Reduced False Positives in Autism Screening Via Digital Bio-markers Inferred from Deep Co-morbidity Patterns

Dmytro Onishchenko[a], Yi Huang[a], James van Horne[a], Peter J. Smith[d,g], Michael M. Msall[e,f], Ishanu Chattopadhyay[a,b,c,*]

[a]*Department of Medicine, University of Chicago, Chicago, IL, USA*

[b]*Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL, USA*

[c]*Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL, USA*

[d]*Department of Pediatrics, Section of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL, USA*

[e]*Department of Pediatrics, Section Chief of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL, USA*

[f]*Joseph P. Kennedy Research Center on Intellectual and Neurodevelopmental Disabilities, University of Chicago, Chicago, IL, USA*

[g]*Executive Committee Chair, American Academy of Pediatrics' Section on Developmental and Behavioral Pediatrics*

Abstract

Autism spectrum disorder (ASD) is a developmental disability associated with significant social, communication, and behavioral challenges. There is a need for tools that help identify children with ASD as early as possible[1,2]. Our current incomplete understanding of ASD pathogenesis, and the lack of reliable biomarkers hampers early detection, intervention, and developmental trajectories. In this study we develop and validate machine inferred digital biomarkers for autism using individual diagnostic codes already recorded during medical encounters. Our risk estimator identifies children at high risk with a corresponding area under the receiver operating characteristic curve (AUC) exceeding 80% from shortly after two years of age for either sex, and across two independent databases of patient records. Thus, we systematically leverage ASD co-morbidities - with no requirement of additional blood work, tests or procedures - to predict elevated risk during the earliest childhood years, when intervention is the most effective. Our methodology has superior performance to common questionnaires-based screenings such as the M-CHAT/F[3], and has the potential to reduce socio-economic, ethnic and demographic biases. Further, by conditioning on the individual M-CHAT/F scores, we can either halve the false positives or boost sensitivity by over 50%, while maintaining specificity above 95%. Translated into practice, our algorithmic approach could significantly reduce the median diagnostic age for ASD, and also reduce long post-screen wait-times[4] currently experienced by families for confirmatory diagnoses and access to evidence based interventions.

INTRODUCTION

Autism spectrum disorder is a developmental disability associated with significant social, and behavioral challenges. Even though ASD may be diagnosed as early as the age of two[5], children frequently remain undiagnosed until after the fourth birthday[6]. At this time, there are no laboratory tests for ASD, so a careful review of behavioral history, and a direct observation of symptoms is necessary[7,8] for a clinical diagnosis. Starting with a positive initial screen, a confirmed diagnosis of ASD is a multi-step process that often takes 3 months to 1 year, delaying entry into time-critical intervention programs. While lengthy evaluations[9], cost of care[10], lack of providers[11], and lack of comfort in diagnosing ASD by primary care providers[11] are all responsible to varying degrees[12], one obvious source of this delay is the number of false positives produced in the initial ASD-specific screening tools in use today. For example, the M-CHAT/F used commonly as a screening tool[8,13], has an estimated sensitivity of 38.8%, specificity of 94.9% and Positive Predictive Value (PPV) of 14.6%[3]. Thus, currently out of every 100 children with ASD, M-CHAT/F flags about 39, and out of every 100 children it flags, about 85 are false positives, exacerbating wait times and queues[12]. Automated screening that might be administered with no specialized training, requires no behavioral observations, and is functionally independent of the tools employed in current practice, has the potential for immediate transformative impact on patient care.

*To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

While the neurobiological basis of autism remains poorly understood, a detailed assessment conducted by the US Centers for Disease Control and Prevention (CDC) demonstrated that children with ASD experience higher than expected rates of many diseases[5]. These include conditions related to dysregulation of immune pathways such as eczema, allergies, asthma, as well as ear and respiratory infections, gastrointestinal problems, developmental issues, severe headaches, migraines, and seizures[14,15]. In the present study, we exploit these co-morbidities to estimate the risk of childhood neuropsychiatric disorders on the autism spectrum. Using sequences of diagnostic codes from past doctor's visits, our risk estimator reliably predicts an eventual clinical diagnosis — or the lack thereof — for individual patients. Thus, the key clinical contribution of this study is the formalization of subtle co-morbidity patterns as a reliable screening tool, and potentially improve wait-times for diagnostic evaluations by significantly reducing the number of false positives encountered in initial screens in current practice.

A screening tool that tracks the risk of an eventual ASD diagnosis, based on the information already being gathered during regular doctor's visits, and which may be implemented as a fully automated background process requiring no time commitment from providers has the potential to reduce avoidable diagnostic delays at no additional burden of time, money and personnel resources. While still lacking the certainty of a diagnostic blood test, use of patterns emergent in the diagnostic history to estimate risk might help reduce the subjective component in questionnaire-based screening tools, resulting in 1) reduced effect of potential language and cultural barriers in diverse populations, and 2) possibly better identify children with milder symptoms[8]. Furthermore, being functionally independent of the M-CHAT/F, we show that there is clear advantage to combining the outcomes of the two tools: we can take advantage of any population stratification induced by the M-CHAT/F scores to significantly boost combined screening performance (See Materials & Methods, and Supplementary text, section 9).

