BMJ Health & Care Informatics

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 [2,3,4,5]

To cite: Chen Y-H, Chen Q, Kong L, *et al*. Early detection of autism spectrum disorder in young children with machine learning using medical claims data. *BMJ Health Care Inform* 2022;**29**:e100544. doi:10.1136/bmjhci-2022-100544

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100544).

Received 06 January 2022
Accepted 19 August 2022

For numbered affiliations see end of article.

Correspondence to
Dr Qiushi Chen;
q.chen@psu.edu

ABSTRACT

Objectives Early diagnosis and intervention are keys for improving long-term outcomes of children with autism spectrum disorder (ASD). However, existing screening tools have shown insufficient accuracy. Our objective is to predict the risk of ASD in young children between 18 months and 30 months based on their medical histories using real-world health claims data.

Methods Using the MarketScan Health Claims Database 2005–2016, we identified 12 743 children with ASD and a random sample of 25 833 children without ASD as our study cohort. We developed logistic regression (LR) with least absolute shrinkage and selection operator and random forest (RF) models for predicting ASD diagnosis at ages of 18–30 months, using demographics, medical diagnoses and healthcare service procedures extracted from individual's medical claims during early years postbirth as predictor variables.

Results For predicting ASD diagnosis at age of 24 months, the LR and RF models achieved the area under the receiver operating characteristic curve (AUROC) of 0.758 and 0.775, respectively. Prediction accuracy further increased with age. With predictor variables separated by outpatient and inpatient visits, the RF model for prediction at age of 24 months achieved an AUROC of 0.834, with 96.4% specificity and 20.5% positive predictive value at 40% sensitivity, representing a promising improvement over the existing screening tool in practice.

Conclusions Our study demonstrates the feasibility of using machine learning models and health claims data to identify children with ASD at a very young age. It is deemed a promising approach for monitoring ASD risk in the general children population and early detection of high-risk children for targeted screening.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Growing evidence has shown that existing autism spectrum disorder (ASD)-specific screening tools (eg, Modified Checklist for Autism in Toddlers) may not yield sufficient accuracy for early detection of children with ASD in clinical practice.

⇒ Previous clinical and health service research has identified clinical risk factors associated with ASD, but the clinical factors from an individual's prior medical history have not been used comprehensively to assess the risk of ASD in young children.

WHAT THIS STUDY ADDS

⇒ This study demonstrated the feasibility of predicting ASD diagnosis with promising accuracy based on an individual's medical record from health claims data using machine learning models.

⇒ Our prediction models were clinically interpretable, which systematically identified key predictors in line with known risk factors and symptoms among ASD children in the literature.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study may serve as a basis for integrating predictive modelling into the health information system and the clinical workflow to enhance the current ASD screening practice.

INTRODUCTION

Autism spectrum disorder (ASD) is a developmental disorder that involves persistent challenges in social interaction, speech and nonverbal communication, and restricted and repetitive behaviours.[1] In the USA, the prevalence of ASD has increased substantially in the past two decades, with an estimate of every 1 in 44 children to be identified with ASD by age 8 in 2016.[2] Although there exist evidence-based interventions which improve core symptoms in children with ASD, many children with ASD still experience long-term challenges with daily life, education and employment.[3]

Early diagnosis is the key to early intervention for improving the long-term outcomes of children with ASD. However, despite the growing evidence shows that accurate and stable diagnoses can be made by 2 years,[4] in real-world settings, the median age of ASD diagnosis is 50 months.[2] To improve early diagnosis, the American Academy of Pediatrics (AAP) has recommended universal screening among all children at 18-month and 24-month well-child visits in the primary care settings using the Modified Checklist for Autism in Toddlers (M-CHAT),[5] a questionnaire that assesses children's behaviour for toddlers.[6] However, growing evidence has shown that using M-CHAT alone may

not yield sufficient accuracy in detecting ASD cases, with a sensitivity below 40% and a positive predictive value (PPV) under 20%.[7 8]

