

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

Author One<sup>a,c,1</sup>, Author Two<sup>b,1,2</sup>, and Author Three<sup>a</sup>

<sup>a</sup>Affiliation One; <sup>b</sup>Affiliation Two; <sup>c</sup>Affiliation Three

This manuscript was compiled on Wednesday 17<sup>th</sup> June, 2020

1 **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**  
2 **with ASD as early as possible (1, 2). Our current incomplete understanding of ASD pathogenesis, and the lack of reliable biomarkers**  
3 **hampers early detection, intervention, and developmental trajectories. In this study we develop and validate machine inferred digital**  
4 **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**  
5 **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**  
6 **80% from shortly after 2 years of age for either sex, and across**  
7 **two independent databases of patient records. Thus, we systematically leverage ASD co-morbidities | with no requirement of additional**  
8 **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**  
9 **approach represents an orthogonal methodology with superior performance. Our methodology compared to screening questionnaires**  
10 **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**  
11 **positives or boost sensitivity by over 50%, while maintaining specificity above 95%. Translated into practice, our algorithmic approach**  
12 **could significantly reduce the median diagnostic age for ASD, and also reduce long post-screen wait-times (4) currently experienced by**  
13 **families for confirmatory diagnoses and access to evidence based interventions.**

Keyword 1 | Keyword 2 | Keyword 3 | ...

1 Autism spectrum disorder is a developmental disability  
2 associated with significant social, and behavioral challenges.  
3 Even though ASD may be diagnosed as early as the age of  
4 two (5), children frequently remain undiagnosed until after  
5 the fourth birthday (6). At this time, there are no laboratory  
6 tests for ASD, so a careful review of behavioral history, and a  
7 direct observation of symptoms is necessary (7, 8) for a clinical  
8 diagnosis. Starting with a positive initial screen, a confirmed  
9 diagnosis of ASD is a multi-step process that often takes 3  
10 months to 1 year, delaying entry into time-critical intervention  
11 programs. While lengthy evaluations (9), cost of care (10),  
12 lack of providers (11), and lack of comfort in diagnosing ASD  
13 by primary care providers (11) are all responsible to varying  
14 degrees (12), one obvious source of this delay is the number of  
15 false positives produced in the initial ASD-specific screening  
16 tools in use today. For example, the M-CHAT/F used com-

monly as a screening tool (8, 13), has an estimated sensitivity of 38.8%, specificity of 94.9% and Positive Predictive Value (PPV) of 14.6% (3). Thus, currently out of every 100 children with ASD, M-CHAT/F flags about 39, and out of every 100 children it flags, about 85 are false positives, exacerbating wait times and queues (12). Automated screening that might be administered with no specialized training, requires no behavioral observations, and is functionally independent of the tools employed in current practice, has the potential for immediate transformative impact on patient care.

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

A screening tool that tracks the risk of an eventual ASD diagnosis, based on the information already being gathered during regular doctor's visits, and which may be implemented as a fully automated background process requiring no time commitment from providers has the potential to reduce avoidable diagnostic delays at no additional burden of time, money and personnel resources. While still lacking the certainty of a

## Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the significance statement is to explain the relevance of the work in broad context to a broad readership. The significance statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any competing interests here.

<sup>1</sup>A.O.(Author One) contributed equally to this work with A.T. (Author Two) (remove if not applicable).

<sup>2</sup>To whom correspondence should be addressed. E-mail: author.two@email.com

diagnostic blood test, use of patterns emergent in the diagnostic history to estimate risk might help reduce the subjective component in questionnaire-based screening tools, resulting in 1) reduced effect of potential language and cultural barriers in diverse populations, and 2) possibly better identify children with milder symptoms (8). Furthermore, being functionally independent of the M-CHAT/F, we show that there is clear advantage to combining the outcomes of the two tools: we can take advantage of any population stratification induced by the M-CHAT/F scores to significantly boost combined screening performance (See Materials & Methods, and Supplementary text, section ??).

## 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. 2 and 3). Our AUC is shown to improve approximately linearly with patient age: Fig. 3A illustrates that the AUC reaches 90% in the Truven dataset at the age of four. Importantly, we train our pipeline on 50% of the Truven dataset, and use held back data from Truven, and the entirety of the UCM dataset for validation: *No new training is done in the UCM dataset.* Good performance on these independent datasets lends strong evidence for our claims. Furthermore, applicability in new datasets *without local re-training* makes it readily deployable in clinical settings.

What are the inferred patterns that elevate risk? Enumerating the top 15 predictive features (Fig. 2B), 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. 3F). Importantly, predictability is relatively agnostic to the number of local cases across US counties (Fig. 2C-D) which is important in light of the current uneven distribution of diagnostic resources (12, 16) 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. 3B), with the risk for the positive cohort rapidly rising. Also, we see a saturation of the risk after  $\approx 3$  years, which corresponds to the median diagnosis age in the database (See 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 sex (Fig. 3D), distinct categories modulate the risk differently, *e.g.*, for a single random patient illustrated in Fig. 3F 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  $\approx 188$  weeks in the Truven dataset, and  $\approx 129$  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 ( $\approx 112$  weeks). Fig. 4A and Table 4 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. 2C, 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) (Table ??), 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 5, Fig. 4B, and SI-Fig. ?? in the supplementary text). Comparing with standalone M-CHAT/F performance (Fig. 4C), 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  $\geq 94\%$ .

## Discussion

In this study, we operationalize a documented aspect of ASD symptomatology in that it has a wide range of co-morbidities (14, 15, 17) occurring at above-average rates (8). Association of ASD with epilepsy (18), gastrointestinal disorders (19–24), mental health disorders (25), insomnia, decreased motor skills (26), allergies including eczema (19–24), immunologic (17, 27–33) and metabolic (23, 34, 35) disorders are widely reported. These studies, along with support from large scale exome sequencing (36, 37), have linked the disorder to putative mechanisms of chronic neuroinflammation, implicating immune dysregulation and microglial activation (29, 32, 38–41) 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 (42).

