A Pilot Study To Evaluate The Autism Co-Morbid Risk Score

Department of Pediatrics, University of Chicago
Department of Medicine, University of Chicago

CONTENTS

ABBREVIATIONS

ASD	Autism Spectrum Disorder
MCHAT-F	Modified Checklist for Autism in Toddlers with Followup
ADOS / ADOS-2	Autism Diagnostic Observation Schedule
ABA	Applied Behavior Analysis
ACoR	Autism Comorbid Risk

1. ABSTRACT

1.1. Context: Early diagnosis of Autism Spectrum Disorder (ASD) and the timely intervention is widely recognized as critical for achieving improved developmental outcomes.[1] With no laboratory tests for ASD, and despite advances from widespread adoption of screening with standardized checklists at 18 and 24 months of age, the median age of diagnosis remains over 4 years. Starting with a positive initial screen, a clinical diagnosis of ASD is a frustrating multi-step process spanning 3 months to 1 year, often delaying time-critical intervention. While a diversity of factors are implicated,[2]–[5] one obvious source of these delays is the vast number of false positives encountered in the current initial screening tools. For example, the M-CHAT/F, the most widely used screen,[1], [6] produces about over 85 false positives out of every 100 flagged for diagnostic evaluation, significantly inflating wait times[5] especially in rural and underserved communities. Further, current screening tools are sensitive to language barriers and cultural issues, and are particularly ineffective for children with milder symptoms with average or above-average cognitive abilities until about school age,[1], [7] often due to a "wait and see" approach adopted at the primary care.

The possibility of precise prediction of neuropsychiatric disorders from patterns in individual medical histories is supported by the high co-morbidity levels in children with ASD[8] that span dysregulations of immune pathways such as eczema, allergies, asthma, as well as ear and respiratory infections, gastrointestinal problems, developmental issues, severe headaches, migraines, and seizures.[9], [10]

1.2. Objectives: In this pilot study, we plan to prospectively collect data to aid in the validation of machine inferred **digital biomarkers** for autism, mined automatically from Electronic Health Record (EHR) databases. Using individual diagnostic codes already recorded during regular doctor's visits, we have already engineered and retrospectively validated a risk estimator (**ASD Co-morbid Risk: ACoR**) enabled by novel machine learning algorithms. Orthogonal to questionnaire based detection of behavioral signals, the proposed tool potentially reduces socio-economic, ethnic and demographic biases to elicit more objective and stable results — with zero administrative burden on clinicians and parents. With a team comprising machine learning experts, pediatric clinicians and developmental experts, we plan to carry out parallel tests with M-CHAT/F and ACoR in the pediatric primary care setting, and patients screening positive under MCHAT-F will receive diagnostic evaluation with Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [1] to ascertain ASD status.

In addition to direct comparison, we aim to validate our strategy of combining ACoR with MCHAT-F scores to substantially bring down the current false positive rates.

Thus, the principal aims this study are the following:

□ **Aim 1: Reduce false positives in current screening protocols.** The current high false positive rate exacerbates post screen wait-time, which is often the main source of diagnostic delay. To evaluate the *hypothesis: ACoR can cut down upto 50% of false positives*, we plan to track the number of cases in which MCHAT-F triggers a flag, but our tool does not. Our goal here is to evaluate the positive predictive value (PPV) of ACoR, under high specificty conditions ($> 95\%$).

□ **Aim 2. Evaluate the statistical relationship between the ACoR score and M/CHAT-F, and formalize a joint or conditional operational protocol.** We will characterize statistical association, if any, between the test scores. *Hypothesis: The uncertainties or errors in the two tests are are statistically independent.* Additionally, we will evaluate our ability to boost performance by conditioning the sensitivity-specificity trade-offs on the M-CHAT/F score of individual patients.

□ **Aim 3. Evaluate the effectiveness of ACoR in a demographically diverse population with a range of socio-economic confounders.** *Hypothesis: A questionnaire-free approach has the potential to mitigate biases that arise from limitation of language, cultural barriers, and demographic diversity, e.g.* disproportionately failing to diagnose children with average to above-average intelligence in diverse populations,[11] and under-reporting of symptoms by parents or primary care-givers due to cultural differences.[12]

□ **Aim 4. Characterize heterogeneity of ASD presentation by relating it to patterns in medical history, and predictive co-morbidities.** Heterogeneous presentation is a key barrier in the mechanistic understanding of ASD pathobiology. *Hypothesis: We can characterize the distinct classes and/or hierarchies of co-morbidities, by leveraging our ability to disambiguate them from individual medical histories*. This will shed light into the potentially intrinsic classes of the underlying disease processes, and refine/inform intervention design.

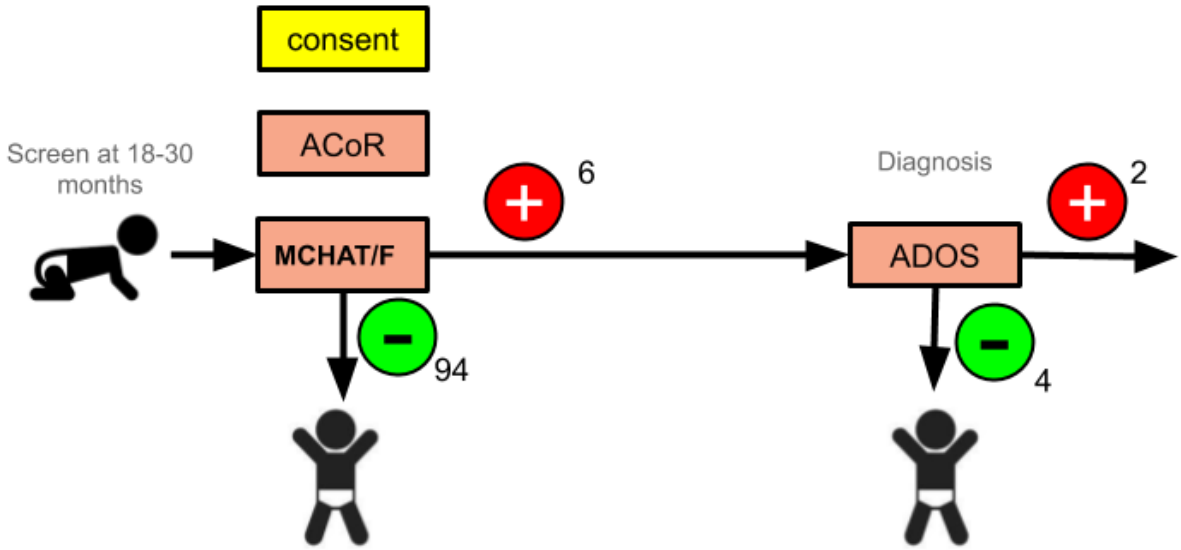

Fig. 1. **Patient flow logic.** Consenting families with children in appropriate age group (18 to 30 months) are subjected to both M-CHAT/F and ACoR screens, and a MCHAT-F flag is scheduled for an ADOS evaluation for ascertaining ASD status. Note that by our calculations, for every 100 children, we expect to 6 MCHAT-F flags, out of which 4 on average will get a negative diagnosis by ADOS.

