Early Identification of Young Children with Autism Through Automated Pattern Recognition From Primary Care Encounter History: Reducing False Positives In Autism Screening With 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,
[2]Committee on Genetics, Genomics & Systems Biology,
[3]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences,
[4]Department of Pediatrics, Section of Developmental and Behavioral Pediatrics,
[5]Department of Pediatrics, Section Chief of Developmental and Behavioral Pediatrics,
[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@uchicago.edu.

Abstract

Autism spectrum disorder (ASD) is a developmental disability associated with significant social, communication, and behavioral challenges. There is a distinct 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 primary care and medical encounters from two independent databases of patient records. We engineer a reliable risk estimator based on stochastic learning algorithms. Our predictive algorithm identifies children at high risk with a corresponding area under the receiver operating characteristic curve (AUC) exceeding 80% from shortly after 2 years of age for either gender, 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 with clinically useful reliability during the earliest childhood years, when intervention is the most effective. Compared with M-CHAT/F[3], a common screening tool used during primary care encounters, this new approach represents an orthogonal methodology with superior performance. Our methodology compared to screening questionnaires has the potential to has the potential to reduce socio-economic, ethnic and demographic biases, and allows for the possibility of tailoring the operating parameters to individual patients. By conditioning on the individual M-CHAT/F scores, we demonstrate personalized sensitivity/specificity trade-offs, to either halve the number of 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.

MAIN

AUTISM spectrum disorder is a developmental disability associated with significant social, communication, 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, the most widely used screen[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.

In this study, we operationalize a documented aspect of ASD symptomology in that it has a wide range of co-morbidities[14–16] occurring at above-average rates[8]. Association of ASD with epilepsy[17], gastrointestinal disorders[18–23], mental health disorders[24], insomnia, decreased motor skills[25], allergies including exzema[18–23], immunologic[16,26–32] and metabolic[22,33,34] disorders are widely reported. These studies, along with support from large scale exome sequencing[35,36], have linked the disorder to putative mechanims of chronic neuroinflammation, implicating immune dysregulation and microglial activation during important brain developmental periods of myelination and synaptogenesis[28,31,37–40]. 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[41].

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, intellectual disability, language impairment, other neuropsychiatric disorders and, also typical development[42]. Additionally, no single gene can be considered "causal" for more than 1% of cases of idiopathic autism[43].

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 sequences 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.

We base our analysis on two independent electronic databases of diagnostic histories: 1) a claims database for private health insurance (Truven Marketscan, the Truven dataset), tracking over 5.6 million children between 2003 and 2012, and 2) set of de-identified diagnostic records for nearly 70 thousand children under 5 years of age treated at the University of Chicago Medical Center between 2006 and 2018 (UCM dataset). Our datasets agree largely with documented prevalence: there is no significant geospatial prevalence variation (Extended Data Fig. 1D) and infections and immunological disorders have differential representation in the positive and control groups (Extended Data Fig. 1C). The median diagnosis age is just over 3 years in the claims database (Extended Data Fig. 1B) versus 3 years 10 months to 4 years in US[44]. Cohort details are given in Table I and discussed in Methods. Importantly, for the positive cohort, we only consider diagnostic history up to the first ASD code.

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. 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 Extended Data Table I). To address these issues, we proceed by partitioning the disease spectrum into 17 broad medical diagnostic categories, *e.g.* infectious diseases, immunologic disorders, endocrine 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 (Extended Data Table II) 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 (See Methods).

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. 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. This novel two step learning algorithm outperforms standard tools, and achieves stable performance across datasets.

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[12,45] 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 (See Extended Data Fig. 1B). 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 gender (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 Extended Data Table III show the out-of-sample PPV vs sensitivity curves for the two databases, stratified by gender, 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] (Extended Data Table VII), 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 Extended Data Table IV, Fig. 3B, and Extended Data Fig. 6). Comparing with standalone M-CHAT/F performance (Fig. 3C), 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 $\geqq 94\%$.

Going beyond screening performance, this approach provides a new tool to uncover clues to ASD pathobiology. Charting individual disorders in the co-morbidity burden 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 Extended Data Fig. 2 and Fig. 3 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) *gendered impact:* there are gender-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 (Extended Data Fig. 2A vs B).

Some of the disorders that show up in Extended Data Fig. 2, 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[46,47], 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, Extended Data Fig. 3A 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[20,34] 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 (Extended Data Fig. 7A), 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 (Extended Data Fig. 7B).

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 Extended Data Fig. 7, panel D). 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. 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.

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[48]. 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.

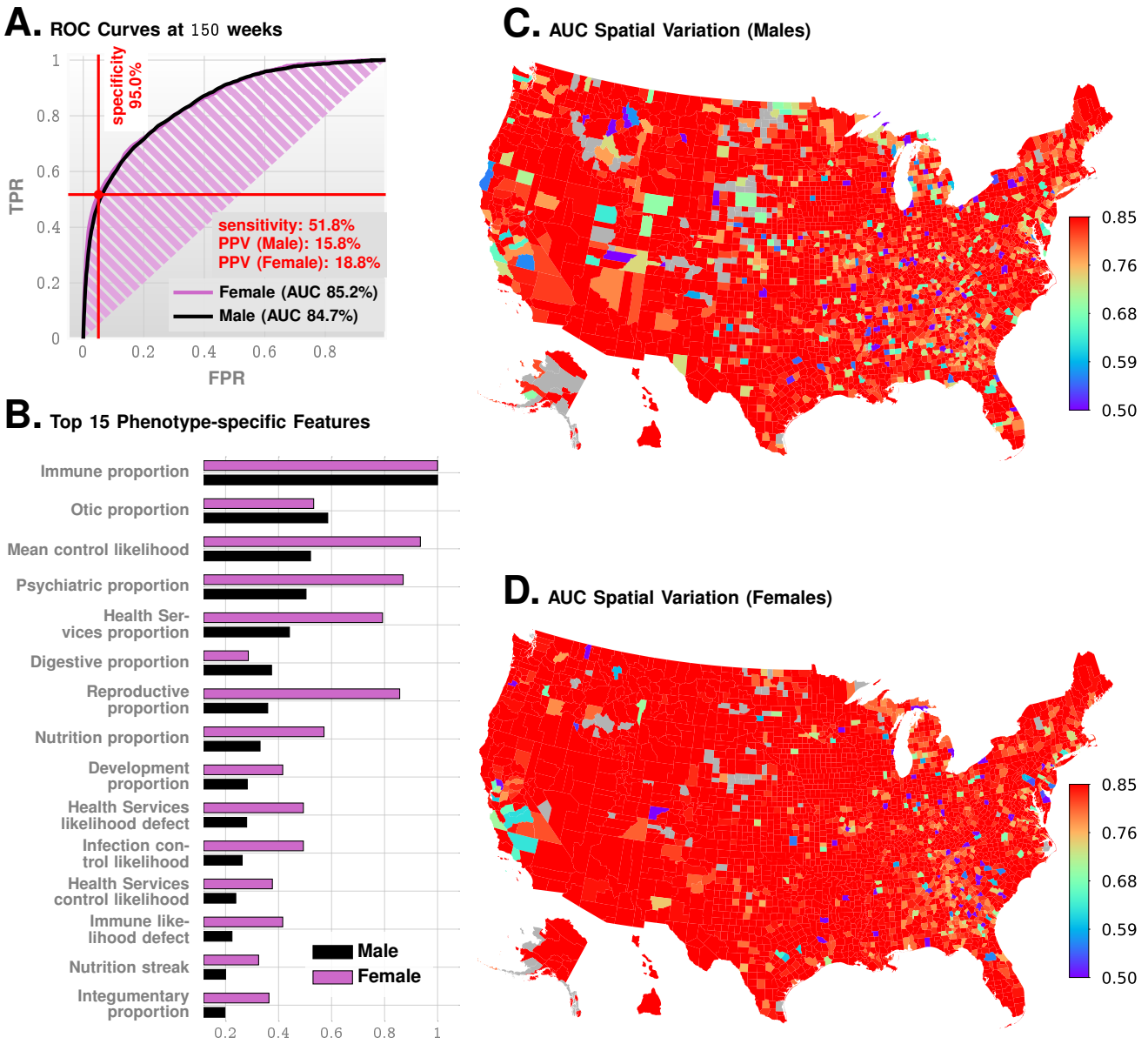


Fig. 1. Predictive Performance. Panel A shows the ROC curves for males and females. Panel B shows the feature importance inferred by our prediction pipeline. The detailed description of the features is given in Extended Data Table I. The most import feature is related to immunologic disorders, and we 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.