In addition to ASD-specific behavioural questionnaires, general clinical and healthcare records may also contain meaningful signals to differentiate the ASD risks among very young children. Studies have found that children with ASD are oftentimes accompanied by certain symptoms and medical issues such as gastrointestinal problems,[9] infections[10 11] and feeding problems.[12] This implies that past diagnosis and healthcare encounter information, commonly available from health insurance claims or Electronic Healthcare Record (EHR), could potentially be used for ASD risk prediction. In fact, medical claims and EHR data have been widely used in the health informatics literature for identifying disease-specific early phenotypes even before the hallmark symptoms start to manifest, such as for chronic diseases like heart failures,[13] diabetes[14] and Alzheimer's disease.[15] In the context of ASD, health record data has been used to identify the ASD subtypes[16 17] and to predict the suicidal risk in adolescents with ASD[18]; however, its use for predicting ASD diagnosis in young children has remained limited. To fill this gap, the objective of this study is to examine the feasibility of using large-scale real-world medical claims data to develop a prediction model for ASD diagnosis in young children, which can be used to support effective ASD screening strategies and facilitate early detection.

METHODS

Data source

We used the deidentified individual-level longitudinal healthcare claims data from the IBM MarketScan Commercial Claims and Encounters Database from 2005 to 2016. This database includes over 273 million unique individuals for both privately and publicly insured people in the USA.[19] The claims data include baseline demographics (eg, sex, birth year, postal region), service providers, insurance plans, medical diagnoses (in international Classification of Diseases (ICD)-9/10 codes) and procedures (in Healthcare Common Procedure Coding System (HCPCS) and Current Procedural Terminology-4 codes) at each encounter of healthcare services.

Study population

We constructed an initial cohort consisting of young children with and without ASD (figure 1). The inclusion criteria of the ASD cohort are as follows: (1) having at least 2 outpatient or 1 inpatient ASD diagnosis encounters (299 for ICD-9 and F84 for ICD-10) throughout the existing records[20 21]; and (2) having continuous enrolment from 4 months to 30 months to ensure the completeness of health records from the claims data that can be used for diagnosis prediction at up to 30 months (online supplemental figure S1). To create the non-ASD cohort, we first identified individuals without any ASD diagnosis throughout their health records, then downsampled 5% of the population to obtain a computationally manageable yet sufficiently large subset of samples. To ensure patients had adequate follow-up time to receive confirmed ASD diagnosis in the database, we restricted our selection of non-ASD patients by requiring a full enrolment period from 4 months to 60 months (online supplemental table S1).

Predictor variables for ASD diagnosis

We examined all diagnosis and procedure codes of a child's medical encounters available from as early as within 4 months after birth up to the age for prediction of ASD. We applied the Clinical Classifications Software (CCS),[22] a commonly used tool in health informatics research, to aggregate the large number of distinct diagnosis and procedure codes into clinically meaningful groups (figure 1). The single-level CCS maps the ICD-9/10 and HCPCS codes to a substantially smaller yet

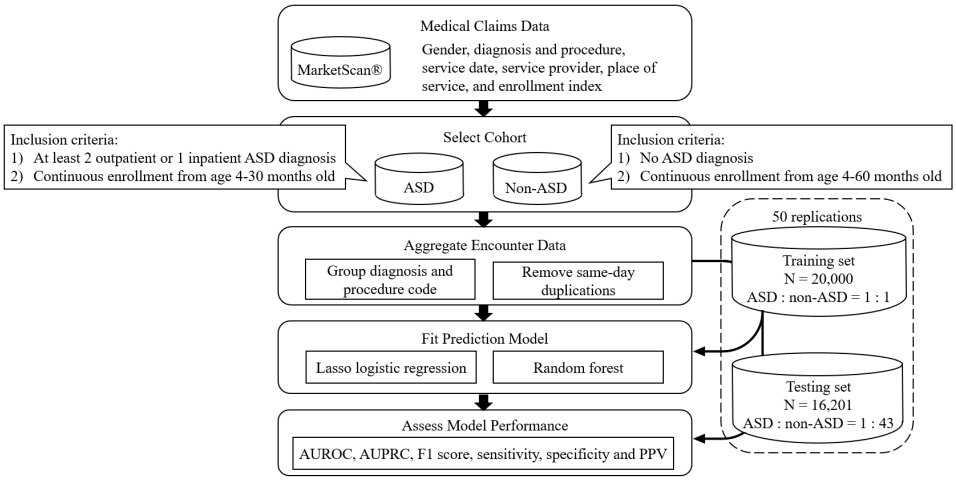

Figure 1 Overview of study design for the predictive analysis. ASD, autism spectrum disorder; AUROC, area under receiver operating characteristic curve; AUPRC, area under precision-recall curve; LASSO, least absolute shrinkage and selection operator; PPV, positive predictive value.

practical set that includes 285 diagnosis and 231 procedure categories.[22] We further removed the same-day duplications of CCS codes after the mapping by counting at most one encounter of a specific CCS category for each person on each day.