MATERIALS & METHODS

Source of Electronic Patient Records

We view the task of predicting ASD diagnosis 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[16] (referred to as the Truven dataset). This US national database contains data contributed by over 150 insurance carriers, large self-insuring companies, and is a culmination of over 4.6 billion inpatient and outpatient service claims and almost six billion diagnosis codes. We extracted histories of patients within the age of $0-9$ years, and excluded patients for whom: 1) At least one code of any available phenotypes is present, 2) Lag between first and last available record for a patient should be at least 15 weeks. These exclusion criteria ensure that we are not considering patients who have too few observations to either train on. Additionally, during validation runs, we restricted the control set to patients observable in the databases to those whose last record is not before the first 150 weeks of life. Characteristics of excluded patients is shown in Table 1a. 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 consisting of 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 $0-5$ years, and applied the same exclusion criteria as the Truven dataset. The number of patients used from the two databases is shown in Table 1a.

Our datasets are consistent with documented prevalence. The median diagnosis age is just over 3 years in the claims database versus 3 years 10 months to 4 years in US[17]. Cohort details are given in Table 1a and discussed in Methods. For the positive cohort, we only consider diagnostic history up to the first ASD code.

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, where standard deep learning approaches do not suffice (See Table **??**). To address these issues, 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. With these inferred patterns included as features (Table 1b) we train a second level predictor that learns to map individual patients

Table 1: Patient Numbers, Inclusion-exc;usion Criteria and Features Used In Analysis

(a) Patient Counts In De-identified Data & The Fraction of Datasets Excluded By Our Exclusion Criteria*

Distinct Patients	Truven		UCM	
	115,805,687		69,484	
	Male	Female	Male	Female
ASD Diagnosis Count†	12,146	3,018	307	70
Control Count†	2,301,952	2,186,468	20,249	17,386
AUC at 125 weeks	82.3%	82.5%	83.1%	81.37%
AUC at 150 weeks	84.79%	85.26%	82.15%	83.39%
Excluded Fraction of the Data sets				
Positive Category	0.0002	0.0	0.0160	0.0
Control Category	0.0045	0.0045	0.0413	0.0476
Average Number of Diagnostic Codes In Excluded Patients (corresponding number in included patients)				
Positive Category	4.33 (35.93)	0.0 (36.07)	2.6 (9.75)	0.0 (10.18)
Control Category	1.57 (17.06)	1.48 (15.96)	2.32 (6.8)	2.07 (6.79)

† Cohort sizes are smaller than the total number of distinct patients due to the following exclusion criteria: 1) At least one code within our complete set of tracked diagnostic codes is present in the patient record, 2) Time-lag between first and last available record for a patient is at least 15 weeks.
* Dataset sizes are after the exclusion criteria are applied

(b) Engineered Features (Total Count: 165)

Feature Type‡	**Description**	**No. of Features**
[Disease Category] $_{\triangle}$	Likelihood Defect (See Methods section)	17
[Disease Category] $_{0}$	Likelihood of control model (See Methods section)	17
[Disease Category] $_{\text{proportion}}$	Occurrences in the encoded sequence / length of the sequence	17
[Disease Category] $_{\text{streak}}$	Maximum Length of adjacent occurrences of [Disease Category]	51
[Disease Category] $_{\text{prevalence}}$	Maximum, mean and variance of Occurrences in the encoded sequence / Total Number of diagnostic codes in the mapped sequence	51
Feature Mean, Feature Variance, Feature Maximum for difference of control and case models	Mean, Variance, Maximum of the [Disease Category] $_{\triangle}$ values	3
Feature Mean, Feature Variance, Feature Maximum for control models	Mean, Variance, Maximum of the [Disease Category] $_{0}$ values	3
Streak	Maximum, mean and variance of the length of adjacent occurrences of [Disease Category]	3
Intermission	Maximum, mean and variance of the length of adjacent empty weeks	3

‡ Disease categories are described in Table **??**

to the control or the positive groups based on their similarity to the identified Markov models of category-specific diagnostic histories (See Methods).