Attempts at curating genetic biomarkers has also met with limited success. ASD genes exhibit extensive phenotypic variability, with identical variants associated with diverse individual outcomes not limited to ASD, including schizophrenia, intellectual disability, language impairment, epilepsy, neuropsychiatric disorders and, also typical development (43). Additionally, no single gene can be considered “causal” for

more than 1% of cases of idiopathic autism (44).

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

Going beyond screening performance, this approach provides a new tool to uncover clues to ASD pathobiology. Perhaps this vulnerability to diverse immunological, endocrinological and neurological impairments reflects how allostatic loads of medical stress get under the skin and disrupt key regulators of CNS organization and synaptogenesis. Charting individual disorders in the co-morbidity burden further reveals novel associations in normalized prevalence | the odds of experiencing a specific disorder, particularly in the early years (age < 3 years), 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. 5 and SI-Fig. ?? 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. 5A vs B).

Some of the disorders that show up in Fig. 5, 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 (45, 46), and with the prevalence of CP significantly lower (1 in 352 vs 1 in 59 for autism), it follows that only a small number of children (approximately 1.17%) with autism have co-occurring CP. Thus, with significantly higher prevalence in children diagnosed with autism compared to the general population (1.7% vs 0.28%), CP codes show up with higher odds in the true positive set. Also, SI-Fig. ??A 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 (21, 35) as a risk factor.

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

that our pipeline is indeed largely specific to ASD.

We carried out a battery of tests to ensure that our results are not significantly impacted by class imbalance (since our control cohort is orders of magnitude larger) or systematic coding errors (See Methods), *e.g.*, we verified that restricting the positive cohort to children with at least two distinct ASD diagnostic codes in their medical histories instead of one, has little impact on out-of-sample predictive performance (SI-Fig. ??B).

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

As a key limitation to our approach, automated pattern recognition might not reveal true causal precursors. The relatively uncurated nature of the data does not correct for coding mistakes by the clinician and other artifacts, *e.g.* a bias towards over-diagnosis of children on the borderline of the diagnostic criteria due to clinicians' desire to help families access service, and biases arising from changes in diagnostic practices over time (47). Discontinuities in patient medical histories from change in provider-networks can also introduce uncertainties in risk estimates, and socio-economic status of patients which impact access to healthcare might skew patterns in EHR databases. Despite these limitations, the design of a questionnaire-free component to ASD screening that systematically leverages co-morbidities has far-reaching consequences, by potentially slashing the false positives and wait-times, as well as removing systemic under-diagnosis issues amongst females and minorities.

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

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

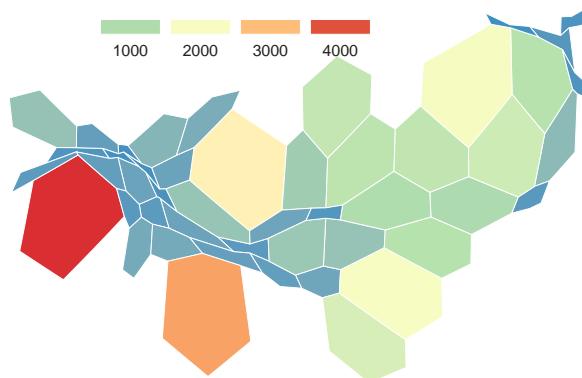
## Materials and Methods

**Source of Electronic Patient Records.** We view the task of predicting ASD diagnosis as a binary classification problem: sequences of diagnostic codes are classified into positive and control categories, where "positive" refers to children eventually diagnosed with ASD, as indicated by the presence of a clinical diagnosis (ICD9 code 299.X) in their medical records. Of the two independent sources of clinical incidence data used in this study, the primary source used to train our predictive pipeline is the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database for the years 2003 to 2012 (50) (referred to as the Truven dataset). This US national database contains data contributed by over 150 insurance carriers, large self-insuring companies, and Medi and is a culmination of over 4.6 billion inpatient and outpatient service

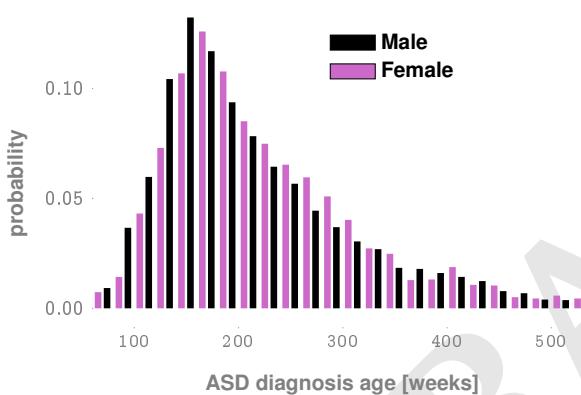
### A. Population-level Prevalence Differences between Positive vs Control Populations



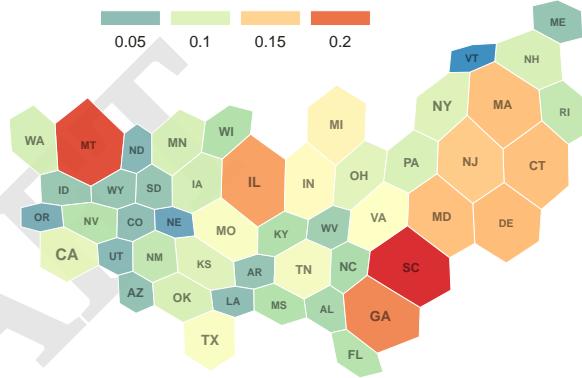
### C. Autism Insurance Claims 2003-2013 (source: Truven Marketscan)



### B. ASD Clinical Diagnosis Age Across Genders



### D. Autism Prevalence in US (Population Normalized)



**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 (48).

claims and almost six billion diagnosis codes. We extracted histories of patients within the age of 0 – 9 years, and excluded patients for whom: 1) At least one code of any available phenotypes is present, 2) Lag between first and last available record for a patient should be at least 15 weeks. These exclusion criteria ensure that we are not considering patients who have too few observations to either train on. Additionally, during validation runs, we restricted the control set to patients observable in the databases to those whose last record is not before the first 150 weeks of life. Characteristics of excluded patients is shown in Table 1. We trained with over 30M diagnostic records (16,649,548 for males and 14,318,303 for females with 9,835 unique codes).

