

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

Dmytro Onishchenko<sup>1</sup>, Yi Huang<sup>1</sup>, James van Horne<sup>1</sup>, Peter J. Smith<sup>4,7</sup>, Michael M. Msall<sup>5,6</sup> and Ishanu Chattopadhyay<sup>1,2,3\*</sup>

<sup>1</sup>Department of Medicine, University of Chicago, Chicago, IL USA

<sup>2</sup>Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA

<sup>3</sup>Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA

<sup>4</sup>Department of Pediatrics, Section of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL USA

<sup>5</sup>Department of Pediatrics, Section Chief of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL USA

<sup>6</sup>Joseph P. Kennedy Research Center on Intellectual and Neurodevelopmental Disabilities, University of Chicago, Chicago, IL USA

<sup>7</sup>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 and behavioral challenges. There is a need for tools that help identify children with ASD as early as possible (1, 2). Our current incomplete understanding of ASD pathogenesis, and the lack of reliable biomarkers hampers early detection, intervention, and developmental trajectories. In this study we develop and validate machine inferred digital biomarkers for autism using individual diagnostic codes already recorded during medical encounters. Our risk estimator identifies children at high risk with a corresponding area under the receiver operating characteristic curve (AUC) exceeding 80% from shortly after two years of age for either sex, and across two independent databases of patient records. Thus, we systematically leverage ASD co-morbidities - with no requirement of additional blood work, tests or procedures - to compute the Autism Co-morbid Risk Score (ACoR) which predicts elevated risk during the earliest childhood years, when interventions are the most effective. By itself, ACoR has superior performance to common questionnaires-based screenings such as the M-CHAT/F (3), and has the potential to reduce socio-economic, ethnic and demographic biases. In addition to standalone performance, independence from questionnaire based screening allows us to further boost performance by conditioning on the individual M-CHAT/F scores – we can either halve the false positive rate of current screening protocols or boost sensitivity to over 60%, while maintaining specificity above 95%. Adopted in practice, ACoR could significantly reduce the median diagnostic age for ASD, and reduce long post-screen wait-times (4) experienced by families for confirmatory diagnoses and access to evidence based interventions.

## INTRODUCTION

Autism spectrum disorder is a developmental disability associated with significant social and behavioral challenges. Even though ASD may be diagnosed as early as the age of two (5), children frequently remain undiagnosed until after the fourth birthday (6). With genetic and metabolomic tests (7–10) still in their infancy, a careful review of behavioral history and a direct observation of symptoms is currently necessary (11, 12) for a clinical diagnosis. Starting with a positive initial screen, a confirmed diagnosis of ASD is a multi-step process that often takes 3 months to 1 year, delaying entry into time-critical intervention programs. While lengthy evaluations (13), cost of care (14), lack of providers (15), and lack of comfort in diagnosing ASD by primary care providers (15) are all responsible to varying degrees (16), one obvious source of this delay is the number of false positives produced in the initial ASD-specific screening tools in use today. For example, a large-scale study of the M-CHAT/F (3) ( $n=20,375$ ), which is used commonly as a screening tool (12, 17), is demonstrated to have an estimated sensitivity of 38.8%, specificity of 94.9% and Positive Predictive Value (PPV) of 14.6%. Thus, currently out of every 100 children with ASD, M-CHAT/F flags about 39, and out of every 100 children it flags, about 85 are false positives, exacerbating wait times and queues (16). Automated screening that might be administered with no specialized training, requires no behavioral observations, and is functionally independent of the tools employed in current practice, has the potential for immediate transformative impact on patient care.

While the neurobiological basis of autism remains poorly understood, a detailed assessment conducted by the US Centers for Disease Control and Prevention (CDC) demonstrated that children with ASD experience

higher than expected rates of many diseases (5). These include conditions related to dysregulation of immune pathways such as eczema, allergies, asthma, as well as ear and respiratory infections, gastrointestinal problems, developmental issues, severe headaches, migraines, and seizures (18, 19). In the present study, we exploit these co-morbidities to estimate the risk of childhood neuropsychiatric disorders on the autism spectrum. We refer to the risk estimated by our approach as the Autism Co-morbid Risk Score (ACoR). Using only sequences of diagnostic codes from past doctor's visits, our risk estimator reliably predicts an eventual clinical diagnosis – or the lack thereof – for individual patients. Thus, the key clinical contribution of this study is the formalization of subtle co-morbidity patterns as a reliable screening tool, and potentially improve wait-times for diagnostic evaluations by significantly reducing the number of false positives encountered in initial screens in current practice.