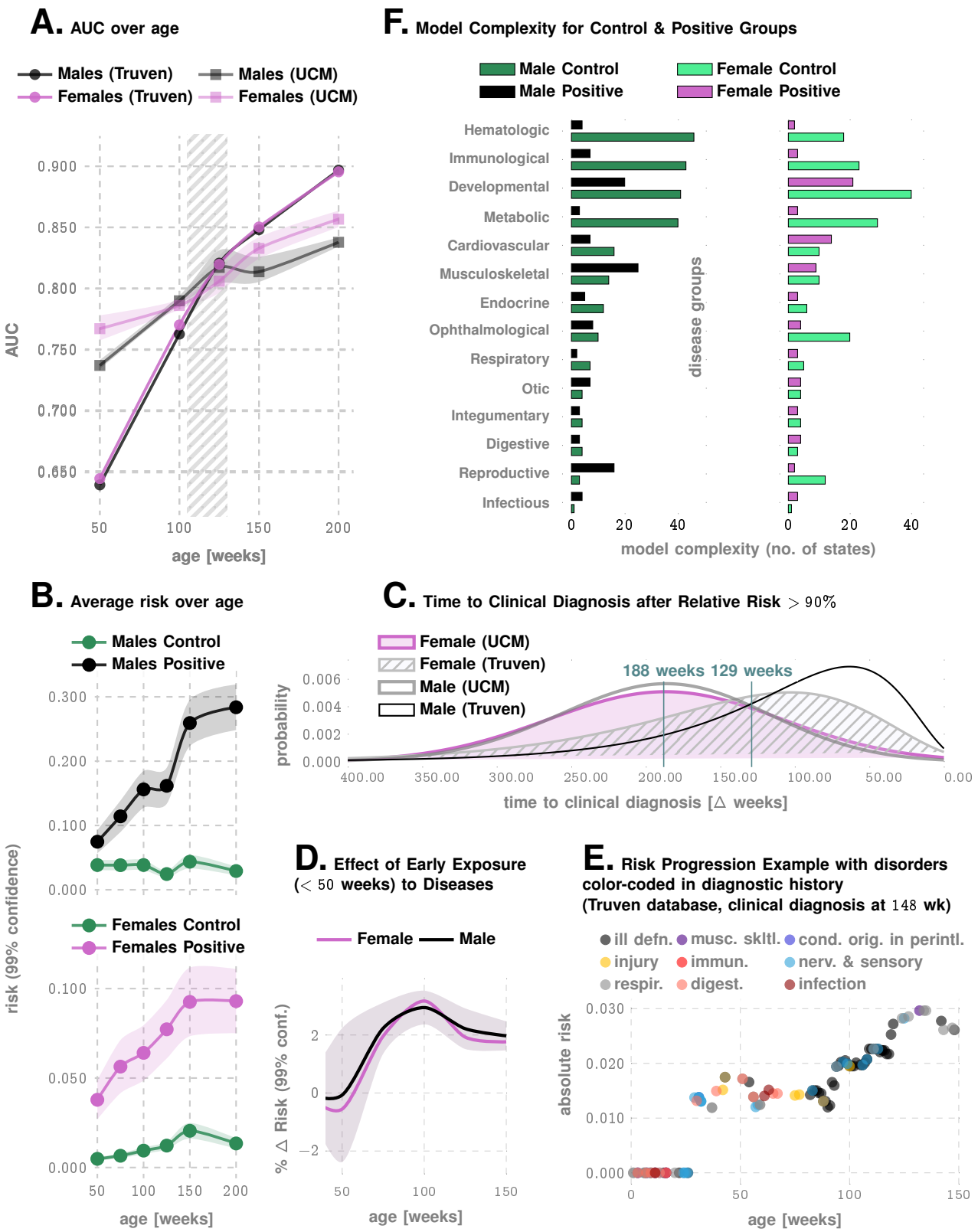


Fig. 2. **More details on Predictive Performance and Variation of Inferred Risk.** Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve $> 80\%$ AUC for either gender from shortly after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

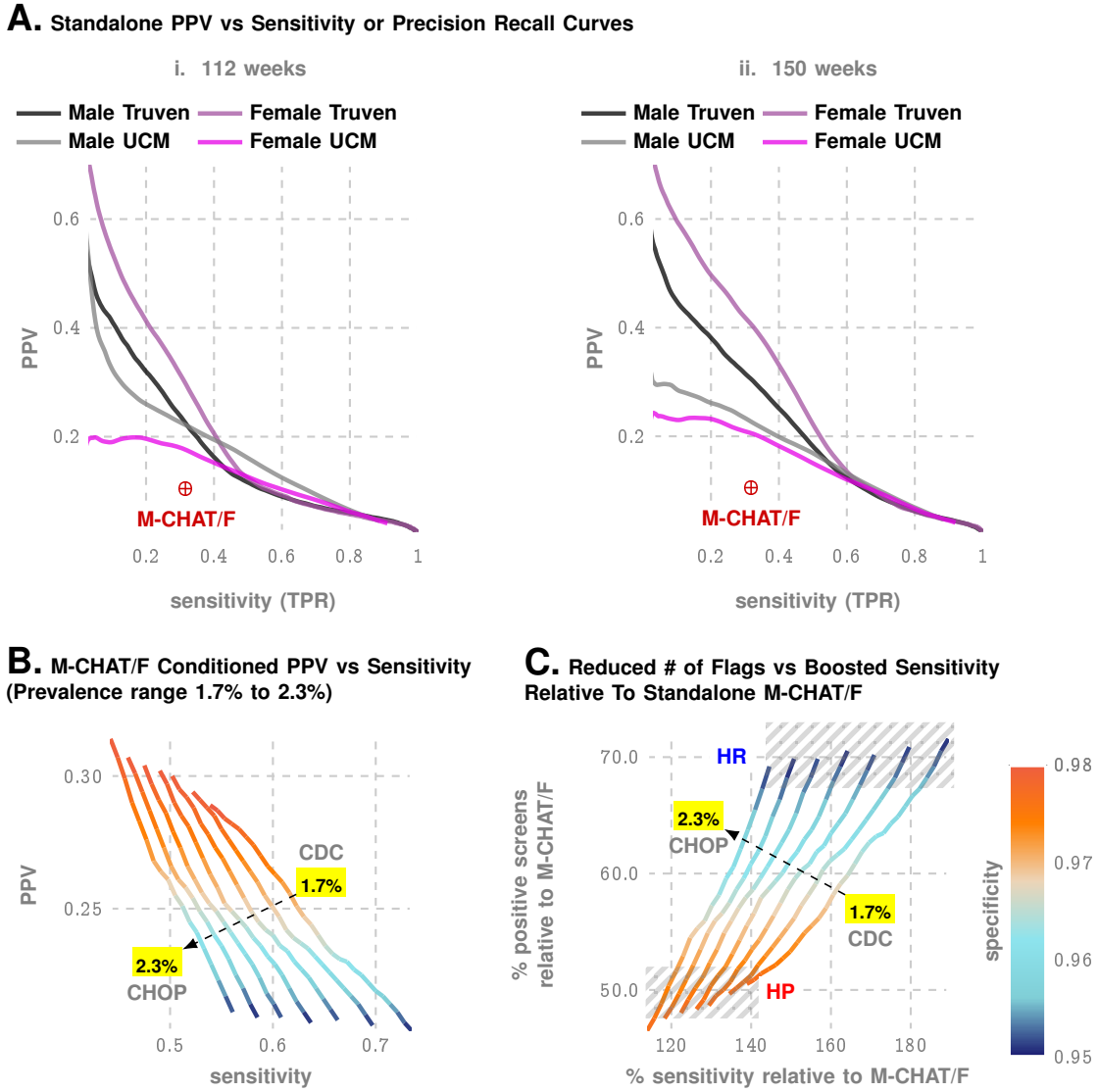


Fig. 3. **Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs.** Panel A shows the precision/recall curves, *i.e.*, the trade-off between PPV and sensitivity. Panel B shows how we can boost performance using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study[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.

TABLE I
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

METHODS

Source of Electronic Patient 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[49] (referred to as the Truven dataset). This US national database contains data contributed by over 150 insurance carriers and large, self-insuring companies, and is a culmination of over 4.6 billion inpatient and outpatient service claims and almost six billion diagnosis codes. For our analysis, we extracted histories of patients within the age of $0-9$ years, and excluded patients who do not satisfy the following criteria: 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, or predict outcomes from. 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. Details on the characteristics of excluded patients is shown in Table I. For training, we analyzed over 30M diagnostic records (16,649,548 for males and 14,318,303 for females with 9,835 unique diagnostic codes).

For practical reasons, this study did not query the records of the mothers of the patients, and therefore does not include analysis of potential pregnancy-related influences. While this is certainly an important question, we delegate such investigations to future work, given that there are barriers in automatically pulling in records of familial members in implementation, due to privacy regulations in the US.

While the Truven database is used for both training and out-of-sample cross-validation with held-back patient data, our second independent dataset (referred to as the UCM dataset) consisting of de-identified diagnostic records for children treated at the University of Chicago Medical Center between the years of 2006 to 2018, 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 I.

Time-series Modeling of Diagnostic History

Individual diagnostic histories can have long-term memory[50], implying that the order, frequency, and comorbid interactions between diseases are potentially important for assessing the future risk of our target phenotype. Our approach to analyzing patient-specific diagnostic code sequences consists of representing the medical history of each patient as a set of stochastic categorical time-series — one each for a specific group of related disorders — followed by the inference of stochastic generators for these individual data streams. These inferred generators are from a special class of Hidden Markov Models (HMMs), referred to as Probabilistic Finite State Automata (PFSA)[51]. The inference algorithm we use is distinct from classical HMM learning, and has important advantages related to its ability to infer structure, and its sample complexity (See Supplementary text, Section **??**). We infer a separate class of models for the positive and control cohorts, and then the problem reduces to determining the probability that the short diagnostic history from a new patient arises from the positive as opposed to the control category of the inferred models. Importantly, the individual histories are typically short, often have large randomly varying gaps, and we have no guarantee that model-structural assumptions[52,53] (linearity, additive noise structure, etc.) often used in the standard time-series analysis is applicable here. Also, the categorical observations are drawn from a large alphabet of possible diagnostic codes, which degrades statistical power. Perhaps most importantly, using patterns emergent at the population level to make individual risk assessments is challenged by the ecological fallacy[54–56] — the fact that group statistics might be neither reflective nor predictive of patterns at the individual level.

Step 1: Partitioning The Human Disease Spectrum

To address the idiosyncrasies of the problem at hand, we begin by partitioning the human disease spectrum into 17 non-overlapping categories, as shown in Extended Data Table I, which remain fixed throughout the analysis. Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9) (See Extended Data Table I in the main text and Table SI-**??** in the Supplementary text for description of the categories used in this study). For this study, we considered $9,835$ distinct ICD9 codes (and their ICD10 General Equivalence Mappings (GEMS)[57] equivalents). We came across 6,089 distinct ICD-9 codes and 11,522 distinct ICD-10 codes in total in the two datasets we analyzed. Transforming the diagnostic histories to report only the broad categories reduces the number of distinct codes that the pipeline needs to handle, thus improving statistical power. The trade-offs for this increased power consist of 1) the loss of distinction between disorders in the same category, and 2) some inherent subjectivity in determining the constituent ICD9 codes

that define each category, $e.g.$ an ear infection may be classified either an otic disease or an infectious one.

Our categories largely align with the top-level ICD9 categories, with small adjustments, $e.g.$ bringing all infections under one category irrespective of the pathogen or the target organ. We do not pre-select the phenotypes; we want our algorithm to seek out the important patterns without any manual curation of the input data. The limitation of the set of phenotypes to 9835 unique codes arises from excluding patients from the database who have very few and rare codes that will skew the statistical estimates. As shown in Table I, we exclude a very small number of patients, and who have very short diagnostic histories with a very small number of codes.

Next, we process raw diagnostic histories to generate data streams that report only the categories instead of the exact codes. For each patient, his or her past medical history is a sequence $(t_1, x_1), \cdots, (t_m, x_m)$, where t_i are timestamps and x_i are ICD9 codes diagnosed at time t_i. We map individual patient history to a three-alphabet categorical time series z^k corresponding to the disease category k, as follows. For each week i, we have:

$$z_i^k = \begin{cases} 0 & \text{if no diagnosis codes in week } i \\ 1 & \text{if there exists a diagnosis of category } k \text{ in week } i \\ 2 & \text{otherwise} \end{cases} \tag{1}$$

The time-series z^k is terminated at a particular week if the patient is diagnosed with ASD the week after. Thus for patients in the control cohort, the length of the mapped trinary series is limited by the time for which the individual is observed within the 2003 – 2012 span of our database. In contrast, for patients in the positive cohort, the length of the mapped series reflect the time to the first ASD diagnosis. Patients do not necessarily enter the database at birth, and we prefix each series with 0s to approximately synchronize observations to age in weeks. In summary, each patient is now represented by 18 mapped trinary series, which we use next to infer population-level PFSA models.