While the Truven database is used for both training and out-of-sample cross-validation with held-back data, our second independent dataset consisting of de-identified diagnostic records for children treated at the University of Chicago Medical Center between the years of 2006 to 2018 (the UCM dataset), aids in further cross-validation. We considered children between the ages of 0 – 5 years, and applied the same exclusion criteria as the Truven dataset. The number of patients used from the two databases is shown in Table 1.

Our datasets are consistent with documented prevalence, with no significant geospatial prevalence variation (Fig. 1D) and reveal that infections and immunological disorders have differential representation in the positive and control groups (Fig. 1C). The median diagnosis age is just over 3 years in the claims database (Fig. 1B) versus 3 years 10 months to 4 years in US (51). Cohort details are

given in Table 1 and discussed in Methods. Importantly, for the positive cohort, we only consider diagnostic history up to the first ASD code.

The significant diversity of diagnostic codes (6,089 distinct ICD-9 codes and 11,522 distinct ICD-10 codes in total in the two datasets), along with the sparsity of codes per sequence and the need to make good predictions as early as possible, makes this a difficult learning problem, where standard deep learning approaches do not suffice (See Table 2). To address these issues, we proceed by partitioning the disease spectrum into 17 broad categories, e.g. infectious diseases, immunologic disorders, endocrinial disorders etc. Each patient is then represented by 17 distinct time series, each tracking an individual disease category. At the population level, these disease-specific sparse stochastic time series are compressed into specialized Markov models (separately for the control and the treatment cohorts) to identify the distinctive patterns pertaining to elevated ASD risk. With these inferred patterns included as features (Table 3) 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).

**Time-series Modeling of Diagnostic History.** Individual diagnostic histories can have long-term memory (52), implying that the order, frequency, and comorbid interactions between diseases are important for assessing the future risk of our target phenotype. We analyze

**Table 1. Patient Counts In De-identified Data & The Fraction of Datasets Excluded By Our Exclusion Criteria\***

Distinct Patients	Truven		UCM	
	Male	Female	Male	Female
ASD Diagnosis Count <sup>†</sup>	12,146	3,018	307	70
Control Count <sup>†</sup>	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)

<sup>†</sup> 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

patient-specific diagnostic code sequences by first 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 models 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) (53). 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.

disease category  $k$ , as follows. For each week  $i$ , we have:

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

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. Each patient is now represented by 17 mapped trinary series.

395  
396  
397  
398  
399  
400  
401  
402  
403  
404

**Step 1: Partitioning The Human Disease Spectrum.** We begin by partitioning the human disease spectrum into 17 non-overlapping categories, as shown in Table 2. Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9) (See Table 2 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) (49) 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. 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 1, we exclude a very small number of patients, and who have very short diagnostic histories with a very small number of codes.

For each patient, the past medical history is a sequence  $(t_1, x_1), \dots, (t_m, x_m)$ , 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

**Step 2: Model Inference & The Sequence Likelihood Defect.** The mapped series, stratified by sex, disease-category, and ASD diagnosis-status are considered to be independent sample paths, and we want to explicitly model these systems as specialized HMMs (PFSA). We model the positive and the control cohorts for each sex, and in each disease category separately, ending up with a total of 68 HMMs at the population level (17 categories, 2 sexes, 2 cohort-types: positive and control, SI-Fig. ?? in the supplementary text 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. Eq. (1)) corresponding to each sex, disease category, and cohort-type (See Section ?? in the Supplementary text for background on PFSA inference).

To reliably infer the cohort-type of a new patient, i.e, the likelihood of a diagnostic sequence being generated by the corresponding cohort model, we generalize the notion of Kullbeck-Leibler (KL) divergence (54, 55) between probability distributions to a divergence  $D_{KL}(G||H)$  between ergodic stationary categorical stochastic processes (56)  $G, H$  as:

$$D_{KL}(G||H) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{x:|x|=n} p_G(x) \log \frac{p_G(x)}{p_H(x)} \quad [2]$$

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.

405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419

**Table 2. 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 <sup>†</sup>	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 musculoskeletal)	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

<sup>†</sup> 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 (49).

Defining the log-likelihood of  $x$  being generated by a process  $G$  as :

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

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):

$$\lim_{|x| \rightarrow \infty} L(x, G) = \mathcal{H}(G) \quad [4a]$$

$$\lim_{|x| \rightarrow \infty} L(x, H) = \mathcal{H}(G) + \mathcal{D}_{KL}(G||H) \quad [4b]$$

where  $\mathcal{H}(\cdot)$  is the entropy rate of a process (54). Importantly, Eq. 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  $G_+^j, G_0^j$  respectively, we can compute the *sequence likelihood*

defect (SLD,  $\Delta^j$ ) as:

$$\Delta^j \triangleq L(G_0^j, x) - L(G_+^j, x) \rightarrow \mathcal{D}_{KL}(G_0^j || G_+^j) \quad [5]$$

With the inferred PFSA models and the individual diagnostic history, we estimate the SLD measure on the right-hand side of Eqn. Eq. (5). The higher this likelihood defect, the higher the similarity of diagnosis history to that of children with autism.

**Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules.** The risk estimation pipeline operates on patient specific information limited to the sex 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 used to train a gradient-boosting classifier (57). The complete list of 165 features used is provided in Table 3.