A screening tool that tracks the risk of an eventual ASD diagnosis, based on the information already being gathered during regular doctor's visits, and which may be implemented as a fully automated background process requiring no time commitment from providers has the potential to reduce avoidable diagnostic delays at no additional burden of time, money and personnel resources. Use of patterns emergent in the diagnostic history to estimate risk might help reduce the subjective component in questionnaire-based screening tools, resulting in 1) reduced effect of potential language and cultural barriers in diverse populations, and 2) possibly better identification of children with milder symptoms (12). Furthermore, being functionally independent of the M-CHAT/F, we show that there is clear advantage to combining the outcomes of the two tools: we can take advantage of any population stratification induced by the M-CHAT/F scores to significantly boost combined screening performance (See Materials & Methods, and Supplementary text, section VIII).

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

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

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

## MATERIALS & METHODS

We view the task of predicting ASD diagnoses as a binary classification problem: sequences of diagnostic codes are classified into positive and control categories, where “positive” refers to children eventually diagnosed with ASD, as indicated by the presence of a clinical diagnosis (ICD9 code 299.X) in their medical records. Of the two independent sources of clinical incidence data used in this study, the primary source used to train our predictive pipeline is the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database for the years 2003 to 2012 (30) (referred to as the Truven dataset). This US national database merges data contributed by over 150 insurance carriers and large self-insurance companies, and comprises over 4.6 billion inpatient and outpatient service claims and almost six billion diagnosis codes. We extracted histories of patients within the age of 0 – 6 years, and excluded patients for whom one or more of the following criteria fails: 1) At least one code pertaining to one of the 17 disease categories we use (See later for discussion of disease categories) is present in the diagnostic history, 2) The first and last available record for a patient are at least 15 weeks apart. These exclusion criteria ensure that we are not considering patients who have too few observations. Additionally, during validation runs, we restricted the control set to patients observable in the databases to those whose last record is not before the first 200 weeks of life. The characteristics of excluded patients is shown in Table Ia. We trained with over 30M diagnostic records (16,649,548 for males and 14,318,303 for females with 9,835 unique codes).

While the Truven database is used for both training and out-of-sample cross-validation with held-back data, our second independent dataset comprising de-identified diagnostic records for children treated at the University of Chicago Medical Center between the years of 2006 to 2018 (the UCM dataset) aids in further cross-validation. We considered children between the ages of 0 – 6 years, and applied the same exclusion criteria as the Truven dataset. The number of patients used from the two databases is shown in Table Ia. Our datasets are consistent with documented ASD prevalence and median diagnostic age (3 years in the claims database versus 3 years 10 months to 4 years in the US (31) with no significant geospatial prevalence variation (See SI-Fig. 1).

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. We proceed by partitioning the disease spectrum into 17 broad categories, *e.g.* infectious diseases, immunologic disorders, endocrinological disorders etc. Each patient is

TABLE I: Patient Numbers, Inclusion-exclusion Criteria and Features Used In Analysis

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

Distinct Patients	Truven		UCM	
	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.