1.3. Study Design: This observational cohort study will assess the applicability of ACoR for augmenting ASD screening in children within the age range of 18-24 months.

1.4. Setting/Participants: Participants will be approximately 200 children who will be evaluated via both the MCHAT-F screening during wellness visits at the 1 year, 1.5 year and 2 year mark, via the standard questionnaire completed by their primary caregivers, and the ACoR algorithm applied to their diagnostic history on file. Inclusion criteria: 1) Child is between 18 and 30 months, 2) Child has diagnostic history on record with at least 5 diagnostic codes, and the first code is at least from 15 weeks in the past. Exclusion criteria: Diagnostic history only consists of health service contact codes.

1.5. Study Interventions and Measures: No intervention is planned within this study. Main outcomes are efficacy and applicability of ACoR compared to MCHAT-F.

1.6. Innovation:

Paradigm Shift in ASD Screening: Despite extensive documentation of co-morbidities, a risk estimator that makes reliable predictions for individuals — based purely on co-morbidity patterns — has never been reported to the best of our knowledge. The sparsity of available diagnostic codes corresponding to individual subjects, and the general absence of physiological disorders that would uniquely signal the eventual emergence of symptoms indicative of a clinical ASD diagnosis, combined with the heterogeneity of ASD presentation, make such an endeavor challenging. In this study we leverage our preliminary work on the formulation of a *first-of-its-kind* framework to make predictions based on models of statistically curated patterns of diagnostic code sequences automatically learned from sufficiently large databases of electronic health records (EHR), that achieves an out-of-sample AUC exceeding 80% for either gender from just over 2 years of age. The machine learning tools that make this possible are also fundamentally novel, designed to address the specific issues in handling sparse, noisy categorical diagnostic sequences.

2. Background Information & Research Rationale

2.1. Significance: Autism spectrum disorder is a developmental disability associated with significant social, communication, and behavioral challenges. The prevalence of ASD has risen dramatically in the United States from 1 in 10,000 in 1972 to 1 in 59 children in 2014, with males diagnosed at nearly four times the rate of females.[8], [13] There is a current lack of consensus on whether increased awareness and recent changes in diagnostic practices[1] can fully explain this trend.[14] Nevertheless, with possibly over 1% of individuals affected worldwide,[15] ASD is a human condition with potentially serious negative impacts on individuals, families, and communities. Early detection can and does improve outcomes,[1] and is of paramount importance

when designing interventions, and is aligned with the envisioned goal of this initiative, as supported by the National Advisory Mental Health Council (NAMHC) (https://www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2018/early-screening-for-autism-spectrum.shtml).

Even though ASD may be reliably diagnosed as early as the age of two,[8] children frequently remain undiagnosed until after the fourth birthday.[16] At this time, there are no laboratory tests for ASD, so a careful review of behavioral history and social interactions is necessary for a clinical diagnosis.[1], [17] Starting with being flagged by an initial screen based on standardized checklists presented to parents at the ages between 1.5 and 2 years, a confirmed ASD diagnosis is a multi-step process that very often spans 3 months to 1 year. Most of this time is spent waiting to see qualified providers who can carry out the evaluation necessary for a clinical diagnosis. This extended wait is stressful to families, and impacts patient outcomes by delaying entry into time-critical intervention programs. While lengthy evaluations,[2] cost of care,[3] lack of providers,[4] and lack of comfort in diagnosing ASD by primary care providers[4] are all responsible to varying degrees,[18] one obvious factor responsible is the number of false positives produced in the initial ASD-specific screening tools in use today. For example, the M-CHAT/F, the most widely used screen,[1], [6] produces about over 85 false positives out of every 100 people flagged for further diagnostic evaluation, contributing to extended queues.[18] The impact from an excessive number of false positives is exacerbated by the current limited access to care and sparse availability of resources except near urban academic centers.[18], [19]

The standardized questionnaires attempt to measure risk by direct observation of behavioral symptoms, as reported by untrained observers (parents). Hence the current screening tests are only as good as the ability of the questions to discern and disambiguate behavior in infants and toddlers on casual observation, and on the ability of parents and caregivers to correctly interpret and answer the items without bias. This has lead to possibility of under-diagnosis in diverse communities as reflected by the lower apparent prevalence among African-American and Hispanic children. Also, children with average or higher-than-average cognitive abilities seem to have been under-diagnosed as reported is large scale population studies.[1] Borderline cases are typically problematic to screen for due to the possibility of subjective interpretation that is built into questionnaire based risk assessment. Responses to checklists are clearly confounded by a host of socio-economic (SES) variables, potential interpretive biases, and cultural differences. The heterogeneity of presentation also causes issues, since a potential plurality of symptom classes makes it harder for clinicians to recognize borderline cases, or on-the-fly combine observed co-morbidities with scores from standardized screening tools.

In this study, we operationalize a documented aspect of ASD symptomology in that it has a wide range of co-morbidities occurring at much higher rates than in the general population.[1] The ACoR methodology we propose in this grant can address the aforementioned complicated challenges of ASD screening, by distilling incipient predictive patterns of elevated risk from past medical history of individual patients, gleaned by machine learning algorithms from large de-identified databases of retrospective patient records. Powered by novel stochastic learning algorithms, we reverse-engineer sparse noisy uncertain diagnostic code sequences into actionable signatures; in effect giving us a fundamentally new approach to ASD screening and risk evaluation.

Potential Impact of ACoR In Science and Health: An automated diagnostic or screening capability that might be administered with little or no specialized training, requires no behavioral observations, and is functionally independent of the current tools has the potential for immediate transformative impact on patient care.

That such predictions of neuropsychiatric disorders might be possible from analyzing patterns in individual medical histories is suggested by the fact that parents of children with ASD often notice a diagnosable developmental problem before their child's first birthday; while vision and hearing problems are not uncommon in the first year, differences in social, communication, and fine motor skills have been reported to be evident from about 6 months of age.[20]–[22]

3. Study Objectives

Specific aims of the study has been enumerated in Section 1.2. We outline their significance below.

Significance of Specific Aim 1: A significant contribution to teh current diagnostic delay arises from families waiting in queue for diagnostic evaluationsr.[18] We expect ACoR to be able to cut down the number of false positives significantly, thus reducing the number of children currently flagged for diagnostic evaluation, which potentially cuts down the wait-time, thereby reducing diagnostic delays. Under Specific Aim 1, our goal is to track the fraction of cases where ACoR correctly signals no-risk while MCHAT-F produces a positive flag.

Significance of Specific Aim 2: In our preliminary studies, we established that ACoR outperforms M-CHAT/F by considering the average reported performance of M-CHAT/F in a recent study[23] on a large cohort with near-universal screening carried out at the Children's Hospital of Philadelphia (CHOP). However, not having observed the ACoR and the M-CHAT/F scores jointly for individual patients, our preliminary studies lack objective assessment of statistical dependence between the two scores. While the very nature of the methodologies suggest functional independence, specific Aim 2 will investigate and establish this rigorously.

