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

Dmytro Onishchenko[1], Yi Huang[1], James van Horne[1], Peter J. Smith[4,7], Michael M. Msall[5,6] and Ishanu Chattopadhyay[1,2,3]★

[1]Department of Medicine, University of Chicago, Chicago, IL USA
[2]Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA
[3]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA
[4]Department of Pediatrics, Section of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL USA
[5]Department of Pediatrics, Section Chief of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL USA
[6]Joseph P. Kennedy Research Center on Intellectual and Neurodevelopmental Disabilities, University of Chicago, Chicago, IL USA
[7]Executive Committee Chair, American Academy of Pediatrics' Section on Developmental and Behavioral Pediatrics

★To whom correspondence should be addressed: e-mail: ishanu@u chicago.edu.

Abstract

Autism spectrum disorder (ASD) is a developmental disability associated with significant social 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 compute the Autism Co-morbid Risk Score (ACoR) which predicts elevated risk during the earliest childhood years, when interventions are the most effective. By itself, ACoR 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. In addition to standalone performance, independence from questionnaire based screening allows us to further boost performance by conditioning on the individual M-CHAT/F scores – we can either halve the false positive rate of current screening protocols or boost sensitivity to over 60%, while maintaining specificity above 95%. Adopted in practice, ACoR could significantly reduce the median diagnostic age for ASD, and reduce long post-screen wait-times (4) 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). With genetic and metabolomic tests (7–10) still at their infancy, a careful review of behavioral history and a direct observation of symptoms is currently necessary (11, 12) 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 (13), cost of care (14), lack of providers (15), and lack of comfort in diagnosing ASD by primary care providers (15) are all responsible to varying degrees (16), 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, a large-scale study of the M-CHAT/F (3) (n=20,375), which is used commonly as a screening tool (12, 17), is demonstrated to have an estimated sensitivity of 38.8%, specificity of 94.9% and Positive Predictive Value (PPV) of 14.6%. 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 (16). 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.

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 (18, 19). In the present study, we exploit these co-morbidities to estimate the risk of childhood neuropsychiatric disorders on the autism spectrum. We refer to the risk estimated by our approach as the Autism Co-morbid Risk Score (ACoR). Using only 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. 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 (12). 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 XVI).

Use of sophisticated analytics to identify children at high risk is a topic of substantial current interest, with independent progress being made by several groups (20–26). Many of these approaches focus on analyzing questionnaires, with recent efforts demonstrating the use of automated pattern recognition in video clips of toddler behavior. However, the inclusion of older kids above the age of 5 years and small cohort sizes might limit the immediate clinical adoption of these approaches for universal screening.

Laboratory tests for ASD have also begun to emerge, particularly leveraging detection of abnormal metabolites in plasma (7, 10), and salivary poly-omic RNA (9). However, as before, inclusion of older children, limits applicability in screening, where we need a decision at $18-24$ months. In addition, such approaches – while instrumental in deciphering ASD pathophysiology – might be too expensive for universal adoption at this time.

In contrast to ACoR, the above approaches require additional data or tests. Use of comorbidity patterns derived from past Electronic Health Records has been either limited to establishing correlative associations (27, 28), or has substantially underperformed (29)(AUC $\leqq 65\%$) compared to our results.

MATERIALS & METHODS

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

TABLE I: Patient Numbers, Inclusion-exclusion 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 SI-Table I in Supplementary Text.

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. The number of patients used from the two databases is shown in Table Ia. 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 (31)) with no significant geospatial prevalence variation (See SI-Fig. 1).