Importantly, to eliminate the possibility that any predictions we get are somehow confounded by codes from the category of "mental, behavioral, and neurodevelopmental diseases" (ICD9 range: 290-319), we 1) carried out a parallel analysis with high out-of-sample predictive performance where we ignored codes from this category, except those for identifying ASD diagnosis, and the category reflecting general health status and contact with health services (ICD9 V0-V91). The results of this analysis are shown in Extended Data Fig. 8 in the Supplementary text, which illustrates that we get just marginally lower performance. This minimizes the possibility that our predictions are somehow informed by the knowledge of prior diagnoses of neurodevelopmental anomalies alone. And 2) verified that if we can distinguish well between ASD and unrelated psychiatric phenotypes (See Results and Extended Data Fig. 7A).

Step 2: Model Inference & The Sequence Likelihood Defect

The mapped series, stratified by gender, disease-category, and ASD diagnosis-status are considered to be independent realizations or sample paths from relatively invariant stochastic dynamical systems, and we want to explicitly model these systems as specialized HMMs (PFSAs) from the observed variations in each subpopulation of patients. We model the positive and the control cohorts for each gender, and in each disease category separately, ending up with a total of 68 HMMs at the population level (17 categories, 2 genders, 2 cohort-types: positive and control, Extended Data Fig. 9 provides some examples). Each of these inferred models is a PFSA; a directed graph with probability-weighted edges, and acts as an optimal generator of the stochastic process driving the sequential appearance of the three letters (as defined by Eq. (1)) corresponding to each gender, disease category, and cohort-type. These models are very nearly assumption-free beyond the requirement that the processes be statistically stationary or slowly varying. (See Section **??** in the Supplementary text for detailed technical background on PFSA inference). In particular, these models are not *a priori* constrained by any structural motifs, complexity, or size, and are compact representations of patterns emerging in the mapped time series. Additionally, when learning models for sets of diagnostic histories corresponding to a patient cohort, the histories can be of different lengths. The modeling objective here is to exploit the relative differences in these probabilistic models to reliably infer the cohort-type of a new patient from their individual sequence of past diagnostic codes.

To that effect, we generalized the well-known notion of Kullbeck-Leibler (KL) divergence[58,59] between probability distributions to a divergence $\mathcal{D}_{\mathsf{KL}}(G||H)$ between ergodic stationary categorical stochastic processes[60] G, H as:

$$\mathcal{D}_{\mathsf{KL}}(G||H) = \lim_{n\to\infty} \frac{1}{n} \sum_{x:|x|=n} p_G(x) \log \frac{p_G(x)}{p_H(x)} \tag{2}$$

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 :

$$L(x, G) = -\frac{1}{|x|} \log p_G(x) \tag{3}$$

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 **??**, Theorem 6 and 7):

$$\lim_{|x|\to\infty} L(x, G) = \mathcal{H}(G) \tag{4a}$$

$$\lim_{|x|\to\infty} L(x, H) = \mathcal{H}(G) + \mathcal{D}_{\mathsf{KL}}(G||H) \tag{4b}$$

where $\mathcal{H}(\cdot)$ is the entropy rate of a process[58]. Importantly, Eq. (4) 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 series corresponding to her diagnostic history be modeled by the PFSA in the positive cohort. Denoting the PFSA 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:

$$\Delta^j \triangleq L(G^j_0, x) - L(G^j_+, x) \to \mathcal{D}_{\mathsf{KL}}(G^j_0||G^j_+) \tag{5}$$

With the inferred population-level PFSA models and the individual diagnostic history, we can now estimate the SLD measure on the right-hand side of Eqn. (5). The higher this likelihood defect, the higher the similarity of the patient's history to ones that have an eventual ASD diagnosis with respect to the disease category being considered. SLD is the core novel analytic tool used in this study to tease out information relevant to the risk estimator and is key to the design of our risk estimation pipeline.

Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules

Ultimately, the risk estimation pipeline operates on patient specific information limited to the gender and available diagnostic history from birth, and produces an estimate of the relative risk of ASD diagnosis at a specific age, with an associated confidence value. To learn the parameters and associated model structures of this pipeline, we transform the patient specific data to a set of engineered features, and the feature vectors realized on the positive and control sets are then used to train a gradient-boosting classifier[61]. Of the set of engineered features, the most important are the disease-category-specific SLD described above. For example, if $SLD > 0$ for a specific patient for every disease category, then he or she is likely to have an ASD diagnosis eventually. However, not all disease categories are equally important for this decision; parametric tuning of the classifier allows us to infer the optimal combination weights, as well as compute the relative risk with associated confidence. In addition to category-specific SLDs, we use a range of other derived quantities as features, including the mean and variance of the defects computed over all disease categories, the occurrence frequency of the different disease groups, etc. The complete list of 165 features used by the estimation pipeline is provided in Extended Data Table II.

Since we need to infer the HMM 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 classifier in the pipeline with features derived from the inferred models. Thus, the analysis proceeds by first carrying out a random 3-way split of the set of unique patients into *feature-engineering* (25%), *training* (25%) and *test* (50%) sets. We use the feature-engineering set of ids first to infer our PFSA models *(unsupervised model inference in each category)*, which then allows us to train the gradient-boosting classifier using the training set and PFSA models *(classical supervised learning)*, and we finally execute out-of-sample validation on the test set. The approximate sizes of the three sets are as follows: $\approx 700K$ each for the feature-engineering and the training sets, and $\approx 1.5M$ for the test set. The top 15 features used in our pipeline may be ranked in order of their relative importance (See Fig. 1B), by estimating the loss in performance when dropped out of the analysis.

Calculating Relative Risk

Our pipeline maps medical histories to a score, which is interpreted as a raw indicator of risk — higher this value, higher the probability of a future diagnosis. However, to make crisp predictions, we must choose a decision threshold for this raw score. Conceptually identical to the notion of Type 1 and Type 2 errors in classical statistical analyses, the choice of a threshold trades off false positives (Type 1 error) for false negatives (Type 2 error): choosing a small threshold results in predicting a larger fraction of future diagnoses correctly, $i.e.$ have a high true positive rate (TPR), while simultaneously suffering from a higher false positive rate (FPR), and vice versa. The receiver operating characteristic curve (ROC) is the plot of the FPR vs the TPR, as we vary this decision threshold. If our predictor is good, we will consistently achieve high TPR with small FPR resulting in a high

area under the ROC curve (AUC); AUC measures intrinsic performance, independent of the threshold choice. More importantly, the AUC is immune to class imbalance, $i.e.$, the fact that the control cohort is several orders of magnitude larger than the positive cohort (See Supplementary text, Section **??** for a brief discussion). An AUC of 50% indicates that the predictor does no better than random, and an AUC of 100% would imply perfect prediction of future diagnoses, with zero false positives. Our reported AUCs, as shown in Fig. 2A, are all computed on out-of-sample data, $i.e.$, on held back subset from the Truven dataset, and on the entirety of the UCM samples (the latter being never used in training and pipeline design). A flowchart of the computational steps is shown in Supplementary Fig. SI-**??**.

Therefore, a choice of a specific decision threshold — necessary for making individual predictions and meaningful risk assessments — 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. Other possible strategies for selecting thresholds are maximizing accuracy (the fraction of correct predictions on the presence or the absence of a future diagnosis, also known as the classification rate), or even simply maximizing the true positives rate or the recall of the decision-maker. However, with a severe class imbalance in our application, the F_1-score is recommended, being independent of the number of true negative samples. (See Supplementary text, Section **??**).

The *relative risk* is then defined as the ratio of the raw pipeline score to the chosen decision threshold. Thus, a relative risk > 1 implies that we are predicting an eventual ASD diagnosis, and on average our decisions maximize the F_1-score of our pipeline. While the raw score does not give us actionable information, the relative risk being close to or greater than 1.0 for a specific child signals the need for intervention.

Calculating PPV, Sensitivity & Specificity Trade-offs

The sensitivity vs PPV plots, also known as the precision-recall curves (See Fig. 3A) are constructed in a similar fashion as the ROC curves by varying the decision threshold. These curves allow direct comparison with the state of the art screening tests,$e.g.$, M-CHAT/F, in a manner that is most relevant to clinical practitioners.

Boosting Performance Via Leveraging Population Stratification Induced By Existing Tests

In this study, 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. Assume that there are m sub-populations such that: the sensitivities, specificities achieved, and the prevalences in each sub-population are given by s_i, c_i and ρ_i respectively, with $i \in \{1, \cdots, m\}$. Let β_i be the relative size of each sub-population. Then, we have (See Supplementary text, Section **??**):

$$s = \sum_{i=1}^{m} s_i \gamma_i \tag{6a}$$

$$c = \sum_{i=1}^{m} c_i \gamma_i' \tag{6b}$$

where we have denoted:

$$\gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma_i' = \beta_i \frac{1-\rho_i}{1-\rho} \tag{6c}$$

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 Extended Data Table VI). The “follow-up” in the context of M-CHAT/F refers to the re-evaluation of responses by qualified personnel. We use published data on the relative sizes and the prevalence statistics in these sub-populations[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.

Importantly, designing the rules for conditional operation only requires average population characteristics, *i.e.*, an estimate of ASD prevalence in the sub-populations defined by the relevant brackets of M-CHAT/F scores, and the prevalence of these score brackets in the general population. In particular, M-CHAT/F scores of individual patients are unnecessary for designing the rules themselves, or evaluating the overall expected performance.

Perturbation Analysis

Since our pipeline maps any sequence of time-stamped diagnostic codes to a predicted ASD risk, we can investigate how small perturbations of the patient histories modulate risk, which is simulated by injecting a single additional code randomly chosen from our disease categories anywhere within the first year of life. We know that children who meet the criteria for ASD experience higher rates of co-morbid disorders, and conversely, our perturbation analysis indicates that each new diagnostic code within the first year of life quantifiably increases the risk burden by approximately 2% irrespective of gender (See in Fig. 2D).

Inferred Model Complexities

Our inferred HMM models aim to capture the variation in the underlying dynamical processes driving disease processes between children in the positive and the control cohorts. Fig. 2F, illustrates the variation in the statistical complexity of the inferred models amongst different disease categories for males and females. Model complexity is the number of states that the corresponding probabilistic machines are inferred to have. With no *a priori* imposed constraints on the model structure in our approach, the inferred size of the state space reflects the intrinsic statistical complexity of the stochastic process[62] it models. In particular, independent identically distributed (i.i.d.) processes have no dependence on past history, and therefore have a single state. And the more complex the historical dependence or the process-memory, the larger the number of model states. In light of this interpretation, the enumeration of model complexity in Fig 2F shows that disease groups have different degrees of process-memory, *e.g.*, immunologic disorders have a significantly larger memory compared to infections in the control cohorts for both males and females. We also note that our analysis indicates that such process-memory is gender-specific, and more importantly, the memory degrades on average for the positive cohorts. In other words, patients who eventually get an ASD diagnosis appear to have a relatively more random experience of disorders, on average, across the disease spectrum.

