Final Project Report

Enabling large-scale research on autism spectrum disorders
through automated processing of EHR using natural language understanding

PI: Gondy Leroy, PhD, University of Arizona

Team Members: Mihai Surdenau, PhD, Sydney Pettyrgrove, PhD, Maureen Galindo, RN

Organization: University of Arizona

Project Period: 09/01/2017 - 08/31/2020

Project Officer: Janey Hsiao

Grant No.: R21 HS24988

Note: This report contains the highlight of our works. Further details can be found in our publications.

1. STRUCTURED ABSTRACT

Purpose: With increasing use of electronic medical records, a large investment is made in a resource still vastly underused. In mental health, extracting information from free text can create precious data sources to inform existing research and lead to new insights and treatments.

Scope: The long-term goal is processing free text in EHR with a focus on Autism Spectrum Disorders (ASD) because prevalence is increasing but not well understood. The project brings a data-driven approach for analysis of large datasets and facilitate review of individual EHR.

Methods: The methods include analysis of EHR free text, gold standard creation, and machine learning and natural language processing (NLP) for automated annotation of diagnostic patterns and labeling EHR. A database of 6000 records was leveraged to design algorithms and demonstrate usefulness of the models for a 10 year period.

Results: Rule-based algorithms showed higher precision and recall than deep learning machine algorithms for criteria extraction. Case labeling was high with a variety of algorithm combinations. Application of the algorithms was demonstrated in a case study showing changes over a 10-year period in diagnostic criteria documented for individual children and prototype interface.

Key Words: ASD, EHR, Natural Language Processing, NLP, machine learning, deep learning

2. PURPOSE

This project addresses an existing and aggravating problem and underused opportunity in healthcare. A treasure of information is available in electronic medical and health records (EHR) but it is not used and not taken into account. The purpose of this project is to address current shortcomings in extracting existing data from EHR free text and leveraging this data for new insights. Autism Spectrum Disorders (ASD) serves as an excellent demonstration area because of the unexplained increase in prevalence. With new research focusing on neural, genetic, or environmental causes for understanding ASD and developing treatments, much new data is generated.

Our purpose is to address the lack of tools to leverage existing data, i.e., for large-scale use of the detailed ASD patient information from the EHRs and leverage the opportunity brought by increasing use of EHRs for a variety of patients. The EHRs represent a large investment and often result in a precious resource for improving service as well as research, understanding, treatments and cures. While it is an immense opportunity, the information is still vastly underused. The portion of the EHR ignored most is the free text because it requires advanced natural language processing (NLP) to transform the unstructured information into a structured form for use at a large scale and for integration with other data. The project comprises a self-contained health IT research project focused on a design of NLP algorithms for extracting ASD specific criteria automatically from EHR as well as assigning case labels. The research of interest is the design of a health IT that comprises algorithms and their combination in models for future use.

The purpose is achieved through the design of NLP algorithms to create human-interpretable models that automatically annotate free text in electronic records and match to criteria in the Diagnostic and Statistical Manual of Mental Disorders (DSM) for ASD as well as case labeling using machine learning. We include a demonstration of the feasibility and usefulness of our models. Currently, autism surveillance is a manual, costly, and slow process that provides basic information about autism cases to the CDC and surveillance investigators. The algorithms created can provide more efficient (time & cost) surveillance techniques for tracking ASD across the country. Furthermore, the text processing tools go beyond discovery of single entities, such as genes or proteins, and provide comprehensive matching to more complex patterns, such as the DSM criteria. Finally, we show new research opportunities through secondary analysis of data.

3. SCOPE

3.1 Background

This project addresses two broad changes in mental health and medicine. One change is the increasing prevalence of Autism Spectrum Disorders (ASD) diagnoses and the resulting increase in research, treatments and interventions but lack of large-scale (phenotype) data use to understand, prevent, and cure. The second change is the increasing use and availability of electronic health records (EHR) containing an abundance of text but little further use of the information contained in them for new research.

EHR bring the opportunity of identifying causes of ASD through large-scale data use. But this requires health IT to extract this information. Much current work has been interesting but narrowly focused and limited to structured data. For example, counts of the presence of conditions in populations (1) or evaluations of highly specific decision support systems (DSS), e.g., a template with ADHD diagnostic information (2). Increasingly, the free text from these EHR is being utilized. This text contains rich information that is often complementary, more detailed and explanatory to the data. However, there has been little focus on ASD with one exception, a research project where ICD-9 codes were combined with concepts to classify case status (3). In contrast, other work has focused on analyzing language created by people on the spectrum (4, 5). In other fields, NLP for EHR has already been shown to be valuable, for example NLP of EHR for safety surveillance for postoperative complications (6), extraction of adverse drug effects from psychiatric records (7), identification of patients needing colonoscopy (8), or the creation of data such as veterans' employment information (9).

3.2 Context

We work with EHR in mental health where free text is of enormous importance due to the complexity of diagnosis and treatment. We believe that the opportunity for leveraging the text is great in mental health, and especially ASD, since diagnosis and treatments are highly individualized, resulting in rich records containing free text detailing approaches and results.

Children with ASD demonstrate drastically variable behaviors that qualify for the same DSM criteria. We work on automated extraction of the Diagnostic and Statistical Manual of Mental Disorders (10) (DSM) criteria for ASD since the DSM specifies the combination of criteria needed to assign ASD case status. Analysis of the results may lead to new insights in the condition, change over time or regions. It may also facilitate early detection and treatment which have been demonstrated to improve outcomes (11) through automated detection in EHR. To the best of our knowledge, no current works leverage DSM diagnostic criteria in the free text in the records. With the widespread use of the DSM, this is a missed opportunity.

Table 1 shows example diagnostic criteria for DSM-IV. During the first part of our project, DSM-IV criteria were used in our EHR. Later in the project, the diagnostic criteria were updated to DSM-5 which is the current standard.

Table 1: Sample rules (and numbering) from DSM-IV-TR to diagnose Autistic Disorder

Rule	Description
	DSM-IV
	A. A total of six or more items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3):
	(1) Qualitative impairment in social interaction, as manifested by at least two of the following:
A1a	(a) Marked impairment in the use of multiple nonverbal behaviors such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction
A1b	(b) Failure to develop peer relationships appropriate to developmental level
	…
	(2) Qualitative impairments in communication as manifested by at least one of the following:
A2a	(a) Delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gesture or mime)
A2b	(b) In individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others

Table 2 shows example criteria as they are found in EHR free text. The difficulty and diversity of examples is similar for DSM-IV and DSM-5 criteria.

Table 2: Example DSM-5 Criterion Labels and EHR Examples

DSM-5 Criteria	Example EHR Text Snippets
A1	He does not yet initiate a turn-taking game or social routine
	He would not answer when his name was called
B1	presented with lots of immediate and delayed echolalia interspersed with some spontaneous language
	Clicks camera repeatedly

3.3 Settings and Participants

The Centers for Disease Control and Prevention (CDC) established the Autism and Developmental Disabilities Monitoring Network (ADDM) to monitor ASD in 4- and 8-year-olds. We work with the data collected as part of ADDM which include autism evaluation text files from special education and medical records coupled with a clinician review and coding of statements meeting DSM 5 criteria as well as ASD diagnosis (12).

3.4 Incidence and Prevalence

ASD is of particular interest because of the increasing number of people affected and the lack of breakthroughs in treatments. In the second half of the 20th century, ASD prevalence was estimated at 5 cases per 10,000 people. Since the 1990s, prevalence estimates were increasing(13) and ranged from 4.5-9.9 cases per 1,000 children in 2000, 1 in 110 children in 2006(14), 1 in 68(15) in 2010, and 1 in 54 children in the US in 2020 (16).

The reasons are uncertain, but factors such as increased public awareness, changing diagnostic criteria and substitution of ASD eligibility for other special education eligibilities have been proposed, as well possible true increasing prevalence of ASD (14, 15). Regardless, data on long-term trends, symptoms and interventions are important for planning interventions and educational and health services. Furthermore, without biological lab test, diagnosing requires observation of complex behavior descriptions usually recorded in text.