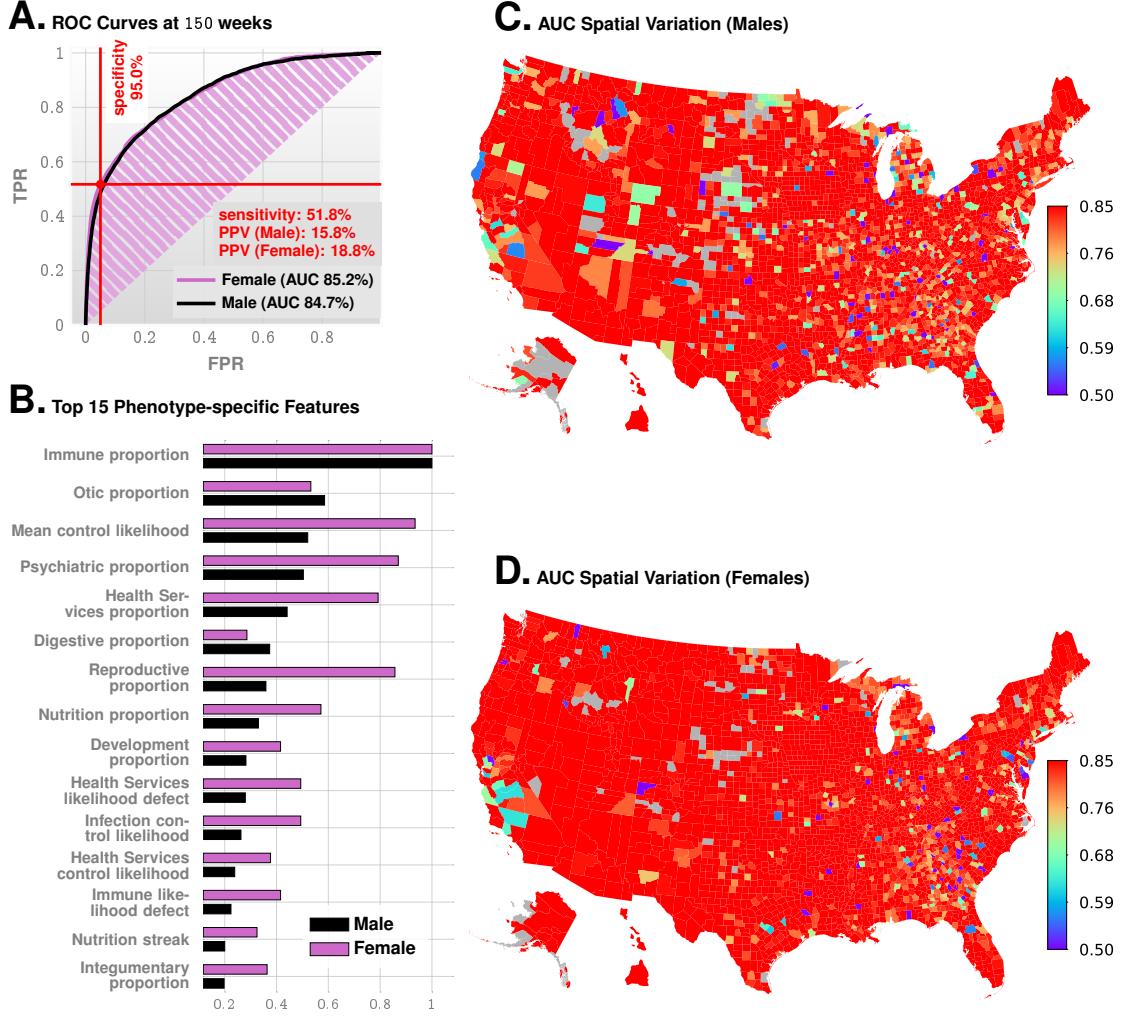
\* Dataset sizes are after the exclusion criteria are applied

(b) 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 SI-Table I in Supplementary Text.

then represented by 17 distinct time series, each tracking an individual disease category. At the population level, these disease-specific sparse stochastic time series are compressed into specialized Markov models (separately for the control and the treatment cohorts) to identify the distinctive patterns pertaining to elevated ASD risk. Each of these inferred models is a Probabilistic Finite State Automaton (PFSA) (32). See Supplementary Text Section X for details on PFSA inference (33, 34). Importantly, the number of states and connectivities in the PFSA models are explicitly inferred from data; with the algorithms generating higher resolution models when needed, and opting for a lower resolution otherwise automatically. This results in a framework with fewer nominal