Measures Relevant To Clinical Practice: PPV, Sensitivity & Specificity Trade-off

For our results to be deemed useful, clinicians need more information than simply the AUC or the relative risk. Of far more importance is the sensitivity of the tool at some high value of specificity (typically 95% or higher), and the associated precision or the PPV. Specificity is the ratio of the number of true negatives to the size of the control set, and indicates the fraction of patients indicated to be at low risk are indeed so. Thus a high specificity indicates a smaller number of false positives and vice versa. Sensitivity or recall is the true positive rate, *i.e.*, the ratio of the number of true positives to the size of the positive set. The higher the sensitivity, the lower the fraction of "misses". PPV is the ratio of the true positives to the total number of patients *indicated to be positive by the tool*. Panels A(i) and A(ii) in Fig. 3 show the out-of-sample PPV vs sensitivity curves for the two databases, stratified by gender, computed at 150 and 100 weeks respectively. A single illustrative 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%. A more detailed picture of this trade-off at 100, 112 (≈ 26 months), and 150 weeks is given in Extended Data Table III, and the PPV vs Sensitivity curves at ages of 112 and 150 weeks are shown in Fig. 3A.

To be adopted in clinical practice, any new tool must compete with the screening tools being deployed and used today. The existing tools are based on questionnaires that help flag early manifestations of symptoms of core deficits related to social communication[8], and are designed to help caregivers identify symptoms observed in children at high risk for ASD. In primary pediatric care, the M-CHAT/F is the most studied and widely used tool for screening toddlers for ASD[8,13]. Guthrie *et al.*[3] from Children's Hospital of Philadelphia (CHOP) has recently demonstrated that when applied as a nearly universal screening tool, M-CHAT/F has a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%, implying that out of every 100 children who in fact ave ASD, the M-CHAT/F flags about 39, and out of every 100 children it flags, about 85 are false positives. The PPV is affected by the prevalence of the disease (See Supplementary text, Section **??**). This work is the only large-scale study of M-CHAT/F (n=20,375) we are aware of with sufficient follow-up after the age of four years to provide a reasonable degree of confidence in the reported performance values.

Comparing the performance metrics achieved at different age groups across data sets and genders for our pipeline (See Extended Data Table III), we conclude that our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at 14% (14.1-33.6%) when sensitivity and specificity are held at comparable values around the age of 26 months (≈ 112 weeks). We cannot compare at other operating points due to a lack of M-CHAT/F performance characterization anywhere else.

Conditioning on Existing Screens To Reduce False Positives & Boost Sensitivity

Using the population stratification induced by M-CHAT/F scores calculated from the CHOP study[3] (See Extended Data Table VII in the Supplementary text), we can compute a conditional choice of sensitivity and specificity for our tool, in each sub-population. This ultimately yields an overall performance significantly superior to M-CHAT/F alone. We carry out a four-dimensional search at the age of 26 months (≈ 112 weeks) to identify the feasible region with PPV $> 14.6\%$ and sensitivity $> 38.8\%$ simultaneously while keeping specificity $> 94\%$. These four dimensions reflect the independent choice of sensitivities in the corresponding sub-populations. For each set of choices, the associated specificities are read-off from our fixed pre-computed ROC curve corresponding to 112 weeks, and then the overall sensitivity, specificity, and PPV are calculated using Eq. (**??**) in the Supplementary text, Section **??**. Since the CHOP study does not report gender stratified data, we averaged our male and female ROC curves. This is reasonable since the curves are similar across the genders. We computed the choices for out-of-sample data in the Truven dataset, and verified in the UCM dataset that those choices yield similar performance (See Extended Data Table IV for results at 26 months, and Extended Data Table V in the Supplementary text for results at 150 weeks with same population stratification statistics).

Importantly, we assume here that the two tests are independent. Since M-CHAT/F is based on the detection of behavioral signals of developmental delay via questionnaires completed by the primary care-givers, while our pipeline is based on the diagnoses of physical co-morbidities, independence is reasonable.

We show the trade-offs between PPV and sensitivity for operation conditioned on M-CHAT/F scores in Fig. 3B (See also Extended Data Fig. 6 for the explicit feasible regions computed). We get a PPV close to or exceeding 30% at the high precision (HP) operating point across datasets ($> 33\%$ fro Truven, $> 28\%$ for UCM), or a sensitivity close to or exceeding 50% for the high recall (HR) operating point ($> 58\%$ for Truven, $> 50\%$ for UCM), when we restrict specificities to above 95% (See Extended Data Table IV).

Importantly, by Eq. (6) the optimization results are dependent on the population prevalence ρ. We report results for population prevalence varying between 1.7% (CDC estimate[8]), and 2.23% (CHOP estimate[3]) (See Fig. 3, panels B and C). We find that if the global prevalence is lower, we can achieve significantly higher sensitivities at the HR point (reaching $> 73\%$ in the Truven dataset, and $> 62\%$ in the UCM dataset), and if the prevalences is higher then we can achieve slightly higher PPVs at the HP operating point (reaching $> 33\%$ in the Truven dataset and $> 28\%$ in the UCM dataset).

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

Sanity Checks: Uncertainty in EHR Records & Baseline Approaches in Machine Learning

Recent changes in diagnostic practice, *e.g.* increased diagnoses from individual clinicians versus prior eras that only allowed diagnosis from the gold-standard multi-disciplinary teams can increase observed prevalence, and raises the possibility that some diagnostic codes pertaining to ASD in medical history databases could be arising from less restrictive workflows, and are susceptible to increased uncertainty. In our study, we verified that restricting the positive cohort to children with at least two distinct ASD diagnostic codes in their medical histories instead of one, has little impact on out-of-sample predictive performance (See Extended Data Fig. 7B).

We also verified that class imbalance is not inappropriately enhancing our performance, by replacing the control cohort with a random sample of size equal to that of the positive cohort in out-of-sample tests (See Extended Data Fig. 7C).

We found that off-the-shelf machine learning tools are unable to deliver good performance when applied directly (See Supplementary text, Section **??**). Closer performance is achieved when we use our pre-processing of diagnostic histories (See Methods), followed by the application of different standard tools. We also compared our optimized pipeline to analyses using only a subset of our features, *e.g.*, using only features derived from

sequence statistics and excluding the ones derived from learning PFSAs, or using only the PFSA-based features, or using simply the density of diagnostic codes (See Extended Data Fig. 7, panel D). In all these cases we analyzed, our pipeline has a clear advantage (See Extended Data Fig. 7, panel D) that is stable across databases, under reductions in sample sizes, and in balanced re-sampling experiments (See Extended Data Fig. 7, panel C).

Notably, the PFSA based features by themselves are comparable to those engineered manually from sequence statistics (See Extended Data Fig. 7, panel D) when used separately. The latter include features such as the proportion of codes in a patient's history corresponding to specific disease categories, mean and variance of adjacent empty weeks etc. (See Extended Data Table II). Nevertheless, the combined feature set produces significantly superior results.

We also found that the density of diagnostic codes in a child's medical history by itself is somewhat predictive of a future ASD diagnosis, with the AUC $\approx 75\%$ in the Truven database at 150 weeks (See Extended Data Fig. 7, panel D). This is expected, since children with autism do indeed have higher rates of co-morbidities. However, it does not have stable performance across databases. We did not include code density in our combined feature set since it has no effect once the rest of the features are combined.

End Notes

A fully functional demonstration pipeline is available at https://pypi.org/project/ehrzero/, which can be installed easily on any python enabled system as per instructions in Section **??** in the Supplementary text.

Acknowledgements

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 necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

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

[15] 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).

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

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

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

[19] 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).

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

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

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

[23] 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).

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

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

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

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

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

[29] 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).

[30] 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).

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

[32] 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).

[33] 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).

[34] Fiorentino, M. *et al.* Blood-brain barrier and intestinal epithelial barrier alterations in autism spectrum disorders. *Mol Autism* **7**, 49 (2016).

[35] 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.

[36] Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).

[37] 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).

[38] 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).

[39] 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).

[40] 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).

[41] 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.

[42] Murdoch, J. D. & State, M. W. Recent developments in the genetics of autism spectrum disorders. *Curr. Opin. Genet. Dev.* **23**, 310–315 (2013).

[43] Hu, V. W. The expanding genomic landscape of autism: discovering the 'forest' beyond the 'trees'. *Future Neurol* **8**, 29–42 (2013).

[44] 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).

[45] 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).

[46] Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning (2020). URL https://www.cdc.gov/ncbddd/cp/features/prevalence.html.

[47] 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).

[48] Rdgaard, E.-M., Jensen, K., Vergnes, J.-N., Soulires, 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.

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

[50] Granger, C. W. J. & Joyeux, R. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.

[51] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).

[52] Stoyanov, S. V., Racheva-Iotova, B., Rachev, S. T. & Fabozzi, F. J. Stochastic models for risk estimation in volatile markets: a survey. *Annals of Operations Research* **176**, 293–309 (2010). URL https://doi.org/10.1007/s10479-008-0468-1.

[53] Shumway, R. H. & Stoffer, D. S. *Time Series Regression and ARIMA Models*, 89–212 (Springer New York, New York, NY, 2000). URL https://doi.org/10.1007/978-1-4757-3261-0_2.

[54] Freedman, D. A. *Ecological Fallacy*, 293–294 (SAGE Publications, Thousand Oaks, CA, 2004).

[55] Rao, B. R., Day, R. D. & Marsh, G. M. Estimation of relative risks from individual and ecological correlation studies **21**, 241–268 (1992).

[56] Bendel, R. B. & Carlin, B. P. Bayes methods in the ecological fallacy context: estimation of individual correlation from aggregate data **19**, 2595–2623 (1990).

[57] General equivalence mappings. URL https://www.cms.gov/Medicare/Coding/ICD10/downloads/ICD-10_GEM_fact_sheet.pdf.

[58] Cover, T. M. & Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).

[59] Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86 (1951). URL https://doi.org/10.1214/aoms/1177729694.

[60] Doob, J. *Stochastic Processes*. Wiley Publications in Statistics (John Wiley & Sons, 1953). URL https://books.google.com/books?id=KvJQAAAAMAAJ.

[61] Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002). URL http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

[62] Crutchfield, J. P. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena* **75**, 11 – 54 (1994). URL http://www.sciencedirect.com/science/article/pii/0167278994902739.