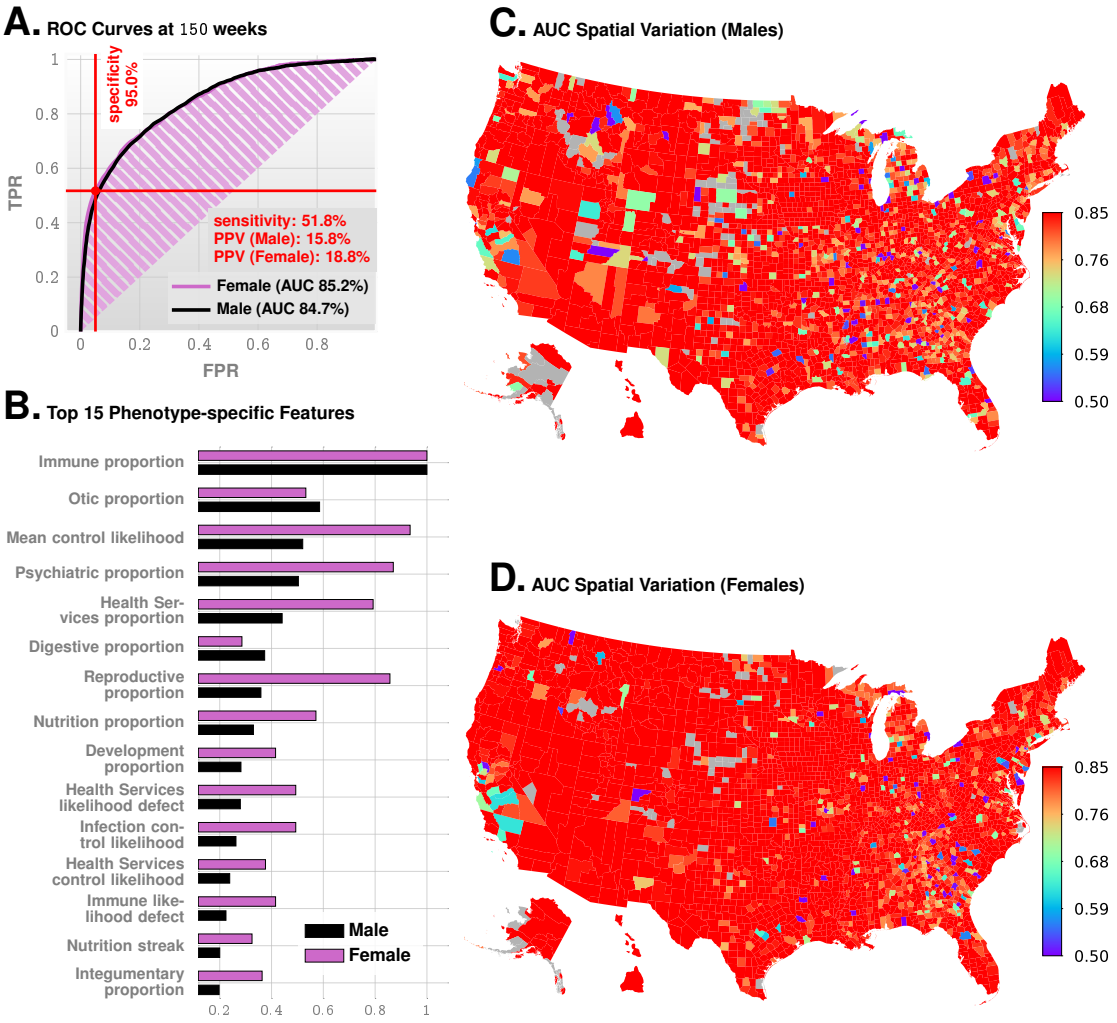

Figure 1: Predictive Performance. Panel A shows the ROC curves for males and females. Panel B shows the feature importance inferred by our prediction pipeline. The detailed description of the features is given in Table **??**. The most import feature is related to immunologic disorders, and we note that 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.

Calculating Relative Risk

Our pipeline maps medical histories to a raw indicator of risk. 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. Therefore, a choice of a specific decision threshold reflects a choice of the maximum FPR and minimum TPR, and is driven by the application at hand. In this study, we base our analysis on 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 (See Supplementary text, Section 5). The *relative risk* is then defined as the ratio of the raw risk to the decision threshold, and a value > 1 predicts a future ASD diagnosis.

Boosting Performance Via Leveraging Population Stratification Induced By Existing Tests

We leverage the population stratification induced by an existing independent screening test (M-CHAT/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. We consider 4 sub-populations defined by M-CHAT/F score brackets[3], and if the screen result is considered a positive

