Online Supplemental Materials Accompanying

Early Detection of Autism Spectrum Disorder in Young Children with Machine Learning Using Medical Claims Data

Yu-Hsin Chen[1], Qiushi Chen[1], Lan Kong[2], Guodong Liu[2, 3, 4]

[1] The Pennsylvania State University, The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, University Park, PA 16802, USA

[2] The Pennsylvania State University, College of Medicine, Department of Public Health Sciences, Hershey, PA 17033, USA

[3] The Pennsylvania State University, College of Medicine, Department of Psychiatry and Behavioral Health, Hershey, PA 17033, USA

[4] The Pennsylvania State University, College of Medicine, Department of Pediatrics, Hershey, PA 17033, USA

Corresponding Author:

Qiushi Chen, PhD

Department of Industrial and Manufacturing Engineering

The Pennsylvania State University

302 Leonhard Building

University Park, PA 16802

Phone: 814-863-4562

Email: q.chen@psu.edu

Chen Y-H, *et al. BMJ Health Care Inform* 2022; 29:e100544. doi: 10.1136/bmjhci-2022-100544

Figure S1. Flow diagram for determining autism spectrum disorder (ASD) and non-ASD cohorts.

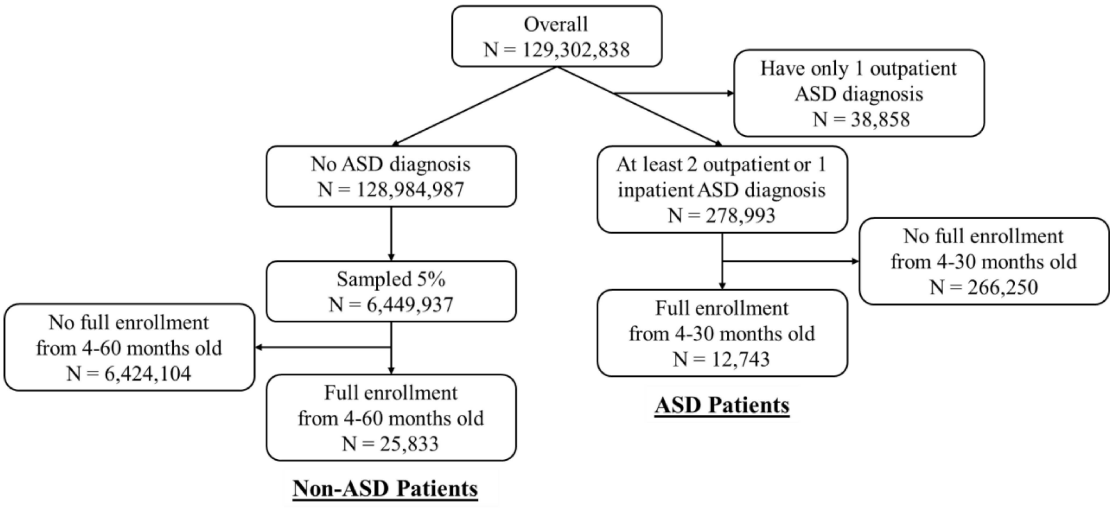

Figure S2. The area under the receiver operating characteristic curve (AUROC) of random forest models with variables sequentially included following the order of the median of Gini importance index from 50 replications.