To predict the ASD diagnosis at the age of 24 months in our base case model, in line with the age when a diagnosis can possibly be made by an experienced professional,[4] we defined the predictor variables as the total number of encounters for each CCS category up to the age for prediction of 24 months. We also included sex and the encounters of emergency department visits, which are well-known clinically relevant factors associated with the autism population.[23] Variables that were present in <1% of both ASD and non-ASD cohorts were excluded.[24] A total of 170 input predictor variables were included for prediction at the age of 24 months. Having considered that the course of clinical events may be following a different pattern after an encounter with ASD diagnosis, we excluded any children who had at least one encounter with ASD diagnosis code prior to the age for prediction in our analysis.

Prediction model development and validation

We employed two machine learning methods, logistic regression (LR) and random forest (RF), which have been widely used for developing risk prediction models in various clinical settings. LR assumes that the independent variables are linearly related to the log odds and that the effects of multiple variables are additive, whereas RF is particularly suitable for exploiting nonlinear interactive effects in high-dimensional data. For the LR model, we also applied the least absolute shrinkage and selection operator (LASSO) as a feature selection technique to enforce the coefficients of weak predictors to be zero. The RF model was limited to up to 100 decision trees in the base case setting (other choices of the maximum number of trees were tested in sensitivity analysis).

To train our model, we sampled 10 000 ASD and 10 000 non-ASD subjects (N=20 000) from the initial cohort to build a large balanced training sample for maximising the discriminatory power learnt by the prediction model. To evaluate the model prediction performance, we created an independent imbalanced testing set (N=16 201) comprised of ASD and non-ASD patients from the remaining cohort that were mutually exclusive from the training set. The testing set resembled the real-world estimates for ASD prevalence of 2.3% (ie, 1 in every 44) in the general population.[2]

We measured the prediction performance with sensitivity (also known as true positive rate or recall), specificity (or true negative rate) and PPV (or precision)[25] at various selected risk thresholds. The model's overall discrimination ability was measured using the area under the receiver operating characteristic curve (AUROC). We also calculated the area under the precision-recall curve (AUPRC) where the precision-recall curve represents the relationship between PPV and sensitivity, and F1 score is defined as the harmonic mean of PPV and sensitivity, which are suited for evaluating the prediction performance for the imbalanced testing sample.[26 27] To assess the stability and the uncertainty of prediction performance, we repeated the training and testing set sampling, model training, testing and performance evaluation with 50 independent replications. The 95% CIs of all performance measures were reported.

Predicting ASD diagnosis at different ages

In addition to the base case prediction model where the risk of ASD diagnosis was assessed based on clinical information up to 24 months, we compared the accuracy of ASD prediction with varying lengths of available medical history at (1) a younger age, 18 months, considering that the universal ASD screening is recommended for children at both 18 months and 24 months[5]; and (2) an older age, 30 months, which is still a critical time point for monitoring the developmental delays and consideration of early intervention.[28] We followed the same approach in the base case to exclude predictor variables of low frequency (resulting in 150 and 180 predictor variables in total for prediction at 18 and 30 months, respectively) and the children with ASD diagnosis prior to the age for prediction.

Identifying key predictor variables

We further explored how many and which key predictive variables had the most impact on the prediction performance using the Gini importance index from the RF model. We added variables incrementally following the order of Gini Index (ie, starting with the most important variable) and evaluated how the prediction accuracy changed as more variables were included. Selected key predictive variables were then compared with those identified by alternative strategies using (1) the absolute value of coefficients from the LASSO LR model and (2) the prevalence of each variable in the identified ASD cohort.

Separating inpatient and outpatient visits

Considering that the underlying severity of the symptoms could potentially differ by inpatient hospitalisations and outpatient visits,[29] we split the number of encounters for each diagnosis and procedure by inpatient and outpatient visit separately and augmented the prediction model with more detailed encounter variables. We compared the prediction performance of the models using the augmented variables with our base case models.

Sensitivity analysis

We performed sensitivity analysis on several modelling assumptions to assess the robustness of our prediction models. Specifically, we strengthened the inclusion criteria for non-ASD subjects by requiring one additional year of enrollment, that is, increased from 4–60 months to 4–72 months. Furthermore, we assessed the potential loss of information due to excluding variables with <1% prevalence, to verify that such a variable prescreening

procedure would not miss out on rare but crucial predictive information.