[63] Jarquin, V. G., Wiggins, L. D., Schieve, L. A. & Van Naarden-Braun, K. Racial disparities in community identification of autism spectrum disorders over time; metropolitan atlanta, georgia, 2000–2006. *Journal of*

Extended Data Tab. I

DISEASE CATEGORIES (A FEW ICD9 CODES SHOWN FROM THE COMPLETE SET OF 9,835 UNIQUE ICD9 CODES CONSIDERED. SEE SI-TABLE **??** IN SUPPLEMENTARY TEXT FOR COMPLETE LIST)

Category†	**Description**	**Examples of ICD9* Codes**
ASD*	Diagnostic Target	299 299.0 299.00 299.01 299.9 299.8 299.91 299.90 299.80 299.81 299.1 299.10 299.11
Immunologic	Diseases related to dys-regulation of the Immune system	580.81 580.89 580.0 580.8 461 461.8 461.0 477.9 477.2 477 477.8
Infectious	Diseases Caused By Pathogens	487.8 488.12 488.0 488.01 487.0 487.1 488.09 464.4 466 466.11 466.1
Nutrition	Symptoms concerning nutrition, metabolism and development	783.0 783.21 783.3 783.40 783.42 783.7 783.9
Mental Disorders	Psychiatric phenotypes other than ASD	290 - 319 (except 299.x)
Health Services	Contact With Health Services and Classification Of Factors Influencing Health Status	V01.0 V01.1 V01.2 V01.3 V01.4 V09.70 V09.71 V88.02 V88.03 V89.01 V89.02 V89.03 V89.04 V89.05 V89.09
Digestive	Diseases Of The Digestive System	540.0 540.1 541.0 542 540 541 543.0 562.03 562.01 562.00 562.10
Otic	Diseases Of The Ear And Mastoid Process	381.51 381.50 381.81 381.89 381.61 381.62 381 381.7 385.82 383.32 380.30
Musculoskeletal	Congenital musculoskeletal anomalies	756.52 756.53 733.02 733.0 733.09 737.43 737.41 737.20 737.29 737.4 737.2
Developmental	Congenital anomalies (Non-overlapping with muskuloskeletal)	755.55 743.45 743.11 743.10 743.00 743.03 743.44 743.22 743.20 743.21 758.4
Reproductive	Diseases Of The Genitourinary System	611.79 611.71 611.89 611.81 676.64 611 676.60 611.6 611.4 611.3 611.2
Integumentary	Diseases Of Skin And Subcutaneous Tissue	706.0 706.1 704.00 704.02 704.09 680.9 680.1 680.5 680.7 680.6 680
Ophthalmologic	Disorders Of The Eye And Adnexa	362.8 362.9 362.6 362.1 362.3 362.18 362.17 362.13 362.11 363.33 363.32
Hematologic	Diseases Of The Blood And Blood-Forming Organs	286.9 286.6 283.19 283.11 283.9 283.1 284.0 284.09 284 284.01
Metabolic	Metabolic Disorders (Non-overlapping with respiratory, digestive and immunological conditions)	273.4 270 270.3 712.11 712.13 712.12 712.14 712.18 712.30 712.37 712.36
Cardiovascular	Diseases Of Arteries, Arterioles, And Capillaries	442.89 441.6 442.82 442.83 441.03 441.02 441.00 442 414.11 447.70 447.71
Respiratory	Diseases Of The Respiratory System (non-overlapping with Infectious)	516.31 516.30 516.32 516.35 516.37 516.36 516.8 516.0 277.0 277.00 277.01
Endocrine	Disorders Of Thyroid and other Endocrine Glands	244 244.9 244.2 255.41 255.5 255.4 259.51 255 259.4 255.11 242.2

† Categories inferred to be important for risk modulation are proportionately highlighted.

* ICD10 codes when present were mapped back to closest ICD9 matches using published General Equivalence Mappings[57].

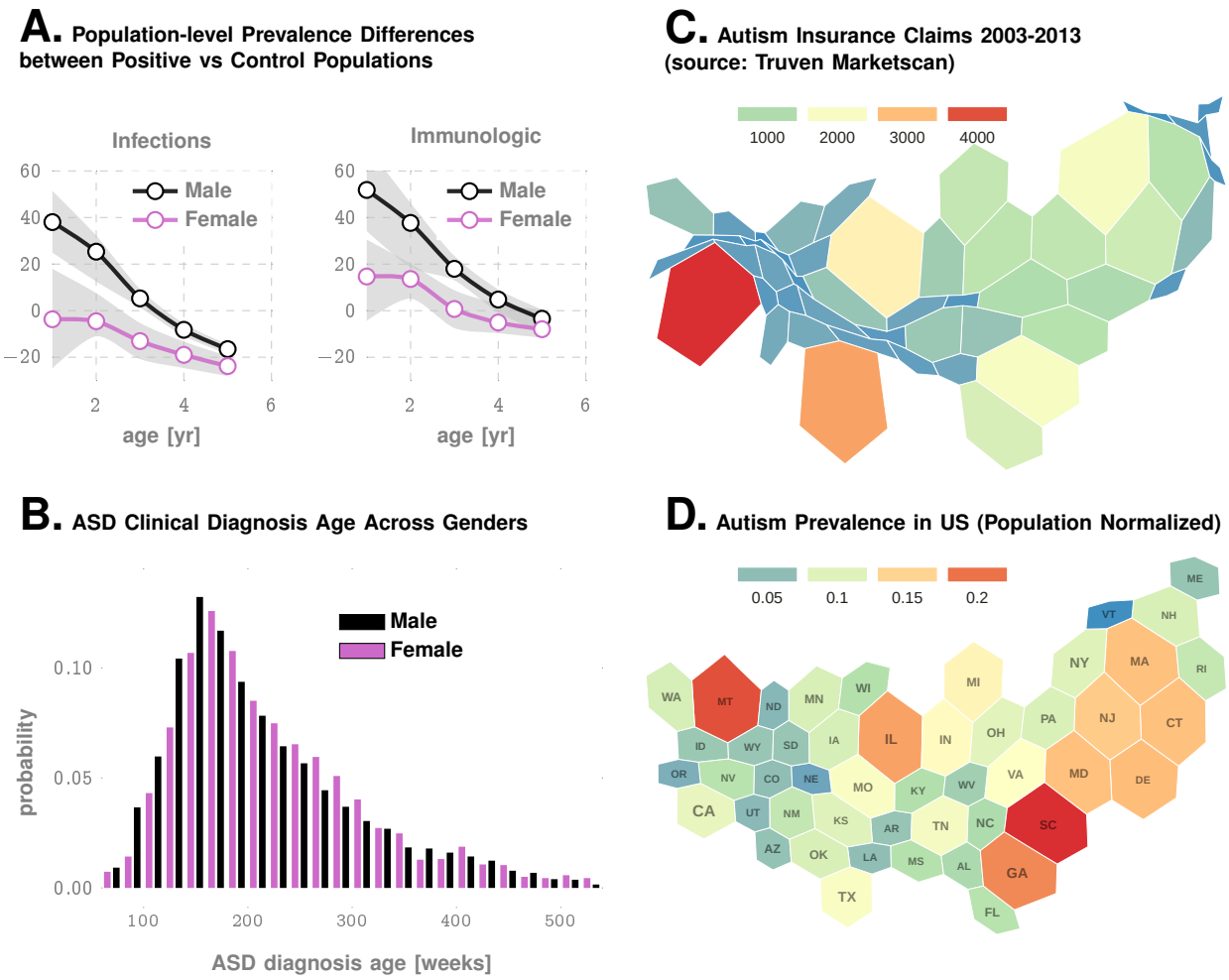
Extended Data Tab. II
ENGINEERED FEATURES (TOTAL COUNT: 165)

Feature Type‡	**Description**	**No. of Features**
[Disease Category] $_{\Delta}$	Likelihood Defect (See Methods section)	17
[Disease Category] $_{0}$	Likelihood of control model (See Methods section)	17
[Disease Category] $_{proportion}$	Occurrences in the encoded sequence / length of the sequence	17
[Disease Category] $_{streak}$	Maximum Length of adjacent occurrences of [Disease Category]	51
[Disease Category] $_{prevalence}$	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] $_{\Delta}$ 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 I

