Title: Prediction of autism spectrum disorder diagnosis using nonlinear measures of language-related EEG at 6 and 12 months

Authors: Fleming C. Peck^{1,2,3}, Laurel J. Gabard-Durnam^{1,4†}, Carol L. Wilkinson^{1†*}, William Bosl^{5,6}, Helen Tager-Flusberg⁷, Charles A. Nelson^{1,8}

[†]Authors made equal contributions to this work, listed alphabetically

Affiliations:

¹Division of Developmental Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA.

²Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA.

³Princeton Neuroscience Institute, Princeton University, Princeton NJ 08544, USA.

⁴Department of Psychology, Northeastern University, Boston, MA 02118, USA.

⁵Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA

⁶Health Informatics Program, University of San Francisco, San Francisco, CA 94117, USA.

⁷Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215, USA

⁸Harvard Graduate School of Education, Cambridge, MA 02138, USA

*To whom correspondence should be addressed: Carol Wilkinson
(Carol.Wilkinson@childrens.harvard.edu)

Abstract: Early identification of autism spectrum disorder (ASD) provides an opportunity for early intervention and improved outcomes. Use of electroencephalography (EEG) in infants has shown promise in predicting later ASD diagnoses and in identifying neural mechanisms underlying the disorder. Given the high co-morbidity with language impairment in ASD, we and others have speculated that infants who are later diagnosed with ASD have altered language learning, including phoneme discrimination. Phoneme learning occurs rapidly within the first postnatal year, so altered neural substrates either during or after the first year may serve as early, accurate indicators of later autism diagnosis. Using longitudinal EEG data collected during a passive phoneme task in infants with high familial risk for ASD, we compared predictive accuracy at 6-months (during phoneme learning) versus 12-months (after phoneme learning). Samples at both ages were matched in size and diagnoses (n=14 with later ASD; n= 40 without ASD). Using Pearson correlation feature selection and support vector machine with radial basis function classifier, 100% predictive diagnostic accuracy was observed at both ages. However, predictive features selected at the two ages differed and came from different scalp locations. We also report that performance across multiple machine learning algorithms was highly variable and declined when the 12-month sample size and behavioral heterogeneity was increased. These results demonstrate that speech processing EEG measures can facilitate earlier identification of ASD but emphasize the need for age-specific predictive models with large sample sizes in order to develop clinically relevant classification algorithms.

Introduction

The past decade has witnessed a dramatic increase in the prevalence of autism spectrum disorder (ASD), a neurodevelopmental disorder that is characterized by social communication and repetitive and restrictive behaviors (1). The Centers for Disease Control and Prevention estimate that one in 54 children has an ASD diagnosis (2). Currently, ASD is diagnosed using behavioral measures, so a diagnosis cannot be made until toddlerhood or later when behavioral symptoms are reliably observable (3). However, studies show that introducing intervention to children earlier in development leads to better intellectual and behavioral outcomes (4, 5). Early identification of ASD is essential to early intervention. Therefore, a central focus for the field has been to develop identification tools using biological markers to facilitate earlier detection of ASD.

Neuroimaging measures provide strong candidate tools for early identification as they can be obtained from prenatal ages onwards. For example, several recent studies have used magnetic resonance imaging (MRI) data collected in infancy to predict ASD diagnoses (6, 7). However, MRI has several notable drawbacks that make it less feasible as a general screening tool: it is expensive, infants must be scanned while asleep, and it is unsafe for certain populations (e.g., those with metal implants). Electroencephalography (EEG), on the other hand, may prove to be a more scalable tool, given its low cost, ease of acquisition in awake or sleeping infants, and the fact that there are no exclusion criteria for its use. Moreover, EEG is known to be sensitive to brain-related changes in ASD before behavioral symptoms are observable (8-12). Initial efforts to predict ASD diagnoses using baseline (i.e. resting-state) EEG early in life have shown promise (13-16). However, ASD diagnostic prediction using EEG recorded during tasks related to ASD

symptoms has not yet been attempted, and it is possible that task-related EEG measures outperform prior baseline EEG-based classification.

In particular, language is frequently delayed or impaired in ASD (17-21). Theories of delays in language acquisition associated with ASD postulate that atypical peak synaptic sensitivity (22) and cortical excitatory and inhibitory imbalance (23) disrupt the neural circuits that typically allow for language development. Therefore, focusing on the brain's electrical activity during language processing may facilitate improved diagnostic prediction accuracy relative to baseline conditions. Notably, EEG has been used to measure differences in language processing in children with ASD who are older than 12 months (24-26), providing evidence that EEG is sensitive to atypical neural processing of language stimuli in ASD.

In this context of predicting ASD diagnosis, focusing on language acquisition within the first 12 months may yield the earliest indicators. One such critical stage in language acquisition is learning to discriminate native from non-native speech sounds, which occurs during a sensitive period over the first year of life. That is, very young infants can discriminate between native and non-native phonemes better than adults, but they lose this ability over the first year of life as their phoneme perception is tuned to the language(s) that they experience in their daily lives (27). This perceptual narrowing of phoneme discrimination is necessary for speech perception and language learning (28, 29) and may be processed differently in infants with ASD (30-32). Therefore, this study focused on the phoneme learning window over the first year of life as a potential source of early task-related EEG indicators of subsequent ASD diagnosis.

There were two overarching goals of the present study. First, we aimed to evaluate whether EEG data collected during a language phoneme task at either 6- or 12-months of age in infants with

familial risk for ASD can accurately predict later ASD diagnosis. To do this, we utilized EEG data collected from high familial risk infant siblings as part of a prospective longitudinal study, where diagnosis of ASD was determined at 2-3 years of age. Specifically, power and nonlinear features of the EEG were tested as potential predictors of subsequent diagnosis. Though power analysis of EEG is most common, electrical activity recorded from the brain has complex dynamical properties that power analysis is not able to quantify. Moreover, these nonlinear features in adult EEG have accurately classified other clinical conditions, including depression (33-35), schizophrenia (36-39), and epilepsy (40-42). Therefore, a combination of nonlinear measures that quantify signal complexity and regularity can describe different properties of the neural time series and may improve ASD prediction. Our lab has previously found that these measures computed from resting state EEG are useful in predicting ASD outcome (15). Second, given that expected perceptual narrowing of phoneme discrimination occurs between 6 and 12 months of age, we next aimed to identify and compare the EEG features most predictive of diagnosis at each age and determine whether there are developmental differences in which features are most important during versus after the language phoneme learning period. To do this, 6- and 12-month datasets were matched in number as well as demographic and ASD outcome characteristics.

Results

Participants and data characteristics

Three sets of EEG data were evaluated in predictive models: 6-month sample, full 12-month sample, and a matching 12-month sample. The 6-month sample included a total of 54 high-risk infants, 40 of whom did not meet criteria for ASD (HR-NoASD) and 14 of whom were later

diagnosed with ASD (HR-ASD). The full 12-month sample included a total of 67 high-risk infants (40 HR-noASD; 27 HR-ASD). In order to more directly compare 6- and 12-month data sets, we created a “matching” 12-month dataset with the same ratio of HR-noASD and HR-ASD infants (40 HR-NoASD vs 14 HR-ASD). These matched datasets included EEG from 39 infants who provided data at both 6- and 12-month times points (24 HR-noASD; 13 HR-ASD). 12-month EEG data from one additional HR-ASD infant was included in this matched dataset and was selected a priori based on matching demographic information of the one 6-month participant who developed ASD but did not participate at 12 months. Demographic characteristics of the infants in all three groups are provided in Table 4.

Autism prediction at 6 months

We evaluated the ability of multiple feature selection methods and machine learning algorithms to predict ASD diagnosis. Three distinct feature selection methods were included: selection based on (1) the features most correlated with autism outcomes (Pearson correlation coefficient), (2) the features with most significant F ratio of mean square variances by group (F-test), and (3) the features selected using recursive feature elimination with cross validation based on a linear support vector machine (recursive feature elimination). We evaluated the prediction ability of four machine learning algorithms with different classification strategies: support vector machine (SVM) with radial basis function (discriminative classifier with a hyperboundary), k -nearest neighbors (voting method), linear discriminant analysis (linear combination of features), and Naïve Bayes (probabilistic classifier). Each of these methods uses a different procedure to select features or model outcomes, so we anticipated variability in the predictive capacity with different combinations.