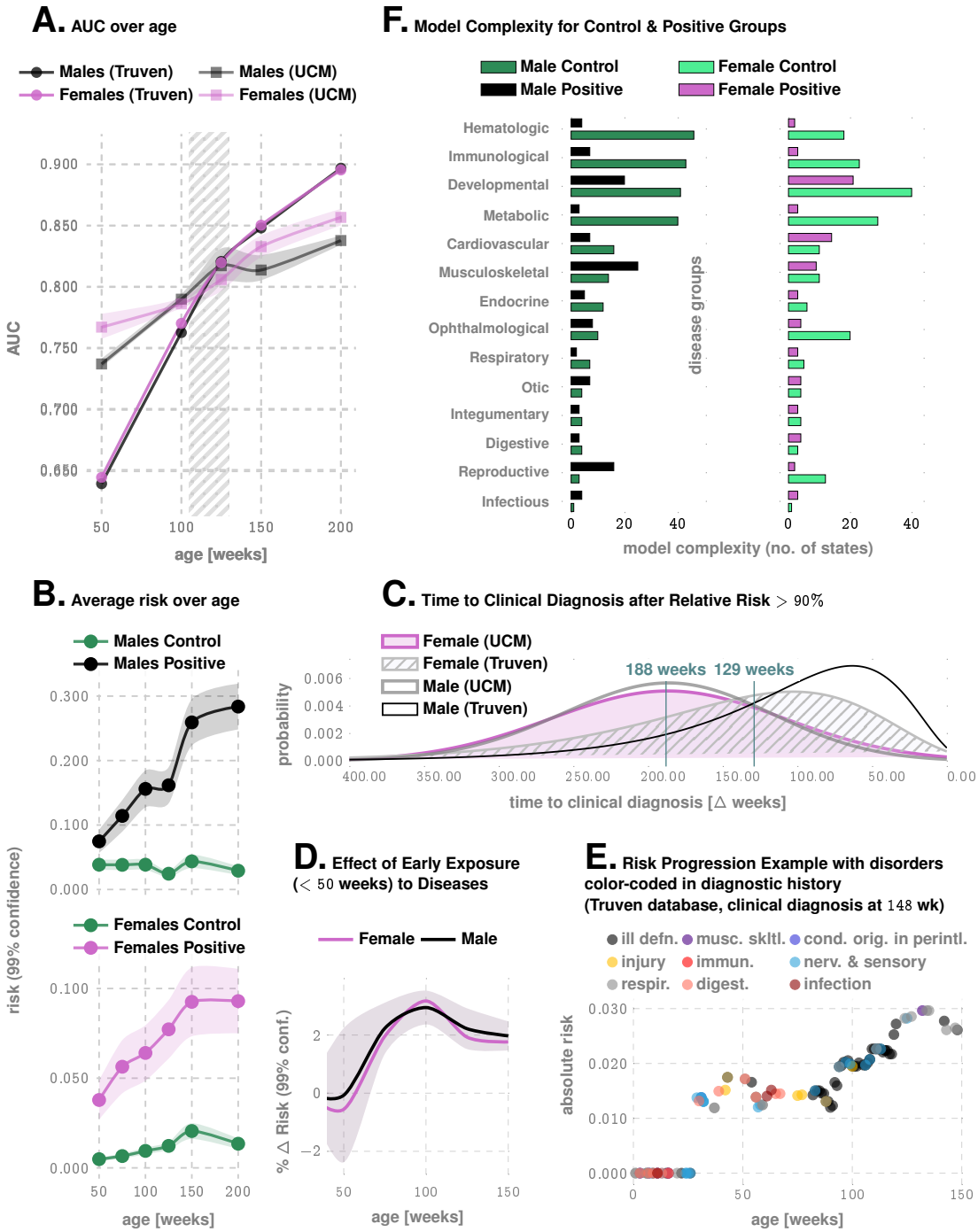

Figure 2: **More details on Predictive Performance and Variation of Inferred Risk.** Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve $> 80\%$ AUC for either sex from shortly after 2 years. Panel B illustrates how the average risk changes with time for the control and the positive cohorts. Panel C shows the distribution of the prediction horizon: the time to a clinical diagnosis after inferred relative risk crosses 90%. Panel D shows that for each new disease code for a low-risk child, ASD risk increases by approximately 2% for either sex. Panel E illustrates the risk progression of a specific, ultimately autistic male child in the Truven database. Abbreviations in the legend: ill defn. (Symptoms, Signs, And Ill-Defined Conditions), musc. skltl. (Diseases Of The Musculoskeletal System And Connective Tissue), cond. orig. in perintl. (Certain Conditions Originating In The Perinatal Period), immun. (Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders), nerv. & sensory (Diseases Of The Nervous System And Sense Organs), respir. (Respiratory Disorders), and digest. (Digestive Disorders). Panel F illustrates how inferred models differ between the control vs. the positive cohorts. On average, models get less complex, implying the exposures get more statistically independent.

(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 SI-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

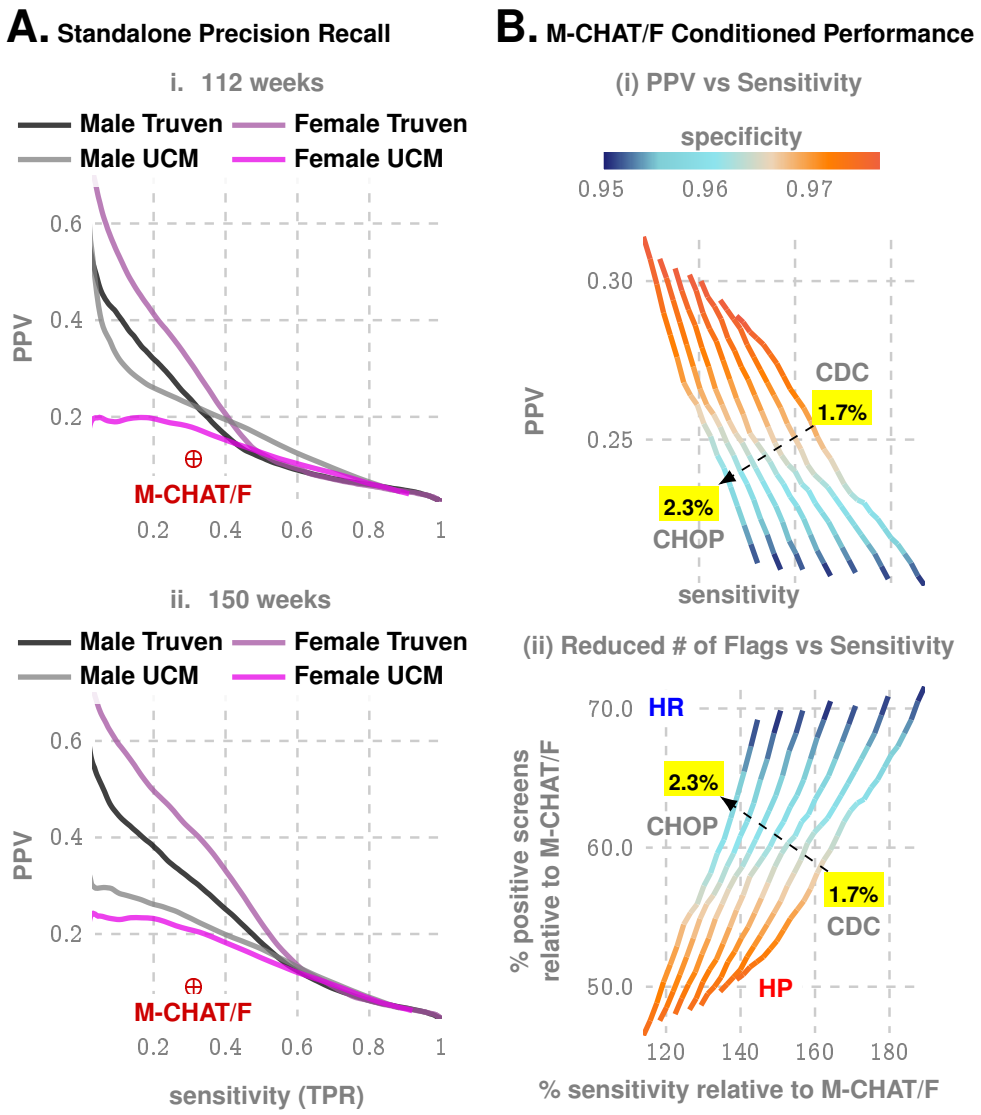

Figure 3: **Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs.** Panel A shows the precision/recall curves, *i.e.*, the trade-off between PPV and sensitivity. Panel B shows how we can boost performance using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study[3]. Panel C illustrates the boosted performance compared to M-CHAT/F alone, measured by the relative percentage increase in sensitivity, and percentage decrease in positive 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 sensitivity 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.

the prevalence statistics in these sub-populations[3] 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. The results are shown in Fig. 3B.