4. METHODS

4.1 Study Design

This project comprises the development of three types of algorithms: 1) natural language processing algorithms to recognize and label DSM diagnostic criteria in EHR free text (both rule-based and machine learning algorithms are used for this), 2) machine learning algorithms to label an entire records as ASD or not, and 3) a prototype interface to highlight DSM criteria automatically in text.

Components 1 and 2 have been extensively tested using standard evaluation metrics. The usefulness of Component 1 was demonstrated with a study of 10 years of EHR. Component 3 was additional work (out of scope) and testing has been limited to use-cases with local clinicians.

4.2 Data Sources Created

A. EHR

We have access to EHR on 6357 children collected as part of the ADDM. Records from educational sources for each child are matched to their records from any of four clinical sources. The combined records are loaded into the tracking system and at each data source abstractors review them to identify those with any of the 32 social behavioral triggers listed in the ARCHE Abstraction Manual (ARCHE is provided by CDC for their ADDM Network surveillance projects) consistent with ASD. Such identified

records are abstracted: demographic and services information is collected with verbatim descriptions of behaviors exhibited by the child that are consistent with or contradict a diagnosis of ASD.

B. EHR Gold Standard

Diagnostic Criteria Extraction

Abstracted information for each child is printed as a report with no identifiers. This report is evaluated by clinical reviewers who apply standardized criteria to determine case status. The annotations by the clinicians include the DSM-IV-TR criteria in all years and will add the DSM-V criteria from this year on. As of now, we have access to 1986 annotated records for which all DSM criteria have been entered in electronic format. However, these reports were printed and so we needed to create an electronic gold standard with all information combined. For a subset of the EHR (about 200), there are also printed versions on which clinicians have underlined the text segments that are an expression of the 12 DSM-IV (Diagnostic and Statistical Manual on Mental Health Disorders) criteria that were used to diagnose ASD. Using an annotation tool (WebAnno) to load the electronic version of these records, we manually created an electronic version. This is a time consuming process. In some cases, the annotation on the printed record was unclear (e.g., missing label or label difficult to read). For these annotations, we consulted with the clinician on our team to ensure we add the correct label. Once records were annotated and stored, they could be reused in every design and test round.

These data are used by the NLP parser for criteria extraction.

Case Labels

For each record, we have the ASD label (ASD or not).

These data are used by the classification algorithms.

C. Lexicons

Identifying ASD diagnostic criteria in text requires recognizing important trigger words, i.e., words describing typical behaviors of ASD. For our first version of our parser, we capture these words in lexicons. Approximately 90 lexicons with about 20 terms each were manually created. Table 3 provides an overview with examples of lexicons and the terms they contain. We used a lexical lookup for each noun found in the text and annotate it with the lexicon's label. These labels form part of the patterns used to describe DSM criteria. Multiple patterns are needed to capture the different free text expressions for each DSM criterion.

The lexicons are optimized for patterns for each DSM criterion, so the same terms may appear in multiple lexicons. However, a few lexicons are shared by all patterns and used for different DSM criteria. Currently, there are eleven lexicons commonly shared by all patterns, e.g., the lexicons containing body parts. In addition, the patterns for the A1, A2 and A3 criteria share respectively seven, three and two lexicons. For example, DSM rules A1a, A1b, A1c and A1d all require identification of "impairment in social interaction," and the relevant terms for this trigger are combined in the lexicon "A1_interact." In addition to these shared patterns, each DSM pattern requires additional individual lexicons optimized for that pattern.

Table 3. Lexicon Overview

Pattern use of Lexicons	Lexicons	Nr of terms	Example Lexicon	Example Terms
All Rules	11	345	Body_parts	arm, eye, hair, teeth, toe, tongue, finger, fingers, nose
Group A1	7	105	A1_interact	interact, interactions, communicate, relationship
Group A2	3	72	A2_positive	severe, significant, pervasive, marked
Group A3	2	72	A3_object	door, toys, vacuum, blocks, book, television, lights
A1a	4	42	A1a_nonVerbalBehavior	eye contact, eye to eye gaze, gestures, nonverbal cues
A1b	2	11	A1b_consistent	good, consistent, appropriately, satisfactory
A1c	5	61	A1c_affect	excitement, feelings, satisfaction, concerns
A1d	12	159	A1d_engage	recognize, recognizes, reacts, respond, regard attend
A2a	4	117	A2a_gained	gained, used, had, obtained, said, spoke
A2b	8	240	A2b_recepLang	direction, instructions, questions, conversations
A2c	7	145	A2c_idiosyncratic	breathy, echolalic, jargon, neologism, reduced
A2d	7	83	A2d_actions	actions, routines, play, signs, gestures, movements
A3a	7	106	A3a_obsess	obsessed, obsessive, perseverates, preoccupation
A3b	7	119	A3b_nonFunctionalPlay	stack, stacks, lines, lined, nonfunctional, arrange
A3c	3	67	A3c_abnormal	grind, grinds, rocks, twirls, spin, tap, clap, flap
A3d	3	43	A3d_sensitive	defensiveness, sensitivity, hypersensitivities
Total	92	1,787		

D. Word Embeddings:

While lexicons are a good starting point, they require the manual addition of all variants of a word, e.g., synonyms, plural. To automate this process, we started working with word embeddings.

Word embeddings are a continuous dense representation of words in a corpus. Instead of representing a word with one value in a large sparse matrix the size of an entire vocabulary, each word is represented by a dense vector of a size pre-determined by the user (usually 50 to 300). Word embeddings are based on the distributional hypothesis, that is, similar words appear in similar contexts. The underlying algorithm "learns" meanings of words based on their occurrence in a large, unlabeled corpus and word embedding vectors can encode semantic meaning, allowing users to programmatically determine semantic similarity of words based on cosine similarity of the vectors.

We created a vocabulary to be used by the machine learning algorithms. Using existing software libraries (i.e., Word2Vec) we created word embeddings using our EHR. We evaluated different versions of word embeddings and compared their quality.

Table 4 shows an overview of the word embeddings created using different datasets. The quality of these was compared and the EHR based word embeddings found to be of the highest quality.

Table 4. Descriptive statistics of training corpora and embedding

Embedding	Tokens	Documents	Words Vectorized
EHR-5M	5,004,165	2,082	11,570
EHR-ALL	10,745,674	4,482	15,663
PubMed-5M	5,007,328	23,249	20,393
PubMed-ALL	6,703,109	31,171	23,481
PsychInfo-5M	5,051,296	27,049	19,251
PsychInfo-ALL	13,316,489	69,601	30,964

4.3 Interventions

A. Diagnostic Criteria Extraction: Parser Development

Rule-based Version (Initial Version)

The parser combines open source libraries, e.g., the General Architecture on Text Engineer (GATE) (17, 18) for standard pre-processing of text: tokenizer, sentence splitter, and the Stanford tagger for the part-

of-speech Tagger (19). After processing all free text, terms are annotated using gazetteer lookup (i.e., the lexicons listed above).

Using 43 annotated records from the ADDSP containing 4732 sentences, we developed 12 sets of patterns (total 104 patterns) for the 12 DSM criteria (see Table 1).