We need two training sets: one to infer the models, and one to train the classifier with features derived from the inferred models. Thus, we do 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

**Table 3. Engineered Features (Total Count: 165)**

Feature Type <sup>‡</sup>	Description	No. of Features
[Disease Category] $\Delta$	Likelihood Defect (See Methods section)	17
[Disease Category] $o$	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] $o$ 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

<sup>‡</sup> Disease categories are described in Table 2

**Table 4. 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 ( $\approx$ 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

441 PFSA models (*classical supervised learning*), and we finally execute  
442 out-of-sample validation on the test set. Fig. 2B shows the top 15  
443 features ranked in order of their relative importance (relative loss  
444 in performance when dropped out of the analysis).

445 **Calculating Relative Risk.** Our pipeline maps medical histories to a  
446 raw indicator of risk. However, to make crisp predictions, we must  
447 choose a decision threshold for this raw score. Conceptually identi-  
448 cal to the notion of Type 1 and Type 2 errors in classical statistical  
449 analyses, the choice of a threshold trades off false positives (Type 1  
450 error) for false negatives (Type 2 error): choosing a small threshold  
451 results in predicting a larger fraction of future diagnoses correctly,  
452 *i.e.* have a high true positive rate (TPR), while simultaneously  
453 suffering from a higher false positive rate (FPR), and vice versa.  
454 Therefore, a choice of a specific decision threshold reflects a choice  
455 of the maximum FPR and minimum TPR, and is driven by the  
456 application at hand. In this study, we base our analysis on maxi-  
457 mizing the  $F_1$ -score, defined as the harmonic mean of sensitivity  
458 and specificity, to make a balanced trade-off between the two kinds  
459 of errors (See Supplementary text, Section ??). The *relative risk* is  
460 then defined as the ratio of the raw risk to the decision threshold,  
461 and a value  $> 1$  predicts a future ASD diagnosis.

462 **Boosting Performance Via Leveraging Population Stratification In-  
463 duced By Existing Tests.** We leverage the population stratification  
464 induced by an existing independent screening test (M-CHAT/F)  
465 to improve combined performance. Here a combination refers to  
466 the conditional choice of the sensitivity/specifity trade-offs for our  
467 tool in each sub-population such that the overall performance is  
468 optimized with respect to whether we wish to maximize the PPV  
469 or the sensitivity at a specified minimum level of specificity. As-  
470 sume that there are  $m$  sub-populations such that: the sensitivities,  
471

472 specificities achieved, and the prevalences in each sub-population  
473 are given by  $s_i, c_i$  and  $\rho_i$  respectively, with  $i \in \{1, \dots, m\}$ . Let  $\beta_i$   
474 be the relative size of each sub-population. Then, we have (See  
475 Supplementary text, Section ??):

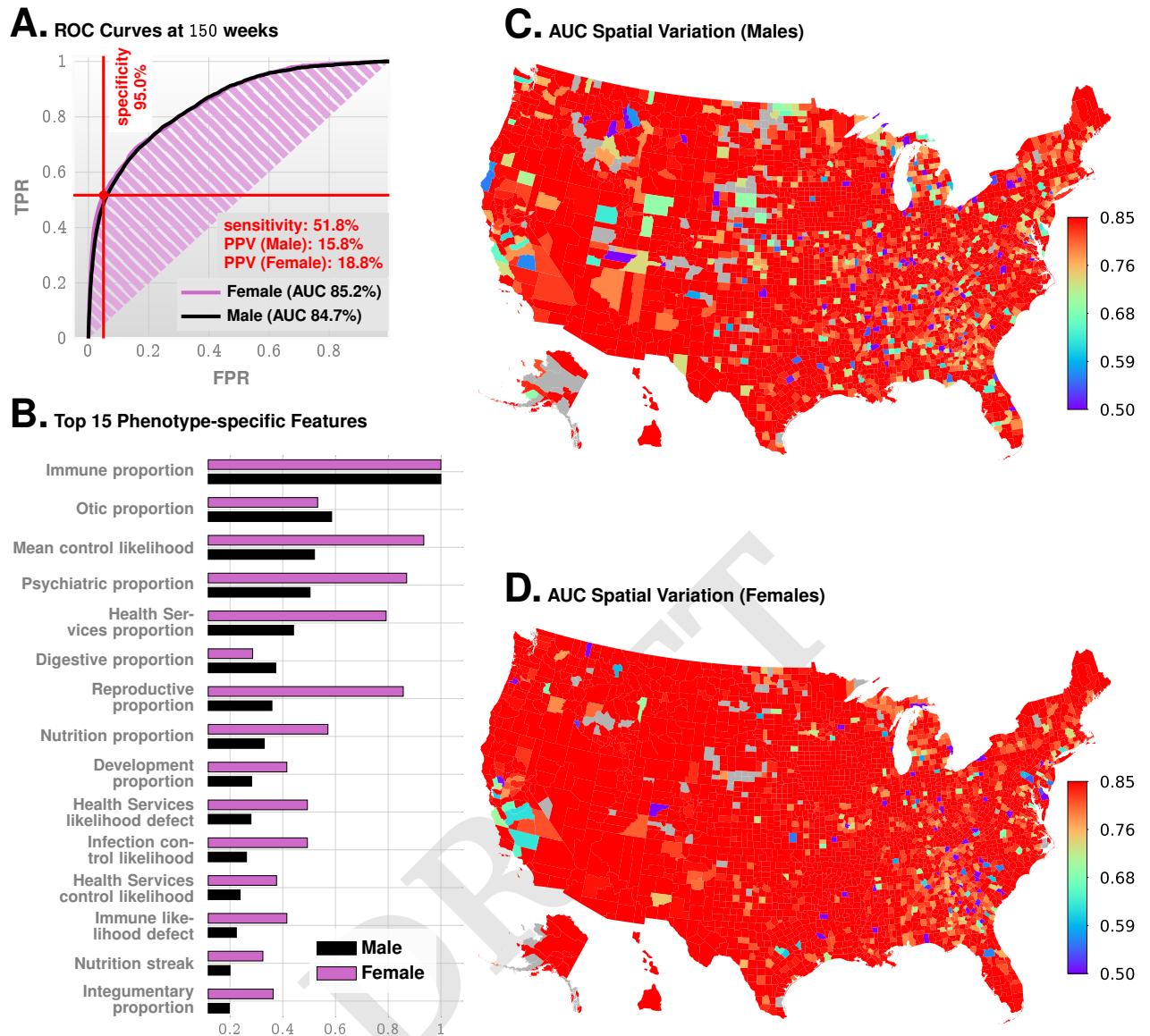
$$s = \sum_{i=1}^m s_i \gamma_i \quad [6a]$$