RESULTS

Predicting ASD diagnosis at age of 24 months

We identified the study cohort consisting of 12 743 ASD subjects and 25 833 non-ASD subjects (more details in online supplemental table S1). When predicting the ASD diagnosis at the age of 24 months in independent testing samples, the LR and RF models achieved the AUROC of 0.758 (95% CI 0.755 to 0.762) and 0.775 (95% CI 0.771 to 0.779), respectively (table 1, figure 2). Compared with the LR model, RF model also showed a higher AUPRC (LR 0.101 (95% CI 0.098 to 0.104); RF 0.143 (95% CI 0.138 to 0.148)) and F1 score (LR 0.193 (95% CI 0.188 to 0.197); RF: 0.246 (95% CI 0.240 to 0.251)). The limit of up to 100 trees in the RF model was deemed sufficient to achieve stable performance. Further increasing the model complexity did not translate to an improvement in prediction accuracy (online supplemental table S2).

Predicting ASD diagnosis at different ages

Comparing the prediction models at the ages of 18, 24 and 30 months, we found that the prediction performance increased substantially with the age. Specifically for the RF model, the AUROC increased from 0.717 (0.714–0.721) at age of 18 months to 0.832 (0.828–0.835) at 30 months (table 1). Similarly, the AUPRC increased from 0.067 (0.065–0.069) to 0.234 (0.227–0.240) (figure 3), and F1 score increased from 0.130 (0.125–0.134) to 0.326 (0.322–0.331) from age of 18–30 months. The LR model, although with a lower prediction accuracy compared with the RF model in general, also showed a consistently increasing prediction performance as the age increased.

Identifying key predictive variables

As the RF model included more variables following the importance order by the Gini index, it showed higher AUROC (online supplemental figure S2). For prediction at age of 24 and 30 months, 30–40 most important variables were sufficient to achieve stable prediction performance with AUROC, whereas for an earlier age of 18 months, the top 50 important variables contributed to most of the prediction performance, while including additional variables could continue to marginally improve the prediction performance. We closely examined the 50 most important variables of the RF model (ranked by Gini index) and the LR model (ranked by the median absolute value of the coefficient) for prediction at age of 24 months (online supplemental figure S3). The identified important variables included sex, developmental and nervous system disorders, psychological and psychiatric services, respiratory system infections and symptoms, gastrointestinal-related diagnosis, ear and eye infections, perinatal conditions, and ED visits, which have also been seen as separate risk factors associated with ASD cases in the clinical literature. The key predictors of the RF model

Table 1 Performance of LASSO logistic regression and random forest models in prediction of autism spectrum disorder

Settings and prediction model	AUROC (95% CI)	AUPRC (95% CI)	F1 (95% CI)	Sensitivity target, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)
At age of 24 months (base case)						
LASSO logistic regression	0.758 (0.755 to 0.762)	0.101 (0.098 to 0.104)	0.193 (0.188 to 0.197)	40*	90.0 (89.7 to 90.4)	8.7 (8.4 to 9.0)
				50	83.7 (83.2 to 84.2)	6.7 (6.5 to 6.9)
				70	66.1 (65.4 to 66.8)	4.6 (4.5 to 47)
Random forest	0.775 (0.771 to 0.779)	0.143 (0.138 to 0.148)	0.246 (0.240 to 0.251)	40	93.0 (92.6 to 93.5)	12.1 (11.4 to 12.8)
				50	87.3 (86.7 to 87.9)	8.4 (8.1 to 8.8)
				70	69.6 (68.7 to 70.4)	5.1 (4.9 to 5.2)
At age of 18 months (younger)						
LASSO logistic regression	0.720 (0.716 to 0.723)	0.066 (0.064 to 0.068)	0.128 (0.124 to 0.132)	50	78.4 (77.9 to 78.9)	5.1 (5.0 to 5.2)
Random forest	0.717 (0.714 to 0.721)	0.067 (0.065 to 0.069)	0.130 (0.125 to 0.134)	50	78.8 (78.3 to 79.4)	5.1 (5.0 to 5.2)
At age of 30 months (older)						
LASSO logistic regression	0.800 (0.797 to 0.803)	0.148 (0.143 to 0.153)	0.255 (0.249 to 0.261)	50	90.4 (90.0 to 90.8)	11 (10.6 to 11.5)
Random forest	0.832 (0.828 to 0.835)	0.234 (0.227 to 0.240)	0.326 (0.322 to 0.331)	50	95.6 (95.3 to 95.8)	21.0 (20.2 to 21.8)

*The sensitivity threshold of 40% was selected to be comparable with the estimated sensitivity of 33%–39% for the existing autism-specific screening tools from real-world clinical settings.[7 8]
AUPRC, area under precision-recall curve; AUROC, area under receiver operator characteristic curve; LASSO, least absolute shrinkage and selection operator; PPV, positive predictive value.