Extended Data Tab. III
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

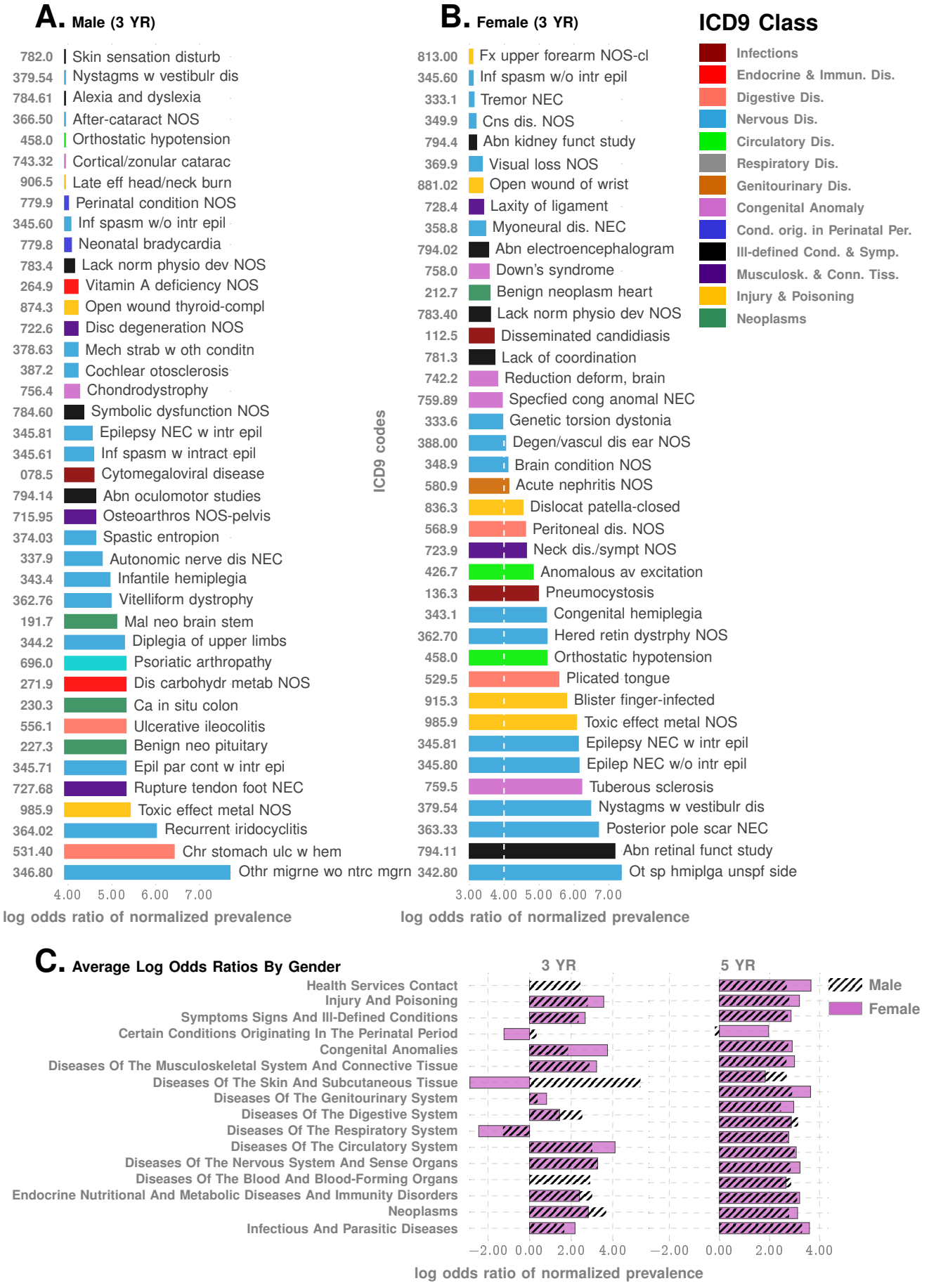


Extended Data Fig. 1. ASD Occurrence Patterns. Panel A illustrates the differential representation of different disease categories in the positive and control cohorts, and panel B shows the distribution of the age of diagnosis for males and females in the Truven dataset. Panel C illustrates the spatial distribution of ASD insurance claims, and panel D shows the same data after population normalization, illustrating the relatively small demographic skew to ASD prevalence within the general population with access to medical insurance, which is consistent with the suggestion that prevalence variation might be linked to regional and socioeconomic disparities in access to services[63].

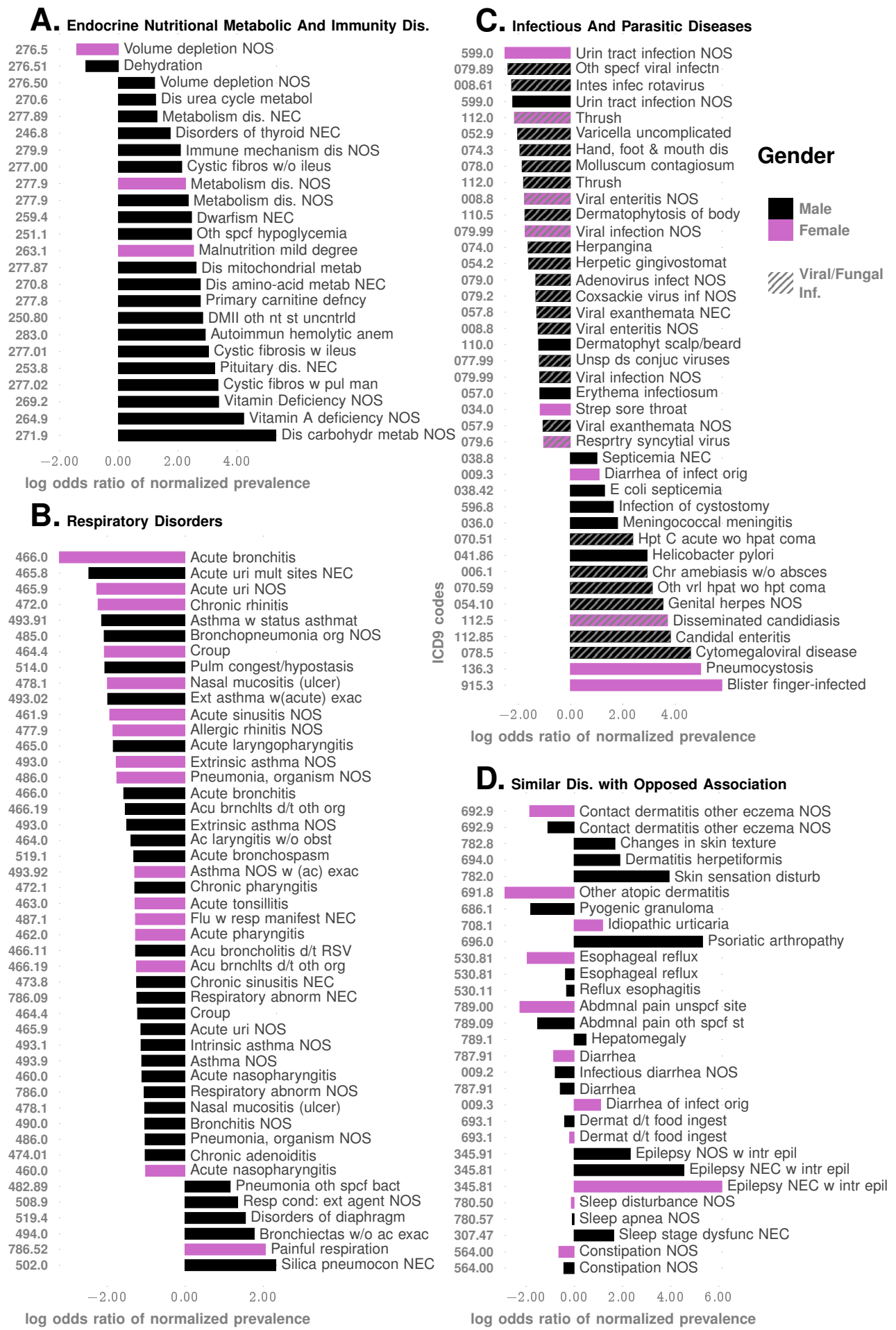
Extended Data Tab. IV
PERSONALIZED OPERATION CONDITIONED ON M-CHAT/F SCORES AT 26 MONTHS

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

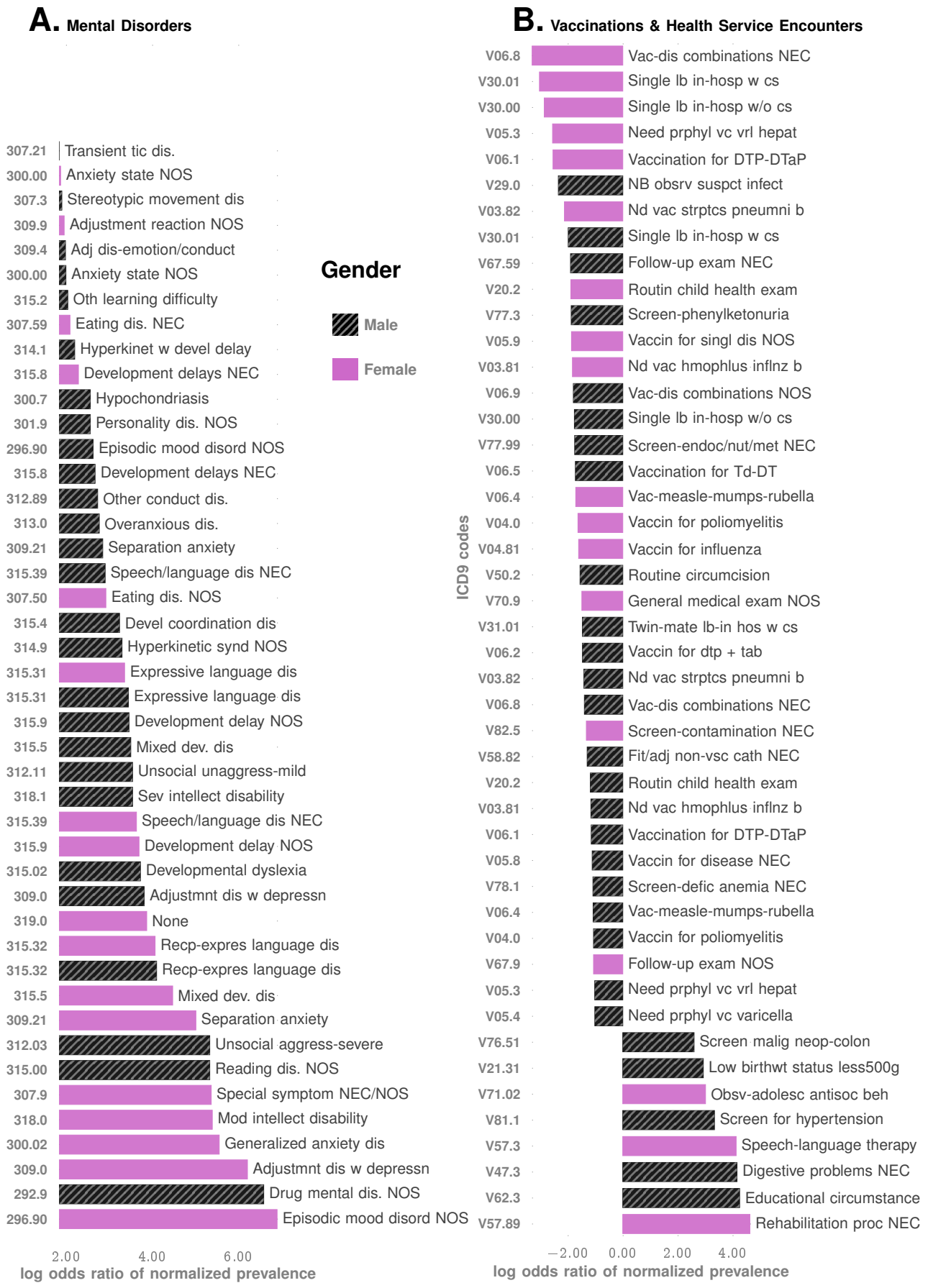
*Prevalence reported by CDC is 1.7%, while the CHOP study reports a value of 2.23%. The results of our optimization depend on the prevalence estimate.