$$c = \sum_{i=1}^m c_i \gamma'_i \quad [6b]$$

476 where we have denoted:

$$\gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma'_i = \beta_i \frac{1 - \rho_i}{1 - \rho} \quad [6c]$$

477 and  $s, c, \rho$  are the overall sensitivity, specificity, and prevalence.  
478 Knowing the values of  $\gamma_i, \gamma'_i$ , we can carry out an  $m$ -dimensional  
479 search to identify the feasible choices of  $s_i, c_i$  pairs for each  $i$ , such  
480 that some global constraint is satisfied, *e.g.* minimum values of  
481 specificity, sensitivity, and PPV. We consider 4 sub-populations  
482 defined by M-CHAT/F score brackets (3), and if the screen result  
483 is considered a positive (high risk, indicating the need for a full  
484 diagnostic evaluation) or a negative, *i.e.*, low risk: 1) score  $\leq 2$   
485 screening ASD negative, 2) score [3 – 7] screening ASD negative  
486 on follow-up, 3) score [3 – 7] and screening ASD positive on follow-  
487 up, and 4) score  $\geq 8$ , screening ASD positive. (See Table ??).  
488 The “follow-up” in the context of M-CHAT/F refers to the re-  
489

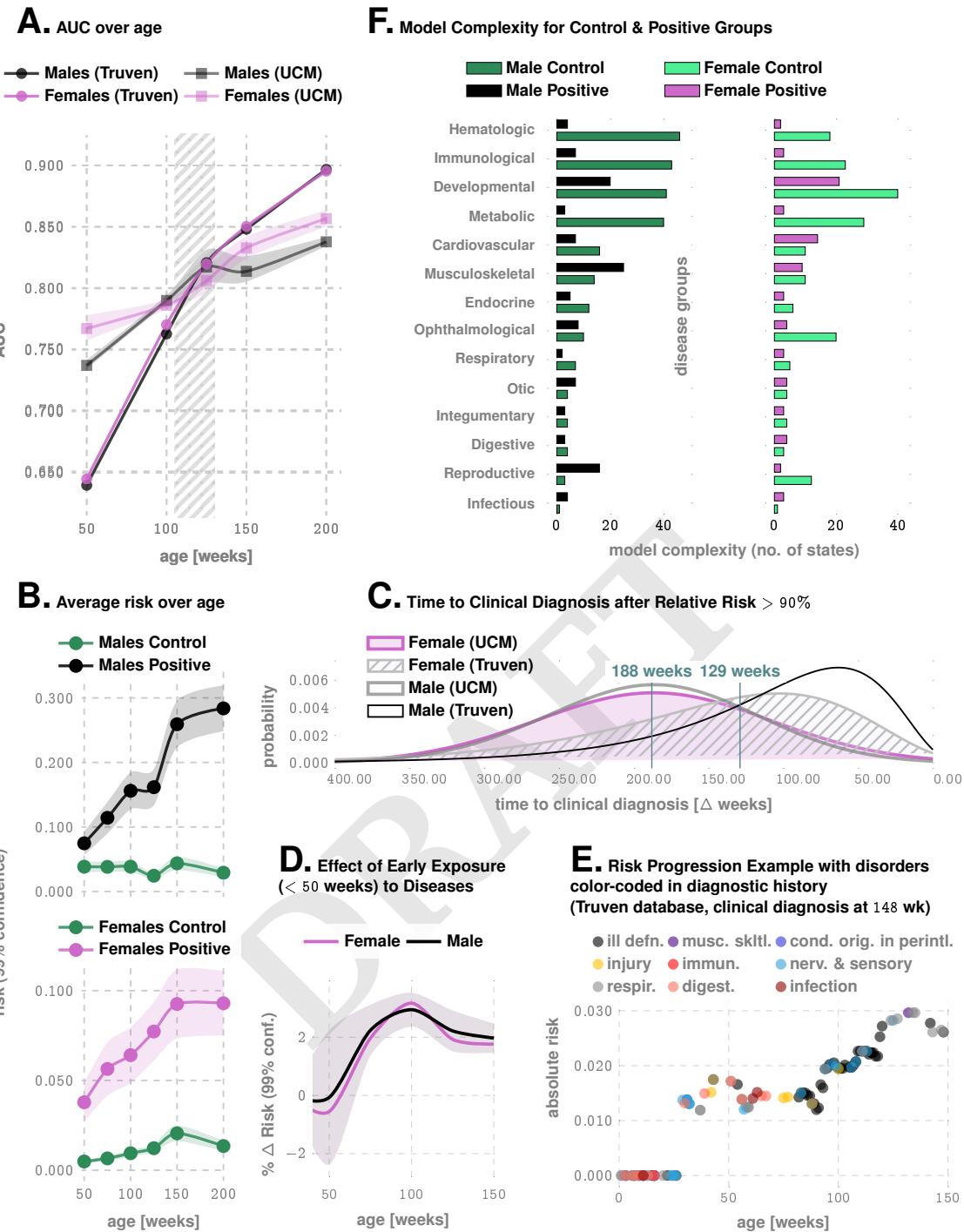


**Fig. 2.** Predictive Performance. Panel A shows the ROC curves for males and females. Panel B shows the feature importance inferred by our prediction pipeline. The detailed description of the features is given in Table 2. The most import feature is related to immunologic disorders, and we note that 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.