Functional independence from existing tools implies we can combine the scores; especially leveraging the population stratification induced by the M-CHAT/F scores as reported by the CHOP study to significantly boost combined screening performance. In particular, since patients in the lower M-CHAT/F score bracket have a smaller chance of an ASD diagnosis compared to the high risk upper brackets, we can tailor the sensitivity/specificity trade-offs in the ACoR to maximize either the global PPV or the global sensitivity without losing specificity. Our preliminary results suggest that the expected gains are substantial, with the possibility of doubling the PPV, or increasing the sensitivity by over 50% while keeping the specificity above 95%. Specific aim 2 will investigate the e viability of the preliminary results in a pediatric primary care setting.

Significance of Specific Aim 3: While still lacking the certainty of a diagnostic blood test, use of subtle patterns emergent in the diagnostic history to estimate risk might help reduce the subjective component in questionnaire-based screening tools, resulting in reduced effect of potential language and cultural barriers in diverse populations.[1] With a significant portion of the cohort expected to be African Americans and Hispanics in our primary care clinic, our comparative investigations will be able to explicitly answer these questions.

Significance of Specific Aim 4: Despite unprecedented advances in charting the numerous genetic variations,[24] and established to be highly heritable,[25] the etiology of autism is still unclear.[26], [27] Despite tremendous recent progress, efforts to identify causal biomarkers for ASD have had limited success. Currently, over one hundred genes have been shown to contribute to autism risk,[24], [28], [29] and it is estimated that up to 1000 genes might be involved in ASD pathogenesis.[30] Still, genetic interactions and mechanisms have accounted for a limited number of ASD cases,[31] potentially implicating environmental triggers that work alongside genetic predispositions. The plausible sources of risk are estimated to range from prenatal factors such as maternal infection and inflammation, diet, and household chemical exposures, to autoimmune conditions and localized inflammation of the central nervous system after birth.[26], [27], [32]–[37] The heterogeneity of ASD presentation admits the possibility of a plurality of etiologies with converging pathophysiological pathways, making the investigation of the etiology of future risk modulation extremely challenging. Specific Aim 4 will aim to unravel clues to mechanistic drivers by charting and categorizing the heterogeneous presentation by the nature of past diagnostic pattern in individual medical histories.

4. Investigational Plan

Parallel Screening in Pediatric Primary Clinic: To achieve the specific goals outlined in the specific aims of this study, we will gather data from both the child and the primary caregiver of all patients in the participating UCM primary pediatric primary care clinic, and test the sensitivity and specificity of ACoR against the most commonly used screening tool M-CHAT/F, by an ADOS-2 based (near) gold-standard evaluation of patients flagged by either tool. Using these data and the inferred statistical dependency (or the lack thereof) properties between the screening tools, ACoR can tailor the selection of sensitivity/specificity trade-offs to the particular informant and to the age of the child, with the view to optimizing global characteristics such as maximizing the PPV or the sensitivity of the screening process. Additionally, the ACoR algorithm will identify categories of heterogeneity that will lead to mechanistic insights into ASD pathobiology.

Thus, the ACoR methodology is potentially able to screen instantaneously every child in primary care, for whom past medical history is available, with zero administrative and resource burden. Our results suggest that ACoR has superior predictive performance to existing screening tools, along with a host of other advantages that directly relate to the stated specific aims of this study.

The key steps in this project are as follows (See Fig. 1):

□ 1. Pediatric clinic team will adminster M-CHAT/F to incoming children with 16-30 months.

□ 2. The University of Chicago Research Informatics Support team will work with the PI and his team to compute the ACoR score corresponding to individual consenting patients

□ 3. If flagged by MCHAT-F, the patients will be scheduled for ADOS-2 evaluation (overseen by Dr. Smith).

□ 4. The evaluation scores will be analyzed by the PI and his team to address the specific aims.

Cohort Selection: We plan to establish the result as a universal procedure, that is applicable irrespective of population characteristics. Nevertheless, it is clearly conceivable that co-morbidities have demographic dependencies. Unfortunately, the key training dataset (Truven) does not have demographic information. Hence, the ACoR algorithm at this point does not take such variables into consideration. With a diverse patient population at the University of Chicago, we plan to *not* pre-select population characteristics, but analyze the performance of ACoR on the different demographic and ethnic groups in the post-processing.

In this pilot study we plan to have a cohort size of ≈ 200 children, which should result in nearly 12 MCHAT-F flags, out of which about 8 should be cleared by ADOS. Due to the very small sample size, we ignore statistical power calculations, but will analyze if our predicted numbers acceptably match with what we observe in Fig. 1.

Risk To Patients: The design of the study guarantees that patients suffer no negative impact from the added ACoR screen. Indeed, patients who are flagged are to be immediately scheduled for ADOS evaluation which eliminates their wait-times. For some borderline cases, which would have been missed by M-CHAT/F, might get flagged by ACoR, and be scheduled for ADOS, which they would not have had to do with just M-CHAT/F. But this is a positive outcome. The possibility that ACoR might have some false positives that are different from that of M-CHAT/F and hence schedule some children for ADOS, who do not have autism is a possibility, and might cause some stress in parents and families. The potential societal benefit gained in lieu of this discomfort is the validation of the expected performance boost for ASD screening at the population level PPV by up to 100%.

Procedures: Eligible patients at the Department of Pediatrics, University of Chicago (patients who present for a well-child visit or any other non-emergency reason) will be asked for consent for carrying out the ACoR screen in the background. The algorithm will be already integrated with EPIC, such that indicating this consent on the screen will automatically trigger pulling up the patient medical history if available, carrying out the calculations, and storing the results to be analyzed in post-processing. If there is a flag either in M-CHAT/F, the attending pediatrician will inform parents of a potential elevated risk of ASD diagnosis, and offer to schedule for an ADOS-2 evaluation.

All study procedures and consent forms will be approved by the University of Chicago Institutional Review Board. For all assessments, basic demographic information, recruitment site, medications and diagnoses assigned by the current clinical treatment team, will be obtained from the parent/caregiver and medical record.

The technical approach of this study consists of the development and extension of the ACoR methodology from our preliminary studies, and application in a primary care setting with view to carrying out objective comparisons between M-CHAT/F and ACoR. Furthermore, we plan to assess our ability to boost performance from a conditional combination of the two scores.

5. Preliminary Retrospective Results

We describe the formulation of the ACoR score, and the underlying principles that distill the estimated risk from EHR databases.