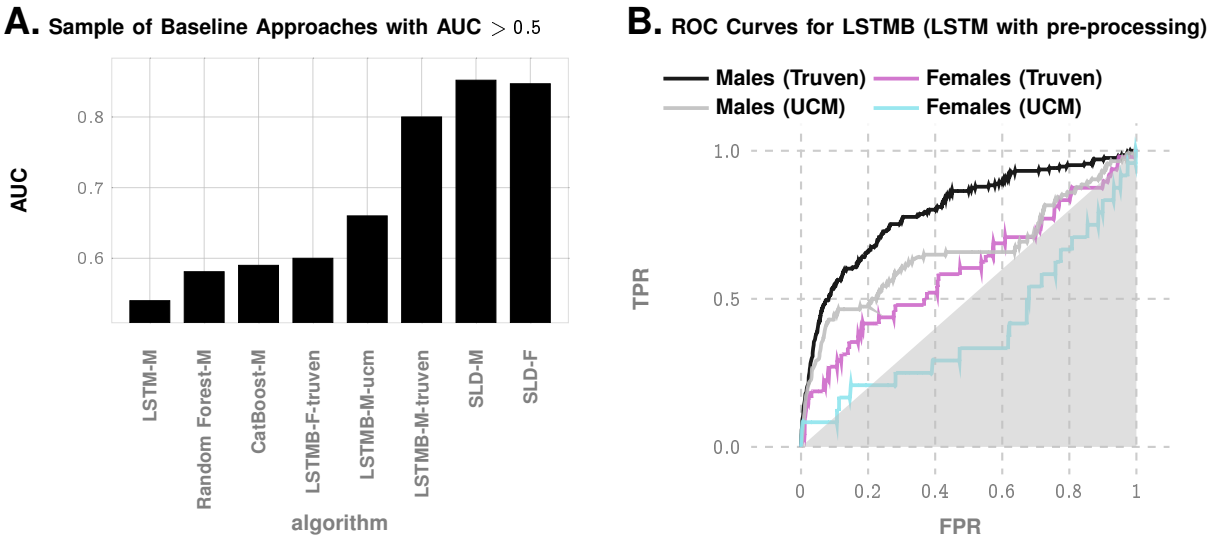
Extended Data Fig. 2. **Co-morbidity Patterns** Panel A and B. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions. The dotted line on panel B shows the abscissa lower cut-off in Panel A, illustrating the lower prevalence of codes in 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.



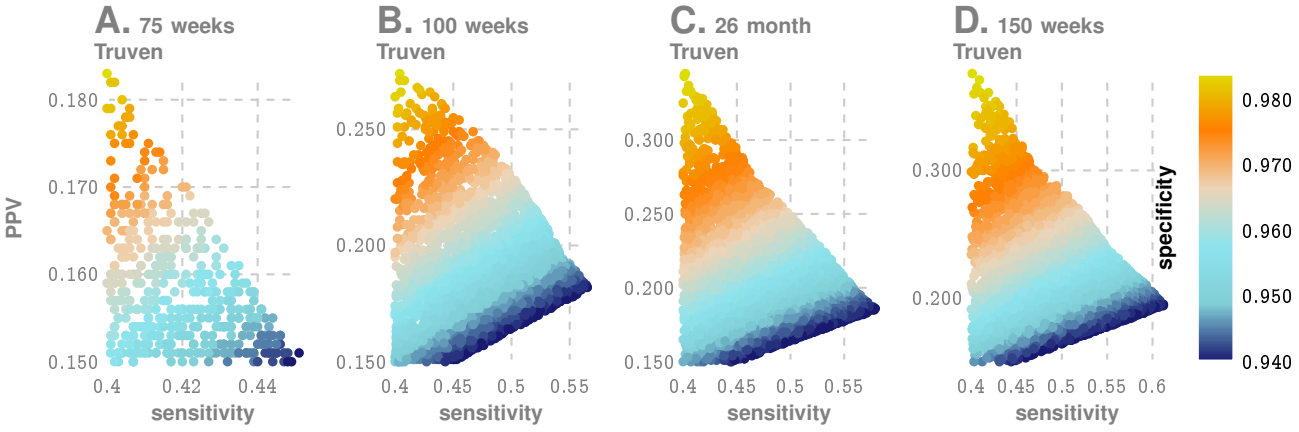
Extended Data Fig. 3. **Details of Co-morbidity Patterns (at age <3 years)** for immunologic (panel A), respiratory (panel B), infections (panel C), and disorders with similar pathobiology manifesting opposing association with autism (panel D).



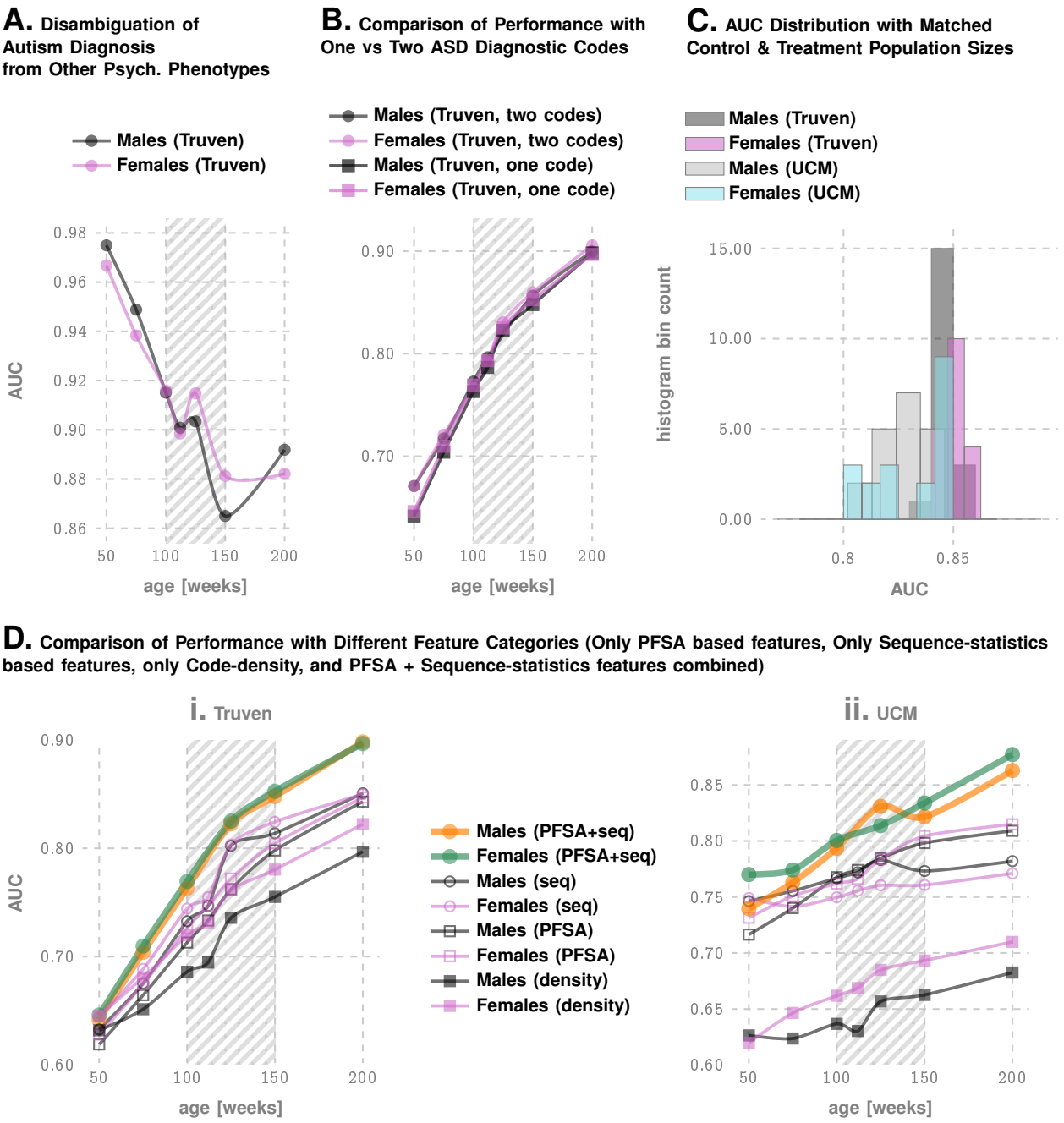
Extended Data Fig. 4. **Co-morbidity Patterns** for mental disorders, vaccinations and health-service encounters.



Extended Data Fig. 5. Performance of standard tools on correctly predicting eventual ASD diagnosis, computed at age 150 weeks of age. Long-short Term Memory (LSTM) networks are the state of the art variation of recurrent neural nets, and Random Forests and Gradient Boosting classifiers (CatBoost) are generally regarded as a representative state of the art classification algorithms. Sequence Likelihood Defect (SLD) is the approach developed in this study. LSTMB denotes LSTM with identical pre-processing as in our pipeline (instead of using raw diagnostic codes). We get much better performance with LSTMB with males in the Truven dataset, but the performance is sensitive to the sizes of the training set, and degrades for smaller samples available for females and in the UCM database, as shown in Panel B.



Extended Data Fig. 6. **4D Search To Take Advantage of Data on Population Stratification (Using Prevalence of 2.23% as reported by CHOP[3]).** While as a standalone tool our approach is comparable to M-CHAT/F at around the 26 month mark (and later), we can take advantage of the independence of the tests to devise a conditional choice of the operating parameters for the new approach. In particular, taking advantage of published estimated prevalence rates of different categories of M-CHAT/F scores, and true positives in each sub-population upon stratification, we can choose a different set of specificity and sensitivity in each sub-population to yield significantly improved overall performance across databases, and much earlier. Additionally, we can choose to operate at a high recall point, where we maximize overall sensitivity, or a high precision point, where we maximize the positive predictive value.



Extended Data Fig. 7. **Evaluations of Feature Subsets, Class Imbalance, Code Density, Coding Uncertainty, & Disambiguation from Other Psychiatric Phenotypes.** Panel A illustrates that the pipeline performance where the control group is restricted to children to have at least one psychiatric phenotype other than ASD. It is clear that we have very good discrimination between ASD and non-ASD phenotypes. Panel B illustrates the situation where we restrict the treatment cohort to children to have at least 2 AD diagnostic codes, to see whether the pipeline performance is markedly different in populations where the coding errors/uncertainty is smaller. We see that such restrictions have no appreciable effect on pipeline performance. Panel C illustrates the AUC distributions obtained by using sampled control cohorts that are of the same size as the treatment cohort, to evaluate the effect of class imbalance. Again we see that such restrictions do not appreciably change performance. Panel D explores the performance changes when we use a restricted set of features, or simply use code density as the sole feature. We conclude that the combined feature set used in our optimized pipeline is superior to using the subsets individually. Code density is the least performant feature, and is not stable across databases.