1
2 Considering the sample size and imbalanced nature of the dataset, nested leave-one-out cross
3 validation was used to evaluate the ability of power and nonlinear measures of language-related
4 EEG signal to predict later ASD diagnosis. At each iteration of cross validation, one participant's
5 data was removed from the dataset, features were selected based on the properties of the
6 remaining samples, and the machine learning model trained using those predictive features from
7 the remaining samples to predict the outcome of the left-out sample. Thus, the feature selection
8 procedures were repeated entirely at every iteration without dependence on the left-out sample.
9 Models were limited to 20 features to prevent over-fitting to the available data samples given the
10 sample size. Table 1 shows the classification results for each of the feature selection methods and
11 classifiers detailed above. Using 20 features selected by the Pearson correlation ranking method,
12 the SVM classifier achieved 100% diagnostic prediction accuracy. Given the imbalanced nature
13 of the sample, permutation testing was used to assess the significance of observed prediction
14 accuracy. A null distribution of predictive accuracy was generated by repeating the following
15 procedure over 1000 iterations: diagnostic labels were shuffled and cross validation analysis
16 using features selected with true labels was performed. In the context of this null distribution,
17 predictive accuracy using true labels was significantly better than chance (z-score 7.35, p
18 <0.0001). Classification performance varied across combinations of the three feature selection,
19 and the four machine learning methods, with an overall accuracy range of 20.4% to 100%. SVM
20 performance was the most variable of all machine learning algorithms considered when used
21 with different feature selection methods.
22
23

Table 1

		Feature selection method				Pearson correlation coefficient				F-test				Recursive feature elimination			
		Machine learning algorithm															
		SVM	KNN	LDA	NB	SVM	KNN	LDA	NB	SVM	KNN	LDA	NB	SVM	KNN	LDA	NB
Dataset	Evaluation metrics (%)																
6 month	PPV	100	0	27.3	50	23.1	55.6	21.6	83.3	22.2	66.7	13	30.3				
	NPV	100	73.6	74.4	75	0	80	0	81.3	72.2	76.5	64.5	81				
	SENS	100	0	21.4	7.1	85.7	35.7	78.6	35.7	28.6	14.3	21.4	71.4				
	SPEC	100	97.5	80	97.5	0	90	0	97.5	65	97.5	50	42.5				
	ACC	100	72.2	64.8	74.1	22.2	75.9	20.4	81.5	55.6	75.9	42.6	50				
Matching 12 month	PPV	100	60	29.4	57.1	43.8	50	21.7	24.5	23.5	27.3	25	31				
	NPV	100	77.6	75.7	78.7	81.6	75	50	60	73	74.4	73.3	80				
	SENS	100	21.4	35.7	28.6	50	7.1	71.4	85.7	28.6	21.4	42.9	64.3				
	SPEC	100	95	70	92.5	77.5	97.5	10	7.5	67.5	80	55	50				
	ACC	100	75.9	61.1	75.9	70.4	74.1	25.9	27.8	57.4	64.8	51.9	53.7				
Full 12 month	PPV	11.1	41.7	15.4	9.1	100	47.1	33.3	36.8	12.8	23.5	35.5	31.3				
	NPV	0	60.5	43.9	44.4	60.6	62	57.7	58.3	21.4	42.4	55.6	51.4				
	SENS	18.5	37	14.8	7.4	3.7	29.6	18.5	25.9	18.5	29.6	40.7	37				
	SPEC	0	65	45	50	100	77.5	75	70	15	35	50	45				
	ACC	7.5	53.7	32.8	32.8	61.2	58.2	52.2	52.2	16.4	32.8	46.3	41.8				

Table 1: Evaluation metrics for machine learning classification results reported in percentages.

All three datasets are reported: 6-month dataset (n = 54; 74.1% HR-noASD), matching 12-month dataset (n = 54; 74.1% HR-noASD), and full 12-month dataset (n = 67; 59.7% HR-noASD). Machine learning classification was evaluated in terms of positive predictive value (PPV), negative predictive value (NPV), sensitivity (SENS), specificity (SPEC), and accuracy (ACC). Evaluation metrics are included for four algorithms: nonlinear support vector machine (SVM), K-nearest neighbors with 7 neighbors (KNN), linear discriminant analysis (LDA), and Naïve Bayes (NB). 20 features were selected by the noted selection methods before nested leave-one-out cross validation using each of the listed

1
2
3

Autism prediction at 12 months: matched sample

In order to compare ASD prediction accuracy at 6 months (during the phoneme learning period) and 12 months (the end of the phoneme learning period), and to compare the nature of the features selected for diagnosis prediction at these two ages, we next analyzed our “matched” 12-month dataset with the same total number of infants and the same proportion of HR-noASD and HR-ASD infants as the 6-month dataset. We evaluated the same feature selection methods and classifiers as in the previous analysis, and the nested cross validation results are reported in Table 1. 100% classification accuracy was achieved using feature selection based on correlation to ASD outcomes and an SVM classifier—the same feature selection method and machine learning algorithm that achieved the most successful prediction accuracy at 6-months. Evaluation metrics based on features selected by the Pearson correlation coefficient were higher than the other two feature selection methods for all evaluation metrics except sensitivity across all machine learning algorithms. Overall, classification accuracies ranged from 25.9 to 100%. Using permutation testing described above, the 100% classification accuracy was significantly better than chance (z-score 7.5, $p < 0.0001$). As with the 6-month prediction results, sensitivity rates were generally lower across models than specificity rates.

Autism prediction at 12 months: full sample

To test the prediction consistency across samples, we evaluated the prediction accuracy of models in the full 12-month dataset, which had 13 additional HR-ASD participants (40 HR-noASD and 27 HR-ASD) who only participated at 12 months of age, almost doubling the HR-ASD sample. The classification results are presented in Table 1. There was a considerable decrease in almost all evaluation metrics and especially overall prediction accuracy from the

previous two classification reports. Only one of the 12 classification schemes presented in Table 1—F-test feature selection with SVM algorithm—resulted in accuracy marginally above chance (61.2%) at the severe expense of sensitivity (3.7%), the measure evaluating the percentage of infants with ASD who were predicted correctly.

Sensitivity was the lowest evaluation metric for the full 12-month dataset for nearly all of the classification schemes that we evaluated. Like the 6-month and matched 12-month analyses, sensitivity was almost always lower than specificity. However, the poor prediction across models resulted from poor performance across all evaluation metrics, rather than a consistent under- or over-prediction of ASD diagnosis.

To investigate possible reasons for the differences in classification accuracy between the matched and full 12-month analyses, we assessed whether there were differences in behavioral measures of the HR-ASD participants. We evaluated five different behavioral assessments and used a *t*-test to compare performance of HR-ASD infants included in the matched data set versus HR-ASD only included in the 12-month full dataset (Table 2). We found that the 12-month participants who participated at the 6-month timepoint had significantly lower severity scores on the Autism Diagnostic Observation Schedule (ADOS) at 36 months compared to the non-matching 12-month infants who only contributed EEG data at 12 months of age. In addition, there were no significant differences between samples in terms of EEG length or quality using HAPPE metrics (all $p > 0.1$).

Table 2

	Full HR-ASD	Matching HR-ASD	Nonmatching HR-ASD	p-value
Behavioral assessments (mean $\pm$ SD (n))				
ADOS				
24mo Severity Score	4.81 $\pm$ 2.5 (25)	4.21 $\pm$ 2.29 (14)	5.81 $\pm$ 2.79 (11)	0.15
36mo Severity Score	4.9 $\pm$ 2.19 (23)	3.85 $\pm$ 1.68 (13)	6.3 $\pm$ 2.93 (10)	0.006
MSEL	n = 23	n = 12	n = 11	
Composite Scaled Score	97.5 $\pm$ 15.6	101.5 $\pm$ 13.9	93.3 $\pm$ 16.9	0.21
Verbal Developmental Quotient	88.9 $\pm$ 21.8	92.7 $\pm$ 21.5	84.7 $\pm$ 22.3	0.39
Non-verbal Developmental Quotient	114.9 $\pm$ 12	118.7 $\pm$ 10.4	110.7 $\pm$ 14.4	0.14

Table 2: Behavioral assessments of 12-month HR-ASD group by participation timepoints. P-value of t-test comparing scores of each behavioral assessment between the matching and nonmatching 12-month HR-ASD infants (significant p-value is emboldened). ADOS: Autism Diagnostic Observation Schedule. MSEL: Mullen Scales of Early Learning.

1 *Predictive features at 6 and 12 months*

2 Finally, the nature and spatial distribution of features selected in the successful 6-month and 12-

3 month predictive models were extracted in order to compare EEG features most predictive of

4 ASD diagnosis either during (6 months of age) or after (12 months of age) perceptual narrowing

5 of phoneme discrimination. Importantly, the same feature selection method and machine learning

6 algorithm (Pearson correlation coefficient feature selection and an SVM classifier) achieved

7 100% predictive accuracy for both the 6-month and matched 12-month datasets, allowing for

8 direct comparison of the features selected at each iteration of nested cross validation across the

9 two ages. Given poor prediction of full 12-month sample, we do not report the features for this

10 sample.