We view the task of predicting ASD diagnoses as a binary classification problem: sequences of diagnostic codes are classified into positive and control categories, where “positive” refers to children eventually diagnosed with ASD, as indicated by the presence of a clinical diagnosis (ICD9 code 299.X) in their medical records. Of the two independent sources of clinical incidence data used in this study, the primary source used to train our predictive pipeline is the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database for the years 2003 to 2012[38] (referred to as the Truven dataset). This US national database merges data contributed by over 150 insurance carriers and large self-insurance companies, and comprises over 4.6 billion inpatient and outpatient service claims and almost six billion diagnosis codes. We extracted histories of patients within the age of $0-6$ years, and excluded patients for whom one or more of the following criteria fails: 1) At least one code pertaining to one of the 17 disease categories we use (See later for discussion of disease categories) is present in the diagnostic history, 2) The first and last available record for a patient are at least 15 weeks apart. These exclusion criteria ensure that we are not considering patients who have too few observations. Additionally, during validation runs, we restricted the control set to patients observable in

TABLE 1

Standalone ACoR PPV Achieved

(For Comparison M-CHAT/F Performance: sensitivity=38.8%, specificity=95%, PPV=14.6% between within 26 months (≈112 weeks))

weeks	specificity	sensitivity	PPV	gender	dataset
100	0.92	0.39	0.14	F	UCM
100	0.95	0.39	0.19	M	UCM
100	0.93	0.39	0.13	F	Truven
100	0.91	0.39	0.10	M	Truven
112	0.93	0.39	0.16	F	UCM
112	0.95	0.39	0.20	M	UCM
112	0.96	0.39	0.22	F	Truven
112	0.95	0.39	0.17	M	Truven
150	0.94	0.39	0.19	F	UCM
150	0.98	0.39	0.34	F	Truven
150	0.97	0.39	0.26	M	Truven
150	0.97	0.39	0.26	M	UCM

the databases to those whose last record is not before the first 200 weeks of life. We trained with over 30M diagnostic records (16,649,548 for males and 14,318,303 for females with 9,835 unique codes).

While the Truven database is used for both training and out-of-sample cross-validation with held-back data, our second independent dataset comprising de-identified diagnostic records for children treated at the University of Chicago Medical Center between the years of 2006 to 2018 (the UCM dataset) aids in further cross-validation. We considered children between the ages of $0-6$ years, and applied the same exclusion criteria as the Truven dataset. Our datasets are consistent with documented ASD prevalence and median diagnostic age (3 years in the claims database versus 3 years 10 months to 4 years in US[13]) with no significant geospatial prevalence variation.

The significant diversity of diagnostic codes (6,089 distinct ICD-9 codes and 11,522 distinct ICD-10 codes in total in the two datasets), along with the sparsity of codes per sequence and the need to make good predictions as early as possible, makes this a difficult learning problem, and standard deep learning approaches do not yield sufficiently high predictive performance or statistical power. Thus, we proceed by partitioning the disease spectrum into 17 broad categories, $e.g.$ infectious diseases, immunologic disorders, endocrinal disorders etc. Each patient is then represented by 17 distinct time series, each tracking an individual disease category. At the population level, these disease-specific sparse stochastic time series are compressed into specialized Markov models (separately for the control and the treatment cohorts) to identify the distinctive patterns pertaining to elevated ASD risk. Each of these inferred models in a Probabilistic Finite State Automaton (PFSA).

We use a novel approach to evaluate subtle deviations in stochastic observations known as the sequence likelihood defect (SLD), to quantify similarity of observed time-series of diagnostic events to the control vs the positive cohorts for individual patients.

Calculating Relative Risk: Our pipeline maps medical histories to a score, which is interpreted as a raw indicator of risk — higher this value, higher the probability of a future diagnosis. However, to make crisp predictions, we must choose a decision threshold for this raw score. Conceptually identical to the notion of Type 1 and Type 2 errors in classical statistical analyses, the choice of a threshold trades off false positives (Type 1 error) for false negatives (Type 2 error): choosing a small threshold results in predicting a larger fraction of future diagnoses correctly, $i.e.$ have a high true positive rate (TPR), while simultaneously suffering from a higher false positive rate (FPR), and vice versa. We choose thresholds for the standalone ACoR method by maximizing the F_1-score, defined as the harmonic mean of sensitivity and specificity, to make a balanced trade-off between the two kinds of errors. The *relative risk* is then defined as the ratio of the raw pipeline score to the chosen decision threshold. While the raw score does not give us actionable information, the relative risk being close to or greater than 1.0 for a specific child signals the need for intervention.

Standalone Predictive Performance: The standalone performance of our risk estimator is summarized in Fig. 2A. The task of predicting if a patient has a future ASD diagnosis, at an earlier date compared to the