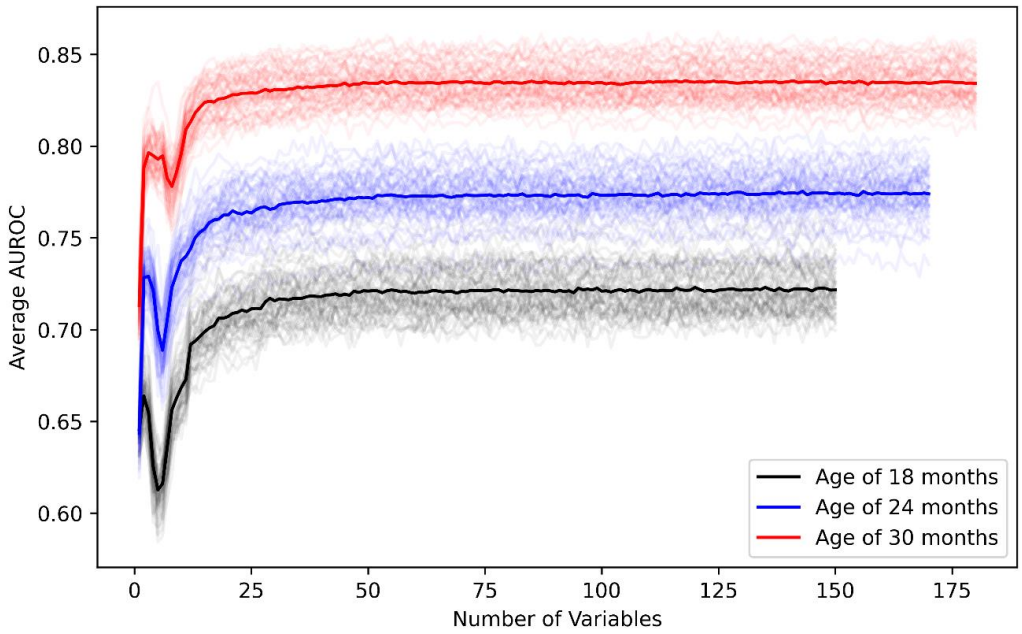

Figure S3. Top 50 important variables in the random forest (A) and Lasso logistic regression (B) models. The variables were ordered by the median of Gini index of random forest and by the absolute values of feature weights of Lasso logistic regression from 50 replications of independent runs. Median values are shown in black markers. For the random forest model (A), blue makers represent the Gini index from 50 replications; for the Lasso logistic regression (B), blue and red markers present positive and negative coefficients, respectively.