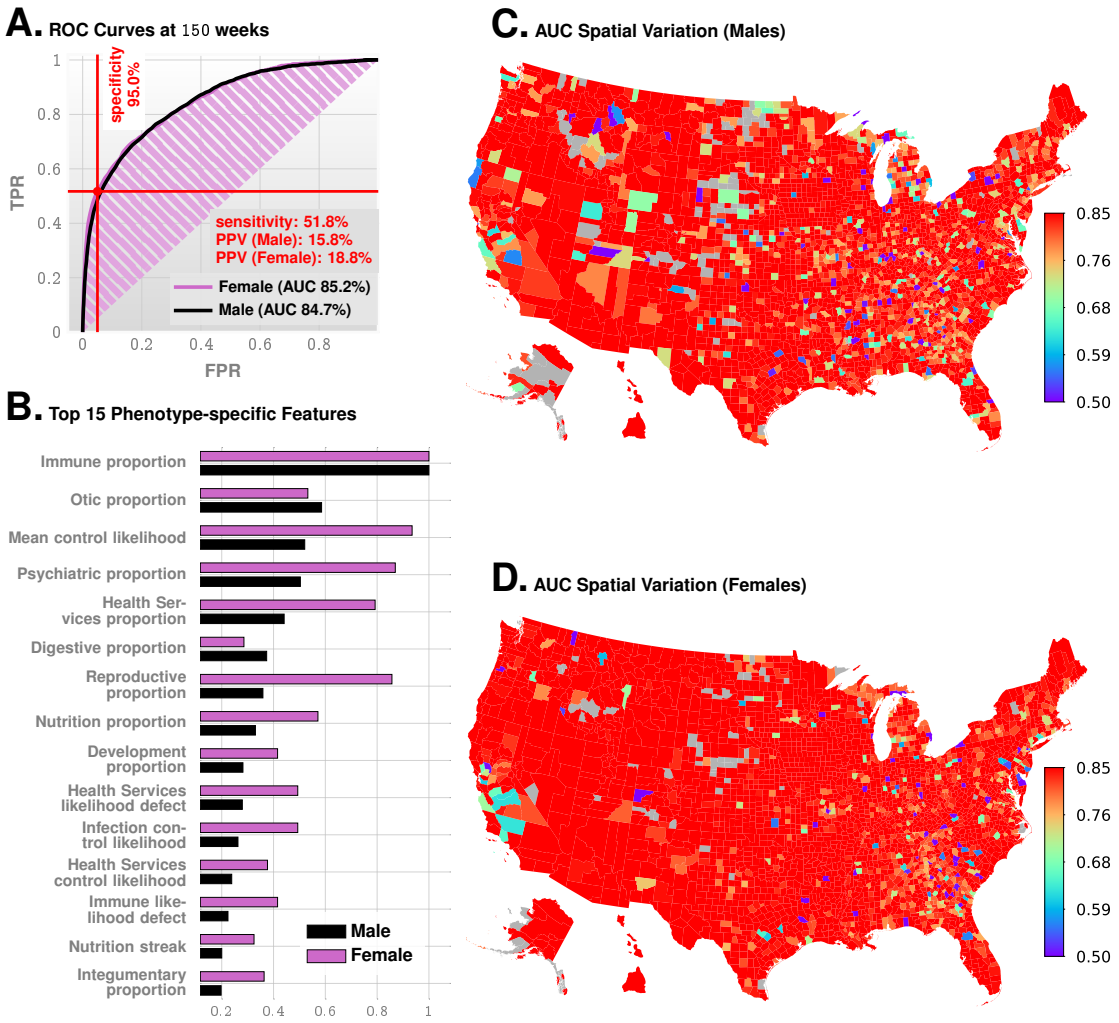

Fig. 1: **Standalone Predictive Performance of ACoR.** Panel A shows the ROC curves for males and females (Truven data shown, UCM is similar, see Fig. 2a). Panel B shows the feature importance inferred by our prediction pipeline. The detailed description of the features is given in Table Ib. The most 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.

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 (See Supp. Fig. 3). 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). See Supplementary Text Section XIX for details on PFSA inference.

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. This novel stochastic inference approach provides significant boost to the overall performance of our predictors; with only state of the art machine learning the predictive performance is significantly worse (See Supplementary Text Sections VII and VI, as well as reported performance in the literature for predicting ASD risk from EHR data with standard algorithms (29)).

We briefly outline the SLD computation: To reliably infer the cohort-type of a new patient, *i.e*, the likelihood of a