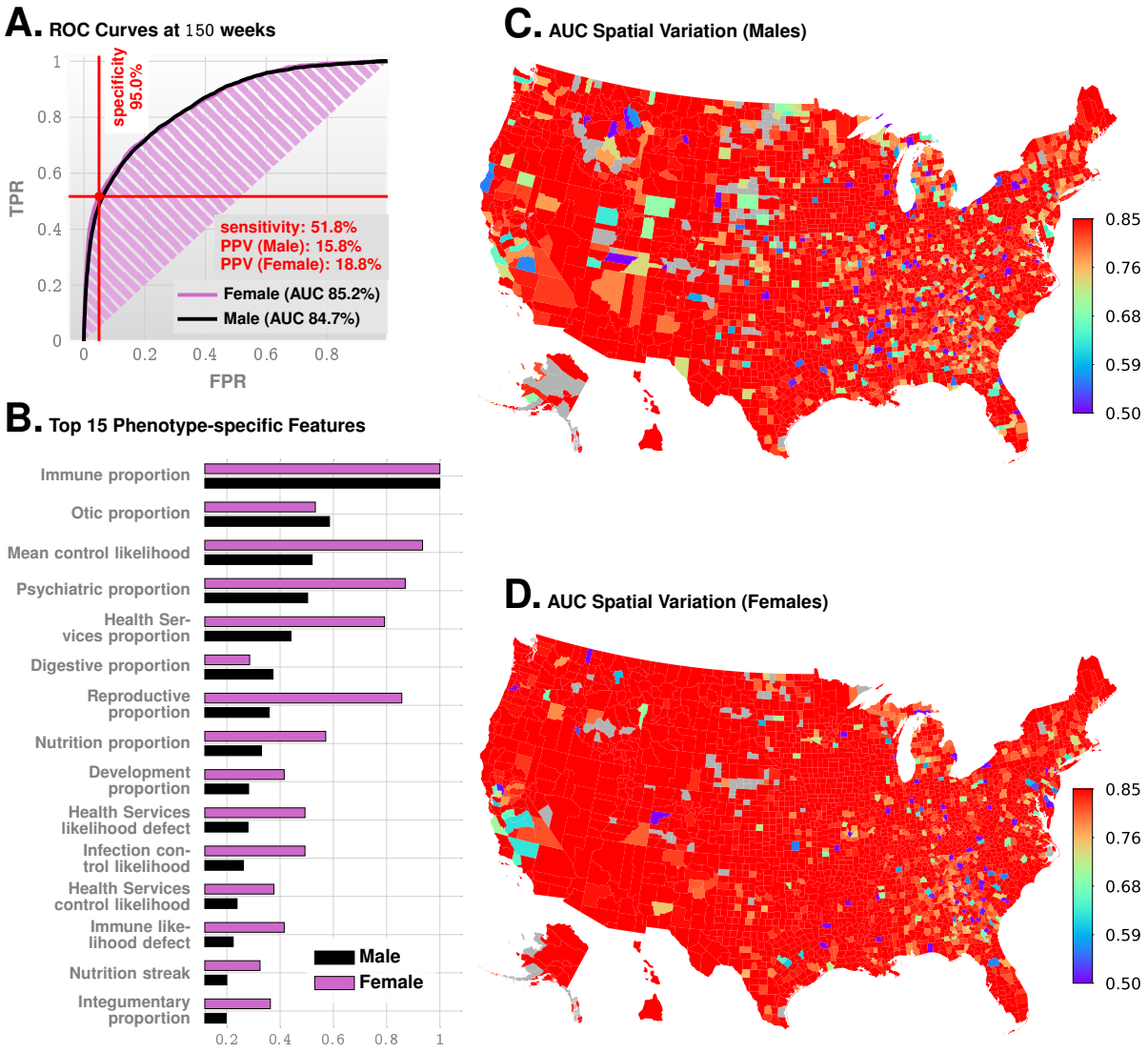

Fig. 2. **Standalone Predictive Performance of ACoR in Retrospective Studies.** Panel A shows the ROC curves for males and females (Truven data shown, UCM is similar). Panel B shows the feature importance inferred by our prediction pipeline. The most import feature is related to immunologic disorders, and we note that in addition to features related to individual disease categories, we also have the mean control likelihood (rank 3), which may be interpreted as the average likelihood of the diagnostic patterns correspond to the control category as opposed to the positive category. Panels C and D show the spatial variation in the achieved predictive performance at 150 weeks, measured by AUC, for males and females, respectively. Gray areas lack data on either positive or negative cases. These county-specific AUC plots show that the performance of the algorithm has relatively weak geospatial dependence, which is important in the light of current uneven distribution of diagnostic resources. Importantly, not all counties has nonzero number of ASD patients; a high performance in those counties reflects a small number of false positives with zero false negatives.

clinical diagnosis, may be viewed as a binary classification problem. In our case, we achieve an out-of-sample AUC of 82.3% for males and 82.5% for females at 125 weeks of age for the Truven dataset. In the UCM dataset, our performance is comparable: 83.1% and 81.3% for males and females respectively at 125 weeks of age. The good agreement of the out-of-sample performance on these independent datasets lends strong evidence for the claims made in this study. The specificity, sensitivity, PPV trade-offs are shown in Table 1.

We enumerate the top 15 predictive features in Fig. 2B. We also computed the county-specific performance of the risk pipeline for the Truven dataset, and we got nearly uniform performance across the country for both genders, with the exception of few isolated counties lacking patients in the appropriate age groups (See Fig. 2, panels C and D) which prevented us from estimating AUC for those counties. Thus the performance of the algorithm is relatively agnostic to the number of local diagnoses, which is import in light of the fact that crucial diagnostic resources currently have a very uneven distribution (only 7% of developmental pediatricians practice in rural areas, and some states in US do not even have a developmental pediatrician[18], [19]).