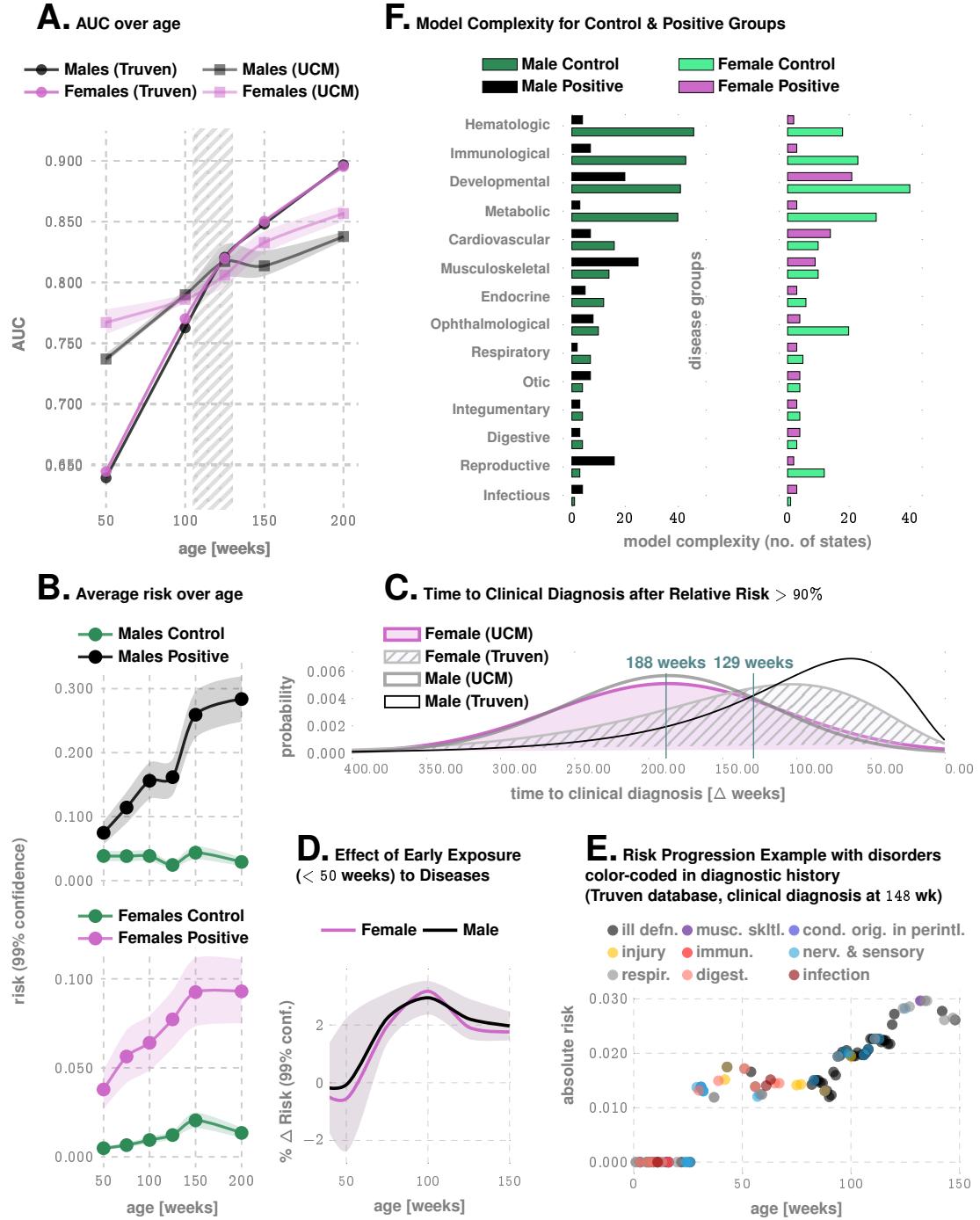


**Fig. 1: Standalone Predictive Performance of ACoR.** Panel A shows the ROC curves for males and females (Truven data shown, UCM is similar, see Fig. 2a). Panel B shows the feature importance inferred by our prediction pipeline. The detailed description of the features is given in Table 1b. The most 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 corresponding 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 the current uneven distribution of diagnostic resources. Importantly, not all counties have nonzero number of ASD patients; high performance in those counties reflects a small number of false positives with zero false negatives.

free parameters compared to standard neural network or deep learning architectures: the simplest deep learning model we investigated has over 185K trainable parameters, while our final pipeline has 13,744. A smaller set of free parameters implies a smaller sample complexity, *i.e.*, can be learned with less data, which, at least in the present application, gives us better performance relative to the standard architectures we investigated (See SI-Fig. 5 in the Supplementary text).

After inferring models of diagnostic history, we need to quantify how such models diverge in the positive cohort vs those underlying the control cohort. We use a novel approach to evaluate subtle deviations in stochastic observations known as the sequence likelihood defect (SLD) (35), to quantify similarity of observed time-series of diagnostic events to the control vs the positive cohorts for individual patients. This novel stochastic inference approach provides significant boost to the overall performance of our predictors; with only state-of-the-art machine learning the predictive performance is significantly worse (See Supplementary Text Sections III and II, as well as reported performance in the literature for predicting ASD risk from EHR data with standard algorithms (29)).