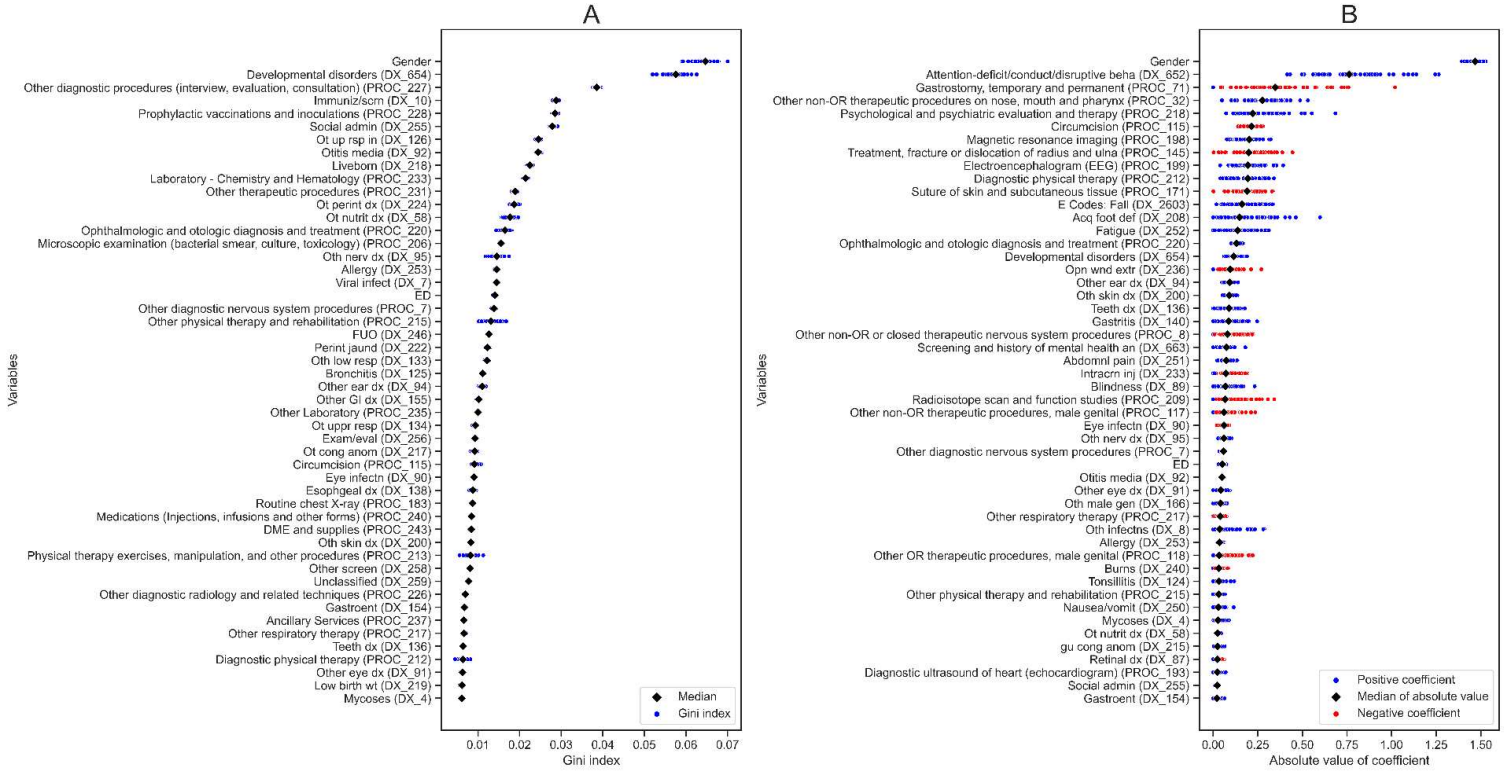

Figure S4. Relationships between 50 of the most important variables in the random forest model (center) and the most frequent variables (i.e., with highest prevalence) in ASD cohort (left), or the most important variables in the logistic regression model (right). Note: The variable importance in random forest model was determined by median of Gini index from 50 replications, whereas that in logistic regression was determined by the absolute coefficients of variables.