RESULTS

We measure our performance using several standard metrics including the AUC, sensitivity, specificity and the PPV. For the prediction of the eventual ASD status, we achieve an out-of-sample AUC of 82.3% and 82.5% for males and females respectively at 125 weeks for the Truven dataset. In the UCM dataset, our performance is

Table 2: Standalone and Combined Performance

(a) Standalone PPV Achieved at 100, 112 and 150 Weeks For Each Dataset and Gender **(M-CHAT/F: sensitivity=38.8%, specificity=95%, PPV=14.6% between 16 and 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

(b) Personalized Operation Conditioned on M-CHAT/F Scores at 26 months

M-CHAT/F Outcome				global performance (Truven)			global performance (UCM)			prevalence*
0-2 NEG	3-7 NEG	3-7 POS	≥ 8 POS	specificity	sensitivity	PPV	specificity	sensitivity	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%. The results of our optimization depend on the prevalence estimate.

comparable: 83.1% and 81.3% for males and females respectively (Fig. 1 and 2). Our AUC is shown to improve approximately linearly with patient age: Fig. 2A illustrates that the AUC reaches 90% in the Truven dataset at the age of four. Importantly, we train our pipeline on 50% of the Truven dataset, and use held back data from Truven, and the entirety of the UCM dataset for validation: *No new training is done in the UCM dataset.* Good performance on these independent datasets lends strong evidence for our claims. Furthermore, applicability in new datasets *without local re-training* makes it readily deployable in clinical settings.

What are the inferred patterns that elevate risk? Enumerating the top 15 predictive features (Fig. 1B), ranked according to their automatically inferred weights (the feature "importances"), we found that while infections and immunologic disorders are the most predictive, there is significant effect from all the 17 disease categories. Thus, the co-morbid indicators are distributed across the disease spectrum, and no single disorder is uniquely implicated (See also Fig. 2F). Importantly, predictability is relatively agnostic to the number of local cases across US counties (Fig. 1C-D) which is important in light of the current uneven distribution of diagnostic resources[12,18] across states and regions.

Unlike individual predictions which only become relevant over 2 years, the average risk over the populations is