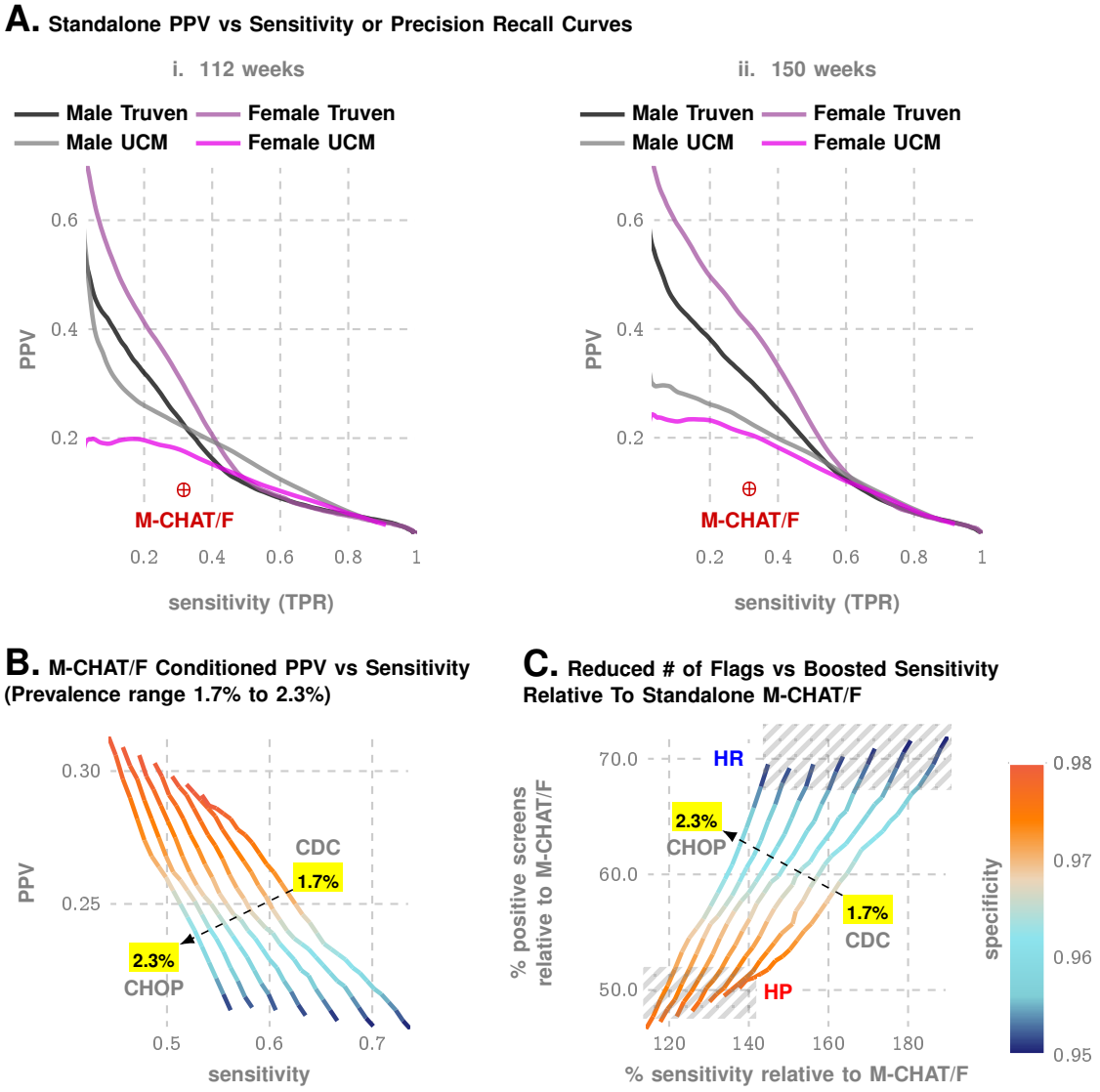

Fig. 3. **Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs.** Panel A shows the precision/recall curves, *i.e.*, the trade-off between PPV and sensitivity. Panel B shows how we can boost performance using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study.[23] Panel C illustrates the boosted performance compared to M-CHAT/F alone, mesured by the relative percentage increase in sensitivity, and percentage decrease in postive screens. Note that the population prevalence impacts this optimization, and hence we have a distinct curve for each prevalence value (1.7% is the CDC estimate, while 2.23% is reported by the CHOP study). The two extreme operating zones marked as High Precision (HP) and High Recall (HR): if we choose to operate in HR, then we do not reduce the number of positive screens by much, but maximize sensitivity, while by operating in HP, we do not increase seinsitivity by much but double the PPV achieved in current practice. Note in all these zones we maintain specificity above 95%, which is the current state of art, implying that by doubling the PPV, we can halve the number of positive screens currently reported, thus potentially sharply reducing the queues and wait-times.

Calculating PPV, Sensitivity & Specificity Trade-offs & M-CHAT/F Comparison: The sensitivity vs PPV plots, also known as the precision-recall curves (See Fig. 3A) are constructed in a similar fashion as the ROC curves by varying the decision threshold. These curves allow direct comparison with the state of the art screening tests,*e.g.*, M-CHAT/F, in a manner that is most relevant to clinical practitioners. Guthrie *et al.*[23] from Children's Hospital of Philadelphia (CHOP) has recently demonstrated that when applied as a nearly universal screening tool, M-CHAT/F has a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%, implying that out of every 100 children who in fact ave ASD, the M-CHAT/F flags about 39, and out of every 100 children it flags, about 85 are false positives. The PPV is affected by the prevalence of the disease. This work is the only large-scale study of M-CHAT/F (n=20,375) we are aware of with sufficient follow-up after the age of four years to provide a reasonable degree of confidence in the reported performance values.

Comparing the performance metrics achieved at different age groups across data sets and genders for our pipeline (See Table 1), we conclude that our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at 14% (14.1-33.6%) when sensitivity and specificity are held at comparable values around the age of

TABLE 2

Population Stratification Results on large M-CHAT/F Study(n=20,375) reproduced from Guthrie *et al.*[23]

Id	Sub-population	Test Result	ASD pos.	ASD Neg.	Total %
A	M-CHAT/F $\geqq 8$	Positive	0.34%	0.64%	0.99%
B	M-CHAT/F $\in [3,7]$	Positive (follow-up)	0.52%	4.39%	4.91%
C	M-CHAT/F $\in [3,7]$	Negative (follow-up)	0.14%	3.1%	3.24%
D	M-CHAT/F $\in [0,2]$	Negative	1.22%	89.63%	90.86%
Total %			2.23%	97.77%	100%

26 months (≈ 112 weeks). We cannot compare at other operating points due to limited availability of M-CHAT/F performance characterization at other time points.

Boosting Performance Via Leveraging Population Stratification Induced By M-CHAT/F: In our retrospective study, we leveraged the population stratification induced by an existing independent screening test (MCHAT-F) to improve combined performance. Here a combination refers to the conditional choice of the sensitivity/specificity trade-offs for our tool in each sub-population such that the overall performance is optimized with respect to whether we wish to maximize the PPV or the sensitivity at a specified minimum level of specificity. Assume that there are m sub-populations such that: the sensitivities, specificities achieved, and the prevalences in each sub-population are given by s_i, c_i and ρ_i respectively, with $i \in \{1, \cdots, m\}$. Let β_i be the relative size of each sub-population. Then, we can show:

$$s = \sum_{i=1}^{m} s_i \gamma_i, \text{ and } c = \sum_{i=1}^{m} c_i \gamma_i' \text{ where we have denoted: } \gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma_i' = \beta_i \frac{1-\rho_i}{1-\rho} \tag{1a}$$

and s, c, ρ are the overall sensitivity, specificity, and prevalence. Knowing the values of γ_i, γ_i', we can carry out an m-dimensional search to identify the feasible choices of s_i, c_i pairs for each i, such that some global constraint is satisfied, *e.g.* minimum values of specificity, sensitivity, and PPV. We consider 4 sub-populations defined by M-CHAT/F score brackets,[23] and if the screen result is considered a positive (high risk, indicating the need for a full diagnostic evaluation) or a negative, *i.e.*, low risk: 1) score ≤ 2 screening ASD negative, 2) score $[3-7]$ screening ASD negative on follow-up, 3) score $[3-7]$ and screening ASD positive on follow-up, and 4) score ≥ 8, screening ASD positive. (See Table 2). The "follow-up" in the context of M-CHAT/F refers to the re-evaluation of responses by qualified personnel. We use published data on the relative sizes and the prevalence statistics in these sub-populations[23] to compute the feasible conditional choices of our operating point to strictly supersede M-CHAT/F performance. Two limiting operating conditions are of special interest here, where we maximize PPV under some minimum specificity and sensitivity (denoted as the High Precision or the HP operating point), and where we maximize sensitivity under some minimum PPV and specificity (denoted as the High Recall or the HR operating point). Taking these minimum values of specificity, sensitivity, and PPV to be those reported for M-CHAT/F, we identify the set feasible set of conditional choices in a four dimensional decision space that would outperform M-CHAT/F in universal screening.

This ultimately yields an overall performance significantly superior to M-CHAT/F alone. We carry out a four dimensional search at the age of 26 months (≈ 112 weeks) to identify the feasible region with PPV $> 14.6\%$ and sensitivity $> 38.8\%$ simultaneously while keeping specificity $> 94\%$. These four dimensions reflect the independent choice of sensitivities in the corresponding sub-populations. For each set of choices, the associated specificities are read-off from our fixed pre-computed ROC curve corresponding to 112 weeks, and then the overall sensitivity, specificity and PPV are calculated using standard relationships.

We get a PPV close to or exceeding 30% at the high precision (HP) operating point across datasets ($> 33\%$ fro Truven, $> 28\%$ for UCM), or a sensitivity close to or exceeding 50% for the high recall (HR) operating point ($> 58\%$ for Truven, $> 50\%$ for UCM), when we restrict specificities to above 95% (See Table 3).

It is important to compare these results directly with M-CHAT/F performance, as shown in Fig. 3, panels C. In panel C, we show that for any stable population prevalence between 1.7% and 2.23%, the conditional operation can achieve double the PPV relative to M-CHAT/F alone without losing sensitivity at $> 98\%$ specificity, or increase the sensitivity by $\sim 50\%$ without sacrificing PPV and not letting the specificity to drop below 94%.