Figure 1 shows the selection rates of features by channel, measure, and frequency. At 6 months, features were selected largely from central and left of the midline locations (Fig 1A), and power was the most frequently selected measure (Fig 1B). Five of the 12 measures (power, approximate entropy, Hurst exponent, Lempel-Ziv complexity, and permutation entropy) were consistently

Fig. 1. Information about features most correlated with autism diagnostic outcome for nearly overlapping 6- and 12-month analyses (n = 54). The bottom row visualizes the values for the 12-month dataset (middle row) minus the 6-month dataset (top row). (**A, D, G**) Average number of features selected from each channel. Color indicates number of features selected from a given channel. (**B, E, H**) Average count of each EEG measure across iterations (orange) and percentage of iterations that each measure was selected at least once (blue). (**C, F, I**) Average count of each wavelet across iterations (orange) and percentage of iterations that each wavelet was selected at least once (blue).

selected across iterations. Within iterations, power was the most frequently selected feature and made up the majority of the 20 selected features. While measures related to all frequency bands were consistently chosen across iterations, the average count per iteration of features related to the 15.6-31.2 Hz range (largely canonical Beta frequencies) was nearly double any other frequency range (Fig 1C). At 12 months, selected channels changed most in the left hemisphere, with increased left-lateralization (i.e. shifts away from midline) and representation from especially dense frontal and temporo-parietal regions (Fig 1D). Seven of the 12 measures were selected in at least 80% of the iterations, and half of the selected measures in each iteration were power or Lyapunov exponent computed at different wavebands and channels (Fig 1E). Additionally, while almost all frequency bands were selected at each iteration, signal filtered to frequency ranges below 16 Hz were more frequently selected at age 12 months (Fig 1F).

We next compared the mean value of each of the 20 most frequently selected features between outcome groups (HR-ASD and HR-noASD; Table 3). After correcting for multiple comparisons, only approximate entropy computed from the F3 electrode in the delta range (1-4Hz) was significantly different between the two groups at 6-months, with HR-ASD demonstrating higher entropy than the HR-noASD infants (HR-ASD: mean 0.815 ± 0.08 ; HR-noASD: mean 0.76 ± 0.037 ; $p < 0.0025$). In contrast, at 12 months of age, the mean values measured for seven of the 20 features most commonly chosen across iterations were significantly different between HR-ASD and HR-noASD groups such that HR-ASD infants had consistently higher values for each of these features than HR-noASD infants. Significant features across model iterations were also those that were most often chosen within a given iteration's set of 20 features (Fig 1E): Lyapunov exponent, Hurst Exponent, sample entropy, and power. Lempel-Ziv complexity was

TABLE 3

	Feature			HR-ASD (mean $\pm$ std)	HR-noASD (mean $\pm$ std)	p-value	% Iterations selected
	Channel	Frequency	Measure				
6-month dataset	F3	1-3.9 Hz	AppE	0.815 $\pm$ 0.08	0.76 $\pm$ 0.037	<0.0025	100
	C3	1-3.9 Hz	Power	0.038 $\pm$ 0.019	0.025 $\pm$ 0.012	<0.05	100
	C3	7.8 - 15.6 Hz	AppE	1.145 $\pm$ 0.035	1.115 $\pm$ 0.033	<0.05	100
	Fz	15.6-31.2 Hz	Power	0.018 $\pm$ 0.008	0.013 $\pm$ 0.006	<0.05	100
	Fz	3.9-7.8 Hz	Power	0.017 $\pm$ 0.007	0.012 $\pm$ 0.005	<0.01	100
	Pz	15.6-31.2 Hz	Power	0.029 $\pm$ 0.016	0.02 $\pm$ 0.009	<0.01	100
	C3	3.9-7.8 Hz	Power	0.026 $\pm$ 0.013	0.018 $\pm$ 0.009	<0.05	98.1
	F3	7.8 - 15.6 Hz	LZW	1.59 $\pm$ 0.027	1.563 $\pm$ 0.036	<0.05	98.1
	Fz	1-3.9 Hz	Power	0.032 $\pm$ 0.017	0.022 $\pm$ 0.01	<0.05	98.1
	Fz	7.8 - 15.6 Hz	Power	0.014 $\pm$ 0.007	0.01 $\pm$ 0.004	<0.05	98.1
	Pz	31.2-62.2 Hz	Power	0.028 $\pm$ 0.014	0.019 $\pm$ 0.009	<0.01	98.1
	Pz	7.8 - 15.6 Hz	Power	0.026 $\pm$ 0.017	0.017 $\pm$ 0.007	<0.01	98.1
	C3	7.8 - 15.6 Hz	Power	0.022 $\pm$ 0.01	0.016 $\pm$ 0.007	<0.05	96.3
	F3	7.8 - 15.6 Hz	PermE	2.575 $\pm$ 0.007	2.566 $\pm$ 0.012	<0.05	96.3
	O1	62.5-100 Hz	HurstE	0.137 $\pm$ 0.027	0.117 $\pm$ 0.025	<0.05	92.6
	Pz	3.9-7.8 Hz	Power	0.032 $\pm$ 0.016	0.023 $\pm$ 0.01	<0.05	92.6
	Fz	31.2-62.2 Hz	Power	0.018 $\pm$ 0.014	0.011 $\pm$ 0.006	<0.05	90.7
	F8	3.9-7.8 Hz	AppE	0.795 $\pm$ 0.038	0.763 $\pm$ 0.048	<0.05	50
	FP1	1-31.2 Hz (Scale 3)	MSE	2.158 $\pm$ 0.058	2.11 $\pm$ 0.072	<0.05	48.1
	FP1	1-15.6 Hz (Scale 4)	MSE	2.135 $\pm$ 0.085	2.084 $\pm$ 0.068	<0.05	48.1
Matching 12-month dataset	F7	7.8 - 15.6 Hz	Power	0.031 $\pm$ 0.015	0.02 $\pm$ 0.008	<0.001	100
	F7	3.9-7.8 Hz	Power	0.044 $\pm$ 0.028	0.022 $\pm$ 0.011	<0.001	100
	F7	1-3.9 Hz	Power	0.07 $\pm$ 0.055	0.032 $\pm$ 0.015	<0.001	100
	T3	1-3.9 Hz	Power	0.107 $\pm$ 0.085	0.049 $\pm$ 0.033	<0.001	100
	Pz	31.2-62.2 Hz	HurstE	0.49 $\pm$ 0.06	0.418 $\pm$ 0.052	<0.001	100
	F4	62.5-100 Hz	LyapE	0.119 $\pm$ 0.009	0.112 $\pm$ 0.006	<0.0025	100
	O1	15.6-31.2 Hz	SampE	2.2 $\pm$ 0.038	2.139 $\pm$ 0.065	<0.0025	100
	F8	62.5-100 Hz	LyapE	0.116 $\pm$ 0.007	0.111 $\pm$ 0.005	<0.01	100
	Fz	15.6-31.2 Hz	LZC	1.547 $\pm$ 0.021	1.53 $\pm$ 0.017	<0.01	100
	F7	31.2-62.2 Hz	Power	0.039 $\pm$ 0.028	0.023 $\pm$ 0.013	<0.01	98.1
	F7	15.6-31.2 Hz	Power	0.037 $\pm$ 0.019	0.024 $\pm$ 0.014	<0.01	98.1
	Fz	7.8 - 15.6 Hz	LyapE	0.049 $\pm$ 0.006	0.043 $\pm$ 0.008	<0.01	98.1
	O2	3.9-7.8 Hz	HurstE	0.563 $\pm$ 0.102	0.475 $\pm$ 0.098	<0.01	98.1
	T3	3.9-7.8 Hz	Power	0.064 $\pm$ 0.049	0.036 $\pm$ 0.024	<0.01	98.1
	T5	1-3.9 Hz	LyapE	0.04 $\pm$ 0.008	0.031 $\pm$ 0.011	<0.01	98.1
	O1	15.6-31.2 Hz	AppE	1.469 $\pm$ 0.029	1.448 $\pm$ 0.024	<0.01	94.4
	P3	15.6-31.2 Hz	LyapE	0.054 $\pm$ 0.005	0.049 $\pm$ 0.006	<0.05	83.3
	F7	15.6-31.2 Hz	PermE	2.583 $\pm$ 0.002	2.58 $\pm$ 0.004	<0.05	81.5
	T6	3.9-7.8 Hz	HurstE	0.539 $\pm$ 0.101	0.453 $\pm$ 0.115	<0.05	53.7
	Pz	31.2-62.2 Hz	SVDE	1.564 $\pm$ 0.017	1.546 $\pm$ 0.026	<0.05	44.4

Table 3. Descriptions of the 20 most-frequently selected features using the Pearson Correlation Coefficient features selection method. Scale in the frequency band feature category refers to the level of coarse-graining procedure (further described in Methods). Significance evaluated with paired sample t-test corrected for 20 comparisons. Parameters that survived Bonferroni correction are in bold

the only measure selected in all iterations that was not significantly different between groups after Bonferroni correction.