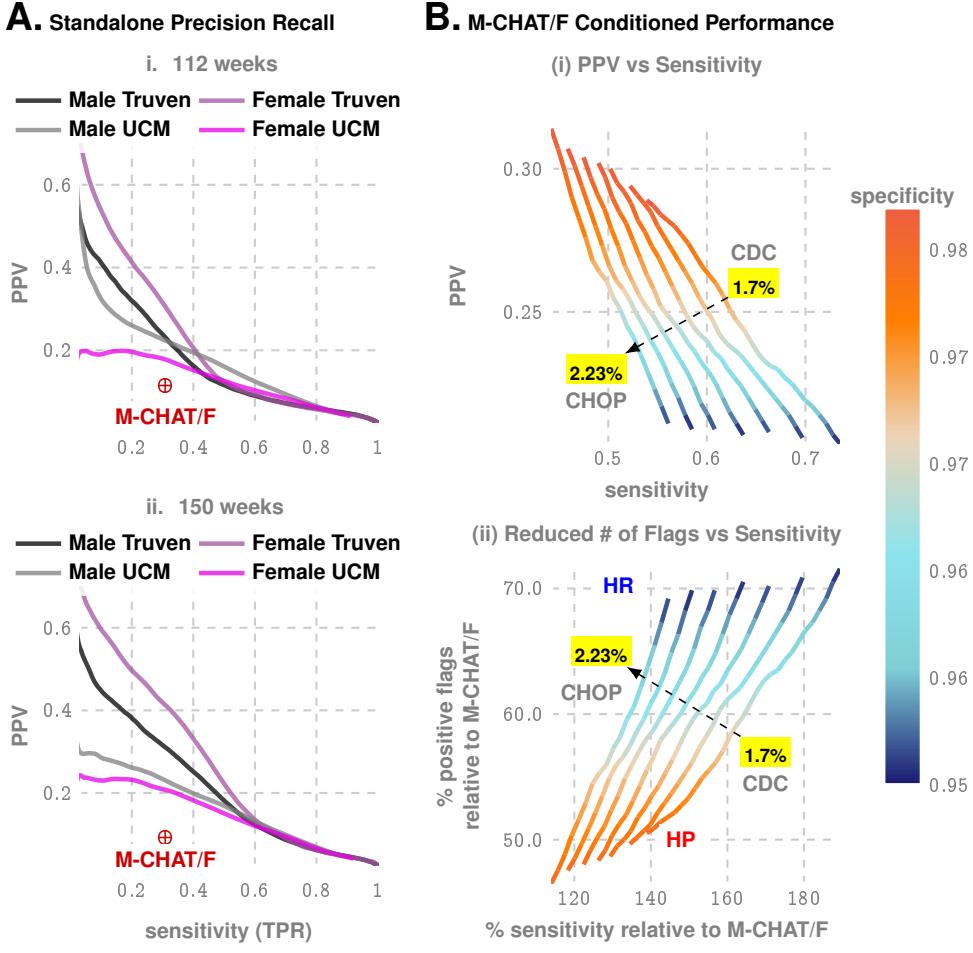
We briefly outline the SLD computation: To reliably infer the cohort-type of a new patient, *i.e.*, the likelihood of a diagnostic sequence being generated by the corresponding cohort model, we generalize the notion of Kullbeck-Leibler (KL) divergence (36, 37) between probability distributions to a divergence  $\mathcal{D}_{\text{KL}}(G||H)$  between ergodic



**Fig. 2: More Details on Standalone Predictive Performance of ACoR and Variation of Inferred Risk.** Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve > 80% AUC for either sex from shortly after 2 years. Panel B illustrates how the average risk changes with time for the control and the positive cohorts. Note that the risk progressions are somewhat monotonic, and the computed confidence bounds suggest that the odds of a child with low risk upto 100 or 150 weeks of age abruptly shifting to a high risk trajectory is low. Panel C shows the distribution of the prediction horizon: the time to a clinical diagnosis after inferred relative risk crosses 90%. Panel D shows that for each new disease code for a low-risk child, ASD risk increases by approximately 2% for either sex. Panel E illustrates the risk progression of a specific, ultimately autistic male child in the Truven database. Abbreviations in the legend: ill defn. (Symptoms, Signs, And Ill-Defined Conditions), musc. skltl. (Diseases Of The Musculoskeletal System And Connective Tissue), cond. orig. in perintl. (Certain Conditions Originating In The Perinatal Period), immun. (Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders), nerv. & sensory (Diseases Of The Nervous System And Sense Organs), respir. (Respiratory Disorders), and digest. (Digestive Disorders). Panel F illustrates how inferred models differ between the control vs. the positive cohorts. On average, models get less complex, implying the exposures get more statistically independent.