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

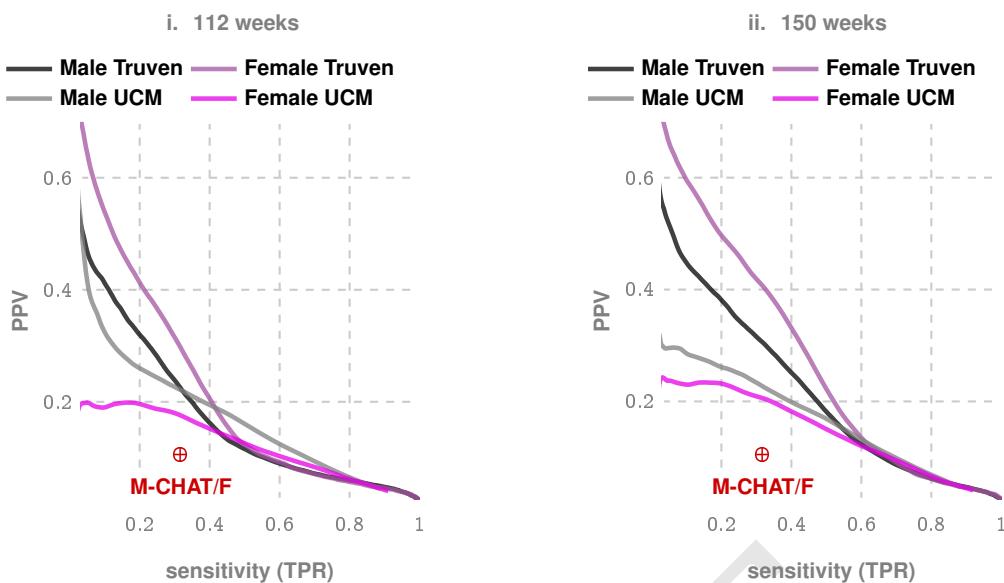
would outperform M-CHAT/F in universal screening. The results are shown in Fig. 4B.

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

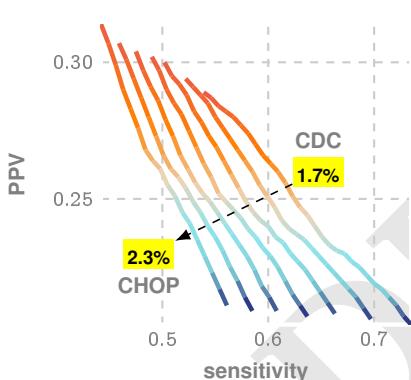


**Fig. 3. More details on Predictive Performance and Variation of Inferred Risk.** Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve > 80% AUC for either sex from shortly after 2 years. Panel B illustrates how the average risk changes with time for the control and the positive cohorts. Panel C shows the distribution of the prediction horizon: the time to a clinical diagnosis after inferred relative risk crosses 90%. Panel D shows that for each new disease code for a low-risk child, ASD risk increases by approximately 2% for either sex. Panel E illustrates the risk progression of a specific, ultimately autistic male child in the Truven database. Abbreviations in the legend: ill defn. (Symptoms, Signs, And Ill-Defined Conditions), musc. sktl. (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.

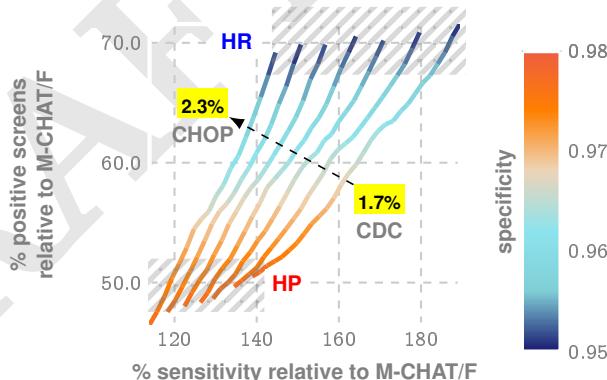
## A. Standalone PPV vs Sensitivity or Precision Recall Curves



## B. M-CHAT/F Conditioned PPV vs Sensitivity (Prevalence range 1.7% to 2.3%)



## C. Reduced # of Flags vs Boosted Sensitivity Relative To Standalone M-CHAT/F



**Fig. 4. 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.

- 498 1. Data & statistics on autism spectrum disorder | cdc (2019).  
499 2. L Gilotty, Early screening for autism spectrum (2019).  
500 3. W Guthrie, et al., Accuracy of Autism Screening in a Large Pediatric Network. *Pediatrics* **144**  
501 (2019).  
502 4. E Gordon-Lipkin, J Foster, G Peacock, Whittling Down the Wait Time: Exploring Models to  
503 Minimize the Delay from Initial Concern to Diagnosis and Treatment of Autism Spectrum  
504 Disorder. *Pediatr. Clin. North Am.* **63**, 851–859 (2016).  
505 5. Data & statistics on autism spectrum disorder | cdc (2019).  
506 6. LA Schieve, et al., Population attributable fractions for three perinatal risk factors for autism  
507 spectrum disorders, 2002 and 2008 autism and developmental disabilities monitoring network.  
508 *Ann Epidemiol* **24**, 260–266 (2014).  
509 7. F Volkmar, et al., Practice parameter for the assessment and treatment of children and adolescents  
510 with autism spectrum disorder. *J. Am. Acad. Child & Adolesc. Psychiatry* **53**, 237–257  
511 (2014).