Figure 1 shows one example patter as a finite state automata (FSA) for DSM criterion A2c. Each label on an arc (e.g., A2c_speech) represents the lexicon of terms (terms indicating 'speak' as relevant to rule A2c). For example, Pattern 1 would match to the text "[often]$_{A2c\_frequent}$ [speaks]$_{word\ token}$ [with]$_{word\ token}$ [reduced]$_{A2c\_idiosyncratic}$ [volume]$_{A2c\_speech.}$"

All patterns are specified in a Java Annotation Pattern Engine (JAPE) file. A JAPE file is a file where patterns to be annotated in text can be described using GATE-specific formatting. GATE 'reads' the JAPE files and applies them to text. When a pattern in the JAPE is recognized in the text, the text matching the pattern is annotated with the labels specified in the JAPE file.
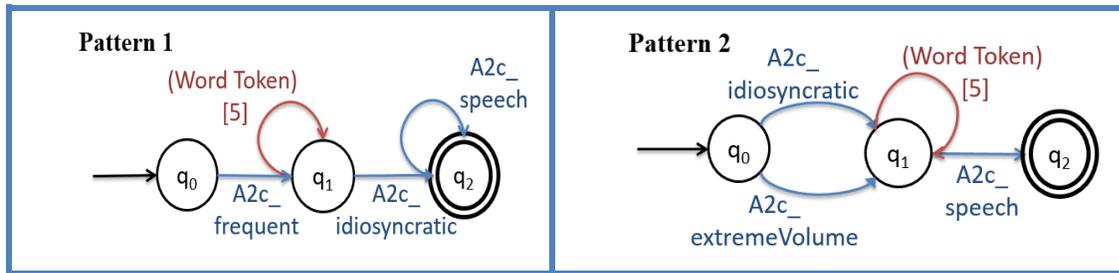
NOTE: The evaluation is in the results section.

Figure 1. Visualization of two patterns (of seven existing) for DSM criteria A2c

Machine Learning Version (Newer Version)

We evaluated the newest deep learning algorithms for our parser. Deep learning delivers good performance in classification tasks, but is suboptimal with small and unbalanced datasets, which are common in many domains.

We created a machine learning version using LSTM. Since some criteria were represented by fewer than one hundred examples, we needed creative methods to optimize the algorithms. To address this limitation, we use conventional machine learning, i.e., support vector machines (SVM) to tune deep learning hyper-parameters.

NOTE: The evaluation is in the results section.

B. ASD Case Assignment

This part on our preliminary work using machine learning algorithms to assign case labels.

We compare the value of different types of features, machine learning algorithms, and ensemble approaches. Each type of feature captures different information about the data, while different algorithms learn different characteristics of the features. An ensemble over features sets combines different types of information, and an ensemble over algorithms combines different ways to analyze information. We found that ensembles over algorithms lead to more accurate and balanced classification than individual classifiers, while ensembles over different features can effectively combine different types of data and give classification that is more accurate compared to the classic combination (concatenation) of the same set of features.

Algorithms

We used support vector machines (SVM), Decision Trees (DT), and Neural Networks (NN). We chose

SVM because it tends to yield good performance and is fast to train. We used a SVM with the traditional linear kernel. We chose DT because they provide interpretable results. We used an optimized version of the CART Algorithm (Classification and Regression Trees) with GINI information gain as the splitting criterion. NN are black-box models and not intended to provide interpretable decisions even though current research exists aimed at deducing rules and automata from trained NN (20, 21). NN provide excellent performance but need large datasets and training is time consuming. For all three algorithms, we used scikit-learn's implementation of these algorithms in Python (22).

We also pilot-tested a deep learning algorithm: convolutional neural network (CNN). However, our datasets is small and results were poor (70.0% accuracy). Therefore, we focused on classic ML algorithms in this study.

Ensemble Methods

Ensemble methods combine results from multiple classifiers, which allow errors to even out and perform better than a single classifier. For transparency and interpretability, we are taking a straightforward approach to combine classifiers: taking a majority vote, which has also been shown to work well in practice (23). The key to ensembles' superior performance is the combination of diverse information from its base classifiers.

In this work, we compare two approaches for creating ensembles with diversity: combining different features and combining different algorithms.

Feature ensembles. Every feature set we generated in this work captures a different aspect of information in the document. Therefore, classifiers trained on different feature sets are learning from different types of information and an ensemble can combine them. Two of our feature sets combine annotation-based features and bag-of-word (BOW) features by concatenating them in the into a single feature vector. In an ensemble, we take a vote over classifiers that use these features sets separately to determine if an ensemble is a more effective way to combine features than simply concatenating the feature vectors.

Algorithm ensembles. One set of ensemble classifiers vote over the output from three different algorithms applied to the same dataset. This scheme leverages the fact that each machine learning algorithm learns something different from the data, which together, should produce more insightful predictions.

4.4 Measures

For our evaluation, we customary calculate four metrics. Precision provides an indication of how correct the annotations made by the parser are, in other words, if the parse annotates sentences with a DSM label what percentage of these labels are correct. Recall (also referred to as sensitivity) provides an indication of how many of the annotations the parser is able to capture, in other words, of all the sentences that received a DSM label by the human annotators what percentage does the parser also label correctly. We also calculate the F-measure which is the harmonic mean of recall and precision. The scores for the F-measure indicate how balanced an approach is: when recall and precision are similarly high, the F-measure will be high, however if one of them is low the F-measure will reflect this with a low F-score. Finally, we also calculate specificity, which indicates how well our parser can ignore sentences that are not an expression of DSM criteria.

- $Precision\ (or\ PPV) = \dfrac{True\ Positive}{True\ Positive + False\ Positive}$

- $Recall\ (or\ Sensitivity)\ = \dfrac{True\ Positive}{True\ Positive + False\ Negatives}$

- $F - measure = 2 * \dfrac{Precision * Recall}{Precision + Recall}$

- $Specificity = \dfrac{True\ Negatives}{True\ Negatives + False\ Positive}$

5. Results and Demonstrations

5.1 Phenotypical Behavior Labeling with DSM-IV Criteria: Rule-based Parser

Our testbed consists of the 50 new EHR records containing 6634 sentences. These are records that were annotated by the clinical experts and the text and annotation stored by us in electronic format. Of the entire set, 1357 sentences (20.45%) contained annotations with some sentences contained more than one annotation. Table 5 shows the number of examples of phenotypical behaviors for each criterion found in the EHR.

Table 5. Gold Standard Overview

DSM Diagnostic Criteria		Gold Standard	
DSM Rule	**Theme**	**Total in Records**	**Average per Record**
A1a	Nonverbal behaviors	126	2.52
A1b	Peer relationships	91	1.82
A1c	Seeking to share	37	0.74
A1d	Emotional reciprocity	165	3.3
A2a	Spoken language	406	8.12
A2b	Initiate or sustain conversation	333	6.66
A2c	Stereotyped or idiosyncratic language	127	2.54
A2d	Social imitative play	66	1.32
A3a	Restricted patterns of interest	62	1.24
A3b	Adherence to routines	135	2.7
A3c	Stereotyped motor mannerisms	68	1.36
A3d	Preoccupation with parts of objects	28	0.56
Total		1644	32.88

Table 6 and Table 7 show the results of our parser and its ability to label these behaviors with the correct DSM label. At the annotation level, we achieved 74% precision and 42% recall on average. We took the micro average, which combines the true and false positive counts across all rules. For individual criteria, precision was higher (75% and higher) for most with the exception of two (Criterion A1d and A3d). Recall was also particularly low for these two criteria, along with A1b and A1c. The best precision and recall were achieved for criterion A1a, with more than half of the annotations (57% recall) identified and with very few errors (96% precision).

Table 6. Annotation-level Results (P= Precision, R= Recall, F= F-Measure)

	Annotation Level		
	(Based on 6634 sentences)		
	Total in GS (#sentences)	Evaluation	
		P	R	F
Annotations	**1644**			
A1a	126	0.96	0.57	0.72
A1b	91	0.63	0.27	0.38
A1c	37	0.78	0.19	0.30
A1d	165	0.62	0.27	0.37
A2a	406	0.69	0.44	0.53
A2b	333	0.79	0.44	0.57
A2c	127	0.68	0.36	0.47
A2d	66	0.79	0.56	0.65
A3a	62	0.83	0.40	0.54
A3b	135	0.75	0.51	0.61
A3c	68	0.82	0.41	0.55
A3d	28	0.53	0.29	0.37
(Micro) Average		*0.74*	*0.42*	*0.53*

The results are very similar for the sentence level evaluation. Both metrics are slightly higher; with average precision at 76% and average recall at 43%. For the A1a criterion, more than half of the required sentences were identified (recall 59%) with minimal errors (97% precision). Using a sentence as a unit of analysis, it is also possible to compute specificity, or true negative rate (which was not possible with annotation level evaluation since we would have to predefine in advance how many possible annotations – i.e., sentence segments – there are in the EHR). However, specificity is not a very interesting metric for this task. We achieve nearly perfect specificity because only 0.5% to 5% of all sentences contain true annotations for each individual rule, and our system reports very few false positives (high precision).