Discussion

In this study we evaluated multiple classification schemes to predict ASD diagnosis in high-familial risk infants using EEG data from infancy. Specifically, EEG signals collected during a language phoneme task were evaluated at either 6- or 12-months of age, timepoints that span a critical early language-learning period in development. 100% diagnostic classification accuracy was achieved for both 6- and matched 12-month data sets using the Pearson correlation coefficient feature selection method with the SVM classifier. That is, the same prediction rate was achieved from phoneme-related EEG regardless of whether infants were within the critical period of language phoneme learning (6 months) or after (12 months). However, the features selected to achieve the 100% prediction rate were quite different between the two ages, both in feature type and spatial distribution. Importantly, across classification schemes at both ages, model performance was highly variable and notably reduced when the sample demographics and size changed. Although we used robust statistical methods to limit overfitting in small samples, these constrained models may be unable to fit the variability present in ASD. The implications of these results in the search for early neuroimaging biomarkers of ASD diagnosis are discussed below.

Developmental shift in predictive features

The longitudinal nature of the study provides the opportunity to assess whether predictive EEG features change across the course of a critical early period of language-learning within the first

year of life. Several differences were identified. First, at 6 months power dominated as the most common measure selected, whereas at 12 months additional nonlinear measures were more consistently selected (Fig 1H). This trend is consistent with previous longitudinal studies that have found power differences between high- and low-risk infants that emerge before 6 months of age but often dissipate by 12 months of age, and can differentiate later ASD diagnostic outcomes (8, 10, 13, 43). We also observed differences between developmental time points topographically, with a noticeable shift leftwards from central-midline locations, including increased involvement of the temporal lobe and lateral frontal cortex at 12 months (Fig 1G). This shift may indicate that infants who go on to have an ASD diagnosis show atypical language processing in the network of left-lateralized regions involved in healthy language perception by 12 months of age (44-48). Moreover, these early differences are robust enough to indicate subsequent diagnostic outcome. Our EEG-related findings corroborate MRI findings of atypical activation related to passive auditory stimuli in adults and children with autism over similar cortical areas (44-51).

Interestingly, the way features contributed to prediction shifted across development as well. At 12 months, more individual features were significantly different between ASD outcome groups (7 features) compared to 6 months (just 1 feature). One possible explanation is different modes of classification are utilized at the two ages, with a combination of feature values taken together at 6 months and independent contributions of individual features at 12 months. For example, power differences between ASD outcome groups were larger and more often statistically significant at 12 months, though these features were selected more frequently at 6 months (Table 2). At 12 months, power computed from the F7 channel was particularly predictive: all five

frequency ranges from which power was computed were in the 20 most commonly selected features, and power was significantly different between HR-ASD and HR-noASD groups for the three frequency ranges below 32 Hz. Previous work suggests a positive correlation between language and cognitive outcomes and low gamma power during both resting-state (45, 52-55) and auditory-task EEG (44, 56, 57). However, the reduced selection of power features and features derived from high frequency signal at 12 months compared to 6 months (Figure 1) suggest that nonlinear measures of low frequency signal better define binary ASD diagnostic outcomes based on language task EEG in the first year of life.

Variable classification performance

The nested leave-one-out cross validation results varied greatly between the different feature selection method and machine learning algorithm combinations. The variability in performance may be attributed to the small sample size of the datasets, which risk overfitting or underfitting to the training data despite our efforts to minimize the dimensionality of the data before classification. Similarly-sized studies using MRI data have previously demonstrated sensitivity lower than 90% with near-perfect specificity (6, 7) of ASD diagnosis prediction in high-familial risk infants in contrast to 100% sensitivity and specificity of the best performing models based on the 6-month and size-matched 12-month datasets of this study. Previously reported MRI studies only presented results from a single classification scheme, so it is unclear whether predictive accuracy was more or less consistent across classification schemes for MRI data.

The decrease in accuracy from 100% with the matched 12-month dataset to 7% with the full 12-month dataset using a SVM classifier suggests an inability to effectively separate the two

diagnostic outcome classes after the HR-ASD group was expanded. We found behavioral phenotypes of the HR-ASD infants added to complete the full 12-month analysis to be more variable and severe than those who participated at both 6 and 12 months. It is possible that enrollment bias influenced the sample characteristics in that high-risk families enrolling at a later age may have had increased concerns about ASD related to observed symptoms. We postulate that the inclusion of more samples resulted in increased feature overlap across outcome classes, preventing accurate hyperboundary separation. Our 12-month results also suggest that more complex modeling will be required at this age to appropriately account for the full range of heterogeneity in ASD at the brain and behavioral levels. It is also possible that classification with these simpler machine learning approaches might do well at 6 months even with a bigger sample size if EEG measures tend to be less variable and better stratified between the ASD diagnostic outcomes at this younger age.

The effect of heterogeneity and sample size on prediction

Autism is a heterogeneous disorder—its defining categories are broad and encompass a spectrum of symptom severity. The ideal model for autism diagnosis prediction will be robust to the degree of variability or heterogeneity of the disorder. Our SVM-based classification results at 6 months and with matched samples at 12 months are highly accurate, but the decline in classification accuracy with increased sample size and important shifts in sample phenotypes (including ADOS severity) highlights factors that must be considered by the greater autism research community for future diagnosis prediction efforts. This within-study classification distinction serves as a case study of poor generalizability to a larger sample.

The HR-ASD infants who were included in both matched and full 12-month classification analyses had significantly different ADOS severity scores at 36 months than the HR-ASD infants who joined the study at the 12-month timepoint. On all five behavioral measures evaluated in Table 3, the 12-month-only HR-ASD group had higher ADOS scores and lower MSEL scores (corresponding to overall lower indexes of development). We hypothesize that the inclusion of a more severe and variable dataset reduced the accuracy of classification since we were predicting ASD as a binary diagnosis. Using resting-state EEG from the same infant-sibling dataset in this analysis, we have previously observed a similar lower accuracy in a 12-month sample, although a similar size sample at 9 months had high accuracy. As discussed in our previous paper, this may be a real neurodevelopmental trend (14, 15). This finding is in line with previous meta-analyses within the brain-disorder field that have found decreasing accuracies reported as sample sizes—and, importantly, heterogeneity within the sample—increase (58-60). However, this may also be a developmental neurophysiology effect that reflects the cross-over of neurodevelopmental trajectories of infants who do and do not go on to develop ASD. More sample data are needed in order to more completely represent the potentially subtle or variable brain activity differences that arise in ASD. Increased sample sizes also permit the use of more complex models that may more appropriately account for the variability and complex associations between brain activity and diagnosis.

Limitations and future directions

In addition to the discussed challenges of sample size and heterogeneity, we acknowledge several limitations of the current study. First, this analysis focused on infants with familial risk of autism, a finding that may generalize to other ASD-risk groups or to the general infant

population. Second, the specificity of our findings to ASD (versus other comorbid conditions such as language delay) is not known. Further testing across clinical populations (eg. Global developmental delay without ASD or isolated Language Delay) would be helpful in understanding the specificity of our findings and whether EEG could also be used to predict comorbidities with significant impact on functional outcomes. This study determines ASD outcome at age 3, which is appropriate for assessing ASD but not for many other developmental conditions that emerge across early childhood. Therefore, questions about comorbidity call for the extension of longitudinal studies across infancy to track participants beyond 3 years in order to capture a more complete and confident clinical description of participants. Additionally, our sample was not diverse ethnically, racially, or in income level. Predictive analyses require not only large sample sizes but must include infants from diverse populations in order to improve clinical applicability for the general population. These results suggest that collaboration across samples is critical to moving forward in developing early predictive models. Future studies of early predictive markers of ASD and other neurodevelopmental disorders need to be acutely aware of participant age, given the dramatic developmental changes in predictive feature profiles over the 6-month age window in the study. Moreover, given the variability of behavioral measures within the ASD outcome group, future studies should consider distinguishing different subpopulations or subprofiles of ASD grouped by biological presentations or phenotype profiles at the behavioral level as opposed to only binary diagnosis.