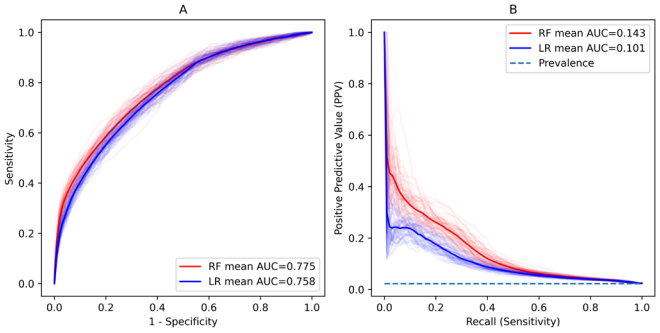

Figure 2 Receiver operating characteristic curves (A) and precision-recall (PR) curves (B) for prediction of autism spectrum disorder (ASD) diagnosis at age of 24 months. The prevalence stands for the baseline 2.27% (ie, 1 in 44) ASD prevalence in the general population. AUC, area under curve; LR, logistic regression; RF, random forest.

were also highly consistent with high prevalence variables, sharing 47 out of 50 most common variables in the ASD cohort (online supplemental figure S4).

Prediction using separated inpatient and outpatient data

Separating inpatient and outpatient encounters further increased the AUROC for prediction at the age of 24 months to 0.766 (95% CI 0.762 to 0.769) in the LR model and 0.834 (95% CI 0.831 to 0.837) in the RF model. At the target sensitivity of 40%, the RF model achieved a higher specificity of 96.4% (95% CI 96.2% to 96.5%) with a PPV of 20.5% (95% CI 19.8% to 21.1%), outperforming the existing screening tool M-CHAT/F (with a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%). We found that using claims data separated by inpatient and outpatient visits improved the prediction performance consistently across all ages (figure 4).

Robustness check and sensitivity analysis

With a more stringent inclusion criterion for non-ASD subjects by requiring a longer full enrollment period up to 72 months (vs 60 months in our base case), we found that the prediction performance had modest improvement (online supplemental table S3). It could be partially attributed to the fact that with longer years to ascertain

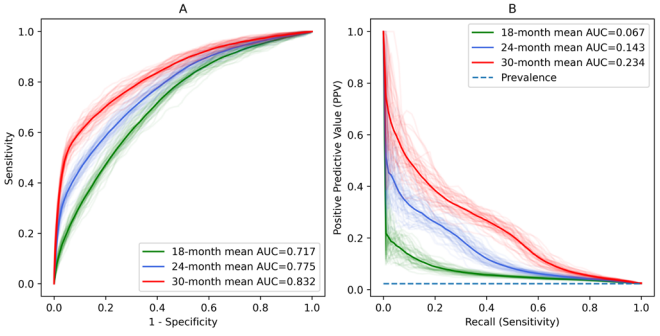

Figure 3 Receiver operating characteristic curves (A) and precision-recall curves (B) for prediction of autism spectrum disorder at ages of 18, 24 and 30 months, respectively, by the random forest model. AUC, area under curve.

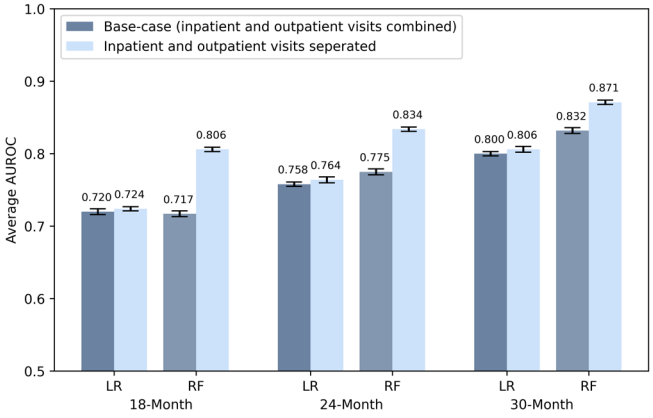

Figure 4 Comparison of area under the receiver operating characteristic curve (AUROC) with combined versus separated inpatient and outpatient encounters by LASSO logistic regression (LR) and random forest (RF) models, at the age of 18, 24 and 30 months, respectively. Error bars in the figure represent the 95% CIs based on results from 50 replications of independent runs. LASSO, least absolute shrinkage and selection operator.