We conducted a final, more lenient approach by evaluating whether the system can identify the relevant sentences for DSM criteria, regardless of which criterion they represent. In this case, we found that our parser achieves 82% precision and 46% recall in identifying the 1,357 sentences that were annotated for autism-like behavior.

Table 7: Sentence-level Results (P= Precision, R= Recall, F= F-Measure, S = Specificity)

| | Sentence Level (Based on 6,634 sentences) | | | |
| | Total in GS (#sentences) | Evaluation | | |
		P	R	F	S
Sentences	1,357				
A1a	120	0.97	0.59	0.74	1.00
A1b	90	0.68	0.30	0.42	1.00
A1c	35	0.78	0.20	0.32	1.00
A1d	158	0.63	0.28	0.39	1.00
A2a	391	0.71	0.45	0.55	0.99
A2b	329	0.83	0.47	0.6	1.00
A2c	121	0.67	0.37	0.48	1.00
A2d	65	0.83	0.58	0.68	1.00
A3a	61	0.73	0.36	0.48	1.00
A3b	123	0.74	0.52	0.61	1.00
A3c	64	0.82	0.42	0.56	1.00
A3d	28	0.53	0.29	0.37	1.00
(Micro) average	**1585**	0.76	0.43	0.55	1.00
All Rules	**1357**	*0.82*	*0.46*	*0.59*	*0.97*

Summary: We created a rule-based parser to extract behaviors matching DSM-IV diagnostic criteria. The parser showed high precision: 76% average over all 12 criteria, ranging from 53% to 97% for individual criteria, and somewhat lower recall with 46% average over all 12 criteria, ranging from 43% to 59% for individual criteria (24).

5.2 Phenotypical Behavior Labeling with DSM-5 Criteria:: Machine Learning-based parser

We evaluated the newest deep learning algorithms for our parser. Deep learning delivers good performance in classification tasks, but is suboptimal with small and unbalanced datasets, which are common in many domains. To address this limitation, we use conventional machine learning, i.e., support vector machines (SVM) to tune deep learning hyper-parameters.

We chose two classification approaches for the task. We use SVM, a reliable, classic machine learning algorithm frequently used with text data, and BI-LSTM, a state-of-the-art deep learning model that is usually applied to text.

SVM

We used scikit-learn's implementation of the SVM in Python (22). Since the SVM naturally has a two-class formulation, we train an independent model to detect the presence of each diagnostic criteria. Our BOW features are the 5000 most frequent tokens from the training data.

BI-LSTM

We used a BI-LSTM with tunable pre-trained embeddings. The input into the BI-LSTM are 200-dimensional pre-trained word embeddings from 4480 ASD EHR from 2000-2010, the complete set of unlabeled EHR text from one ADDM surveillance site during that time. During training, we randomly removed half the cases without any positive labels to adjust for the small proportion of positive cases. Each LSTM Layer has an internal layer size of 350 and was trained with a dropout ratio of 0.5. We use a sigmoid output layer with one unit for each label. The model is set to train for up to 50 epochs with early stopping. In practice, most models in our experiment trained for less than 25 epochs. In this study, we used Keras (2.1.5) (25) to implement the BI-LSTM and Deeplearning4J's word2vec implementation (26) to train the word embeddings.

Tuning Process

On a personal computer, it takes a few minutes to train a SVM on our dataset, compared to approximately two hours needed to train a BI-LSTM. Therefore, we can conduct fairly thorough parameter tuning for the SVM through-grid search. We validated the parameters on 20% of our training examples, and retrained the final model using the entire dataset based on the best set of parameters.

It is less feasible to exhaustively tune the BI-LSTM through grid-search. We selected the baseline architecture based on a manual search, guided by our previous experience working with text data.

The training parameter from the SVM that can be informative for training the BI-LSTM is class weights. Since we have a highly imbalanced dataset, we can increase the weights of the minority class to increase their impact on the model. In addition to a plain BI-LSTM, we also tested a version of an LSTM in which each class is weighted by the best class weights found by the SVM.

In summary, we compare the following three systems:

- SVMs: a set of highly-tuned SVMs, one for each class
- BI-LSTM-1: a regular BI-LSTM.
- BI-LSTM-W: a BI-LSTM trained with class weights 1 from the tuned SVM

We evaluated our approach for DSM-5 annotations (Table 8). A bidirectional LSTM (BI-LSTM) could not learn the labels for the seven scarcest classes, but saw an increase in performance after training with optimal weights learned from tuning SVMs. With these customized class weights, the F1 scores for rare classes rose from 0 to values ranging from 18% to 57%. Overall, the BI-LSTM with SVM customized class weights achieved a micro-average of 47.1% for F1 across all classes, an improvement over the regular BI-LSTM's 45.9%. The main contribution lies in avoiding null performance for rare classes.

Using these optimal values for class weights from tuning the SVM, we were able to improve the overall performance and avoid null values in seven classes that originally showed 0 value for F1 (see above for definition) but then improved to 57.1% (27). However, re-weighting also resulted in small negative impacts on classes with many examples, which showed the need for multiple models and ensembles. Second, we used the rule-based parser to create training data (a type of weak supervision). We were able to improve the recall from 60.5% to 69.8% and precision from 51.5% to 52.3% by training on the extended data (28).

Table 8. Classification Results for DSM-5 Diagnostic Criteria Labeling.

Label	SVMs			BI-LSTM-1			BI-LSTM-W		
-	P	R	F1	P	R	F1	P	R	F1
A1	0.606	0.407	0.487	0.497	0.499	**0.498**	0.450	0.437	0.443
A2	0.577	0.723	**0.642**	0.599	0.621	0.610	0.450	0.813	0.579
A3	0.454	0.460	0.457	0.695	0.327	0.444	0.522	0.475	**0.497**
B1	0.597	0.456	**0.517**	0.446	0.495	0.469	0.462	0.528	0.492
B2	0.733	0.525	**0.612**	0.648	0.488	0.556	0.525	0.593	0.557
B3	0.509	0.329	**0.400**	0.500	0.012	0.023	0.274	0.271	0.272
B4	0.506	0.517	0.512	0.605	0.586	**0.595**	0.497	0.695	0.579
AF1a	0.418	0.583	0.487	0.560	0.292	0.384	0.520	0.542	**0.531**
AF1b	1.000	0.235	0.381	*0.000*	*0.000*	*0.000*	0.435	0.588	**0.500**
AF2	0.618	0.366	**0.460**	0.806	0.314	0.452	0.667	0.349	0.458
AF3	0.167	0.143	0.154	*0.000*	*0.000*	*0.000*	0.143	0.571	**0.229**
AF4	0.532	0.652	**0.586**	0.765	0.146	0.245	0.516	0.562	0.538
AF5	0.343	0.427	**0.381**	0.352	0.173	0.232	0.310	0.246	0.274
AF6	0.467	0.343	**0.396**	0.398	0.353	0.374	0.275	0.588	0.375
AF7	0.567	0.454	0.504	0.470	0.551	0.508	0.509	0.514	**0.512**
AF8a	0.188	0.286	0.226	*0.000*	*0.000*	*0.000*	0.200	0.286	**0.235**
AF8b	0.200	0.125	0.154	*0.000*	*0.000*	*0.000*	0.333	0.125	**0.182**
AF10	0.209	0.409	0.277	*0.000*	*0.000*	*0.000*	0.333	0.409	**0.367**
AF11a	0.013	0.250	0.025	*0.000*	*0.000*	*0.000*	0.500	0.125	**0.200**
AF11b	0.050	0.150	0.075	*0.000*	*0.000*	*0.000*	0.194	0.300	**0.235**
AF12	0.707	0.933	**0.805**	0.813	0.520	0.634	0.725	0.880	0.795
AF13a	0.491	0.274	0.351	0.416	0.416	**0.416**	0.315	0.568	0.405
AF13b	0.539	0.318	0.400	*0.000*	*0.000*	*0.000*	0.769	0.455	**0.571**
AF14	0.282	0.600	**0.384**	*0.000*	*0.000*	*0.000*	0.311	0.350	0.329
Micro-average	*0.481*	*0.453*	*0.467*	*0.526*	*0.407*	*0.459*	*0.426*	*0.531*	*0.472*