Materials and Methods

Study design and participant

1 Participants in this study were recruited to Boston Children’s Hospital/Harvard Medical School
2 and Boston University to participate in a longitudinal study of brain and behavioral development
3 of infant siblings of children with ASD. Institutional review board approval was obtained from
4 Boston University and Boston Children’s Hospital (# X06-08-0374) prior to the start of the
5 study. All infants had a gestational age of at least 36 weeks and no history of seizures, prenatal
6 drug exposure, hearing impairment, or known genetic mutation involved in neurodevelopment.
7 The infants were designated high risk (HR) for ASD based on the confirmed ASD clinical
8 diagnosis of an older sibling. A total of 104 HR infants were enrolled in the longitudinal study.
9 ASD outcome was determined using the Autism Diagnostic Observation Schedule (ADOS) in
10 conjunction with a clinical best estimate. For infants meeting criteria on the ADOS or coming
11 within 3 points of cutoffs, near the clinical diagnosis cutoff, a Licensed Clinical Psychologist
12 reviewed scores and video recordings and provided a best estimate clinical judgment of three
13 categories: typically developing, ASD, or non-spectrum disorder, such as anxiety or language
14 delay. Of the 54 HR infants who met the inclusion criteria at 6 months, 14 infants developed ASD
15 (HR-ASD) and 40 infants did not (HR-noASD). Data from 67 HR infants were included at 12
16 months (27 HR-ASD and 40 HR-noASD). 13 HR-ASD and 24 HR-noASD infants were included
17 at both 6- and 12-month timepoints. In order to assess longitudinal changes in ASD prediction at
18 different ages, a “matching” 12-month dataset was curated by including data from the HR-ASD
19 and HR-noASD infants who participated at 6 and 12 months. One HR-ASD participant from the
20 6-month cohort did not contribute data at 12 months, so a corresponding match was identified
21 based on demographic similarity to replace that case. All 40 HR-ASD 12-month infant samples
22 were included in the “matching” 12-month dataset, resulting in the same sample size as that of the
23 6-month dataset. Demographic and EEG data quality information of each outcome groups

1 contributing data is presented in Table 4. Fisher's exact test was used to evaluate significant
 2 differences of demographic information between groups. The 12-month high-risk group with ASD
 3 had significantly lower mean maternal education than the high-risk group without ASD ($p =$
 4 0.004). No other demographic measures were significantly different between groups.

TABLE 4 Sample demographics of 6- and 12-month-old participants.

	6-month dataset		12-month dataset		
	HR-ASD n = 14	HR-noASD n = 40	HR-ASD n = 27	Matching HR-ASD n = 14	HR-noASD n = 40
Sex	8 M, 6 F	19 M, 21 F	17 M, 10 F	8 M, 6 F	18 M, 22 F
Child ethnicity (%)					
Caucasian	92.9	97.5	77.8	92.9	95
Hispanic/Latinx	21.4	2.5	14.8	21.4	7.5
Asian American	0	0	3.7	0	0
African American	0	0	3.7	0	0
Multirace	7.1	2.5	14.8	7.1	7.1
Mean household income (\$1000s)	65-75	65-75	65-75	65-75	65-75
Mean maternal education (%)					
< 4-year college	28.6	27.5	37	28.6	25
= 4-year college	35.7	15	37	42.9	12.5
> 4-year college	35.7	57.5	25.9	28.5	62.5
EEG HAPPE metrics (mean [SD])					
Length of raw EEG (seconds)	701.3 [212.9]	693.6 [220.7]	763.7 [213]	735.8 [225.2]	764.8 [209.5]
Good channels (%)	93.1 [4.1]	94.2 [4.7]	92 [4.8]	91.6 [4.7]	92.8 [5.5]
Rejected components (%)	47 [13]	44.1 [13]	43.8 [10]	42.6 [8.4]	43.4 [12.2]
EEG variance retained (%)	61.6 [13.3]	64.9 [17.4]	66.5 [12.6]	70.1 [10.5]	69.6 [16]
Mean retained artifact probability	0.16 [0.05]	0.17 [0.03]	0.15 [0.05]	0.16 [0.05]	0.15 [0.04]
Median retained artifact probability	0.13 [0.06]	0.15 [0.06]	0.1 [0.07]	0.11 [0.07]	0.1 [0.06]

Behavioral assessments

The Mullen Scales for Early Learning (MSEL) was administered at each data collection visit of the longitudinal study, including 6 and 12 months. The MSEL provides an index of ability in domains including language, cognition, and motor development. Additionally, the Autism Diagnostic Observation Schedule (ADOS) was administered at 24- and 36-month visits. The MSEL composite score and verbal and non-verbal developmental quotients at 6 and 12 months, and ADOS calibrated severity scores at 24 and 36 months were used to evaluate developmental differences within the HR-ASD group.

EEG paradigm

The phoneme speech task consisted of a stream of consonant-vowel stimuli that were presented using a double-oddball paradigm. Three phonemes were included and presented based on a probability of selection: the English standard “/da/” presented 80% of the time, the native deviant “/ta/” presented 10% of the time, and the non-native deviant “da/” presented 10% of the time. A maximum of 600 stimuli were presented with a variable interstimulus period of minimum 700ms. The language task lasted on average 15 minutes. For this study, only stretches of consecutive standard native phonemes (English ‘da’) presentations were included to ensure consistent stimuli ordering across individuals and so that the features were calculated over a single stimulus condition (no oddball stimuli were analyzed in the study).

EEG data acquisition

EEG data were acquired in a dimly lit, sound-attenuated, electrically-shielded room. A research assistant was present in the room to ensure that the infant remained calm and still during the

language paradigm by blowing bubbles or presenting toys if the infant became distracted or fussy. Assistants did not engage in social interaction with the infant during task completion. EEG data were collected with either a 64-channel Geodesic Sensor Nets or a 128-channel Hydrocel Geodesic Senso Net (Electrical Geodesics, Inc (EGI), Eugene, OR, USA), using a 0.1 Hz high-pass analog filter and online rereferencing to the vertex (channel Cz) through NetStation software (EGI, Eugene, OR, USA). Impedances were kept below 100K Ω in accordance with the connected DC-coupled amplifier (Net Amps 200 or Net Amps 300, Electrical Geodesics, Inc.). Data were sampled at either 250 or 500 Hz.

The EEG data were exported from NetStation to MATLAB format (R2017A). All files were batch processed using the Batch Electroencephalography Automated Processing Platform (BEAPP) that allows for preprocessing, artifact removal, and data quality assessment by the Harvard Automated Processing Pipeline for EEG (HAPPE) with the same empirical criteria for artifact removal applied in the same way across all files in the batch (61). HAPPE is a preprocessing pipeline optimized for developmental EEG data, which has been shown to reject more artifacts and preserve more EEG signal during processing (62). A spatially distributed subset of channels providing whole-head coverage were processed through HAPPE in order to optimize artifact rejection performance given the duration and sampling rates of the EEG data. This consisted of the 10-20 montage channel equivalents for each net type and channels clustered over the bilateral temporal-parietal cortex (64-channel net—2, 3, 6, 8, 9, 11, 12, 13, 15, 16, 17, 21, 24, 25, 27, 28, 34, 37, 40, 46, 49, 50, 52, 53, 54, 57, 58, 61, 62; 128-channel net—3, 4, 9, 11, 13, 19, 20, 22, 23, 24, 27, 28, 33, 36, 40, 41, 45, 46, 47, 52, 58, 62, 70, 75, 83, 92, 96, 98, 102, 103, 104, 108, 109, 112, 117, 118, 122, 123, 124). See Figure 2 for net layouts.

- 1 EEG pre-processing consisted of the following steps for each file. A 1 Hz digital high-pass filter
- 2 and a 100 Hz low-pass filter were applied to each file, and data sampled at 500 Hz were
- 3 resampled with interpolation to 250 Hz (so all files had the same effective sampling rate). Bad
- 4 channels, including channels with high impedances or displacement during recording, were
- 5 identified using joint probability criteria in HAPPE and were removed from further processing.
- 6 Artifact removal steps utilizing HAPPE involved removal of 60 Hz electrical noise using
- 7 CleanLine's multi-taper approach (Mullen, 2012) and removal of participant artifact rejection
- 8 such as eye blinks, movement, and muscle activity using wavelet-enhanced ICA with automated

component rejection via the Multiple Artifact Rejection Algorithm (Winkler et al., 2014). The bad channels that were removed before artifact rejection were repopulated using spherical interpolation of post-artifact rejection data to reduce spatial bias in rereferencing. EEG data were then rereferenced to the average reference and mean signal detrended.