the non-ASD cohort, children would be less likely to be misclassified. We also verified that including the low-prevalence variables would not result in substantial differences but only marginal changes of AUROC within 0.01 across all model specifications.

DISCUSSION

Early identification is vital for children with ASD to ensure their access to timely intervention and to optimise long-term outcomes. In this study, we demonstrated the feasibility of predicting ASD diagnosis at early ages using health claims data and machine learning models. We found that LASSO LR and RF models achieved an overall AUROC above 0.75 when predicting ASD diagnosis at age of 24 months. Our results also showed that prediction performance increased with age at the time of prediction. This is reasonable because more clinical information accumulated over a longer follow-up period since birth may contain more distinctive patterns to effectively differentiate children with ASD. The prediction models developed in our study are clinically interpretable. Key predictors, such as sex (male), developmental delays, gastrointestinal disorders, respiratory system infections and otitis media have shown strong predictive values for ASD diagnosis, which are in line with previous clinical studies that have shown these symptoms being associated with ASD children. Finally, our study showed that separating inpatient and outpatient claims as predictors could further improve the prediction accuracy.

In our study, both LASSO LR and RF models showed promising accuracy in predicting ASD diagnosis based on an individual's medical claims data. This robust finding implies that there may exist distinct patterns in health conditions and health service needs among young children with ASD, well before the onset of most hallmark ASD behavioural symptoms. Such predictive signals can

be easily extracted from the electronic health records or medical claims administrative data, and used for the early identification of ASD cases. We also observed differences in the performance between the two models. The RF model outperformed the LASSO LR model in general, likely because, with its tree-based model structure, the RF model is better at capturing complex interactive effects among the predictor variables to distinguish between the ASD and non-ASD cases, whereas the LR model synthesises the effects of multiple variables additively. The advantage of the RF model became more salient when input variables were separated by inpatient and outpatient claims into a more granular level.

Our study has made an important contribution to applying health informatics in the field of ASD. Although there exists a plethora of literature identifying individual risk factors of ASD, using large healthcare service data and machine learning models to systematically predict ASD diagnosis has remained much less explored. Unlike existing clinical informatics studies that focused on detecting ASD subtypes,[16 17] we aim to detect ASD cases among the general children population, that is, the early detection. This could be particularly challenging due to the low prevalence of ASD in the general population (ie, a highly imbalanced dataset), and the scarcity of information available at such a young age. Nevertheless, our model showed promising prediction performance. The RF model with separated inpatient and outpatient encounters achieved a specificity of 96.4% at a sensitivity of 40% for the ASD prediction at the age of 24 months, outperforming the accuracy of the existing ASD-specific screening tool (sensitivity: 38.8%; specificity: 94.9%) from a clinical observational study.[7] It is worth noting that under a similar ASD prevalence (2.2%), our model showed a higher PPV (20.5% vs 14.6%).

Our prediction model for ASD diagnosis could lead to a significant impact on the screening strategies for ASD in young children. Although the AAP guidelines recommend universal screening in all children, it has been debated that, without the perfect screening tool, universal screening may result in overburdened diagnostic services in the healthcare system as these clinical resources are in extremely short supply.[30] Our prediction models have demonstrated promising improvement over the existing ASD screening tool by using clinical information, which could potentially serve as a 'triaging tool' for identifying high-risk patients for diagnostic evaluation. Moreover, the models only based on health claims data makes it practically feasible to integrate into an EHR system or insurance claims database. It could further enable an automatic screening tool, which can continuously monitor an individual's risk as new diagnosis and procedure information emerges, and send reminders to patients or providers for a timely clinical assessment if necessary. On the other hand, it is possible that some diagnosis and procedure information appear after a concern that the child had autism has already existed, such as following a positive screening event, which could alter the course of subsequent clinical events. As such, our prediction model is not designed to direct the screening decisions, but rather a tool to enhance the screening accuracy. If more detailed electronic health record data were available, the proposed risk prediction model could be further extended by incorporating screening results with clinical information, or by differentiating the clinical information before versus after the screening events, to further improve the accuracy of identifying high-risk ASD cases for further diagnostic evaluation.