Extended Data Tab. V

BOOSTED SENSITIVITY, SPECIFICITY AND PPV ACHIEVED AT **150 WEEKS** PERSONALIZED OPERATION CONDITIONED ON M-CHAT/F SCORES

M-CHAT/F Outcome				global performance (Truven)			global performance (UCM)			prevalence
0-2 NEG	3-7 NEG	3-7 POS	≥ 8 POS	speci-ficity	sensi-tivity	PPV	speci-ficity	sensi-tivity	PPV	
specificity choices										
0.28	0.66	0.93	0.97	0.95	0.64	0.224	0.95	0.577	0.206	0.022
0.31	0.67	0.9	0.97	0.95	0.641	0.223	0.95	0.577	0.205	0.022
0.54	0.86	0.97	0.99	0.98	0.494	0.361	0.98	0.393	0.31	0.022
0.41	0.89	0.96	0.99	0.98	0.493	0.362	0.98	0.391	0.311	0.022
0.31	0.61	0.86	0.98	0.95	0.808	0.219	0.95	0.713	0.198	0.017
0.33	0.6	0.86	0.98	0.95	0.809	0.218	0.95	0.715	0.197	0.017
0.66	0.95	0.98	0.99	0.98	0.574	0.337	0.98	0.417	0.269	0.017
0.53	0.97	0.98	0.99	0.98	0.573	0.337	0.98	0.412	0.267	0.017
0.54	0.91	0.97	0.99	0.978	0.615	0.322	0.978	0.499	0.278	0.017
0.52	0.92	0.97	0.99	0.978	0.612	0.324	0.978	0.492	0.278	0.017

Extended Data Tab. VI

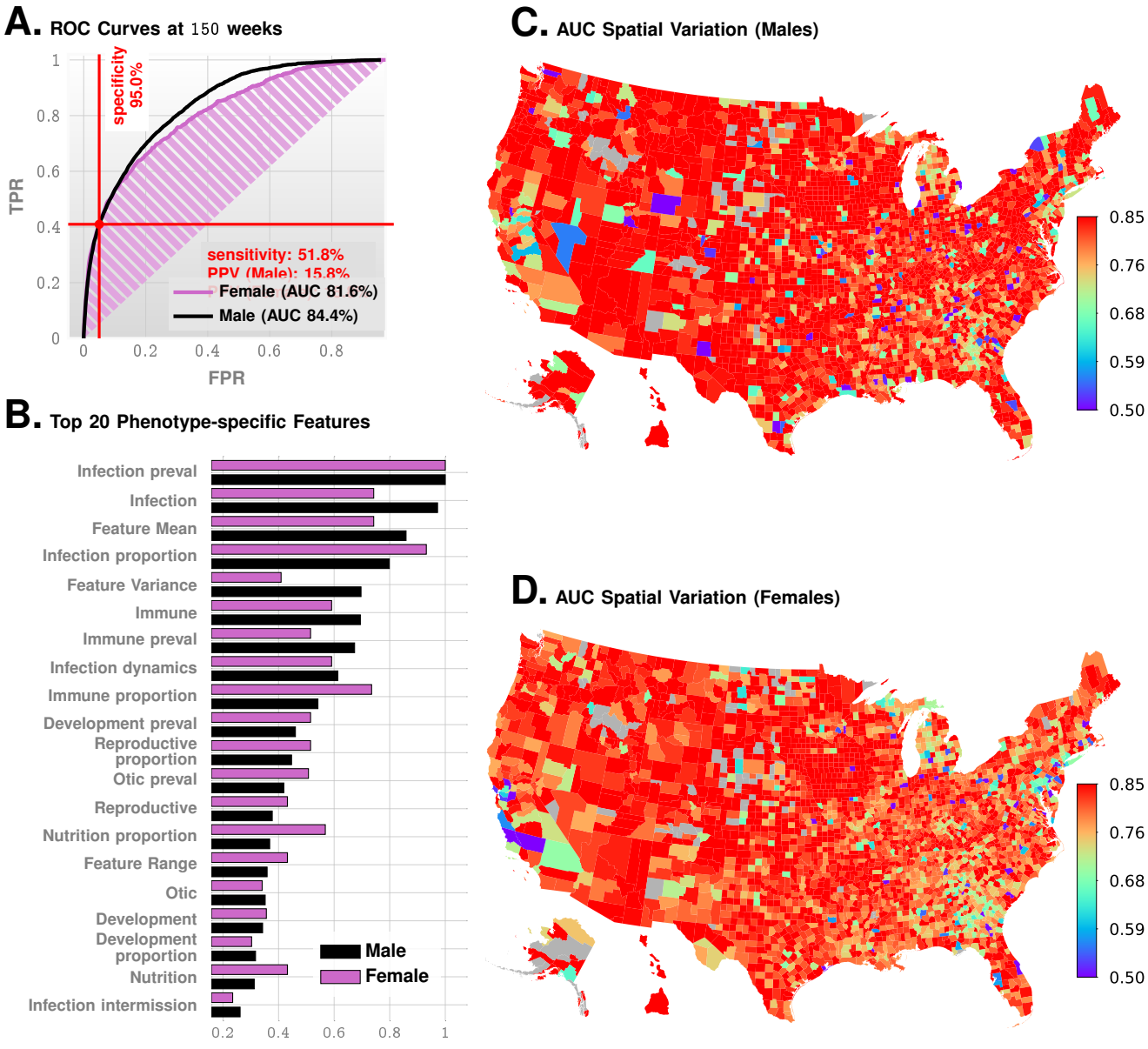
POPULATION STRATIFICATION RESULTS ON LARGE M-CHAT/F STUDY(N=20,375) REPRODUCED FROM GUTHRIE *et al.*[3]

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

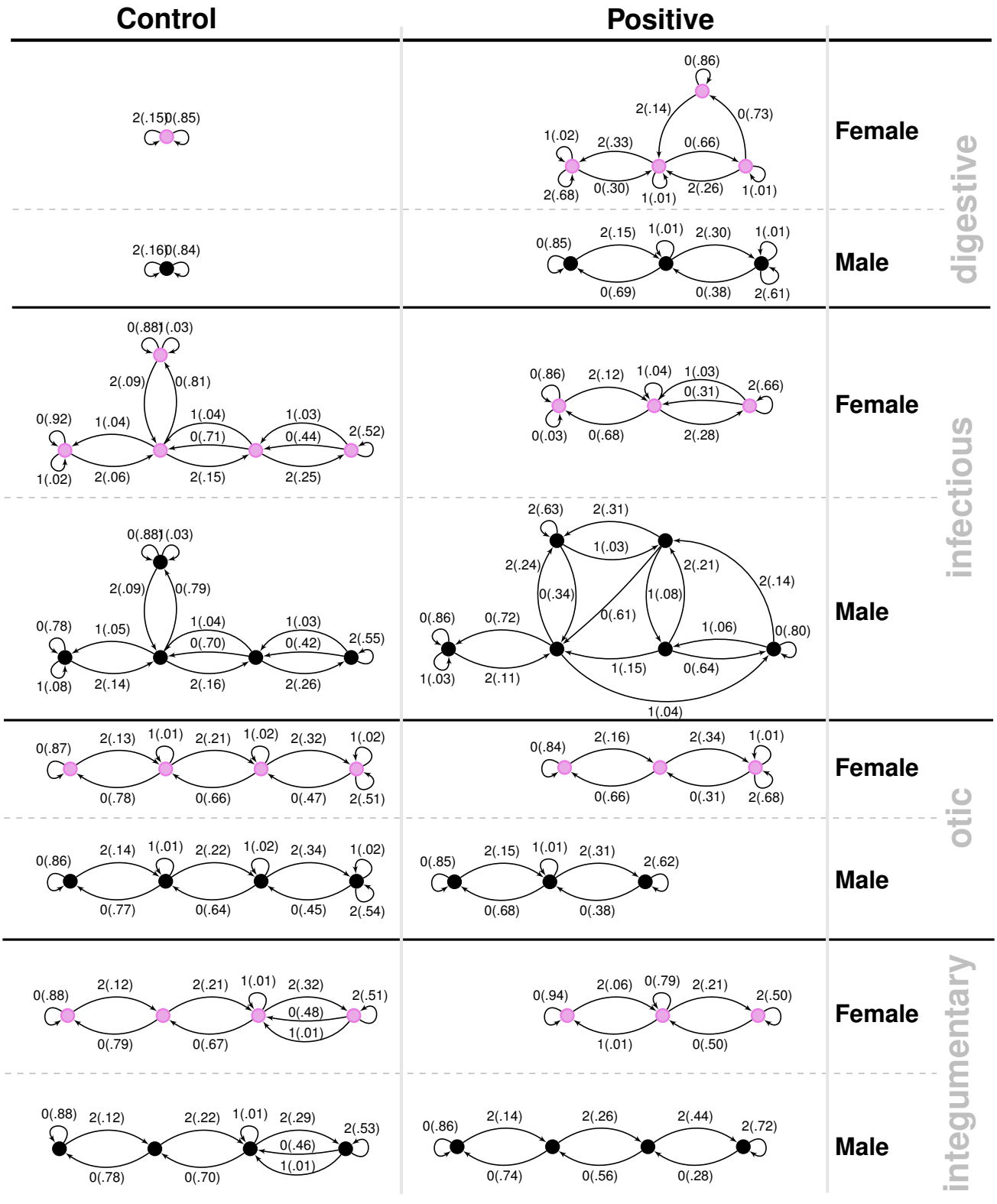
Extended Data Tab. VII

γ, γ' COMPUTED FROM POPULATION STRATIFICATION RECORDED IN M-CHAT/F STUDY[3] ($\rho = 0.0223$)

Id	Sub-population	Test Result	β_i	ρ_i	γ_i	γ'_i
A	M-CHAT/F $\geqq 8$	Positive	.0099	.3469	.1540	.0066
B	M-CHAT/F $\in [3,7]$	Positive (after follow-up)	.0491	.1059	.2331	.0449
C	M-CHAT/F $\in [3,7]$	Negative (after follow-up)	.0324	.0432	.0627	.0317
D	M-CHAT/F $\in [0,2]$	Negative	.9086	.0134	.5471	.9168



Extended Data Fig. 8. Predictive Performance without psychiatric codes (ICD9 290 - 319) and codes for health status and services (ICD9 V0-V91) included. As shown, the performance is comparable at 150 weeks, with the AUC for females marginally lower (compare with Fig. 1 in the main text). The feature importances also are similar, with infectious diseases inferred to have the most importance (or weight) in the pipeline, which is also the case once we add psychiatric phenotypes, and codes for health services in our analysis. As shown in Extended Data Fig. 4A, the psychiatric codes all increase risk, and the vaccination codes (See Extended Data Fig. 4B) all decrease risk when those codes are included. This is why an alternate analysis was carried out to make sure that we are not picking up on psychiatric codes alone. Note in particular that the sensitivity/specificity point highlighted in panel A above is identical after adding the codes. This suggests that our predictive performance arises from patterns learned from co-morbidities, which are not just neuropsychiatric in nature.



Extended Data Fig. 9. Probabilistic Finite State Automata models generated for different disease categories for the control and positive cohorts. We note that in the first cases (digestive disorder), the models get more complex in the positive cohort, suggesting that the disorders become less random. However, in the categories of otic and integumentary disorders, the models become less complex suggesting increased independence from past events of similat nature. In case of infectious diseases, the model gets more complex for males, and less complex for females, suggesting distinct gendered responses associated with high ASD risk.