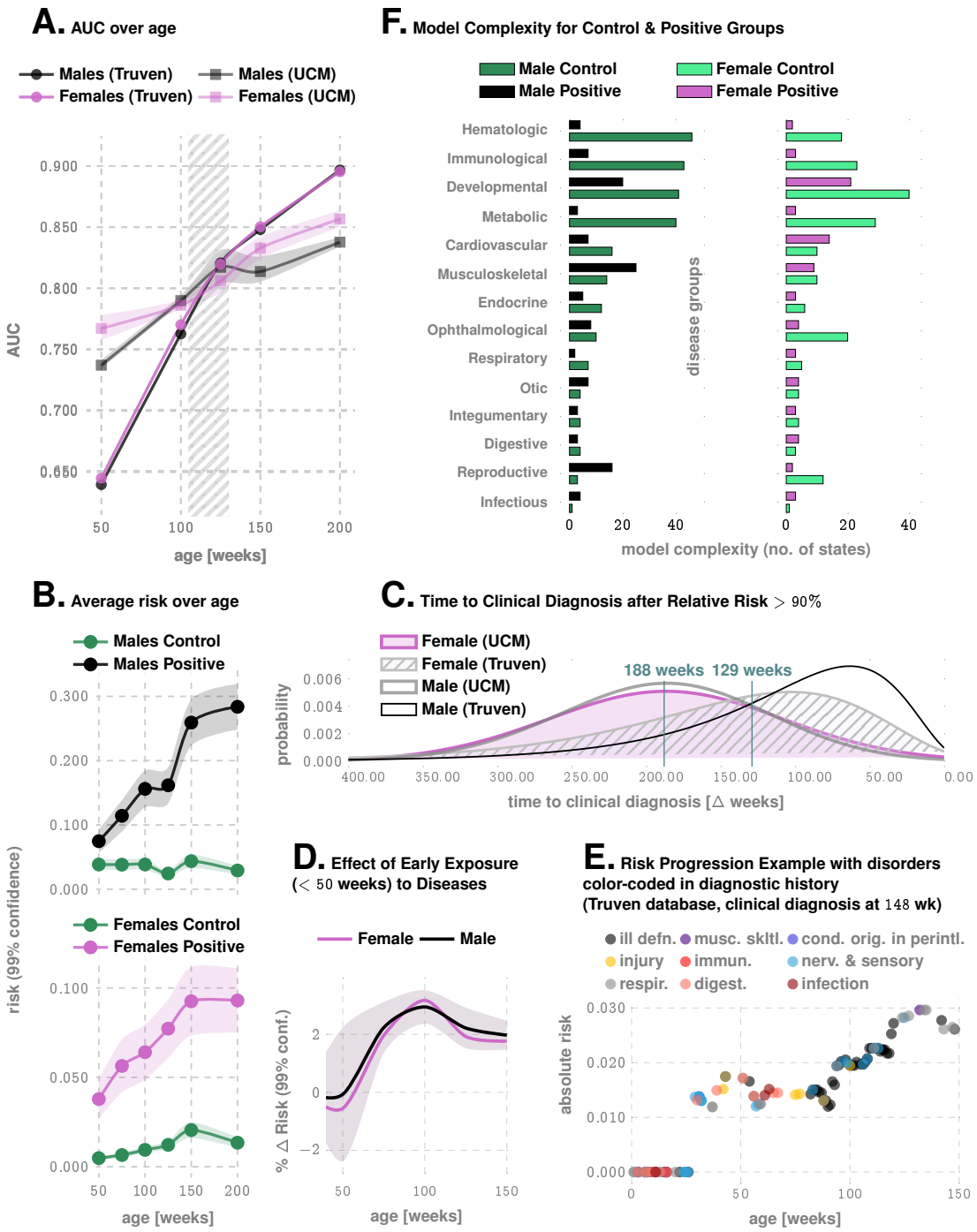

Fig. 2: **More Details on Standalone Predictive Performance of ACoR and Variation of Inferred Risk.** Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve $> 80\%$ AUC for either sex 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.

diagnostic sequnce being generated by the corresponding cohort model, we generalize the notion of Kullbeck-Leibler (KL) divergence (32, 33) between probability distributions to a divergence $\mathcal{D}_{\mathrm{KL}}(G||H)$ between ergodic