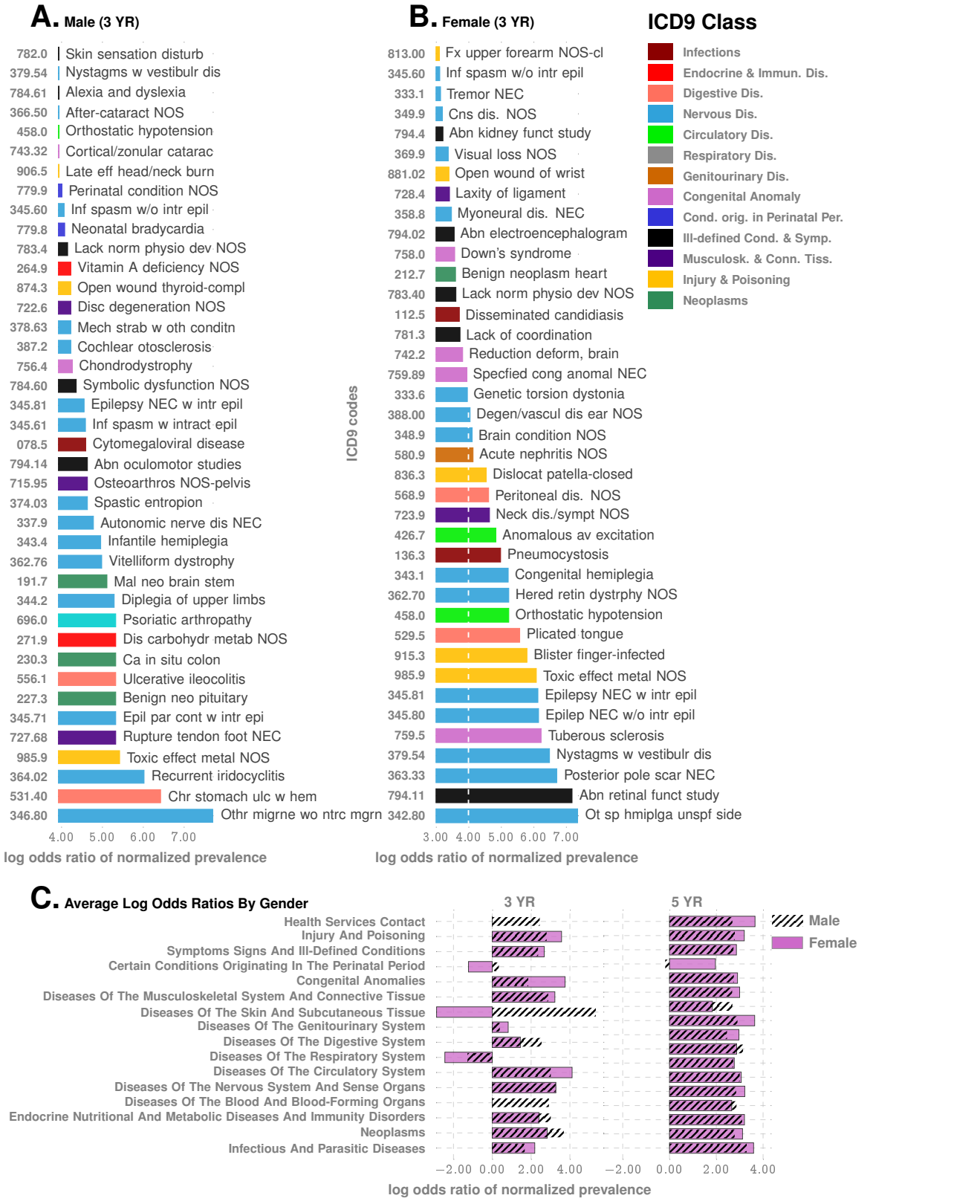

Figure 4: **Co-morbidity Patterns** Panel A and B. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions. The dotted line on panel B shows the abscissa lower cut-off in Panel A, illustrating the lower prevalence of codes in females. Panel C illustrates log-odds ratios for ICD9 disease categories at different ages. Importantly, the negative associations disappear when we consider older children, consistent with the lack of such reports in the literature which lack studies on very young cohorts.

clearly different from around the first birthday (Fig. 2B), with the risk for the positive cohort rapidly rising. Also, we see a saturation of the risk after ≈ 3 years, which corresponds to the median diagnosis age in the database (See Fig. **??**B). Thus, if a child is not diagnosed up to that age, then the risk falls, since the probability of a diagnosis in the population starts to go down after this age. While average discrimination is not useful for individual patients, these reveal important clues as to how the risk evolves over time. Additionally, while each new diagnostic code

within the first year of life increases the risk burden by approximately 2% irrespective of sex (Fig. 2D), distinct categories modulate the risk differently, *e.g.*, for a single random patient illustrated in Fig. 2F infections and immunological disorders dominate early, while diseases of the nervous system and sensory organs, as well as ill-defined symptoms, dominate the latter period.

Given these results, it is important to ask how much earlier can we trigger an intervention? On average, the first time the relative risk (risk divided by the decision threshold set to maximize F1-score, see Methods) crosses the 90% threshold precedes diagnosis by ≈ 188 weeks in the Truven dataset, and ≈ 129 weeks in the UCM dataset. This does not mean that we are leading a possible clinical diagnosis by over 2 years; a significant portion of this delay arises from families waiting in queue for diagnostic evaluations. Nevertheless, since delays are rarely greater than one year[12], we are still likely to produce valid red flags significantly earlier than the current practice.

Our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at least 14% (14.1-33.6%) when sensitivity and specificity are held at comparable values (approx. 38% and 95%) around the age of 26 months (≈ 112 weeks). Fig. 3A and Table 2a show the out-of-sample PPV vs sensitivity curves for the two databases, stratified by sex, computed at 100, 112 and 100 weeks. A single illustrative operating point is also shown on the ROC curve in Fig. 1C, where at 150 weeks, we have a sensitivity of 51.8% and a PPV of 15.8% and 18.8% for males and females respectively, both at a specificity of 95%.

Beyond standalone performance, independence from standardized questionnaires implies that we stand to gain substantially from combined operation. With the recently reported population stratification induced by M-CHAT/F scores[3] (SI-Table 3), we can compute a conditional choice of sensitivity for our tool, in each sub-population (M-CHAT/F score brackets: $0-2$, $3-7$ (negative assessment), $3-7$ (positive assessment), and >8), leading to a significant performance boost. With such conditional operation, we get a PPV close to or exceeding 30% at the high precision (HP) operating point across datasets ($>33\%$ for 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 2b, Fig. 3B(i), and SI-Fig. 4 in the supplementary text). Comparing with standalone M-CHAT/F performance (Fig. 3B(ii)), we show that for any prevalence between 1.7% and 2.23%, we can *double the PPV* without losing sensitivity at $>98\%$ specificity, or increase the sensitivity by $\sim 50\%$ without sacrificing PPV and keeping specificity $\geqq 94\%$.

DISCUSSION

In this study, we operationalize a documented aspect of ASD symptomology in that it has a wide range of co-morbidities[14,15,19] occurring at above-average rates[8]. Association of ASD with epilepsy[20], gastrointestinal disorders[21–26], mental health disorders[27], insomnia, decreased motor skills[28], allergies including eczema[21–26], immunologic[19,29–35] and metabolic[25,36,37] disorders are widely reported. These studies, along with support from large scale exome sequencing[38,39], have linked the disorder to putative mechanisms of chronic neuroinflammation, implicating immune dysregulation and microglial activation[31,34,40–43] during important brain developmental periods of myelination and synaptogenesis. However, these advances have not yet led to clinically relevant diagnostic biomarkers. Majority of the co-morbid conditions are common in the control population, and rate differentials at the population level do not automatically yield individual risk[44].

Attempts at curating genetic biomarkers has also met with limited success. ASD genes exhibit extensive phenotypic variability, with identical variants associated with diverse individual outcomes not limited to ASD, including schizophrenia, intellectual disability, language impairment, epilepsy, neuropsychiatric disorders and, also typical development[45]. Additionally, no single gene can be considered “causal” for more than 1% of cases of idiopathic autism[46].

In the absence of biomarkers, current screening in pediatric primary care visits uses standardized questionnaires to categorize behavior. This is susceptible to potential interpretative biases arising from language barriers, as well as social and cultural differences, often leading to systematic under-diagnosis in diverse populations[8]. In this study we use time-stamped sequence of past disorders to elicit crucial information on the developing risk of an eventual diagnosis, and formulate a screening protocol that is free from such biases, and yet significantly outperforms the tools in current practice.

