Later portions of the record seem to demonstrate more sustained sleep, but with ongoing eye movements. HR: 66 BPM.</td>
</tr>
<tr>
<td><strong>IMPRESSION:</strong></td>
<td>Abnormal EEG due to: 1. Slow and disorganized background. 2. Left occipital sharp waves, at times becoming somewhat periodic in sleep. 3. Some additional epileptiform discharges with more of a mid to posterior temporal localization.</td>
</tr>
<tr>
<td><strong>CLINICAL CORRELATION:</strong></td>
<td>This tracing raises the possibility of a mechanism for seizures outside of the area of the abscess described above. The photoparoxysmal response is unusual and may be accentuated by the previous surgery in the posterior brain regions.</td>
</tr>
</tbody>
</table>
Presentation Outline

1. Introduction
2. Data
3. Methods
4. Evaluation
5. Conclusions
Methods: Deep Section Recovery Model

The **Extractor**:  
- automatically extracts feature vectors representing contextual and background knowledge associated with each word in a given EEG report  
- automatically extracts a feature vector encoding semantic, background, and domain knowledge about the entire report;

The **Generator**:  
- uses the feature vectors extracted by the Extractor  
- produces the most likely clinical correlation section for the given report while also considering the semantics of the natural language it is generating
Methods: Problem Definition

\[ \theta = \arg \max_{\theta'} \Pr(S \mid R; \theta') \]

\[
\Pr(S \mid R; \theta) \approx \frac{\Pr(e, h_1, \ldots, h_N \mid R; \theta)}{\Pr(S \mid e, h_1, \ldots, h_N; \theta)}
\]

where:

- \( R \in \{0,1\}^{N \times V} \) is an EEG report
- \( N \) is number of words in the report
- \( S \in \{0,1\}^{M \times V} \) is a clinical correlation section
- \( M \) is the number of words in the clinical correlation section
- \( V \) is the number of words in the vocabulary.
- \( \theta \) represents the learnable parameters of the model (i.e., the weights of connections between neurons)
- \( e \) is the automatically extracted report-level feature vector
- \( h_1, \ldots, h_N \) are the automatically extracted word-level feature vectors
Methods: The Extractor

Overall Role:

- identify important **neurological findings** and **observations**
  - e.g., “background slowing”
- identify descriptions of the patient’s **clinical picture**
  - e.g., “previous seizure”
- determine the **inferred relationship(s)** between each finding and the clinical picture
  - e.g., “observed epileptiform activity is consistent with head trauma”

Neural Layers:

1. **Embedding**: embed each word in the EEG report $R_i$ into a $K$-length continuous vector (where $K \ll V$)

2. 1\textsuperscript{st} bi-RNN: extract features by processing words reading left-to-right and right-to-left

3. 1\textsuperscript{st} Concatenation: combine “forward” and “backward” features of each word

4. 2\textsuperscript{nd} bi-RNN: refine features for each word by reading left-to-right and right-to-left; use final “forward” memory as $e$

5. 2\textsuperscript{nd} Concatenation: combine final “forward” and “backward” features to produce $h_1, \ldots, h_N$
Methods: The Generator

Training Configuration

Inference Configuration

Key:
- Addition
- Attention
- Concatenation
- Softmax
- Embedding
Methods: The Generator II

To reduce complexity of language generation, we assume each word in the clinical correlation section was generated by considering only: (1) word-level feature vectors; (2) report-level feature vector; and (3) previously generated words

\[
Pr(S'|R) = \prod_{j=1}^{M} Pr(S'_j|S'_{j-1}, \cdots, S'_1, e, h_1, \cdots, h_N; \theta)
\]

Neural Layers:

1. **Embedding: (training only)** embeds each word in the clinical correlation section $S$ into a $L$-length continuous vector (where $L \ll V$)
2. **Concatenation**: combines report-level features with embedding of previously generated word
3. **Gated Recurrent Unit (GRU)**: accumulates memories about each generated word
4. **Attention**: allows the generator to consider word-level feature vectors when generating each word of the clinical correlation
5. **Addition**: combine the outputs of the attention and GRU layers
6. **Softmax Projection: (inference only)** produces a distribution over words in the vocabulary such that $S'_j$ is generated as the word with the highest probability
Methods: Training & Inference

Training Setup

Minimize cross-entropy loss between each word in the generated and gold-standard clinical correlation sections:

$$\mathcal{L}(\theta) \propto \sum_{(R,S) \in \mathcal{T}} \left[ \sum_{j=1}^{M} s'_j \log S_j + (1 - s'_j) \log(1 - S_j) \right]$$

We used Adaptive Moment Estimation (ADAM)

- initial learning rate $$\eta = 0.001$$

Inference Setup

1. Produce each word $$S'_j$$ based on previously produced word $$S'_{j-1}$$.
2. Repeat until the END-OF-SECTION symbol is produced.
3. Associate each Softmax vector $$s'_j$$ to a one-hot $$V$$-length vector associated with a single word.
Presentation Outline

1. Introduction
2. Data
3. Methods
4. Evaluation
5. Conclusions
Evaluation: Experimental Setup

Standard 3:1:1 split for training, validation, and testing sets.