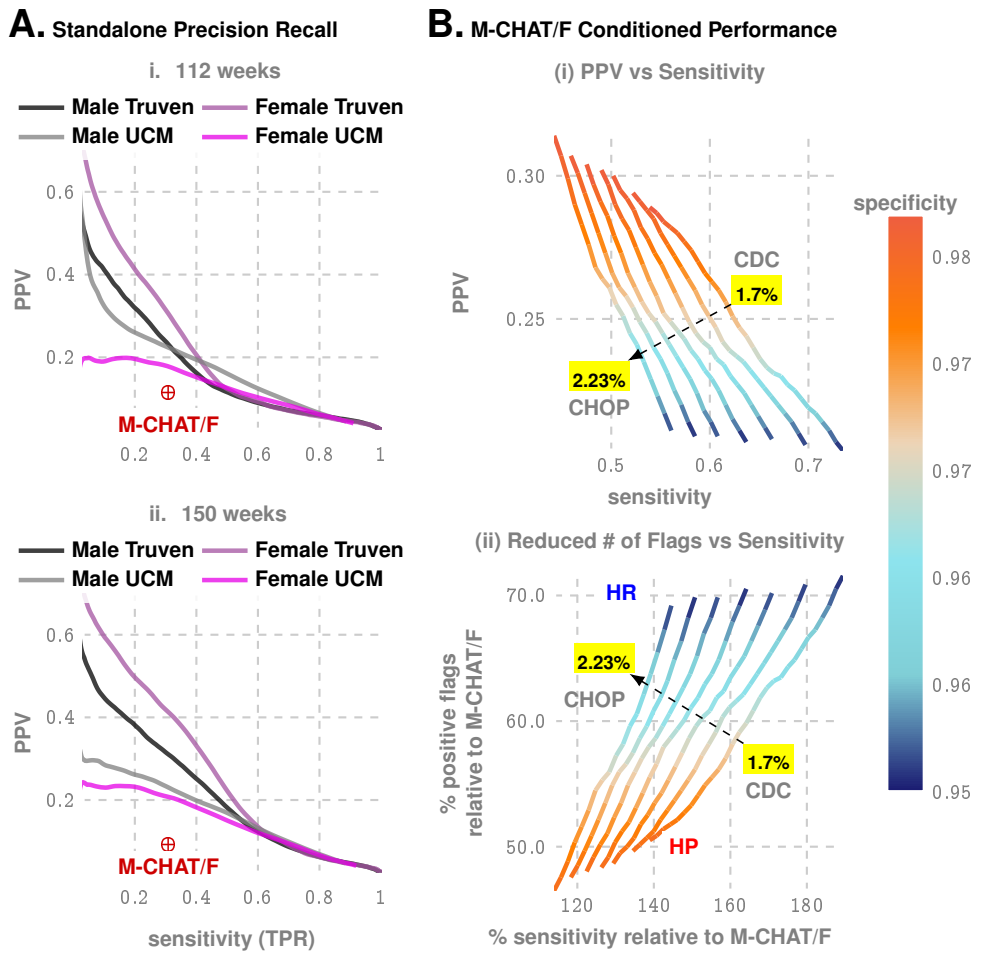

Fig. 3: **Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs.** Panel A shows the precision/recall curves, *i.e.*, the trade-off between PPV and sensitivity for **standalone operation** with ACoR. Panel B shows how we can **boost ACoR performance** using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study (3). This is possible because ACoR and M-CHAT/F use independent information (co-morbidities vs questionnaire responses). 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 increase sensitivity by 20-40% (depending on the prevalence) but double the PPV achieved in current practice. In contrast, when choosing to maximize sensitivity by operating in the HR zone, we only cut down positive flags to about 70% of what we get with M-CHAT/F, but boost sensitivity by $50-90\%$ (Reaching sensitivities over 70%). 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.

stationary categorical stochastic processes (34) G, H as:

$$\mathcal{D}_{\mathsf{KL}}(G||H) = \lim_{n\to\infty} \frac{1}{n} \sum_{x:|x|=n} p_G(x) \log \frac{p_G(x)}{p_H(x)} \tag{1}$$

where $|x|$ is the sequence length, and $p_G(x), p_H(x)$ are the probabilities of sequence x being generated by the processes G, H respectively. Defining the log-likelihood of x being generated by a process G as:

$$L(x, G) = -\frac{1}{|x|} \log p_G(x) \tag{2}$$

The cohort-type for an observed sequence x — which is actually generated by the hidden process G — can be formally inferred from observations based on the following provable relationships (See Suppl. text Section XIX, Theorem 6 and 7):

$$\lim_{|x|\to\infty} L(x, G) = \mathcal{H}(G) \tag{3a}$$

$$\lim_{|x|\to\infty} L(x, H) = \mathcal{H}(G) + \mathcal{D}_{\mathsf{KL}}(G||H) \tag{3b}$$

where $\mathcal{H}(\cdot)$ is the entropy rate of a process (32). Importantly, Eq. (3) shows that the computed likelihood has an additional non-negative contribution from the divergence term when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with ASD, then we expect that the disease-specific mapped

TABLE II: Standalone ACoR Performance and Boosted Performance Conditioned on M-CHAT/F

(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	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%. The results of our optimization depend on the prevalence estimate.

series corresponding to her diagnostic history be modeled by the corresponding model in the positive cohort. Denoting the model corresponding to disease category j for positive and control cohorts as G^j_+, G^j_0 respectively, we can compute the sequence likelihood defect (SLD, Δ^j) as:

$$\Delta^j \triangleq L(G^j_0, x) - L(G^j_+, x) \rightarrow \mathcal{D}_{\mathsf{KL}}(G^j_0 || G^j_+) \tag{4}$$

With the inferred models and the individual diagnostic history, we estimate the SLD measure on the right-hand side of Eqn. (4). The higher this likelihood defect, the higher the similarity of diagnosis history to that of children with autism.

In addition to the category specific Markov models, we use a range of engineered features that reflect various aspects of the diagnostic histories, including the porportion of weeks in which a diagnostic code is generated, the maximum length of consecutive weeks with codes, the maximum length of weeks with no codes (See Tab. Ib for complete description), resulting in a total of 165 different features that are evaluated for each patient. With these inferred patterns included as features, we train a second level predictor that learns to map individual patients to the control or the positive groups based on their similarity to the identified Markov models of category-specific diagnostic histories, and the other engineered features (See Section I on detailed mathematical details in Supplementary text).