Importantly, designing the rules for conditional operation only requires average population characteristics,

TABLE 3

Boosted Sensitivity, specificity and PPV Achieved at 26 months Personalized Operation Conditioned on M-CHAT/F Scores

M-CHAT/F Outcome				global performance (Truven)			global performance (UCM)			prevalence*
0-2 NEG	3-7 NEG	3-7 POS	≥ 8 POS	speci-ficity	sensi-tivity	PPV	speci-ficity	sensi-tivity	PPV	
specificity choices										
0.2	0.54	0.83	0.98	0.95	0.585	0.209	0.95	0.505	0.186	0.022
0.21	0.53	0.83	0.98	0.95	0.586	0.208	0.95	0.506	0.184	0.022
0.42	0.87	0.98	0.99	0.98	0.433	0.331	0.98	0.347	0.284	0.022
0.48	0.87	0.97	0.99	0.98	0.432	0.331	0.98	0.355	0.289	0.022
0.38	0.54	0.94	0.98	0.95	0.736	0.203	0.95	0.628	0.178	0.017
0.3	0.55	0.94	0.98	0.95	0.737	0.203	0.95	0.633	0.179	0.017
0.58	0.96	0.98	0.99	0.98	0.492	0.302	0.98	0.373	0.247	0.017
0.59	0.96	0.98	0.99	0.98	0.491	0.303	0.98	0.372	0.248	0.017
0.46	0.92	0.97	0.99	0.977	0.534	0.291	0.977	0.448	0.256	0.017
0.48	0.92	0.97	0.99	0.978	0.533	0.292	0.978	0.448	0.257	0.017

*Prevalence reported by CDC is 1.7%, while the CHOP study reports a value of 2.23%. Results depend on the prevalence estimate.

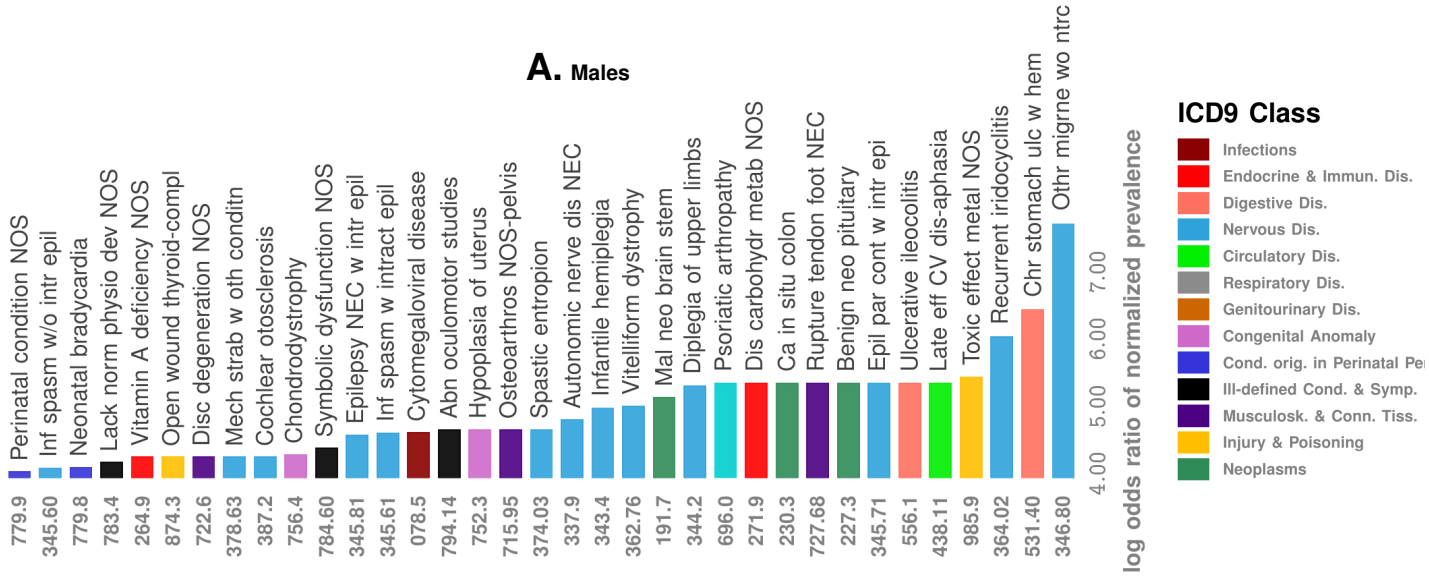

Fig. 4. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions in males. The color coding shows the disease categories of the co-morbidities.

i.e., an estimate of ASD prevalence in the sub-populations defined by the relevant brackets of M-CHAT/F scores, and the prevalence of these score brackets in the general population. In particular, M-CHAT/F scores of individual patients are unnecessary for designing the rules themselves, or evaluating the overall expected performance in the population, provided the stratification statistics (Table 2) remains invariant. However, in the proposed study, we will be able to compute explicit dependency relationships between M-CHAT?F and ACoR scores.

Inferred Co-morbidity Patterns & Normalized Prevalence Comparison: The predictive ability of our pipeline arises from the difference in patterns of co-morbid disorders between the positive and the control cohorts: the diagnostic history of individual patients is not random and hides key signatures to future neuropsychiatric outcomes. While the ASD co-morbidity burden is reported to be high for nearly the entire spectrum of physiological disorders, in our preliminary we find novel association patterns in normalized prevalence — the odds of experiencing a specific disorder, particularly in the early years (age < 3 years), normalized over all unique disorders experienced in the specified time-frame. Additionally, we only focus on the true positives

in the positive cohort and the true negatives in the control cohort. This allows us to investigate patterns that correctly disambiguate future ASD status, $i.e.$, strongly favor one outcome over the other at the individual level (as opposed to population-level prevalence rates), as shown in Fig. 4 for males.

Disambiguation From Unrelated Psychiatric Phenotypes: In our retrospective analyses, we can discriminate between ASD and other unrelated psychiatric phenotypes. Does our pipeline pick up on any psychiatric conditions, or is it specific to ASD? We evaluated this question, by restricting the control cohort in validation to patients with at least one psychiatric code other than ASD. We get very high discrimination reaching AUCs over 90% at $100 - 125$ weeks of age, which establishes that our pipeline is indeed largely specific to ASD.

Sanity Checks: Uncertainty in EHR Records: Recent changes in diagnostic practice, $e.g.$ increased diagnoses from individual clinicians versus prior eras that only allowed diagnosis from the gold-standard multidisciplinary teams can increase observed prevalence, and raises the possibility that some diagnostic codes pertaining to ASD in medical history databases could be arising from less restrictive workflows, and are susceptible to increased uncertainty. In our study, we verified that restricting the positive cohort to children with at least two distinct ASD diagnostic codes in their medical histories instead of one, has little impact on out-of-sample predictive performance.

We also found that the density of diagnostic codes in a child's medical history by itself is somewhat predictive of a future ASD diagnosis, but not at clinically significant levels.