Measured automatically generated clinical correlation sections in terms of:

- **Accuracy:**
  - Word Error Rate (WER): how many “steps” it takes to convert the generated clinical correlation section to the gold-standard
  - Bilingual Evaluation Understudy (BLEU): analogue for Precision commonly used to evaluate NLG systems

- **Completeness:**
  - Recall-Oriented Understudy for Gisting Evaluation (ROUGE): analogue for Recall commonly used to evaluation NLG systems

- **Surface-level coherence:**
  - Measured bigram/trigram-level BLEU and ROUGE
  - Manual evaluation
Evaluation: Baselines

We implemented four competitive baseline systems:

1. **NN: Cosine.** Nearest-neighbor baseline which copies the gold-standard clinical correlation section for the most similar EEG report in the training set
   - Relies on cosine-similarity and bag-of-word representations

2. **NN: LDA.** Nearest-neighbor baseline which copies the gold-standard clinical correlation section for the most similar EEG report in the training set
   - Relies on Latent Dirichlet Allocation (LDA) topic representations of reports and Euclidean similarity

3. **DL: Attn-RNLM.** Deep-learning baseline relying on a Recurrent Neural Language Model (RNLM) operating on word-embeddings
   - Resembles DSRM if the Extractor were removed

4. **DL: Basic-S2S.** Deep-learning baseline relying on a Sequence-to-Sequence model
   - Resembles DSRM if the word-level features were not extracted by the generator
## Evaluation: Quantitative Results

### Table 2: Evaluation of automatically inferred clinical correlation sections.

<table>
<thead>
<tr>
<th>System/Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-3</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN:Cosine</td>
<td>.55334***</td>
<td>.40274***</td>
<td>.32137***</td>
<td>.54284***</td>
<td>.38516***</td>
<td>.31508***</td>
<td>2.521***</td>
</tr>
<tr>
<td>NN:LDA</td>
<td>.51730**</td>
<td>.36316***</td>
<td>.28199***</td>
<td>.52389***</td>
<td>.36863***</td>
<td>.28686***</td>
<td>2.891***</td>
</tr>
<tr>
<td>DL:Attn-RNLM</td>
<td>.57907**</td>
<td>.41619*</td>
<td>.32433*</td>
<td>.58196***</td>
<td>.41960*</td>
<td>.32575***</td>
<td>2.315***</td>
</tr>
<tr>
<td>DL:Basic-S2S</td>
<td>.58992**</td>
<td>.36829***</td>
<td>.26806***</td>
<td>.47487***</td>
<td>.31170***</td>
<td>.23445***</td>
<td>2.658***</td>
</tr>
<tr>
<td>DSRM</td>
<td>.68792</td>
<td>.54686</td>
<td>.46323</td>
<td>.63523</td>
<td>.50459</td>
<td>.42894</td>
<td>1.631</td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; statistical significance against DSRM using the Wilcoxon signed-rank test.
Evaluation: Qualitative Results

• Manually evaluated the generated clinical correlation sections for 100 EEG reports
• 5-point Likert scale
• **Average score of 3.491**
• Inferred clinical correlations are generally accurate, but may omit important information or contain inconsistencies

<table>
<thead>
<tr>
<th>Likert Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>(strongly disagree)</em> clinical correlation section is incomprehensible</td>
</tr>
<tr>
<td>2</td>
<td><em>(disagree)</em> clinical correlation section is not correct</td>
</tr>
<tr>
<td>3</td>
<td><em>(weakly agree)</em> clinical correlation section is generally correct, but</td>
</tr>
<tr>
<td></td>
<td>omits important information or contains additional false or inconsistent</td>
</tr>
<tr>
<td></td>
<td>information</td>
</tr>
<tr>
<td>4</td>
<td><em>(agree)</em> clinical correlation section is correct but omits minor details</td>
</tr>
<tr>
<td>5</td>
<td><em>(strongly agree)</em> clinical correlation section is effectively equivalent to</td>
</tr>
<tr>
<td></td>
<td>the gold-standard</td>
</tr>
</tbody>
</table>
## Evaluation: Error Analysis

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Report: 00005044_s03</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inferred</strong>: No epileptiform features are identified. If epilepsy is an important consideration, a repeat EEG capturing deeper stages or sleep deprivation prior to the EEG may be helpful to identify epileptiform activity.</td>
<td></td>
</tr>
<tr>
<td><strong>Gold</strong>: There are no definitive epileptiform discharges, but there is an amplitude asymmetry and there is an asymmetry of wicket activity. Additional recording capturing more extensive sleep may be helpful to identify epileptiform activity.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 2</th>
<th>Record: 00010462_s01</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inferred</strong>: This EEG supports a severe underlying encephalopathy and diffuse disturbance of cerebral dysfunction involving both gray and white matter. Contributing factors can include some of the renal failure, acute or metabolic processes. The focal features described above should be correlated with imaging.</td>
<td></td>
</tr>
<tr>
<td><strong>Gold</strong>: This abnormal EEG demonstrates a severe, diffuse disturbance of cerebral function involving both gray and subcortical white matter. This EEG pattern was communicated to the primary care team.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3</th>
<th>Report: 000004928_s02</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inferred</strong>: This EEG is not suggestive of a metabolic or intermittent encephalopathy. The rare left with focal feature suggests conforms with underlying metabolic pattern.</td>
<td></td>
</tr>
<tr>
<td><strong>Gold</strong>: As discussed with the team on the date of this recording, this EEG is most compatible with a metabolic encephalopathy.</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Comparisons of inferred and gold-standard clinical correlation sections for three EEG reports.
1. Introduction
2. Data
3. Methods
4. Evaluation
5. Conclusions
Conclusion

• It is important to consider both word- and report-level features
• Deep learning can automatically
  • extract useful word- and report-level features
  • generate reasonably coherent, accurate, and complete natural language

• Current limitations:
  • neural networks struggle to account for inconsistencies
  • a single word (e.g. not) can change the meaning of the entire clinical correlation section while not having a proportional impact on the loss function

We need methods of automatically quantifying the disagreement between a generated clinical correlation and a gold-standard clinical correlation section.
Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Thank you!

Any questions?