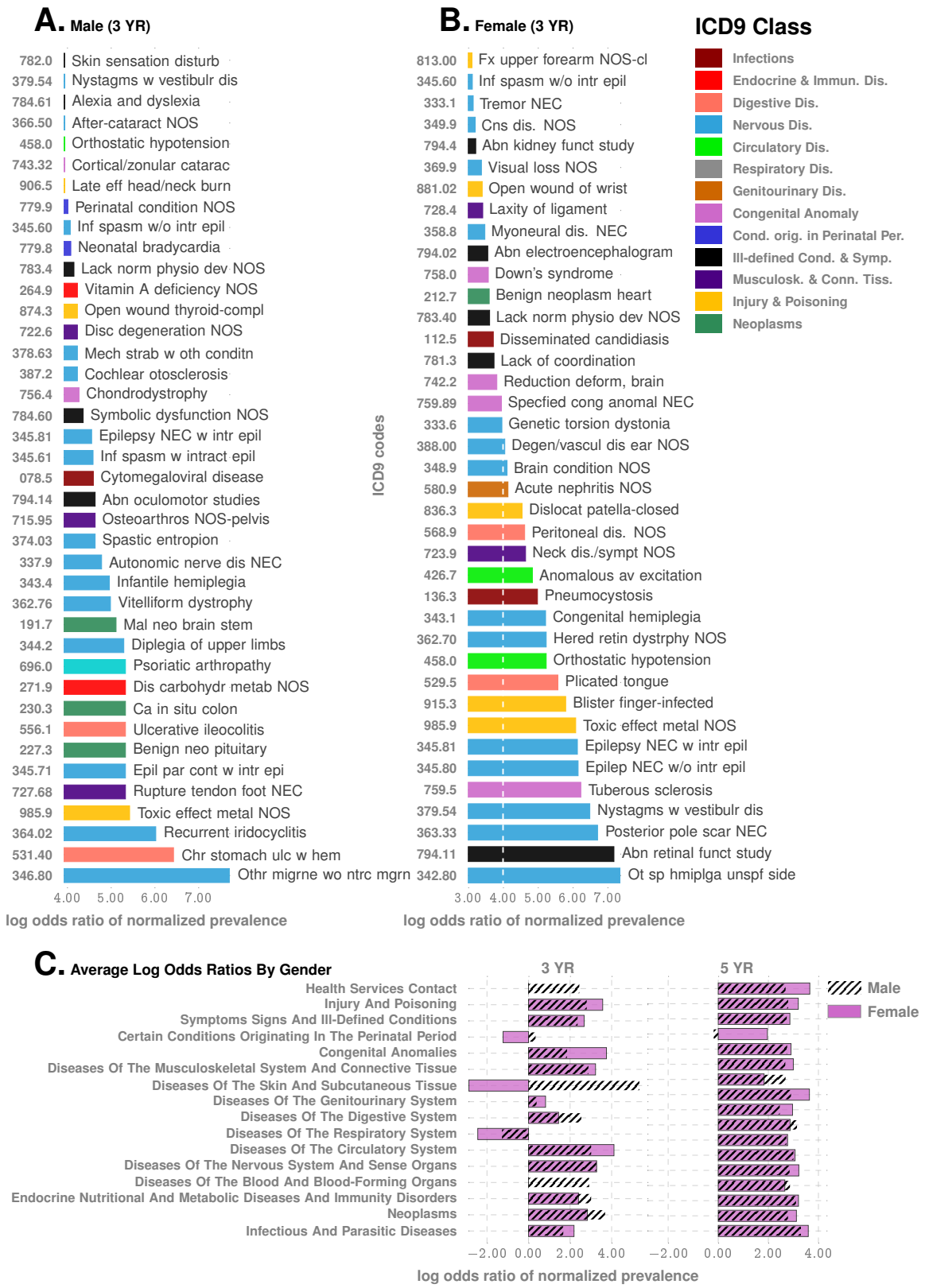

Fig. 4: **Co-morbidity Patterns** Panel A and B. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions. The dotted line on panel B shows the abscissa lower cut-off in Panel A, illustrating the lower prevalence of codes in females. Panel 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.

Since we need to infer the Markov models prior to the calculation of the likelihood defects, we need two training sets: one that is used to infer the models, and one that subsequently trains the final classifier with features derived from the inferred models along with other engineered features. Thus, the analysis proceeds by first carrying out a random 3-way split of the set of unique patients (in the Truven dataset) into *Markov model inference* (25%),

classifier training (25%) and *test* (50%) sets. The approximate sample sizes of the three sets are as follows: $\approx 700K$ for each of the training sets, and $\approx 1.5M$ for the test set. The features used in our pipeline may be ranked in order of their relative importance (See Fig. 1B for the top 15 features), by estimating the loss in performance when dropped out of the analysis. We verified that different random splits do not adversely affect performance. The UCM dataset in its entirety is used as a test set, with no retraining of the pipeline.

Our pipeline maps medical histories to a raw indicator of risk. Ultimately, to make crisp predictions, we must choose a decision threshold for this raw score. 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 Type 1 and Type 2 errors (See Supplementary text, Section IX). 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. Our two step learning algorithm outperforms standard tools, and achieves stable performance across datasets strictly superior to documented M-CHAT/F.

The independence of our approach from questionnaire based screening implies that we can further boost our performance by conditioning the sensitivity/specificity trade-offs on individual M-CHAT/F scores. In particular, we leverage the population stratification induced by M-CHAT/F to improve combined performance. Here a combination of ACoR with M-Chat/F refers to the conditional tuning of the sensitivity/specificity for ACoR in each sub-population such that the overall performance is maximized. To describe this approach briefly, we assume that there are m sub-populations with 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 have (See Supplementary text, Section XVII):

$$s = \sum_{i=1}^{m} s_i \gamma_i \tag{5a}$$

$$c = \sum_{i=1}^{m} c_i \gamma_i' \tag{5b}$$

where we have denoted:

$$\gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma_i' = \beta_i \frac{1-\rho_i}{1-\rho} \tag{5c}$$

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 (3), 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 SI-Table III). The "follow-up" in the context of M-CHAT/F refers to the re-evaluation of responses by qualified personnel. We use published data from the CHOP study (3) on the relative sizes and the prevalence statistics in these sub-populations to compute the feasible conditional choices of our operating point to vastly supersede standalone M-CHAT/F performance. The CHOP study is the only large-scale study of M-CHAT/F we are aware of with sufficient follow-up after the age of four years to provide a reasonable degree of confidence in the sensitivity of M-CHAT/F.