Our study has several limitations. First, diagnosis of ASD established only based on existing diagnosis codes from claims data could be inaccurate and unreliable sometimes in practice. We followed a validated approach in ASD health service research literature to identify the ASD cohort in our study.[31] Second, the absence of ASD diagnosis codes in one's health record may not necessarily indicate an individual not having ASD, especially for children born in later years, due to limited follow-up time prior to the cut-off date in the database. Thus, we required full enrollment up to 60 months without ASD diagnoses to identify the non-ASD cohort, and verified the robustness of our base case results in a sensitivity analysis requiring full enrolment up to 72 months. Third, as autistic children are likely to have a wide range of comorbid conditions with various frequencies, for individuals who do not present comorbid conditions from the past healthcare encounter data, our model may provide limited value. Our risk prediction model can be further augmented by additional information other than information from the health claims database, such as ASD/developmental screening results and behaviour-related information from a more comprehensive EHR dataset in future studies. Lastly, the diagnosis and procedure codes in insurance claims data may be subject to variabilities and irregularities. Instead of the original detailed clinical codes, we used aggregated CCS categories for diagnoses and procedures for more robust clinical measures.

CONCLUSIONS

Using real-world health claims data and machine learning methods, we developed a prediction model that can successfully predict ASD diagnosis for children under 30 months with promising prediction accuracy. Our model also identified the important predictors for the diagnosis prediction, which showed meaningful clinical relevance and intuition. Our predictive modelling approach could potentially be generalised to broader clinical settings for predicting the diseases that may show early signals from past healthcare service encounters in claims or EHR data. Future studies could explore the prediction of ASD diagnosis dynamically over time as new healthcare encounter occurs, and investigate how validated risk prediction models could be integrated and used to inform ASD screening strategies.

Author affiliations
[1]The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA
[2]Department of Public Health Sciences, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania, USA

[3]Department of Psychiatry and Behavioral Health, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania, USA
[4]Department of Pediatrics, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania, USA
[5]The Center for Applied Studies in Health Economics (CASHE), The Pennsylvania State University College of Medicine, Hershey, Pennsylvania, USA

Contributors GL, QC and Y-HC conceived of the presented idea, and developed it with support from GL and QC. Y-HC cleaned and preprocessed the data, developed prediction models, and performed model evaluations. All authors interpreted the model results. Y-HC and QC drafted the manuscript, which was critically revised by all authors. QC is the guarantor of the project.

Funding This work has been supported by Penn State Social Science Research Institute Level 1 Seed Grant (QC, GL), Penn State College of Engineering Multidisciplinary Research Seed Grant (Y-HC, QC, GL) and NIH R21 grant: 1 R21 MH119480-01A1 (GL, LK).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available.

ORCID iDs
Yu-Hsin Chen http://orcid.org/0000-0002-3678-7517
Qiushi Chen http://orcid.org/0000-0003-4031-2669
Lan Kong http://orcid.org/0000-0001-6098-9445
Guodong Liu http://orcid.org/0000-0001-8683-0803