EEG data decomposition

After pre-processing with HAPPE, the middle 20 seconds of the longest stretch of consecutive standard phoneme (English ‘da’) presentations in each file were selected for analysis. A 20-second duration was selected in order to maximize the number of language-task participants for inclusion in the study while maintaining length long enough to ensure that the nonlinear time series measures could be appropriately calculated. EEG data from 10-20 channel equivalents (18 channels) for each net type were then decomposed into frequency sub-bands using a discrete wavelet transform and a coarse-graining procedure (described below).

Discrete wavelet transform localizes a signal in both the time and frequency domains, and it is a useful tool that allows analysis of narrow frequency sub-bands of the original signal thought to

TABLE 5

Wavelet level	Wavelet frequency range	Approximate EEG waveband
D1	62.5 - 100 Hz	high gamma
D2	31.3 - 62.5 Hz	low gamma
D3	15.6 - 31.3 Hz	beta
D4	7.8 - 15.6 Hz	alpha
D5	3.9 - 7.8 Hz	theta
A5	1-3.9 Hz	delta

Table 5. Wavelet decomposition frequency bands for wavelet levels associated with traditional EEG wavebands.

have biological relevance. The discrete wavelet transform is a filter that decomposes the signal into high pass (details) and low pass (approximations) sub-bands, which are down-sampled to half of the sampling rate. Note that an upper and lower bound of 100 Hz and 1 Hz, respectively, were imposed on the signal in the preprocessing signal steps before wavelet decomposition. The discrete wavelet transform was performed using the PyWavelet *dwt* function with the Daubechies 4 wavelet and symmetric signal extension on the approximate signal of the previous level of decomposition, starting with the original signal at Level 0. Table 5 lists the wavebands that correspond to frequency ranges of wavebands used in traditional EEG power spectral analyses. Only these non-overlapping frequency bands resulting from the wavelet decomposition—D1, D2, D3, D4, D5, and A5—were selected for further analysis with nonlinear measures and covered the range from 1-100 Hz.

In addition to the wavelet decomposition, a coarse-grained time series was computed from the original signal using the multiscale entropy analysis methodology introduced by Costa and colleagues (63). The coarse-graining procedure involves dividing the time series into nonoverlapping windows of size t and averaging data points within each window, resulting in a signal that is $1/t$ times the length of the original signal. Each scale corresponds to window size: the second scale coarse grains with t of two, the third with three, and so on. Five scales were computed from the signal of each electrode, and only multiscale entropy was computed from each of the sub-signals. We note that the coarse-graining procedure is equivalent to using the overlapping frequency bands obtained from the approximations in a discrete wavelet transform using a Haar basis (Bosl, 2017).

1 EEG measures

2 12 different measures were computed for each waveband from each of the 18 10-20 electrodes for
 3 each participant. The nonlinear measures quantify the complexity and regularity of the time series
 4 using different approaches, and power quantifies the amplitude of oscillations, both on the
 5 frequency bands described in Table 5. Table 6 provides a summary description of each measure
 6 and credits the software package used for computation.

TABLE 6: Descriptions of measures.

Nonlinear variable	Abbreviation	Description
NOLDS software package (64)		
Detrended fluctuation analysis	DFA	Measures long-range correlation of the physiological time series
Sample entropy	SampE	Measures the irregularity of physiological time series without self-matches
Hurst Exponent	HurstE	Measures the long-term memory processes of a time series
Lyapunov exponent	LyapE	Measures the chaotic or periodic properties of a time series
EntroPy (https://github.com/raphaelvallat/entropy)		
Permutation entropy	PermE	Measures the information content of a given time series based on probability distribution of a set of continuous points
Spectral entropy	SpecE	Measures degree of skewness in the frequency distribution
Singular value decomposition entropy	SVDE	Measures the dimensionality of a time series.
Approximate entropy	AppE	Measures the regularity of times series fluctuations
Higuchi fractal dimension	HFD	Measures self-similarity in time series using increasingly distanced samples in time
Katz fractal dimension	KFD	Measures complexity and self-similarity in time series using consecutive time points
Lempel-Ziv Complexity Reference (65)		
Lempel Ziv Complexity	LZC	Measures the randomness of finite sequences
Linear variable	Abbreviation	Description
SciPy Signal Processing (66)		
Power	Power	Measures the frequencies of oscillations in a time series.

Feature Selection Strategies

Frequently in clinical modeling, the number of features greatly outnumbers the number of samples in the dataset. In this study we included many different measures in order to provide the most comprehensive description of the EEG signal as possible, resulting in over 1,200 features to describe each participant. However, including an excessive number of features relative to the number of samples leads to model overfitting, which would decrease the model's generalizability. Therefore, feature selection, or dimensionality reduction, was used in order to retain only the most discriminative features. We evaluated three methods of feature selection using *scikit-learn* software in order to select only the most relevant, informative features: Pearson correlation coefficient, ANOVA F-value, and recursive feature elimination (RFE). The three methods are distinct in that they are based on feature relationship directly with autism outcomes, with variance associated with autism outcomes, and directly with model development, respectively. Each feature selection method was restricted to selecting 20 of the most salient features based on relatedness to class prediction criteria.

Classification strategies

Multiple classifiers were evaluated in this study, both to identify the best-performing algorithm but also to examine variance across classifier algorithms to see whether they converged on similar prediction performance. There are two main foundational types of supervised learning models, discriminative and generative, that predict the conditional probability of an outcome given input features. Generative models model the distribution of the individual classes with a joint probability distribution. Discriminative models, on the other hand, learn explicit boundaries that best separate the given classes with the conditional probability distribution. We evaluated two generative

classifiers, the Naïve Bayes algorithm and linear discriminant analysis, and one discriminative model, an SVM with radial basis functions. Additionally, we included k-nearest neighbors, which is considered a discriminative classifier as well but really employs a voting method based on the outcomes of the k nearest neighbors in feature space using a simple Euclidean distance metric. The number of neighbors was set to 7, the square root of the sample size of the 6-month and matching 12-month datasets, and we maintained a selection of 7 neighbors for the full 12-month dataset for consistency. All other models were trained using default parameters from the Python open source package *scikit-learn*. All feature selection methods and machine learning algorithms were used for each dataset analysis, resulting in evaluation of 12 different classification schemes.

Evaluation of classification strategies

Nested leave-one-out cross validation was used to test the generalizability and validity of model performance given features selected by one of the three dimensionality reduction methods. This validation method involves removing one sample from the dataset, conducting feature selection with the remaining samples to reduce the dimensionality to 20 features, training a model using these samples based on only the selected features, and using that model to predict the outcome of the sample that was removed. This process is repeated for all samples, resulting in one prediction for each sample by a model trained using all other samples. “Nesting” the feature selection method within each iteration of cross validation ensures that the left-out sample in no way contributes to the model that will be used to predict its outcome. Previous studies have found that including test samples in feature selection biases accuracies (67).

There are four possible prediction outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Evaluation of model performance in predicting the left-out cases is critical to understanding the ability of a particular algorithm to use learned properties of data to predict the outcome of unseen samples. Five performance criteria (defined below) were calculated for each classifier algorithm:

$$(1) \quad \text{Accuracy (ACC)} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$(2) \quad \text{Sensitivity (SEN)} = \frac{TP}{TP+FN}$$

$$(3) \quad \text{Specificity (SPC)} = \frac{TN}{TN+FP}$$

$$(4) \quad \text{Positive predictive value (PPV)} = \frac{TP}{TP+FP}$$

$$(5) \quad \text{Negative predictive value (NPV)} = \frac{TN}{TN+FN}$$

Characterizing Selected Features Across 6 and 12 months

To determine the average number of times characteristics of a feature (i.e. channel, frequency band, EEG measure) were selected during the nested leave-one-out procedure (Figure 1), the 20 features selected over each of the 54 iterations for 6-month and matching 12-month data sets were collated by the above characteristics, summed, and then divided by 54, the number of iterations/participants. To determine which 20 features were directly compared between outcome groups for each of the above datasets (Table 3), we identified the 20 most frequently selected features over all the iterations. Student's t-test with Bonferroni correction for multiple comparisons was used to determine if there were significant differences between mean feature values between HR-noASD and HR-ASD groups.

Acknowledgments: We thank all of the families who participated in this study. We also thank Dr. Casey Lew-Williams for his thoughtful feedback and support of this project. Finally, we thank the entire ISP study team for their role in data collection and study coordination.