Two limiting operating conditions are of particular interest in this optimization scheme, 1) where we maximize PPV under some minimum specificity and sensitivity (denoted as the High Precision or the HP operating point), and 2) 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 significantly outperform M-CHAT/F in universal screening. The results are shown in Fig. 3B.

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, *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. 4B).

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

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.

Recall, that the UCM dataset is used purely for validation, and 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.

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 (16, 35) across states and regions.

Unlike individual predictions which only become relevant over 2 years, the average risk over the populations is 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. 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 (16), 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 IIa 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 IV), 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 IIb, Fig. 3B(i), and SI-Fig. 8 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 (18, 19, 36) occurring at above-average rates (12). Association of ASD with epilepsy (37), gastrointestinal disorders(38–43), mental health disorders (44), insomnia, decreased motor skills (45), allergies including eczema (38–43), immunologic (36, 46–52) and metabolic(42, 53, 54) disorders are widely reported. These studies, along with support from large scale exome sequencing (55, 56), have linked the disorder to putative mechanisms of chronic neuroinflammation, implicating immune dysregulation and microglial activa-

tion (48, 51, 57–60) 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 (61).

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 (62). Additionally, no single gene can be considered "causal" for more than 1% of cases of idiopathic autism (63).

Despite these hurdles, laboratory tests and potential biomarkers for ASD have begun to emerge (7, 9, 10). These tools are still at their infancy, and have not demonstrated performance in the 18-24 month age group. In the absence of clinically useful 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 (12). 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 the autism comorbid risk score. Th ACoR is free from aforementioned 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, potentially linking the observed vulnerability to diverse immunological, endocrinological and neurological impairments to the possibility of allostatic stress load disrupting 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), normalized 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. 5 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. However, this effect is easily explainable: since only about 7% of the children with cerebral palsy (CP) are estimated to have a co-occurring ASD (64, 65), 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. Other patterns are harder to explain. For example, SI-Fig. 5A shows that the immunological, metabolic, and endocrine disorders are almost completely risk-increasing, and 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 (40, 54) 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 (12). 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. 4A), which establishes that our pipeline is indeed largely specific to ASD.

Can our performance be matched by simply asking how often a child is sick? Indeed the code density in a child's medical history is higher for those eventually diagnosed with autism (See Table Ia). However, this turns out to be a rather crude measure. While somewhat predictive, achieving AUC $\approx 75\%$ in the Truven database at 150 weeks (See SI-Fig. 4, panel D in the supplementary text), code density by itself does not have stable performance across the two databases (with particularly poor performance in validation in the UCM database). Additionally adding code density as an additional feature shows no appreciable improvement in our pipeline.

We also investigated the effect of removing all psychiatric codes (ICD9 290 - 319, and corresponding ICD10)

from the patient histories to eliminate the possibility that our performance is simply reflective of prior psychiatric evaluation results. Training and validation on this modified data found no appreciable difference in performance (See SI-Fig. 7). Additionally, we found that including information on prescribed medications and medical procedures in addition to diagnostic codes did not improve results.

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 (66). 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 & Software Availability

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/.

Acknowledgement

This work is funded in part by the Defense Advanced Research Projects Agency (DARPA) project number HR00111890043/P00004. 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. IRB exemption was granted due to de-identified subjects from University of Chicago, IRB Committee: BSD (Contact: Amy Horst ahorst@medicine.bsd.uchicago.edu). IRB#: Predictive Diagnoses IRB19-1040.

Competing Interests: The authors declare that they have no competing interests.

Author Contributions

DO implemented the algorithm and ran validation tests. DO, YH, JH and IC carried out mathematical modeling, and algorithm design. PS, MM and IC interpreted results and guided research. IC wrote the paper.

Supplementary Materials

Supplemental tables, figures and modeling details are included in the Supplementary Text document.

References

[1] Centers for Disease Control and Prevention, Data & statistics on autism spectrum disorder — cdc (2019).
[2] L. Gilotty, Early screening for autism spectrum (2019).
[3] W. Guthrie, *et al.*, *Pediatrics* **144** (2019).
[4] E. Gordon-Lipkin, J. Foster, G. Peacock, *Pediatr. Clin. North Am.* **63**, 851 (2016).
[5] Data & statistics on autism spectrum disorder — cdc (2019).
[6] L. A. Schieve, *et al.*, *Ann Epidemiol* **24**, 260 (2014).
[7] D. P. Howsmon, U. Kruger, S. Melnyk, S. J. James, J. Hahn, *PLoS computational biology* **13**, e1005385 (2017).
[8] G. Li, O. Lee, H. Rabitz, *PloS one* **13**, e0192867 (2018).
[9] S. D. Hicks, *et al.*, *Frontiers in genetics* **9**, 534 (2018).