- 512 8. SL Hyman, SE Levy, SM Myers, , et al., Identification, evaluation, and management of chil-  
513 dren with autism spectrum disorder. *Pediatrics* **145** (2020).  
514 9. LG Kalb, et al., Determinants of appointment absenteeism at an outpatient pediatric autism  
515 clinic. *J. Dev. & Behav. Pediatr.* **33**, 685–697 (2012).  
516 10. J Bisgaier, D Levinson, DB Cutts, KV Rhodes, Access to autism evaluation appointments  
517 with developmental-behavioral and neurodevelopmental subspecialists. *Arch. pediatrics &*  
518 *adolescent medicine* **165**, 673–674 (2011).  
519 11. TS Feniklié, K Ellerbeck, MK Filippi, CM Daley, Barriers to autism screening in family medicine  
520 practice: a qualitative study. *Prim. health care research & development* **16**, 356–366 (2015).  
521 12. E Gordon-Lipkin, J Foster, G Peacock, Whittling down the wait time: exploring models to min-  
522 imize the delay from initial concern to diagnosis and treatment of autism spectrum disorder.  
523 *Pediatr. Clin. North Am.* **63**, 851–859 (2016).  
524 13. DL Robins, et al., Validation of the modified checklist for autism in toddlers, revised with  
525 follow-up (m-chat-r/f). *Pediatrics* **133**, 37–45 (2014).  
526

**Table 5. 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.

- 526 14. C Tye, AK Runicles, AJO Whitehouse, GA Alvares, Characterizing the Interplay Between  
527 Autism Spectrum Disorder and Comorbid Medical Conditions: An Integrative Review. *Front  
528 Psychiatry* **9**, 751 (2018).
- 529 15. IS Kohane, et al., The co-morbidity burden of children and young adults with autism spectrum  
530 disorders. *PLoS ONE* **7**, e33224 (2012).
- 531 16. LA Althouse, JA Stockman, Pediatric workforce: A look at pediatric nephrology data from the  
532 american board of pediatrics. *The J. pediatrics* **148**, 575–576 (2006).
- 533 17. O Zerbo, et al., Immune mediated conditions in autism spectrum disorders. *Brain Behav.  
534 Immun.* **46**, 232–236 (2015).
- 535 18. H Won, W Mah, E Kim, Autism spectrum disorder causes, mechanisms, and treatments:  
536 focus on neuronal synapses. *Front Mol Neurosci* **6**, 19 (2013).
- 537 19. G Xu, et al., Association of Food Allergy and Other Allergic Conditions With Autism Spectrum  
538 Disorder in Children. *JAMA Netw Open* **1**, e180279 (2018).
- 539 20. JB Adams, et al., Nutritional and metabolic status of children with autism vs. neurotypical  
540 children, and the association with autism severity. *Nutr Metab (Lond)* **8**, 34 (2011).
- 541 21. A Fattorusso, L Di Genova, GB Dell'Isola, E Mencaroni, S Esposito, Autism Spectrum Disorders  
542 and the Gut Microbiota. *Nutrients* **11** (2019).
- 543 22. R Diaz Heijtz, et al., Normal gut microbiota modulates brain development and behavior. *Proc.  
544 Natl. Acad. Sci. U.S.A.* **108**, 3047–3052 (2011).
- 545 23. S Rose, et al., Mitochondrial dysfunction in the gastrointestinal mucosa of children with  
546 autism: A blinded case-control study. *PLoS ONE* **12**, e0186377 (2017).
- 547 24. EM Sajdel-Sulkowska, et al., Common Genetic Variants Link the Abnormalities in the Gut-  
548 Brain Axis in Prematurity and Autism. *Cerebellum* **18**, 255–265 (2019).
- 549 25. MS Kayser, J Dalmat, Anti-NMDA Receptor Encephalitis in Psychiatry. *Curr Psychiatry Rev*  
550 **7**, 189–193 (2011).
- 551 26. OI Dadalko, BG Travers, Evidence for Brainstem Contributions to Autism Spectrum Disorders.  
552 *Front Integr Neurosci* **12**, 47 (2018).
- 553 27. Y Yamashita, et al., Anti-inflammatory Effect of Ghrelin in Lymphoblastoid Cell Lines From  
554 Children With Autism Spectrum Disorder. *Front Psychiatry* **10**, 152 (2019).
- 555 28. L Shen, et al., Proteomics Study of Peripheral Blood Mononuclear Cells (PBMCs) in Autistic  
556 Children. *Front Cell Neurosci* **13**, 105 (2019).
- 557 29. K Ohja, et al., Neuroimmunologic and Neurotrophic Interactions in Autism Spectrum Disorders:  
558 Relationship to Neuroinflammation. *Neuromolecular Med.* **20**, 161–173 (2018).
- 559 30. D Gadyz, A Krzywdziska, KK Hozyasz, Immune Abnormalities in Autism Spectrum Disorder-  
560 Could They Hold Promise for Causative Treatment? *Mol. Neurobiol.* **55**, 6387–6435 (2018).
- 561 31. TC Theoharides, I Tsilioni, AB Patel, R Doyle, Atopic diseases and inflammation of the brain  
562 in the pathogenesis of autism spectrum disorders. *Transl Psychiatry* **6**, e844 (2016).
- 563 32. AM Young, et al., From molecules to neural morphology: understanding neuroinflammation  
564 in autism spectrum condition. *Mol Autism* **7**, 9 (2016).
- 565 33. LA Croen, et al., Family history of immune conditions and autism spectrum and developmen-  
566 tal disorders: Findings from the study to explore early development. *Autism Res* **12**, 123–135  
567 (2019).
- 568 34. T Vargason, DL McGuinness, J Hahn, Gastrointestinal Symptoms and Oral Antibiotic Use in  
569 Children with Autism Spectrum Disorder: Retrospective Analysis of a Privately Insured U.S.  
570 Population. *J Autism Dev Disord* **49**, 647–659 (2019).
- 571 35. M Fiorentino, et al., Blood-brain barrier and intestinal epithelial barrier alterations in autism  
572 spectrum disorders. *Mol Autism* **7**, 49 (2016).
- 573 36. FK Satterstrom, et al., Large-scale exome sequencing study implicates both developmental  
574 and functional changes in the neurobiology of autism. *bioRxiv* (2019).
- 575 37. T Gaugler, et al., Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**,  
576 881–885 (2014).
- 577 38. DL Vargas, C Nascimbene, C Krishnan, AW Zimmerman, CA Pardo, Neuroglial activation  
578 and neuroinflammation in the brain of patients with autism. *Ann. Neurol.* **57**, 67–81 (2005).
- 579 39. H Wei, et al., IL-6 is increased in the cerebellum of autistic brain and alters neural cell adhe-  
580 sion, migration and synaptic formation. *J Neuroinflammation* **8**, 52 (2011).
- 581 40. AM Young, E Campbell, S Lynch, J Suckling, SJ Powis, Aberrant NF-kappaB expression  
582 in autism spectrum condition: a mechanism for neuroinflammation. *Front Psychiatry* **2**, 27  
583 (2011).