Funding: Support for this work was provided by: The National Institutes of Health (R01-DC010290 to HTF and CAN; R21 DC 08637 to HTF; 1T32MH112510 to CLW), Autism Speaks (1323 to HTF), Simons Foundation (137186 to CAN) and the Princeton Fifty-Five Fund for Senior Thesis Research (FP).

Author contributions: C.A.N. and H.T.F. designed the longitudinal study. L.G.-D. and C.L.W. processed the electrophysiological data. F.C.P., C.L.W., L.G.-D., and W.B. designed and carried out the analyses. F.C.P. and W.B. authored the code. F.C.P. drafted the paper, and all authors provided critical revisions.

Competing interests: Authors have no competing interests to declare.

Data and materials availability: Python and Jupyter Notebook scripts used for computation of nonlinear measures and machine learning analysis are available. The datasets analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES:

1. American Psychiatric Association, Ed., Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, ed. 5, 2013).
2. M. J. Maenner, K. A. Shaw, J. Baio, Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network (Morbidity and Mortality Weekly Report, Surveillance Summaries, 2002).
3. J. Piven, J. T. Ellison, M. J. Zylka, Toward a conceptual framework for early brain and behavior development in autism. *Molecular Psychiatry* **22**, 1385–1394 (2017).
4. H. Cohen, M. Amerine-Dickens, T. Smith, Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Journal of Developmental & Behavioral Pediatrics* **27**, S145-S155 (2006).
5. P. Howlin, T. Charman, Autism spectrum disorders: interventions and outcome. *The SAGE handbook of developmental disorders*, 307-328 (2011).

6. R. W. Emerson, C. Adams, T. Nishino, H. C. Hazlett, J. J. Wolff, L. Zwaigenbaum, J. N. Constantino, M. D. Shen, M. R. Swanson, J. T. Elison, S. Kandala, A. M. Estes, K. N. Botteron, L. Collins, S. R. Dager, A. C. Evans, G. Gerig, H. Gu, R. C. McKinsty, S. Paterson, R. T. Schultz, M. Styner, IBIS Network, B. L. Schlagger, J. R. Pruett, Jr., J. Piven, Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Sci. Transl. Med.* **9**, eaag2882 (2017).
7. H. C. Hazlett, H. Gu, B. C. Munsell, S. H. Kim, M. Styner, J. J. Wolff, J. T. Elison, M. R. Swanson, H. Zhu, K. N. Botteron, D. L. Collins, J. N. Constantino, S. R. Dager, A. M. Estes, A. C. Evans, V. S. Fonov, G. Gerig, P. Kostopoulos, R. C. McKinsty, J. Pandey, S. Paterson, J. R. Pruett, Jr., R. T. Schultz, D. W. Shaw, L. Zwaigenbaum, J. Piven, IBIS Network, Early brain development in infants at high risk for autism spectrum disorder. *Nature* **542**, 348-351 (2017).
8. A. L. Tierney, L. Gabard-Durnam, V. Vogel-Farley, H. Tager-Flusberg, C. A. Nelson, Developmental trajectories of resting EEG power: an endophenotype of autism spectrum disorder. *PloS One* **7**, e39127 (2012).
9. C. L. Wilkinson, A. R. Levin, L. J. Gabard-Durnam, H. Tager-Flusberg, C. A. Nelson, Reduced frontal gamma power at 24 months is associated with better expressive language in toddlers at risk for autism. *Autism Res.* **12**, 1211-1224 (2019).
10. L. Gabard-Durnam, A. L. Tierney, V. Vogel-Farley, H. Tager-Flusberg, C. A. Nelson, Alpha Asymmetry in Infants at Risk for Autism Spectrum Disorders. *J. Autism. Dev. Disord.* **45**, 473-480 (2015)

11. E. V. Orekhova, M. Elsabbagh, E. J. Jones, G. Dawson, T. Charman, M. H. Johnson, EEG hyper-connectivity in high-risk infants is associated with later autism. *J. Neurodev. Disord.* **6**, 1-11 (2014)
12. G. Righi, A. L. Tierney, H. Tager-Flusberg, C. A. Nelson, Functional connectivity in the first year of life in infants at risk for autism spectrum disorder: an EEG study. *PLoS One* **9**, e105176 (2014).
13. L. J. Gabard-Durnam, C. Wilkinson, K. Kapur, H. Tager-Flusberg, A. R. Levin, C. A. Nelson, Longitudinal EEG power in the first postnatal year differentiates autism outcomes. *Nat. Commun.* **10**, 1-12 (2019).
14. W. Bosl, A. Tierney, H. Tager-Flusberg, C. A. Nelson, EEG complexity as a biomarker for autism spectrum disorder risk. *BMC medicine* **9**, 18 (2011).
15. W. J. Bosl, H. Tager-Flusberg, C. A. Nelson, EEG analytics for early detection of autism spectrum disorder: a data-driven approach. *Scientific reports* **8**, 6828 (2018).
16. A. Dickinson, M. Daniel, A. Marin, B. Gaonkar, M. Dapretto, N. McDonald, S. Jeste, Multivariate neural connectivity patterns in early infancy predict later autism symptoms. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging* (2020).
17. E. Billstedt, C. Gillberg, C. Gillberg, Autism after Adolescence: Population-based 13- to 22-year Follow-up Study of 120 Individuals with Autism Diagnosed in Childhood. *J. Autism Dev. Disord.* **35**, 351-360 (2005).
18. I. Magiati, X. W. Tay, P. Howlin, Cognitive, language, social and behavioral outcomes in adults with autism spectrum disorders: A systematic review of longitudinal follow-up studies in adulthood. *Clinical psychology review* **34**, 73-86 (2014).

19. K. Chawarska, A. Klin, R. Paul, S. Macari, F. Volkmar, A prospective study of toddlers with ASD: Short-term diagnostic and cognitive outcomes. *J. Child Psychol. Psychiatry* **50**, 1235-1245 (2009).
20. P. Howlin, L. Mawhood, M. Rutter, Autism and Developmental Receptive Language Disorder—a Follow-up Comparison in Early Adult Life. II: Social, Behavioral, and Psychiatric Outcomes. *J Child Psychol. Psychiatry* **41**, 561-578 (2000).
21. D. K. Anderson, C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, K. Welch, A. Pickles, Patterns of growth in verbal abilities among children with autism spectrum disorder. *J. Consult. Clin. Psychol.* **75**, 594-604 (2007).
22. J. M. Berger, T. T. Rohn, J. T. Oxford, Autism as the early closure of a neuroplastic critical period normally seen in adolescence. *Biological Systems, open access* **1** (2013).
23. R. Gao, P. Penzes, Common mechanisms of excitatory and inhibitory imbalance in schizophrenia and autism spectrum disorders. *Current molecular medicine* **15**, 146-167 (2015).
24. R. Ferri, M. Elia, N. Agarwal, B. Lanuzza, S. A. Musumeci, G. Pennisi, The mismatch negativity and the P3a components of the auditory event-related potentials in autistic low-functioning subjects. *Clinical Neurophysiology* **114**, 1671–1680 (2003).
25. J. Martineau, B. Garreau, S. Roux, G. Lelord, Auditory evoked responses and their modifications during conditioning paradigm in autistic children. *J. Autism Dev. Disord.* **17**, 525-539 (1987).
26. A. Bidet-Caulet, M. Latinus, S. Roux, J. Malvy, F. Bonnet-Brilhault, N. Bruneau, Atypical sound discrimination in children with ASD as indicated by cortical ERPs. *J. Neurodev. Disord.* **9** (2017).

27. L. Polka, J. F. Werker, Developmental changes in perception of nonnative vowel contrasts. *J. Exp. Psychol: Human perception and performance* **20**, 421 (1994).
28. E. Jansson-Verkasalo, T. Ruusuvirta, P. Alku, E. Kushnerenko, K. Suominen, S. Rytty, M. Luotonen, T. Kaukola, U. Tolonen, M. Hallman, Atypical perceptual narrowing in prematurely born infants is associated with compromised language acquisition at 2 years of age. *BMC neuroscience* **11**, 1-7 (2010).
29. J. F. Werker, C. E. Lalonde, Cross-language speech perception: Initial capabilities and developmental change. *Dev. Psychol.* **24**, 672 (1988).
30. Y. Minagawa-Kawai, N. Naoi, N. Kikuchi, J. I. Yamamoto, K. Nakamura, S. Kojima, Cerebral laterality for phonemic and prosodic cue decoding in children with autism. *Neuroreport* **20**, 1219-1224 (2009).
31. A. M. Seery, V. Vogel-Farley, H. Tager-Flusberg, C. A. Nelson, Atypical lateralization of ERP response to native and non-native speech in infants at risk for autism spectrum disorder. *Dev. Cogn. Neurosci.* **5**, 10-24 (2013).
32. J. J. LeBlanc, M. Fagiolini, Autism: a “critical period” disorder? *Neural plasticity* (2011).
33. B. Hosseinifard, M. H. Moradi, R. Rostami, Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Comput. Meth. Prog. Bio.* **109**, 339-345 (2013).
34. U. R. Acharya, V. K. Sudarshan, H. Adeli, J. Santhosh, J. E. Koh, S. D. Puthankatti, A. Adeli, A novel depression diagnosis index using nonlinear features in EEG signals. *European neurology* **74**, 79-83 (2015).