[10] A. M. Smith, *et al.*, *Autism Research* **13**, 1270 (2020).
[11] F. Volkmar, *et al.*, *Journal of the American Academy of Child & Adolescent Psychiatry* **53**, 237 (2014).
[12] S. L. Hyman, S. E. Levy, S. M. Myers, *et al.*, *Pediatrics* **145** (2020).
[13] L. G. Kalb, *et al.*, *Journal of Developmental & Behavioral Pediatrics* **33**, 685 (2012).
[14] J. Bisgaier, D. Levinson, D. B. Cutts, K. V. Rhodes, *Archives of pediatrics & adolescent medicine* **165**, 673 (2011).
[15] T. S. Fenikilé, K. Ellerbeck, M. K. Filippi, C. M. Daley, *Primary health care research & development* **16**, 356 (2015).
[16] E. Gordon-Lipkin, J. Foster, G. Peacock, *Pediatric Clinics* **63**, 851 (2016).
[17] D. L. Robins, *et al.*, *Pediatrics* **133**, 37 (2014).
[18] C. Tye, A. K. Runicles, A. J. O. Whitehouse, G. A. Alvares, *Front Psychiatry* **9**, 751 (2018).
[19] I. S. Kohane, *et al.*, *PLoS ONE* **7**, e33224 (2012).
[20] K. K. Hyde, *et al.*, *Review Journal of Autism and Developmental Disorders* **6**, 128 (2019).
[21] H. Abbas, F. Garberson, S. Liu-Mayo, E. Glover, D. P. Wall, *Scientific reports* **10**, 1 (2020).
[22] M. Duda, J. Daniels, D. P. Wall, *Journal of autism and developmental disorders* **46**, 1953 (2016).
[23] M. Duda, J. Kosmicki, D. Wall, *Translational psychiatry* **4**, e424 (2014).
[24] V. A. Fusaro, *et al.*, *PLOS one* **9**, e93533 (2014).
[25] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, T. F. DeLuca, *PloS one* **7**, e43855 (2012).
[26] D. P. Wall, J. Kosmicki, T. Deluca, E. Harstad, V. A. Fusaro, *Translational psychiatry* **2**, e100 (2012).
[27] F. Doshi-Velez, Y. Ge, I. Kohane, *Pediatrics* **133**, e54 (2014).
[28] L. Bishop-Fitzpatrick, *et al.*, *Autism Research* **11**, 1120 (2018).
[29] T. Lingren, *et al.*, *PloS one* **11**, e0159621 (2016).
[30] L. Hansen, *Truven Health Ananlytics IBM Watson Health* (2017).
[31] J. Baio, *et al.*, *MMWR Surveill Summ* **67**, 1 (2018).
[32] T. M. Cover, J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).
[33] S. Kullback, R. A. Leibler, *Ann. Math. Statist.* **22**, 79 (1951).
[34] J. Doob, *Stochastic Processes*, Wiley Publications in Statistics (John Wiley & Sons, 1953).
[35] L. A. Althouse, J. A. Stockman, *The Journal of pediatrics* **148**, 575 (2006).
[36] O. Zerbo, *et al.*, *Brain Behav. Immun.* **46**, 232 (2015).
[37] H. Won, W. Mah, E. Kim, *Front Mol Neurosci* **6**, 19 (2013).
[38] G. Xu, *et al.*, *JAMA Netw Open* **1**, e180279 (2018).
[39] J. B. Adams, *et al.*, *Nutr Metab (Lond)* **8**, 34 (2011).
[40] A. Fattorusso, L. Di Genova, G. B. Dell'Isola, E. Mencaroni, S. Esposito, *Nutrients* **11** (2019).
[41] R. Diaz Heijtz, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3047 (2011).
[42] S. Rose, *et al.*, *PLoS ONE* **12**, e0186377 (2017).
[43] E. M. Sajdel-Sulkowska, *et al.*, *Cerebellum* **18**, 255 (2019).
[44] M. S. Kayser, J. Dalmau, *Curr Psychiatry Rev* **7**, 189 (2011).
[45] O. I. Dadalko, B. G. Travers, *Front Integr Neurosci* **12**, 47 (2018).
[46] Y. Yamashita, *et al.*, *Front Psychiatry* **10**, 152 (2019).
[47] L. Shen, *et al.*, *Front Cell Neurosci* **13**, 105 (2019).
[48] K. Ohja, *et al.*, *Neuromolecular Med.* **20**, 161 (2018).
[49] D. Gadysz, A. Krzywdziska, K. K. Hozyasz, *Mol. Neurobiol.* **55**, 6387 (2018).
[50] T. C. Theoharides, I. Tsilioni, A. B. Patel, R. Doyle, *Transl Psychiatry* **6**, e844 (2016).
[51] A. M. Young, *et al.*, *Mol Autism* **7**, 9 (2016).
[52] L. A. Croen, *et al.*, *Autism Res* **12**, 123 (2019).
[53] T. Vargason, D. L. McGuinness, J. Hahn, *J Autism Dev Disord* **49**, 647 (2019).
[54] M. Fiorentino, *et al.*, *Mol Autism* **7**, 49 (2016).
[55] F. K. Satterstrom, *et al.*, *bioRxiv* (2019).
[56] T. Gaugler, *et al.*, *Nat. Genet.* **46**, 881 (2014).
[57] D. L. Vargas, C. Nascimbene, C. Krishnan, A. W. Zimmerman, C. A. Pardo, *Ann. Neurol.* **57**, 67 (2005).
[58] H. Wei, *et al.*, *J Neuroinflammation* **8**, 52 (2011).
[59] A. M. Young, E. Campbell, S. Lynch, J. Suckling, S. J. Powis, *Front Psychiatry* **2**, 27 (2011).
[60] H. K. Hughes, E. Mills Ko, D. Rose, P. Ashwood, *Front Cell Neurosci* **12**, 405 (2018).
[61] N. Pearce, *Journal of epidemiology and community health* **54**, 326 (2000).
[62] J. D. Murdoch, M. W. State, *Curr. Opin. Genet. Dev.* **23**, 310 (2013).
[63] V. W. Hu, *Future Neurol* **8**, 29 (2013).
[64] Centers for Disease Control and Prevention, Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning (2020).