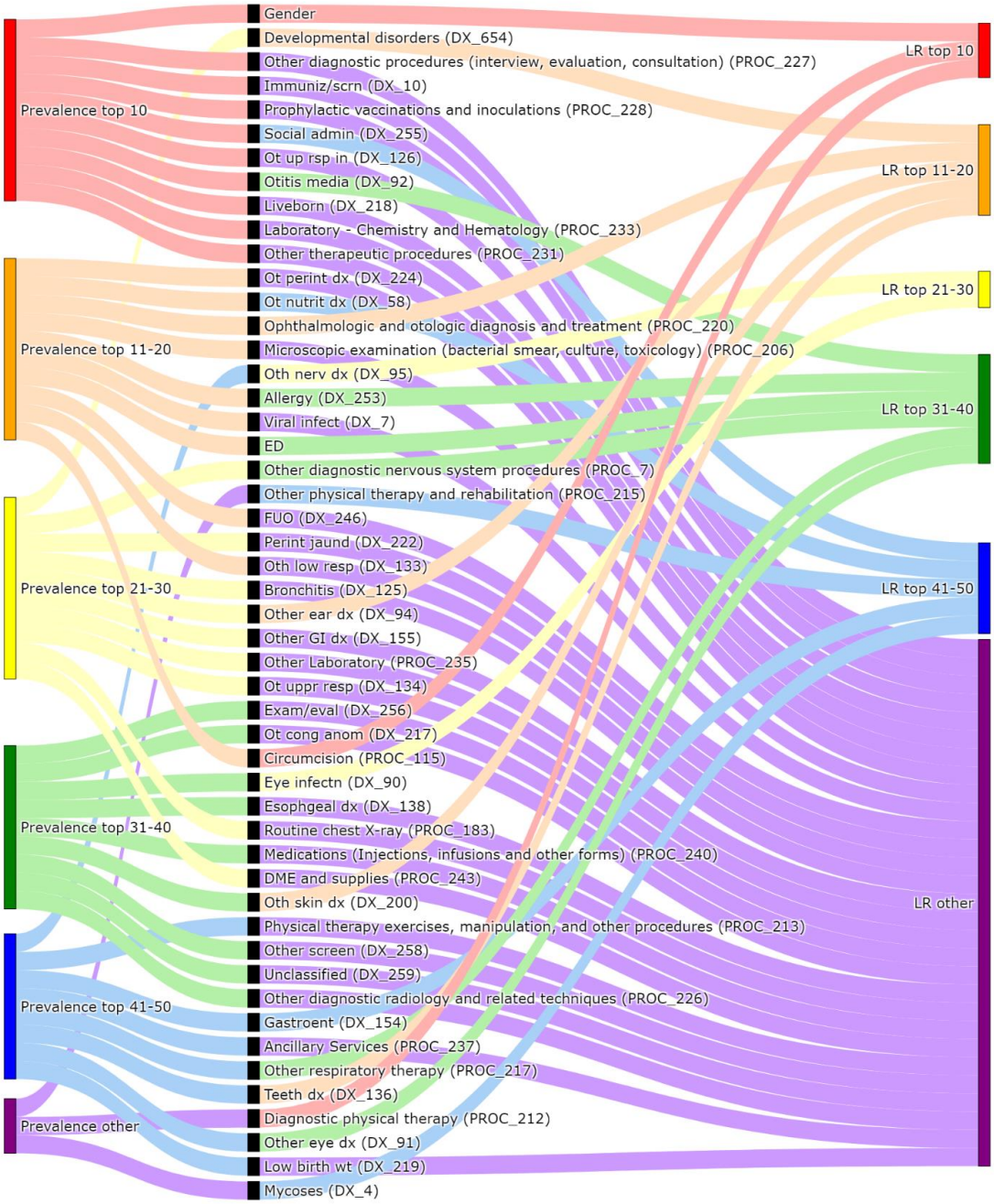

Table S1. Characteristics of autism spectrum disorder (ASD) and non-ASD cohorts.

	ASD cohort (N=12,743)	Non-ASD cohort (N=25,833)
Gender		
Male	10,174 (79.8%)	13,135 (50.8%)
Female	2,569 (20.2%)	12,698 (49.2%)
Diagnosis age (month)	48.3[a] (47.9-48.8[b])	
Before 24 months	804 (6.3%)	-
25-36 months	3,644 (28.6%)	-
37-48 months	3,215 (25.2%)	-
49-60 months	1,927 (15.1%)	-
Above 60 months	3,153 (24.7%)	-
Initial enrollment age		
Month 0	6,298 (49.4%)	12,159 (47.1%)
Month 1	4,156 (32.6%)	8,671 (33.6%)
Month 2	1,561 (12.2%)	3,525 (13.6%)
Month 3	721 (5.7%)	1,478 (5.6%)
Insurance type		
HMO	3,088 (24.2%)	6,538 (25.3%)
PPO	8,837 (69.3%)	18,339 (71.0%)
Others	818 (6.4%)	956 (3.7%)
Number of visits by time		
Before 18 months	27.51 (26.97-28.06)	20.41 (20.22-20.60)
18-24 months	8.72 (8.50-8.95)	4.26 (4.19-4.32)
25-36 months	12.57 (12.26-12.88)	3.89 (3.82-3.95)
The 20 most frequent CCS categories[c] up to age of 24 months in ASD cohort		
Other diagnostic procedures (interview, evaluation, consultation) (Procedure 227)	26.22[d] (99.6%[e])	21.05 (99.4%)[*]
Administrative/social admission (Diagnosis 255)	7.63 (98.8%)	7.35 (98.4%)[**]
Prophylactic vaccinations and inoculations (Procedure 228)	6.91 (96.8%)	6.48 (95.9%)[**]
Liveborn (Diagnosis 218)	3.85 (71.7%)	2.57 (71.8%)
Immunizations and screening for infectious disease (Diagnosis 10)	3.62 (81.3%)	3.29 (78.0%)[**]
Other perinatal conditions (Diagnosis 224)	2.87 (60.5%)	1.29 (49.0%)[**]
Other upper respiratory infections (Diagnosis 126)	2.75 (77.6%)	2.59 (77.4%)
Laboratory - Chemistry and Hematology (Procedure 233)	2.71 (84.2%)	2.10 (79.1%)[**]
Otitis media and related conditions (Diagnosis 92)	2.69 (63.0%)	2.81 (64.2%)[*]
Physical therapy exercises; manipulation; and other procedures (Procedure 213)	2.69 (12.7%)	0.38 (2.4%)[**]
Other nutritional; endocrine; and metabolic disorders (Diagnosis 58)	2.53 (36.3%)	0.57 (19.9%)[**]
Short gestation; low birth weight; and fetal growth retardation (Diagnosis 219)	2.27 (12.8%)	0.76 (7.2%)[**]
Developmental disorders (Diagnosis 654)	2.21 (27.6%)	0.17 (2.7%)[**]
Other therapeutic procedures (Procedure 231)	2.03 (68.0%)	1.46 (60.9%)[**]
Other congenital anomalies (Diagnosis 217)	1.54 (21.2%)	0.46 (11.4%)[**]