35. M. Bachmann, L. Päeske, K. Kalev, K. Aarma, A. Lehtmetts, P. Ööpik, J. Lass, H. Hinrikus, Methods for classifying depression in single channel EEG using linear and nonlinear signal analysis. *Comput. Meth. Prog. Bio.* **155**, 11-17 (2018).
36. M. Sabeti, S. Katebi, R. Boostani, Entropy and complexity measures for EEG signal classification of schizophrenic and control participants. *Artificial Intelligence in Medicine* **47**, 263-274 (2009).
37. R. Boostani, K. Sadatnezhad, M. Sabeti, An efficient classifier to diagnose of schizophrenia based on the EEG signals. *Expert Sys. Appl.* **36**, 6492-6499 (2009).
38. Q. Zhao, B. Hu, L. Liu, M. Ratcliffe, H. Peng, J. Zhai, L. Li, Q. Shi, Q. Liu, Y. Qi, An EEG based nonlinearity analysis method for schizophrenia diagnosis. *Biomedical Engineering* **9**, 136 (2012).
39. V. Jahmunah, S. Lih Oh, V. Rajinikanth, E. J. Ciaccio, K. Hao Cheong, N. Arunkumar, U. R. Acharya, Automated detection of schizophrenia using nonlinear signal processing methods. *Artificial Intelligence in Medicine* **100**, 101698 (2019).
40. W. Bosl, T. Loddenkemper, C. A. Nelson, Nonlinear EEG biomarker profiles for autism and absence epilepsy. *Neuropsychiatr. Electrophysiol.* **3**, 1 (2017).
41. A. Sathyanarayana, R. El Atrache, M. Jackson, A. S. Alter, K. D. Mandl, T. Loddenkemper, W. J. Bosl, Nonlinear Analysis of Visually Normal EEGs to Differentiate Benign Childhood Epilepsy with Centrottemporal Spikes (BECTS). *Sci. Rep.* **10**, 1-12 (2020).
42. P. Fergus, A. Hussain, D. Hignett, D. Al-Jumeily, K. Abdel-Aziz, H. Hamdan, A machine learning system for automated whole-brain seizure detection. *Appl. Comput. Inform.* **12**, 70-89 (2016).

43. A. R. Levin, K. J. Varcin, H. M. O’Leary, H. Tager-Flusberg, C. A. Nelson, EEG power at 3 months in infants at high familial risk for autism. *J. Neurodev. Disord.* **9**, 34 (2017).
44. S. Ortiz-Mantilla, J. A. Hämäläinen, T. Realpe-Bonilla, A. A. Benasich, Oscillatory dynamics underlying perceptual narrowing of native phoneme mapping from 6 to 12 months of age. *J. Neurosci.* **36**, 12095-12105 (2016).
45. C. Cantiani, C. Piazza, G. Mornati, M. Molteni, V. Riva, Oscillatory gamma activity mediates the pathway from socioeconomic status to language acquisition in infancy. *Infant Behav. Dev.* **57**, 101384 (2019).
46. G. Musacchia, S. Ortiz-Mantilla, N. Choudhury, T. Realpe-Bonilla, C. Roesler, A. A. Benasich, Active auditory experience in infancy promotes brain plasticity in theta and gamma oscillations. *Dev. Cogn. Neurosci.* **26**, 9-19 (2017).
47. G. Dehaene-Lambertz, The human infant brain: A neural architecture able to learn language. *Psychonomic bulletin & review* **24**, 48-55 (2017).
48. G. M. Sket, J. Overfeld, M. Styner, J. H. Gilmore, S. Entringer, P. D. Wadhwa, J. M. Rasmussen, C. Buss, Neonatal white matter maturation is associated with infant language development. *Front. Hum. Neurosci.* **13**, 434 (2019).
49. H. Gervais, P. Belin, N. Boddaert, M. Leboyer, A. Coez, I. Sfaello, C. Bathélemy, F. Brunelle, Y. Samson M. Zilbovicius, Abnormal cortical voice processing in autism. *Nat. Neurosci.* **7**, 801-802 (2004).
50. G. Lai, H. D. Schneider, J. C. Schwarzenberger, J. Hirsch, Speech stimulation during functional MR imaging as a potential indicator of autism. *Radiology* **260**, 521-530 (2011).

51. J. Liu, T. Tsang, L. Jackson, C. Ponting, S. S. Jeste, S. Y. Bookheimer, M. Dapretto,
Altered lateralization of dorsal language tracts in 6-week-old infants at risk for
autism. *Dev. Sci.* **22**, e12768 (2019).
52. Z. Gou, N. Choudhury, A. A. Benasich, Resting frontal gamma power at 16, 24 and 36
months predicts individual differences in language and cognition at 4 and 5 years. *Behav.*
Brain Res. **220**, 263-270 (2011).
53. A. R. Tarullo, J. Obradović, B. Keehn, M. A. Rasheed, S. Siyal, C. A. Nelson, A. K.
Yousafzai, Gamma power in rural Pakistani children: Links to executive function and
verbal ability. *Dev. Cogn. Neurosci.* **26**, 1-8 (2017).
54. P. Tomalski, D. G. Moore, H. Ribeiro, E. L. Axelsson, E. Murphy, A. Karmiloff-Smith,
M. H. Johnson, E. Kushnerenko, Socioeconomic status and functional brain
development—associations in early infancy. *Dev. Sci.* **16**, 676-687 (2013).
55. A. A. Benasich, Z. Gou, N. Choudhury, K. D. Harris, Early cognitive and language skills
are linked to resting frontal gamma power across the first 3 years. *Behav. Brain Res.* **195**,
215-222 (2008).
56. C. Cantiani, S. Ortiz-Mantilla, V. Riva, C. Piazza, R. Bettoni, G. Musacchia, M. Molteni,
C. Marino, A. A. Benasich, Reduced left-lateralized pattern of event-related EEG
oscillations in infants at familial risk for language and learning impairment. *NeuroImage:*
Clinical **22**, 101778 (2019).
57. A. Kolesnik, J. B. Ali, T. Gliga, J. Guiraud, T. Charman, M. H. Johnson, E. J. Jones,
Increased cortical reactivity to repeated tones at 8 months in infants with later
ASD. *Transl. Psychiatry* **9**, 1-11 (2019).

58. M. Arbabshirani, S. Plis, J. Sui, V. D. Calhoun, Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* **145**, 137-165 (2017).
59. G. Varoquaux, Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage* **180**, 68-77 (2018).
60. H. G. Schnack, R. S. Kahn, Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Frontiers in psychiatry* **7**, 50 (2016).
61. A. R. Levin, A. S. Méndez Leal, L. J. Gabard-Durnam, H. M. O’Leary, BEAPP: the batch electroencephalography automated processing platform. *Frontiers in neuroscience* **12**, 513 (2018).
62. L. J. Gabard-Durnam, A. S. Mendez Leal, C. L. Wilkinson, A. R. Levin, The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): standardized processing software for developmental and high-artifact data. *Frontiers in neuroscience* **12**, 97 (2018).
63. M. Costa, A. L. Goldberger, C. K. Peng, Multiscale entropy analysis of biological signals. *Physical review E* **71**, 021906 (2005).
64. C. Schölzel, Nonlinear measures for dynamical systems (Version 0.5.2). *Zenodo* (2019).
65. G. Ruffini, D. Ibañez, E. Kroupi, J. F. Gagnon, J. Montplaisir, R. B. Postuma, M. Castellano, A. Soria-Frisch, Algorithmic complexity of EEG for prognosis of neurodegeneration in idiopathic rapid eye movement behavior disorder (RBD). *Ann. Biomed. Eng.* **47**, 282-296 (2019).
66. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J.

Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R.
Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F.
Pedregosa, P. van Mulbregt, sciPu 1.0 Contributors, SciPy 1.0: fundamental algorithms
for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).
67. P. Kassraian-Fard, C. Matthis, J. H. Balsters, M. H. Maathuis, N. Wenderoth, Promises,
pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric
imaging data, with autism as an example. *Frontiers in psychiatry* **7**, 177 (2016).