[65] D. Christensen, *et al.*, *Developmental Medicine & Child Neurology* **56**, 59 (2014).
[66] E.-M. Rødgaard, K. Jensen, J.-N. Vergnes, I. Soulières, L. Mottron, *JAMA Psychiatry* **76**, 1124 (2019).
[67] General equivalence mappings.
[68] J. Baio (2014).
[69] P. F. Bolton, J. Golding, A. Emond, C. D. Steer, *Journal of the American Academy of Child & Adolescent Psychiatry* **51**, 249 (2012).
[70] J. Bondy, U. Murty, *Grad. Texts in Math* (2008).
[71] L. Breiman, *Mach. Learn.* **45**, 5 (2001).
[72] I. Chattopadhyay, A. Ray, *International Journal of Control* **81**, 820 (2008).
[73] I. Chattopadhyay, H. Lipson, *Journal of The Royal Society Interface* **11**, 20140826 (2014).
[74] C. Chlebowski, J. A. Green, M. L. Barton, D. Fein, *Journal of autism and developmental disorders* **40**, 787 (2010).
[75] I. Chattopadhyay, H. Lipson, *Philos Trans A* **371**, 20110543 (2013).
[76] T. M. Cover, J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).
[77] J. Doob, *Stochastic processes*, Wiley publications in statistics (Wiley, 1990).
[78] A. N. Esler, *et al.*, *Journal of Autism and Developmental Disorders* **45**, 2704 (2015).
[79] T. Falkmer, K. Anderson, M. Falkmer, C. Horlin, *European child & adolescent psychiatry* **22**, 329 (2013).
[80] J. H. Friedman, *Comput. Stat. Data Anal.* **38**, 367 (2002).
[81] G. Hardy, *Éditions Jacques Gabay, Sceaux* (1992).
[82] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **9**, 1735 (1997).
[83] J. E. Hopcroft, *Introduction to automata theory, languages, and computation* (Pearson Education India, 2008).
[84] V. G. Jarquin, L. D. Wiggins, L. A. Schieve, K. Van Naarden-Braun, *Journal of Developmental & Behavioral Pediatrics* **32**, 179 (2011).
[85] C. P. Johnson, S. M. Myers, *et al.*, *Pediatrics* **120**, 1183 (2007).
[86] L. C. Kai, *Markov Chains: With Stationary Transition Probabilities* (Springer-Verlag, 1967).
[87] J. M. Kleinman, *et al.*, *Journal of autism and developmental disorders* **38**, 606 (2008).
[88] A. Klenke, *Probability theory: a comprehensive course* (Springer Science & Business Media, 2013).
[89] A. M. Kozlowski, J. L. Matson, M. Horovitz, J. A. Worley, D. Neal, *Developmental neurorehabilitation* **14**, 72 (2011).
[90] C. Lord, *et al.*, *Archives of general psychiatry* **63**, 694 (2006).
[91] C. W. J. Granger, R. Joyeux, *Journal of Time Series Analysis* **1**, 15 (1980).
[92] A. G. d. G. Matthews, J. Hensman, R. Turner, Z. Ghahramani, *Journal of Machine Learning Research* **51**, 231 (2016).
[93] M. Penner, E. Anagnostou, W. J. Ungar, *Molecular autism* **9**, 16 (2018).
[94] A. N. Trahtman, *Proc. of Prague stringology conference* (Citeseer, 2008), vol. 1, p. 12.
[95] M. Vidyasagar, *Hidden markov processes: Theory and applications to biology*, vol. 44 (Princeton University Press, 2014).
[96] L. Zwaigenbaum, *et al.*, *Pediatrics* **136**, S60 (2015).