5.3 EHR Case Labeling with Machine Learning
A. Individual Models

Different input features were compared (Table 9) ranging from bag-of-words (BOW), labels automatically generated with the parser (ANNOT), and labels using a manually created lexicon (MLEX). For BOW, 2,000 or 10,000 terms were chosen based on Term Frequency-Inverse Document Frequency (TF-IDF) or Pointwise Mutual Information (PMI) scores. The features used are as follows:

- BOW: TF-IDF BOW of terms that appears in more than 2 documents
- BOW 2K-F: TF-IDF BOW of the 2,000 terms with the highest cross-document frequency
- BOW 10K-F: TF-IDF BOW of the 10,000 terms with the highest cross-document frequency
- BOW 2K-P: TF-IDF BOW of the 2,000 terms with the highest absolute PMI
- BOW 10K-P: TF-IDF BOW of the 10,000 terms with the highest absolute PMI
- ANNOT: 72 metrics (12 criteria, 6 metrics each) based on lexical overlap with sample annotations for the diagnostic criteria
- ANNOT BOW2K-F: 72 annotation-based metrics and terms in BOW 2K-F
- MLEX: 92 manually developed lexicons
- MLEX BOW2K-F: 92 manually developed lexicons and terms BOW 2K-F

The human interpretable DT algorithm yielded lower classification accuracy than SVM and NN for all except two feature sets: BOW 2K-P and ANNOT. Using only MLEX yield the worst performance for the DT, with only 69.5% accuracy, seven percentage points behind the best performing SVM system.

Comparing the different sets of features, 2,000 most frequent terms, which can be derived without any knowledge of the domain or data, is the most reliable predictor of case status using both DT (78.0% accuracy) and NN (82.5% accuracy). SVM saw its best performance at 80.9% accuracy, using every

terms that occurred more than twice in training. Using PMI-based features resulted in DT models that had much higher precision than recall; the same also occurred in MLEX with SVM and MLEX BOW2K-F with DT.

Table 9. Classification Results – Classic Approach EHR Case Labeling

Feature Set	Support Vector Machine				Decision Tree				Neural Network			
	A	P	R	F1	A	P	R	F1	A	Pc	R	F1
BOW	**0.809**	**0.821**	**0.803**	**0.812**	0.779	0.777	0.798	0.787	0.793	0.792	0.810	0.801
BOW 2K-F	0.782	0.785	0.794	0.789	**0.780**	**0.779**	**0.798**	**0.788**	**0.825**	**0.848**	**0.804**	**0.825**
BOW 10K-F	0.806	0.818	0.801	0.809	**0.780**	**0.779**	**0.798**	**0.788**	0.812	0.834	0.793	0.813
BOW 2K-P	0.731	0.745	0.723	0.734	0.775	*0.824*	*0.715*	0.766	0.773	0.762	0.814	0.787
BOW 10K-P	0.782	0.799	0.771	0.785	0.776	*0.853*	*0.682*	0.758	0.788	0.797	0.789	0.793
ANNOT	0.743	0.759	0.734	0.746	0.713	0.719	0.727	0.723	0.688	0.692	0.707	0.699
ANNOT BOW2K-F	0.783	0.784	0.796	0.790	0.778	0.796	0.763	0.779	**0.825**	**0.848**	**0.803**	**0.825**
MLEX	0.765	*0.826*	*0.687*	0.750	0.695	0.717	0.672	0.694	0.758	0.806	0.806	0.747
MLEX BOW2K-F	0.777	0.777	0.793	0.785	0.773	*0.828*	*0.704*	0.761	0.823	0.845	0.802	0.823

B. Feature Ensembles

Table 10 summarizes the results of majority-vote ensembles based on multiple feature sets and a single base classifier, and compares them against the best metric from any single feature set using the same algorithm. Since the SVM saw its best performance with the large BOW feature set, we added two additional ensemble models (Models 7 and 8) that uses the full BOW as the vocabulary component in the ensemble. The best ensemble model was voting over three NN models that combined BOW 2K-F, MLEX, and MLEX BOW 2KF. This model gave predictive accuracy of 83.3%, precision of 86.1%, recall of 80.6%, and overall F1 of 83.2%. This is an improvement of 0.8% in accuracy and 0.7% in F1 compared the best performance of a single feature set. Every ensemble was able to outperform any single component feature set except for two instances (precision in Model 2 and recall in Model 8).

Table 10. Classification Results – Algorithm Ensemble

#	Base Classifier	Voting Ensemble Model				Best Metric from Single Feature Set			
		A	P	R	F1	A	P	Rec	F1
		Feature sets: BOW 2k-F, ANNOT, ANNOT BOW 2K-F							
1	SVM	**0.790**	0.793	0.800	0.797	0.783	0.785	0.796	0.790
2	DT	**0.786**	0.779	0.816	0.797	0.780	*0.796*	0.798	0.788
3	NN	**0.829**	0.852	0.808	0.829	0.825	0.848	0.804	0.825
		Feature Sets: BOW 2k-F, MLEX, MLEX BOW 2K-F							
4	SVM	**0.793**	0.800	0.795	0.798	0.782	0.826	0.794	0.789
5	DT	**0.783**	0.792	0.785	0.788	0.780	0.828	*0.798*	0.788
6	NN	**0.833**	0.861	0.806	0.832	0.825	0.848	0.806	0.825
		Feature Sets: BOW, ANNOT, ANNOT BOW 2K-F							
7	SVM	**0.817**	0.826	0.816	0.821	0.809	0.821	0.803	0.812
		Feature Sets: BOW, MLEX, MLEX BOW 2K-F							
8	SVM	**0.819**	0.840	0.799	0.819	0.809	0.826	_0.803_	0.812

Table 11 summarizes the results of a majority-vote ensemble using three classification algorithms on each feature set, and compares them against the best metric from any single algorithm. Three features sets (BOW 10K-F, ANNOT BOW2K-F, and MLEX BOW2K-F) tied for highest classification accuracy with 82.7%, higher than accuracy from any single algorithm; the BOW 10K-F feature set gained 1.5% in accuracy, the greatest margin in this work. The ensemble model improved accuracy for every feature set with three exceptions. The best feature set using a single algorithm, BOW 2K-F, had the same accuracy with the ensemble and the NN. The two small features sets, ANNOT and MLEX, saw better performance with a single algorithm. Except for ANNOT and MLEX, the ensemble models produced more balanced predictions and higher F1, while a particular algorithm may slightly favor precision or recall.

Table 11. Classification Results: Feature Ensemble

#	Feature Set	Voting Ensemble Model				Best Metric from Single Algorithm			
		Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
1	BOW	**0.820**	0.829	0.817	0.823	0.809	0.821	0.810	0.812
2	BOW 2K-F	**0.825**	0.839	0.813	0.827	**0.825**	*0.848*	0.804	0.825
3	BOW 10 K-F	**0.827**	0.844	0.815	0.829	0.812	0.834	0.801	0.813
4	BOW 2K-P	**0.785**	0.797	0.780	0.788	0.775	*0.824*	*0.814*	0.787
5	BOW 10K-P	**0.809**	0.848	0.765	0.804	0.788	*0.853*	*0.789*	0.793
6	ANNOT	0.723	0.731	0.730	0.730	**0.743**	*0.759*	*0.734*	*0.746*
7	ANNOT BOW2K-F	**0.827**	0.851	0.804	0.827	0.825	0.848	0.803	0.825
8	MLEX	0.763	0.816	0.697	0.752	**0.765**	*0.826*	*0.806*	0.750
9	MLEX BOW2K-F	**0.827**	0.859	0.792	0.825	0.823	0.845	*0.802*	0.823

5.4 Application: Symptom Prevalence Study

We applied the rule-based parser to demonstrate changes in diagnostic criteria presence over a period of 10 years of ADDM surveillance (24).