Other lower respiratory disease (Diagnosis 133)	1.29 (38.3%)	0.80 (33.4%)**
Microscopic examination (bacterial smear, culture, toxicology) (Procedure 206)	1.24 (51.6%)	1.08 (48.6%)**
Other physical therapy and rehabilitation (Procedure 215)	1.23 (11.6%)	0.10 (1.1%)**
Routine chest X-ray (Procedure 183)	1.15 (30.1%)	0.55 (25.3%)**
DME and supplies (Procedure 243)	1.14 (26.6%)	0.53 (21.4%)**

[a] Average diagnosis age.
[b] The 95% confidence interval for average diagnosis age.
[c] The detailed definition of each CCS category can be found at https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf
[d] Average number of visits with the given CCS diagnosis or procedure code up to age of 24 months per patient.
[e] Proportion of patients who ever had the given CCS diagnosis or procedure code during a healthcare visit prior to the age of 24 months (i.e., the prevalence of the CCS category variable).
* p-value ≤0.05 by a two-proportion z-test.
** p-value ≤0.01 by a two-proportion z-test.
Abbreviations: ASD, autism spectrum disorder; HMO, health maintenance organization; PPO, preferred provider organization; DX, diagnosis; PROC, procedure; CCS, clinical classifications software.

Table S2. The area under the receiver operating characteristic curve (AUROC) of random forest models with various limits on the number of trees for ASD prediction at the age of 24 months.

	AUROC (95% Confidence Interval)	
Number of trees in random forest model	Combined inpatient and outpatient encounters (base case)	Separated inpatient and outpatient encounters
1	0.611 (0.607, 0.615)	0.635 (0.632, 0.639)
25	0.759 (0.756, 0.763)	0.819 (0.816, 0.822)
50	0.767 (0.764, 0.771)	0.830 (0.827, 0.833)
75	0.777 (0.774, 0.781)	0.834 (0.830, 0.837)
100	0.775 (0.771, 0.779)	0.834 (0.831, 0.837)
125	0.778 (0.774, 0.782)	0.835 (0.832, 0.839)
150	0.776 (0.772, 0.780)	0.838 (0.835, 0.842)

Table S3. Sensitivity analysis results for Lasso logistic regression and random forest models.

Prediction Model	**AUROC (95% CI)**		
	Base case[a]	Including low prevalence variables[b]	4-72-month full enrollment[c]
At age of 18-month-old			
Lasso logistic regression	0.720 (0.716, 0.723)	0.724 (0.721, 0.727)	0.792 (0.788, 0.795)
Random forest	0.717 (0.714, 0.721)	0.721 (0.718, 0.724)	0.810 (0.808, 0.813)
At age of 24-month-old			
Lasso logistic regression	0.758 (0.755, 0.762)	0.764 (0.76, 0.768)	0.821 (0.818, 0.823)
Random forest	0.775 (0.771, 0.779)	0.777 (0.773, 0.780)	0.846 (0.844, 0.849)
At age of 30-month-old			
Lasso logistic regression	0.800 (0.797, 0.803)	0.806 (0.802, 0.809)	0.850 (0.848, 0.853)
Random forest	0.832 (0.828, 0.835)	0.832 (0.829, 0.835)	0.886 (0.884, 0.888)

[a] Excluded the variables that are present in less than 1% of both ASD and non-ASD cohorts, and required 4-60-month full enrollment to be included in the non-ASD cohort.
[b] Contained all variables.
[c] Required 4-72-month full enrollment as the inclusion criteria for the non-ASD cohort.
Abbreviations: AUROC, area under receiver operator characteristic curve; CI, confidence interval.