REFERENCES

[1] Hyman, S. L., Levy, S. E., Myers, S. M. et al. Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics **145** (2020).

[2] Kalb, L. G. et al. Determinants of appointment absenteeism at an outpatient pediatric autism clinic. Journal of Developmental & Behavioral Pediatrics **33**, 685–697 (2012).

[3] Bisgaier, J., Levinson, D., Cutts, D. B. & Rhodes, K. V. Access to autism evaluation appointments with developmental-behavioral and neurodevelopmental subspecialists. Archives of pediatrics & adolescent medicine **165**, 673–674 (2011).

[4] Fenikilé, T. S., Ellerbeck, K., Filippi, M. K. & Daley, C. M. Barriers to autism screening in family medicine practice: a qualitative study. Primary health care research & development **16**, 356–366 (2015).

[5] Gordon-Lipkin, E., Foster, J. & Peacock, G. Whittling Down the Wait Time: Exploring Models to Minimize the Delay from Initial Concern to Diagnosis and Treatment of Autism Spectrum Disorder. Pediatr. Clin. North Am. **63**, 851–859 (2016).

[6] Robins, D. L. et al. Validation of the modified checklist for autism in toddlers, revised with follow-up (m-chat-r/f). Pediatrics **133**, 37–45 (2014).

[7] Jashar, D. T., Brennan, L. A., Barton, M. L. & Fein, D. Cognitive and adaptive skills in toddlers who meet criteria for autism in dsm-iv but not dsm-5. Journal of autism and developmental disorders **46**, 3667–3677 (2016).

[8] Data & statistics on autism spectrum disorder — cdc (2019). URL https://www.cdc.gov/ncbddd/autism/data.html.

[9] Tye, C., Runicles, A. K., Whitehouse, A. J. O. & Alvares, G. A. Characterizing the Interplay Between Autism Spectrum Disorder and Comorbid Medical Conditions: An Integrative Review. Front Psychiatry **9**, 751 (2018).

[10] Kohane, I. S. et al. The co-morbidity burden of children and young adults with autism spectrum disorders. PLoS ONE **7**, e33224 (2012).

[11] Christensen, D. L. et al. Prevalence and characteristics of autism spectrum disorder among children aged 4 years—early autism and developmental disabilities monitoring network, seven sites, united states, 2010, 2012, and 2014. MMWR Surveillance Summaries **68**, 1 (2019).

[12] Burkett, K., Morris, E., Manning-Courtney, P., Anthony, J. & Shambley-Ebron, D. African american families on autism diagnosis and treatment: The influence of culture. Journal of Autism and Developmental Disorders **45**, 3244–3254 (2015).

[13] Baio, J. et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. MMWR Surveill Summ **67**, 1–23 (2018).

[14] King, M. & Bearman, P. Diagnostic change and the increased prevalence of autism. Int J Epidemiol **38**, 1224–1234 (2009).

[15] Elsabbagh, M. et al. Global prevalence of autism and other pervasive developmental disorders. Autism Res **5**, 160–179 (2012).

[16] Schieve, L. A. et al. Population attributable fractions for three perinatal risk factors for autism spectrum disorders, 2002 and 2008 autism and developmental disabilities monitoring network. Ann Epidemiol **24**, 260–266 (2014).

[17] Volkmar, F. et al. Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. Journal of the American Academy of Child & Adolescent Psychiatry **53**, 237–257 (2014).

[18] Gordon-Lipkin, E., Foster, J. & Peacock, G. Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. Pediatric Clinics **63**, 851–859 (2016).

[19] Althouse, L. A. & Stockman, J. A. Pediatric workforce: A look at pediatric nephrology data from the american board of pediatrics. The Journal of pediatrics **148**, 575–576 (2006).

[20] Kozlowski, A. M., Matson, J. L., Horovitz, M., Worley, J. A. & Neal, D. Parents' first concerns of their child's development in toddlers with autism spectrum disorders. Dev Neurorehabil **14**, 72–78 (2011).

[21] Herlihy, L., Knoch, K., Vibert, B. & Fein, D. Parents' first concerns about toddlers with autism spectrum disorder: Effect of sibling status. Autism **19**, 20–28 (2015).

[22] Chawarska, K., Macari, S. & Shic, F. Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. Biological psychiatry **74**, 195–203 (2013).

[23] Guthrie, W. et al. Accuracy of Autism Screening in a Large Pediatric Network. Pediatrics **144** (2019).

[24] Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. bioRxiv (2019). https://www.biorxiv.org/content/early/2019/04/24/484113.full.pdf.

[25] Sandin, S. et al. The Heritability of Autism Spectrum DisorderReassessing the Heritability of Autism Spectrum DisordersLetters. JAMA **318**, 1182–1184 (2017). URL https://doi.org/10.1001/jama.2017.12141. https://jamanetwork.com/journals/jama/articlepdf/2654804/jama_sandin_2017_ld_170037.pdf.
[26] Ohja, K. et al. Neuroimmunologic and Neurotrophic Interactions in Autism Spectrum Disorders: Relationship to Neuroinflammation. Neuromolecular Med. **20**, 161–173 (2018).
[27] Gadysz, D., Krzywdziska, A. & Hozyasz, K. K. Immune Abnormalities in Autism Spectrum Disorder-Could They Hold Promise for Causative Treatment? Mol. Neurobiol. **55**, 6387–6435 (2018).
[28] Sanders, S. J. et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron **87**, 1215–1233 (2015).
[29] Gaugler, T. et al. Most genetic risk for autism resides with common variation. Nat. Genet. **46**, 881–885 (2014).
[30] Werling, D. M. The role of sex-differential biology in risk for autism spectrum disorder. Biol Sex Differ **7**, 58 (2016).
[31] Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. Nat. Rev. Genet. **9**, 341–355 (2008).
[32] Yamashita, Y. et al. Anti-inflammatory Effect of Ghrelin in Lymphoblastoid Cell Lines From Children With Autism Spectrum Disorder. Front Psychiatry **10**, 152 (2019).
[33] Shen, L. et al. Proteomics Study of Peripheral Blood Mononuclear Cells (PBMCs) in Autistic Children. Front Cell Neurosci **13**, 105 (2019).
[34] Theoharides, T. C., Tsilioni, I., Patel, A. B. & Doyle, R. Atopic diseases and inflammation of the brain in the pathogenesis of autism spectrum disorders. Transl Psychiatry **6**, e844 (2016).
[35] Young, A. M. et al. From molecules to neural morphology: understanding neuroinflammation in autism spectrum condition. Mol Autism **7**, 9 (2016).
[36] Croen, L. A. et al. Family history of immune conditions and autism spectrum and developmental disorders: Findings from the study to explore early development. Autism Res **12**, 123–135 (2019).
[37] Zerbo, O. et al. Immune mediated conditions in autism spectrum disorders. Brain Behav. Immun. **46**, 232–236 (2015).
[38] Hansen, L. The truven health marketscan databases for life sciences researchers. Truven Health Ananlytics IBM Watson Health (2017).