We analyzed our 4480 records not been used during the development of the parser and contain a minimum of text (40 characters was used as the cutoff). Figure 2 shows the descriptive statistics. Records were collected every two years starting in 2000 and ending (for our analysis) in 2010. In the first three collection periods, fewer records were collected, however, in each of the last 3 collection periods, around 1,000 records were collected. The prevalence of autism in the records is lower the first year (39%). This is associated with the relative inexperience of the data collection team who abstracted more records than necessary to avoid missing cases. In subsequent years data collection was more efficiently focused on records that included information consistent with an autism diagnosis and the proportion children's records included was between 50 and 60% with the one exception of study year 2006 (41.8%).
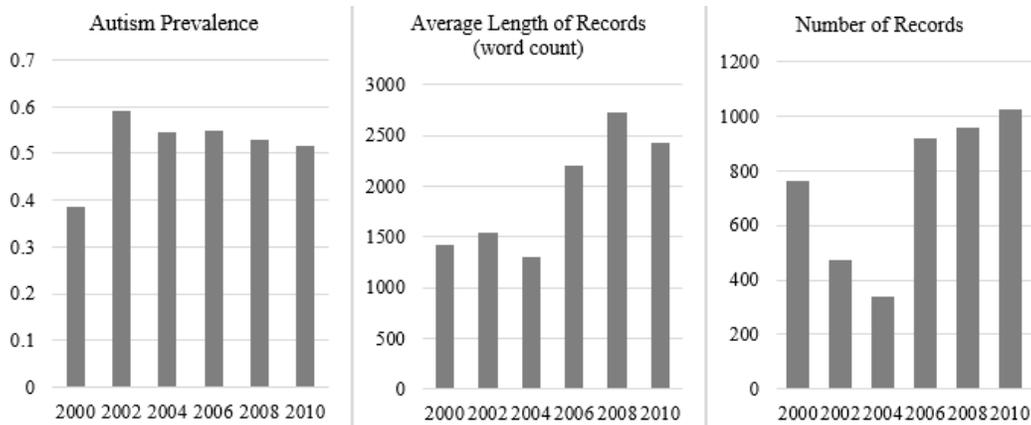
Figure 2: Descriptive Information: Number of records, ASD Prevalence, and Average Text Length

We then applied our rule-based parser to all records and show an overview of phenotypical behaviors and their match to the individual diagnostics criteria as expressed in the records.

The records contained on average 5.76 different DSM criteria. We performed our analysis separately for records of children with and without ASD. All counts are normalized by record length: the number of criteria found is divided by number of words in the document.

A1 DSM criteria: impairments in social interaction (Figure 3). For children with ASD (Panel A) the A1d criterion (social/emotional reciprocity) is the most common criterion found in the records. The least commonly found was the A1c criterion (shared interest). In the last 4 years, the average number of A1a,b, and d criteria described in the records increased, but no similar increase in the average number records containing A1c was observed.

We performed the same analysis for children without ASD (Panel B). The results show, as expected, that there are fewer criteria recorded in their records. The patterns are also different. The number of criteria recorded shows a decreasing trend over the last 4 years of record.

Figure 3: Average A1 Criteria Per Record
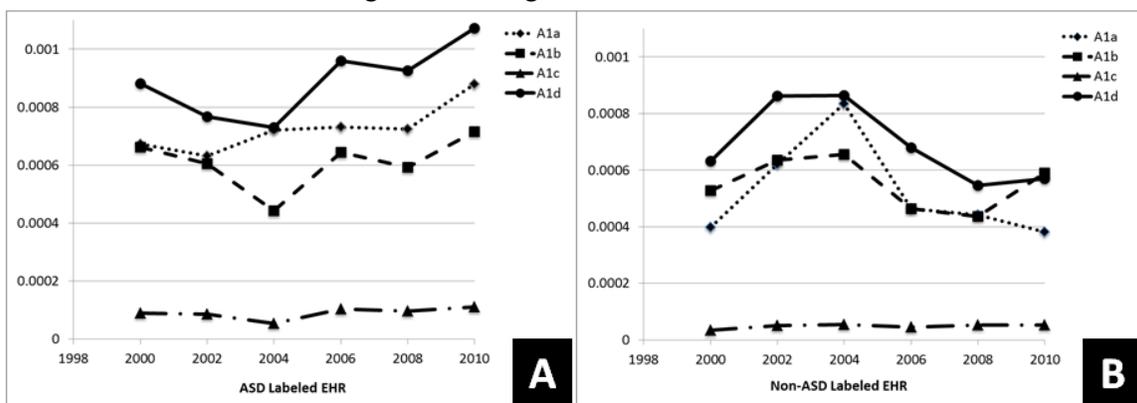
A2 DSM criteria: impairments in communication (Figure 4). The changes for A2 criteria are very small over the years. The most commonly found criterion is A2a (spoken language) and the least commonly found criteria are A2c (stereotyped/repetitive/idiosyncratic language) and A2d (Imaginative play). For the records of children with ASD, there is a slight increase in 2002 and 2004, but few changes over the collection years. The total number of these A2 criteria is higher than for A1 criteria (see axis).

Interesting, there is little difference between the number of criteria found in ASD versus not ASD.

Figure 4: Average A2 Criteria Per Record
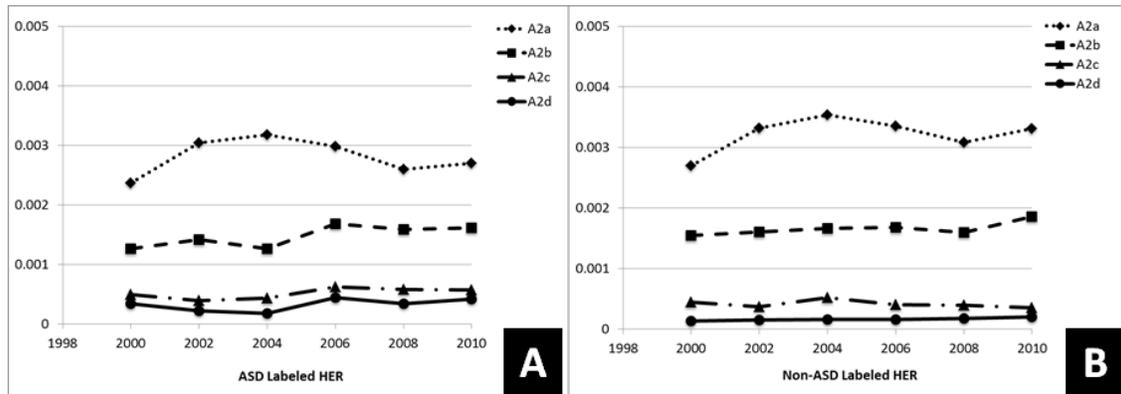
A3 DSM criteria: restricted repetitive and stereotyped behavior patterns (Figure 5). For the records labeled with ASD, the most commonly found criterion is A3b (Adherence to routines) with the other three criteria being less common and comparable to each other.