Going beyond screening performance, this approach provides a new tool to uncover clues to ASD pathobiology. Perhaps this vulnerability to diverse immunological, endocrinological and neurological impairments reflects how allostatic loads of medical stress get under the skin and disrupt key regulators of CNS organization and synaptogenesis. Charting individual disorders in the co-morbidity burden further reveals novel associations in normalized prevalence — the odds of experiencing a specific disorder, particularly in the early years (age <3 years), nor-

malized over all unique disorders experienced in the specified time-frame. We focus on the true positives in the positive cohort and the true negatives in the control cohort to investigate patterns that correctly disambiguate ASD status. On these lines Fig. 4 and SI-Fig. 1 in the supplementary text outline two key observations: 1) *negative associations:* some diseases that are negatively associated with ASD with respect to normalized prevalence, $i.e.$, having those codes relatively over-represented in one's diagnostic history favors ending up in the control cohort, 2) *impact of sex:* there are sex-specific differences in the impact of specific disorders, and given a fixed level of impact, the number of codes that drive the outcomes is significantly more in males (Fig. 4A vs B).

Some of the disorders that show up in Fig. 4, panels A and B are surprising, $e.g.$, congenital hemiplegia or diplegia of the upper limbs indicative of either cerebral palsy (CP) or a spinal cord/brain injury, neither of which has a direct link to autism. Since only about 7% of the children with cerebral palsy (CP) are estimated to have a co-occurring ASD[47,48], and with the prevalence of CP significantly lower (1 in 352 vs 1 in 59 for autism), it follows that only a small number of children (approximately 1.17%) with autism have co-occurring CP. Thus, with significantly higher prevalence in children diagnosed with autism compared to the general population (1.7% vs 0.28%), CP codes show up with higher odds in the true positive set. Also, SI-Fig. 1A shows that the immunological, metabolic, and endocrine disorders are almost completely risk-increasing. In contrast, respiratory diseases (panel B) are largely risk-decreasing. On the other hand, infectious diseases have roughly equal representations in the risk-increasing and risk-decreasing classes (panel C). The risk-decreasing infectious diseases tend to be due to viral or fungal organisms, which might point to the use of antibiotics in bacterial infections, and the consequent dysbiosis of the gut microbiota[23,37] as a risk factor.

Any predictive analysis of ASD must address if we can discriminate ASD from general developmental and behavioral disorders. The DSM-5 established a single category of ASD to replace the subtypes of autistic disorder, Asperger syndrome, and pervasive developmental disorders[8]. This aligns with our use of diagnostic codes from ICD9 299.X as specification of an ASD diagnosis, and use standardized mapping to 299.X from ICD10 codes when we encounter them. For other psychiatric disorders, we get high discrimination reaching AUCs over 90% at $100-125$ weeks of age (SI-Fig. 5A), which establishes that our pipeline is indeed largely specific to ASD.

We carried out a battery of tests to ensure that our results are not significantly impacted by class imbalance (since our control cohort is orders of magnitude larger) or systematic coding errors (See Methods), $e.g.$, 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 (SI-Fig. 5B).

Can our performance be matched by simply asking how often a child is sick? We found that the density of codes in a child's medical history is indeed somewhat predictive of a future ASD diagnosis, with the AUC $\approx 75\%$ in the Truven database at 150 weeks (See SI-Fig. 5, panel D in the supplementary text). This is expected, since children with autism do indeed have higher rates of co-morbidities. However, it does not have stable performance across databases, and has no significant effect once the rest of the features are combined.

As a key limitation to our approach, automated pattern recognition might not reveal true causal precursors. The relatively uncurated nature of the data does not correct for coding mistakes by the clinician and other artifacts, $e.g.$ a bias towards over-diagnosis of children on the borderline of the diagnostic criteria due to clinicians' desire to help families access service, and biases arising from changes in diagnostic practices over time[49]. Discontinuities in patient medical histories from change in provider-networks can also introduce uncertainties in risk estimates, and socio-economic status of patients which impact access to healthcare might skew patterns in EHR databases. Despite these limitations, the design of a questionnaire-free component to ASD screening that systematically leverages co-morbidities has far-reaching consequences, by potentially slashing the false positives and wait-times, as well as removing systemic under-diagnosis issues amongst females and minorities.

Future efforts will attempt to realize our approach within a clinical setting. We will also explore the impact of maternal medical history, and the use of calculated risk to trigger blood-work to look for expected transcriptomic signatures of ASD. Finally, the analysis developed here applies to phenotypes beyond ASD, thus opening the door to the possibility of general comorbidity-aware risk predictions from electronic health record databases.

Data Sharing

Software implementation of the pipeline is available at: https://github.com/zeroknowledgediscovery/ehrzero, and installation in standard python environments may be done from https://pypi.org/project/ehrzero/. A sample of de-identified data from the UCM database is be shared as part of the public software package.

ACKNOWLEDGEMENT

This work is funded in part by the Defense Advanced Research Projects Agency (DARPA) project #FP070943-01-PR. The claims made in this study do not reflect the position or the policy of the US Government. The UCM dataset is provided by the Clinical Research Data Warehouse (CRDW) maintained by the Center for Research Informatics (CRI) at the University of Chicago. The Center for Research Informatics is funded by the Biological Sciences Division, the Institute for Translational Medicine/CTSA (NIH UL1 TR000430) at the University of Chicago.

REFERENCES

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

[2] Gilotty, L. Early screening for autism spectrum (2019). URL https://www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2018/early-screening-for-autism-spectrum.shtml.

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

[4] 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).

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

[6] 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).

[7] 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).

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

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

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] 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).