REFERENCES

1 American Psychological Association. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.
2 Maenner MJ, Shaw KA, Bakian AV, *et al*. Prevalence and characteristics of autism spectrum disorder among children aged 8 Years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2018. *MMWR Surveill Summ* 2021;70:1–16.
3 McPheeters ML, Weitlauf A, Vehorn A. *U.S. preventive services Task force evidence syntheses, formerly systematic evidence reviews. screening for autism spectrum disorder in young children: a systematic evidence review for the US preventive services Task force*. Rockville (MD): Agency for Healthcare Research and Quality (US), 2016.
4 Lord C, Risi S, DiLavore PS, *et al*. Autism from 2 to 9 years of age. *Arch Gen Psychiatry* 2006;63:694–701.
5 Lipkin PH, Macias MM, Council on children with disabilities, section on developmental and behavioral pediatrics. Promoting optimal development: identifying infants and young children with developmental disorders through developmental surveillance and screening. *Pediatrics* 2020;145. doi:10.1542/peds.2019-3449. [Epub ahead of print: 16 Dec 2019].
6 Robins DL, Fein D, Barton ML, *et al*. The modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *J Autism Dev Disord* 2001;31:131–44.
7 Guthrie W, Wallis K, Bennett A, *et al*. Accuracy of autism screening in a large pediatric network. *Pediatrics* 2019;144.
8 Carbone PS, Campbell K, Wilkes J, *et al*. Primary care autism screening and later autism diagnosis. *Pediatrics* 2020;146. doi:10.1542/peds.2019-2314. [Epub ahead of print: 06 07 2020].
9 Chaidez V, Hansen RL, Hertz-Picciotto I. Gastrointestinal problems in children with autism, developmental delays or typical development. *J Autism Dev Disord* 2014;44:1117–27.
10 Rosen NJ, Yoshida CK, Croen LA. Infection in the first 2 years of life and autism spectrum disorders. *Pediatrics* 2007;119:e61–9.
11 Adams DJ, Susi A, Erdie-Lalena CR, *et al*. Otitis media and related complications among children with autism spectrum disorders. *J Autism Dev Disord* 2016;46:1636–42.
12 Ledford JR, Gast DL. Feeding problems in children with autism spectrum disorders. *Focus Autism Other Dev Disabl* 2006;21:153–66.
13 Sideris C, Alshurafa N, Pourhomayoun M, *et al*. A data-driven feature extraction framework for predicting the severity of condition of congestive heart failure patients. *Annu Int Conf IEEE Eng Med Biol Soc* 2015;2015:2534–7.
14 Nguyen BP, Pham HN, Tran H, *et al*. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Comput Methods Programs Biomed* 2019;182:105055.
15 Park JH, Cho HE, Kim JH, *et al*. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ Digit Med* 2020;3:46.
16 Lingren T, Chen P, Bochenek J, *et al*. Electronic health record based algorithm to identify patients with autism spectrum disorder. *PLoS One* 2016;11:e0159621.
17 Vargason T, Frye RE, McGuinness DL, *et al*. Clustering of co-occurring conditions in autism spectrum disorder during early childhood: a retrospective analysis of medical claims data. *Autism Res* 2019;12:1272–85.
18 Downs J, Velupillai S, George G, *et al*. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. *AMIA Annu Symp Proc* 2017;2017:641–9.
19 IBM MarketScan research databases, 2020. Available: https://www.ibm.com/products/marketscan-research-databases
20 Burke JP, Jain A, Yang W, *et al*. Does a claims diagnosis of autism mean a true case? *Autism* 2014;18:321–30.
21 Coleman KJ, Lutsky MA, Yau V, *et al*. Validation of autism spectrum disorder diagnoses in large healthcare systems with electronic medical records. *J Autism Dev Disord* 2015;45:1989–96.
22 Agency for Healthcare Research and Quality R, MD. HCUP clinical classifications software (CCS) for ICD-9-CM Healthcare Cost and Utilization Project (HCUP) 2006-2009; 2020. www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
23 Loomes R, Hull L, Mandy WPL. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *J Am Acad Child Adolesc Psychiatry* 2017;56:466–74.
24 He D, Mathews SC, Kalloo AN, *et al*. Mining high-dimensional administrative claims data to predict early hospital readmissions. *J Am Med Inform Assoc* 2014;21:272–9.
25 Hunink MGM, Weinstein MC, Wittenberg E. *Decision making in health and medicine : Integrating evidence and values*. 2nd ed. Cambridge University Press, 2014.
26 Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data recommendations for the use of performance metrics. *Int Conf Affect Comput Intell Interact Workshops* 2013;2013:245–51.
27 Ozenne B, Subtil F, Maucort-Boulch D. The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol* 2015;68:855–9.
28 Hyman SL, Levy SE, Myers SM, *et al*. Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics* 2020;145:e20193447.
29 Pottick K, Hansell S, Gutterman E, *et al*. Factors associated with inpatient and outpatient treatment for children and adolescents with serious mental illness. *J Am Acad Child Adolesc Psychiatry* 1995;34:425–33.
30 Siu AL, Bibbins-Domingo K, *et al*, US Preventive Services Task Force (USPSTF). Screening for autism spectrum disorder in young children: US preventive services task force recommendation statement. *JAMA* 2016;315:691–6.
31 Liu G, Pearl AM, Kong L, *et al*. Risk factors for emergency department utilization among adolescents with autism spectrum disorder. *J Autism Dev Disord* 2019;49:4455–67.