Overall, fewer criteria are found in the Non-ASD labeled records.
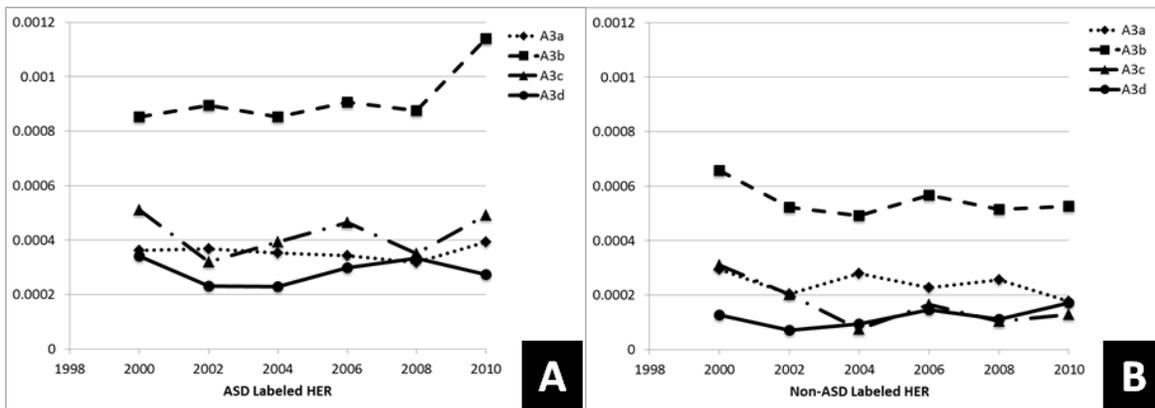

Figure 5: Average A3 Criteria Per Record

5.4 Application: Interface Development (Additional work, not in original aims)

We developed a prototype interface that shows criteria extracted for an individual record. The use of such an interface is review of EHR by clinicians for faster surveillance review or easier clinical decision making.

The following figures show highlighting of individual criteria for DSM-IV and DSM-5 (although the underlying DSM-5 algorithms are tentative) (Figure 6), highlighting of multiple criteria matching the same sentences (Figure 7), and a summary of criteria found in a record (Figure 8).
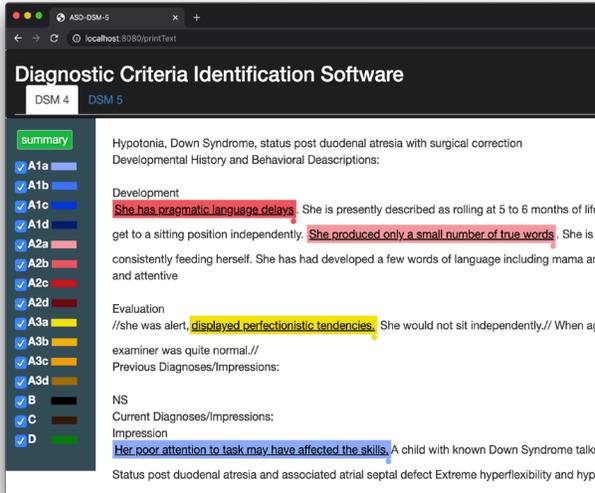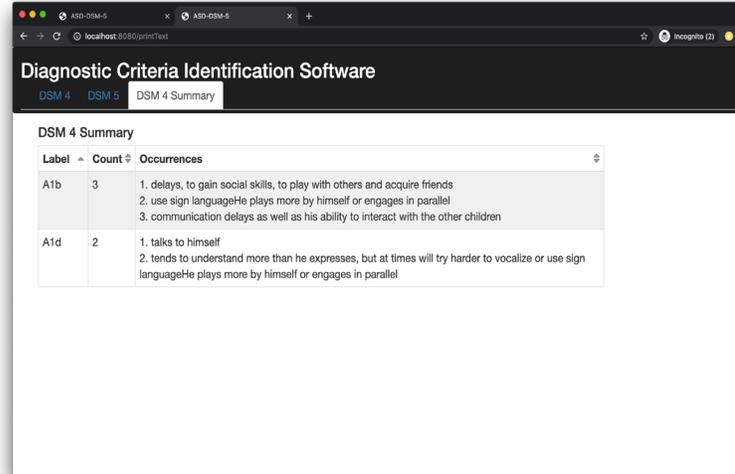
Figure 6: DSM-IV Criteria Highlighted
in Text

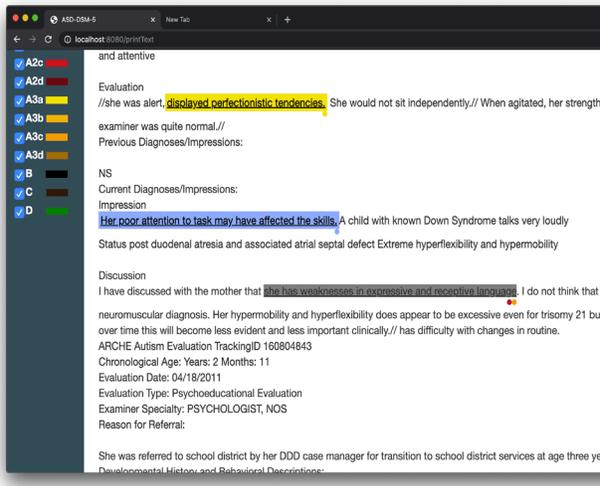Figure 7: DSM-5 Multiple Criteria
Matching One Phrase

Figure 8: Summary of Counts for EHR

5.5 Conclusions

The project focused on developing algorithms to algorithmically detect and label phenotypical expression of ASD behavior in EHR with the correct DSM diagnostic criterion. Precision was high but recall was somewhat lower for recognizing diagnostic criteria for the rule-based approach. Machine learning approaches had somewhat lower performance but significantly less development time. With further tuning and increasing the dataset, these machine learning algorithms can be improved.

In addition, machine learning algorithms were further developed to label an entire EHR with the ASD label or not. Accuracy was high for the best models and can be further improved by adding more different data fields (current work was focused on free text only) and a larger dataset.

Two demonstrations showed the value of the work. First, we processed 10-years' worth of EHR and showed changes over time in phenotypical behaviors expressed and matched to DSM criteria by children with and without ASD as recorded in the EHR. Second, we developed a prototype interface showing the

potential of these algorithms for clinicians in diagnosing or reviewing records as well as the potential for time- and cost-effective surveillance.

5.6 Significance and Implications

The significance of our work lies in the algorithms created as well as in the demonstration of their potential for surveillance efforts, new research, and earlier diagnosing of children.

Application of our algorithm, with or without human oversight, may lead to nationwide surveillance of all relevant EHR for ASD. This would provide a major cost saving as well as much broader base for tracking ASD.

The algorithms can also be applied to extracting ASD phenotypical expressions at a large scale, which has not yet been accomplished by others. This type of data can be added to existing datasets for clinical review and data mining. This may contribute to precision medicine approaches for ASD.

Finally, our algorithms can be integrated in a user-friendly interface or accessed by API which can facilitate diagnosing of children by clinicians with limited expertise. This would improve early diagnosing and treatment of children with ASD leading to better outcomes.

5.7 Limitations

Our work has a first limitation related to the data used for the project: ADDM EHR and mostly DSM-IV. Our project, as proposed, uses EHR collected for the ADDM network which contains rich free text and which are annotated and diagnosed for DSM-IV. Further work should expand to the use of any type of EHR and switching to DSM-5.

There is a second limitation, but also opportunity, due to the data fields used so far. For this work we intentionally only used the free text in the EHR. By using additional structured fields, it is expected that all measures will improve because more information is provided to the algorithms.

6 PUBLICATIONS FROM THIS WORK

Journal Publications:

- G. Leroy, Y. Gu, S. Pettygrove, M.K. Galindo, A. Arora, and M. Kurzius-Spencer, "*Automated Extraction of Diagnostic Criteria from Electronic Health Records for Autism Spectrum Disorders: Development, Evaluation and Case Study*", Journal of Medical Internet Research (JMIR), November 2018. DOI: 10.2196/10497, PMID: 30404767, PMCID: 6249505

Conference Proceedings (Papers):

- Y. Gu and G. Leroy, *"Use of Conventional Machine Learning to Optimize Deep Learning Hyper-parameters for NLP Labeling Tasks"*, Hawaiian International Conference on System Sciences (HICSS), Maui, HI, Jan 7-10, 2020. (http://hdl.handle.net/10125/63867)

- Y. Gu and G. Leroy, "*Mechanisms for Automated Training Data Labelling for Machine Learning* ", International Conference on Information Systems (ICIS), Munich, Germany, Dec 12-15, 2019.