### A. Male (3 YR)

782.0	Skin sensation disturb
379.54	Nystagms w vestibulr dis
784.61	Alexia and dyslexia
366.50	After-cataract NOS
458.0	Orthostatic hypotension
743.32	Cortical/zonular catarac
906.5	Late eff head/neck burn
779.9	Perinatal condition NOS
345.60	Inf spasm w/o intr epile
779.8	Neonatal bradycardia
783.4	Lack norm physio dev NOS
264.9	Vitamin A deficiency NOS
874.3	Open wound thyroid-compl
722.6	Disc degeneration NOS
378.63	Mech strab w oth conditn
387.2	Cochlear otosclerosis
756.4	Chondrodytrophy
784.60	Symbolic dysfunction NOS
345.81	Epilepsy NEC w intr epile
345.61	Inf spasm w intract epile
078.5	Cytomegaloviral disease
794.14	Abn oculomotor studies
715.95	Osteoarthros NOS-pelvis
374.03	Spastic entropion
337.9	Autonomic nerve dis NEC
343.4	Infantile hemiplegia
362.76	Vitelliform dystrophy
191.7	Mal neo brain stem
344.2	Diplegia of upper limbs
696.0	Psoriatic arthropathy
271.9	Dis carbohydr metab NOS
230.3	Ca in situ colon
556.1	Ulcerative ileocolitis
227.3	Benign neo pituitary
345.71	Epil par cont w intr epi
727.68	Rupture tendon foot NEC
985.9	Toxic effect metal NOS
364.02	Recurrent iridocyclitis
531.40	Chr stomach ulc w hem
346.80	Othr migrne wo ntrc mgnr

4.00 5.00 6.00 7.00

log odds ratio of normalized prevalence

### B. Female (3 YR)

813.00	Fx upper forearm NOS-cl
345.60	Inf spasm w/o intr epile
333.1	Tremor NEC
349.9	Cns dis. NOS
794.4	Abn kidney funct study
369.9	Visual loss NOS
881.02	Open wound of wrist
728.4	Laxity of ligament
358.8	Myoneural dis. NEC
794.02	Abn electroencephalogram
758.0	Down's syndrome
212.7	Benign neoplasm heart
783.40	Lack norm physio dev NOS
112.5	Disseminated candidiasis
781.3	Lack of coordination
742.2	Reduction deform, brain
759.89	Specified cong anomal NEC
333.6	Genetic torsion dystonia
388.00	Degen/vascul dis ear NOS
348.9	Brain condition NOS
580.9	Acute nephritis NOS
836.3	Dislocat patella-closed
568.9	Peritoneal dis. NOS
723.9	Neck dis./symp NOS
426.7	Anomalous av excitation
136.3	Pneumocystosis
343.1	Congenital hemiplegia
362.70	Hered retin dysrphy NOS
458.0	Orthostatic hypotension
529.5	Plicated tongue
915.3	Blister finger-infected
985.9	Toxic effect metal NOS
345.81	Epilepsy NEC w intr epile
345.80	Epilep NEC w/o intr epile
759.5	Tuberous sclerosis
379.54	Nystagms w vestibulr dis
363.33	Posterior pole scar NEC
794.11	Abn retinal funct study
342.80	Ot sp hmiplga unspf side

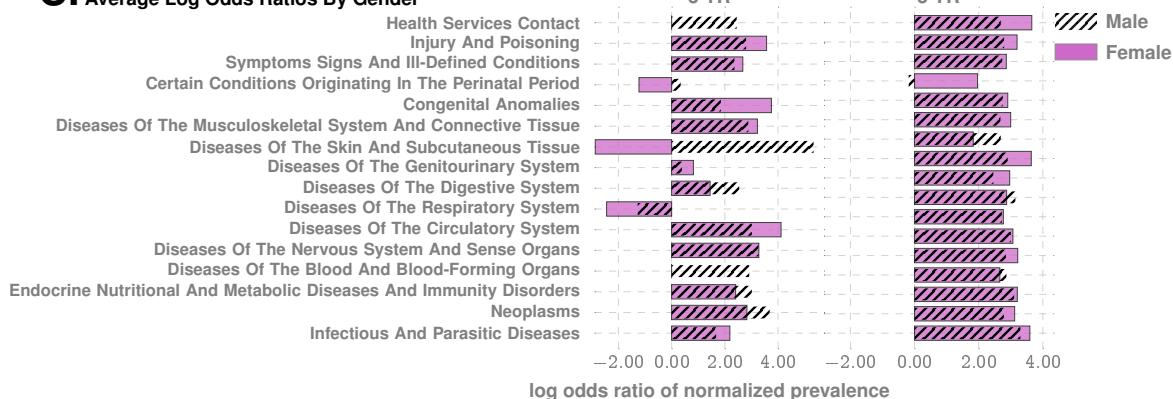
3.00 4.00 5.00 6.00 7.00

log odds ratio of normalized prevalence

### ICD9 Class



### C. Average Log Odds Ratios By Gender



log odds ratio of normalized prevalence

**Fig. 5. 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.