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

[16] Hansen, L. The truven health marketscan databases for life sciences researchers. *Truven Health Ananlytics IBM Watson Health* (2017).

[17] 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).

[18] 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).

[19] Zerbo, O. *et al.* Immune mediated conditions in autism spectrum disorders. *Brain Behav. Immun.* **46**, 232–236 (2015).

[20] Won, H., Mah, W. & Kim, E. Autism spectrum disorder causes, mechanisms, and treatments: focus on neuronal synapses. *Front Mol Neurosci* **6**, 19 (2013).

[21] Xu, G. *et al.* Association of Food Allergy and Other Allergic Conditions With Autism Spectrum Disorder in Children. *JAMA Netw Open* **1**, e180279 (2018).

[22] Adams, J. B. *et al.* Nutritional and metabolic status of children with autism vs. neurotypical children, and the association with autism severity. *Nutr Metab (Lond)* **8**, 34 (2011).

[23] Fattorusso, A., Di Genova, L., Dell'Isola, G. B., Mencaroni, E. & Esposito, S. Autism Spectrum Disorders and the Gut Microbiota. *Nutrients* **11** (2019).

[24] Diaz Heijtz, R. *et al.* Normal gut microbiota modulates brain development and behavior. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3047–3052 (2011).

[25] Rose, S. *et al.* Mitochondrial dysfunction in the gastrointestinal mucosa of children with autism: A blinded case-control study. *PLoS ONE* **12**, e0186377 (2017).

[26] Sajdel-Sulkowska, E. M. *et al.* Common Genetic Variants Link the Abnormalities in the Gut-Brain Axis in Prematurity and Autism. *Cerebellum* **18**, 255–265 (2019).

[27] Kayser, M. S. & Dalmau, J. Anti-NMDA Receptor Encephalitis in Psychiatry. *Curr Psychiatry Rev* **7**, 189–193 (2011).

[28] Dadalko, O. I. & Travers, B. G. Evidence for Brainstem Contributions to Autism Spectrum Disorders. *Front Integr Neurosci* **12**, 47 (2018).

[29] Yamashita, Y. *et al.* Anti-inflammatory Effect of Ghrelin in Lymphoblastoid Cell Lines From Children With Autism Spectrum Disorder. *Front Psychiatry* **10**, 152 (2019).

[30] Shen, L. *et al.* Proteomics Study of Peripheral Blood Mononuclear Cells (PBMCs) in Autistic Children. *Front Cell Neurosci* **13**, 105 (2019).

[31] Ohja, K. *et al.* Neuroimmunologic and Neurotrophic Interactions in Autism Spectrum Disorders: Relationship to Neuroinflammation. *Neuromolecular Med.* **20**, 161–173 (2018).

[32] 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).

[33] 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).

[34] Young, A. M. *et al.* From molecules to neural morphology: understanding neuroinflammation in autism spectrum condition. *Mol Autism* **7**, 9 (2016).

[35] 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).
[36] Vargason, T., McGuinness, D. L. & Hahn, J. Gastrointestinal Symptoms and Oral Antibiotic Use in Children with Autism Spectrum Disorder: Retrospective Analysis of a Privately Insured U.S. Population. *J Autism Dev Disord* **49**, 647–659 (2019).
[37] Fiorentino, M. *et al.* Blood-brain barrier and intestinal epithelial barrier alterations in autism spectrum disorders. *Mol Autism* **7**, 49 (2016).
[38] 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.
[39] Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
[40] Vargas, D. L., Nascimbene, C., Krishnan, C., Zimmerman, A. W. & Pardo, C. A. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann. Neurol.* **57**, 67–81 (2005).
[41] Wei, H. *et al.* IL-6 is increased in the cerebellum of autistic brain and alters neural cell adhesion, migration and synaptic formation. *J Neuroinflammation* **8**, 52 (2011).
[42] Young, A. M., Campbell, E., Lynch, S., Suckling, J. & Powis, S. J. Aberrant NF-kappaB expression in autism spectrum condition: a mechanism for neuroinflammation. *Front Psychiatry* **2**, 27 (2011).
[43] Hughes, H. K., Mills Ko, E., Rose, D. & Ashwood, P. Immune Dysfunction and Autoimmunity as Pathological Mechanisms in Autism Spectrum Disorders. *Front Cell Neurosci* **12**, 405 (2018).
[44] Pearce, N. The ecological fallacy strikes back. *Journal of epidemiology and community health* **54**, 326–7 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10814650http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1731667.
[45] Murdoch, J. D. & State, M. W. Recent developments in the genetics of autism spectrum disorders. *Curr. Opin. Genet. Dev.* **23**, 310–315 (2013).
[46] Hu, V. W. The expanding genomic landscape of autism: discovering the 'forest' beyond the 'trees'. *Future Neurol* **8**, 29–42 (2013).
[47] Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning (2020). URL https://www.cdc.gov/ncbddd/cp/features/prevalence.html.
[48] Christensen, D. *et al.* Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning–a utism and d evelopmental d isabilities m onitoring n etwork, usa, 2008. *Developmental Medicine & Child Neurology* **56**, 59–65 (2014).
[49] Rødgaard, E.-M., Jensen, K., Vergnes, J.-N., Soulières, I. & Mottron, L. Temporal Changes in Effect Sizes of Studies Comparing Individuals With and Without Autism: A Meta-analysis. *JAMA Psychiatry* **76**, 1124–1132 (2019). URL https://doi.org/10.1001/jamapsychiatry.2019.1956. https://jamanetwork.com/journals/jamapsychiatry/articlepdf/2747847/jamapsychiatry_rdgaard_2019_oi_190046.pdf.