- Y. Gu and G. Leroy, "*Theory-driven procedures to analyze the impact of training data on deep learning when resources are scarce and expensive"*. Health Information Technology Symposium (HITS), Munich, Germany, December 15, 2019.

- Y. Gu, G. Leroy, S. Pettygrove, M. K. Galindo, and M. Kurzius-Spencer, "*Optimizing Corpus Creation for Training Word Embedding in Low Resource Domains: A Case Study in Autism Spectrum Disorder (ASD)*", AMIA Annual Symposium (AMIA), San Francisco, CA, November 3-7, 2018. PMID: 30815091 PMCID: PMC6371367.

Conference Proceedings (Posters):

- Y. Gu, G. Leroy, M. Surdeanu, and S. Pettygrove, (Poster) *"Case Status Classification for Mental Health with Electronic Health Records"*. Workshop on Information Technologies and Systems (WITS), Santa Clara, CA, December 16-18, 2018.

Presentations:

- Y. Gu and G. Leroy. (Poster). *"Large-scale Text Analysis of Electronic Health Records for Autism Spectrum Disorder",* Women in Data Science Tucson Conference (WiDS-Tucson), online, April 17, 2020.

- Y. Gu, G. Leroy. *"Automated Training Data Discovery and Labelling for Machine Learning in a Low Resource Domain*", INFORMS General Annual Meeting (INFORMS), Seattle, WA, Oct 22-25, 2019.

- Y. Gu and G. Leroy, "*Classification with Feature and Algorithm Machine Learning Ensembles for Autism Spectrum Disorders"*, Conference on Health IT and Analytics (CHITA), Washington DC, November 14-15, 2019.

- Y. Gu and G. Leroy, "*A Classification Artifact to Support Mental Health Surveillance: A Comparison of Feature and Classifier Ensembles*", Workshop on Information Technology and Systems (WITS), Santa Clara, December 2018.

Work in Progress:

- Data-driven Estimations of the Predictions of Deep Learning for Text in Low-Resource Domain. Targeting: Management Information Systems Quarterly (MISQ), 2021 submission.
- A Systematic Evaluation of Systems and Data Sources for Automatic Training Label Creation for Machine Learning. Targeting: ACM Transaction on Management Information Systems (ACM TMIS) or Journal of the American Medical Informatics Associations (JAMIA), Spring/Summer 2021 submission.
- Case Status Prediction for Autism Spectrum Disorder based on Electronic Health Records: Features, Ensembles, and Population Biases. Targeting: ACM Transaction on Management Information Systems (ACM TMIS), Spring/Summer 2021 submission.

7. BIBLIOGRAPHY

1.   Tyler CV, Schramm S, Karafa M, Tang AS, Jain A. Electronic Health Record Analysis of the Primary Care of Adults With Intellectual and Other Developmental Disabilities. Journal of Policy and Practice in Intellectual Disabilities. 2010;7(3):204-10. PMID: 26113866.
2.   John PT, Johnson SA, Poon EG, Fiski J, Rao SR, Cleave JV, et al. Electronic Health Record Decision Support and Quality of Care for Children With ADHD. Pediatrics. 2010;126(2):239-46. PMID: 20643719.
3.   Lingren T, Chen P, Bochenek J, Doshi-Velez F, Manning-Courtney P, Bickel J, et al. Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. PLOS ONE. 2016;July:1-16. PMID: 27472449.
4.   Luo SX, Shinall JA, Peterson BS, Gerber AJ. Semantic mapping reveals distinct patterns in descriptions of social relations in adults with autism spectrum disorder. Autism Research. 2016;9(8):846-53. PMID: 26613541
5.   Aramaki E, Shikata S, Miyabe M, Usuda Y, Asada K, Ayaya S, et al. Understanding the Relationship between Social Cognition and Word Difficulty. A Language Based Analysis of Individuals with Autism Spectrum Disorder. Methods Inf Med 2015;54(6):522-9. PMID 26391807.
6.   FitzHenry F, Murff HJ, Matheny ME, Gentry N, Elliot M. Fielstein, Brown SH, et al. Exploring the Frontier of Electronic Health Record Surveillance. Medical Care. 2013;6(509-516).
7.   Iqbal E, Mallah R, Jackson RG, Ball M, Ibrahim ZM, Broadbent M, et al. Identification of Adverse Drug Events from Free Text Electronic Patient Records and Information in a Large Mental Health Case Register. PLOS ONE. 2015;10(8):e0134208.
8.   Denny JC, Choma NN, Peterson JF, Miller RA, Bastarache L, Li M, et al. Natural Language Processing Improves Identification of Colorectal Cancer Testing in the Electronic Medical Record. Medical Decision Making. 2012;Jan-Feb:188-97.

9.   Dillahunt-Aspillaga C, Finch D, Massengale J, Kretzmer T, Luther SL, McCart JA. Using Information from the Electronic Health Record to Improve Measurement of Unemployment in Service Members and Veterans with mTBI and Post-Deployment Stress. PLOS ONE. 2014:1-21. PMID: 25541956.

10.   DSM-IV TFo. Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition, Text Revision): DSM-IV-TR. Arlington, VA: American Psychiatric Association; 2000.

11.   Lang R, Hancock TB, Singh NN. Overview of Early Intensive Behavioral Intervention for Children with Autism. arly Intervention for Young Children with Autism Spectrum Disorder: Springer International Publishing; 2016. p. 1-14.

12.   Baio J, Wiggins L, Christensen DL, Maenner MJ, Daniels J, Warren Z, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. Surveillance Summaries. 2018;67(6):1-23. https://www.cdc.gov/mmwr/volumes/67/ss/ss6706a1.htm.

13.   Abney S, Schapire RE, Singer Y, editors. Boosting Applied to Tagging and PP Attachment. Empirical Methods in Natural Language Processing and Very Large Corpora; 1999.

14.   Centers for Disease Control and Prevention. Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network 2006. United States: 2009.

15.   CDC I. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. Morbidity and mortality weekly report Surveillance summaries. 2014;63(2):1-21.

16.   Maenner MJ, Shaw KA, Baio J, Washington A, Patrick M, DiRienzo M, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. Surveillance Summaries 2020;69(4):1-12. Epub March 27, 2020

17.   Cunningham H, Maynard D, Bontcheva K, Tablan V, editors. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02); 2002 July; Philadelphia.

18.   Sheffield Natural Language Processing Group. General Architecture for Text Engineering 3.0 ed.: Sheffield, UK: http://gate.ac.uk/; 2005.

19.   Manning C, Surdeanu M, Bauer J, Finkel J, J. Bethard S, McClosky D, editors. The Stanford CoreNLP Natural Language Processing Toolkit. Annual meeting of the ssociation for computational linguistics; 2014.

20.   Frosst N, Hinton G. Distilling a Neural Network Into a Soft Decision Tree2017.

21.   Weiss G, Goldberg Y, Yahav E. Extracting Automata from Recurrent Neural Networks Using Queries and Counterexamples. arXiv preprint arXiv:171109576. 2017.

22.   Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2012;12.

23.   Surdeanu M, Manning CD, editors. Ensemble models for dependency parsing: cheap and good? Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2010: Association for Computational Linguistics.

24.   Leroy G, Gu Y, Pettyrgrove S, Kelly-Galindo M, Arora A, Kurzius-Spencer M. Automated Extraction of Diagnostic Criteria from Electronic Health Recordes for Autism Spectrum Disorders: Development, Evaluation and Case Study. Journal of Medical Internet Research (JMIR). 2018;20(11).

25.   Chollet F. Keras. https://en.wikipedia.org/wiki/Keras2015.

26.   Team EDjD. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0. Available from: http://deeplearning4j.org.

27.   Gu Y, Leroy G, editors. Use of Conventional Machine Learning to Optimize Deep Learning Hyper-parameters for NLP Labeling Tasks. Proceedings of the 53rd Hawaii International Conference on System Sciences; 2020.

28.   Gu Y, Leroy G. Mechanisms for Automatic Training Data Labeling for Machine Learning. International Conference on Information Systems (ICIS); December 2019; Munich, Germany2019.