stationary categorical stochastic processes (38)  $G, H$  as:

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



**Fig. 3: Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs.** Panel A shows the precision/recall curves, i.e., the trade-off between PPV and sensitivity for **standalone operation** with ACoR. Panel B shows how we can **boost ACoR performance** using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study (3). This is possible because ACoR and M-CHAT/F use independent information (co-morbidities vs questionnaire responses). Note that the population prevalence impacts this optimization, and hence we have a distinct curve for each prevalence value (1.7% is the CDC estimate, while 2.23% is reported by the CHOP study). The two extreme operating zones marked as High Precision (HP) and High Recall (HR): if we choose to operate in HR, then we do not reduce the number of positive screens by much, but maximize sensitivity, while by operating in HP, we increase sensitivity by 20-40% (depending on the prevalence) but double the PPV achieved in current practice. In contrast, when choosing to maximize sensitivity by operating in the HR zone, we only cut down positive flags to about 70% of what we get with M-CHAT/F, but boost sensitivity by 50 – 90% (Reaching sensitivities over 70%). Note in all these zones, we maintain specificity above 95%, which is the current state of art, implying that by doubling the PPV, we can halve the number of positive screens currently reported, thus potentially sharply reducing the queues and wait-times.

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:

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

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

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

$$\lim_{|x| \rightarrow \infty} L(x, H) = \mathcal{H}(G) + D_{KL}(G||H) \quad (3b)$$

where  $\mathcal{H}(\cdot)$  is the entropy rate of a process (36). Importantly, Eq. (3) shows that the computed likelihood has an additional non-negative contribution from the divergence term when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with ASD, then we expect that the disease-specific mapped series corresponding to her diagnostic history be modeled by the corresponding model in the positive cohort. Denoting the model corresponding to disease category  $j$  for positive and control cohorts as  $G_+^j, G_0^j$  respectively, we can compute the sequence likelihood defect (SLD,  $\Delta^j$ ) as:

$$\Delta^j \triangleq L(G_0^j, x) - L(G_+^j, x) \rightarrow D_{KL}(G_0^j || G_+^j) \quad (4)$$

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

(a) Standalone PPV Achieved at 100, 112 and 150 Weeks For Each Dataset and Gender (**M-CHAT/F: sensitivity=38.8%, specificity=95%, PPV=14.6% between 16 and 26 months ( $\approx 112$  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

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

M-CHAT/F Outcome				global performance (Truven)			global performance (UCM)			prevalence*
0-2 NEG	3-7 NEG	3-7 POS	$\geq 8$ POS	speci- ficity	sensi- tivity	PPV	speci- ficity	sensi- tivity	PPV	
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

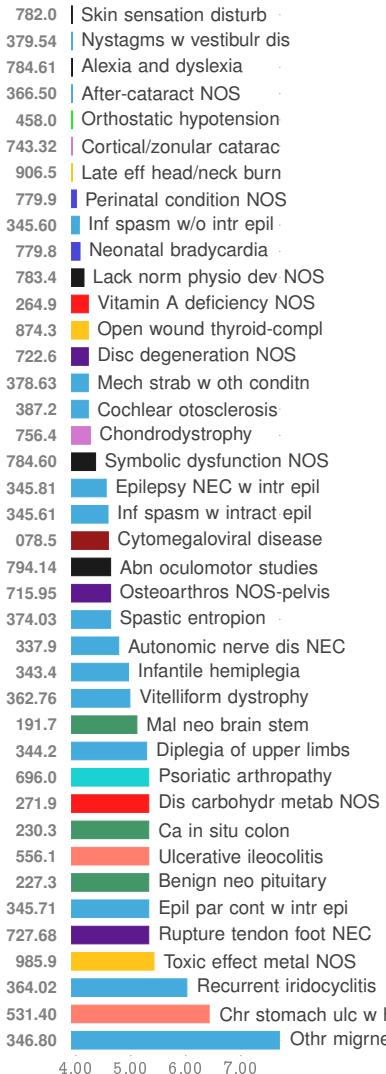
\*Prevalence reported by CDC is 1.7%, while the CHOP study reports a value of 2.23%. The results of our optimization depend on the prevalence estimate.

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

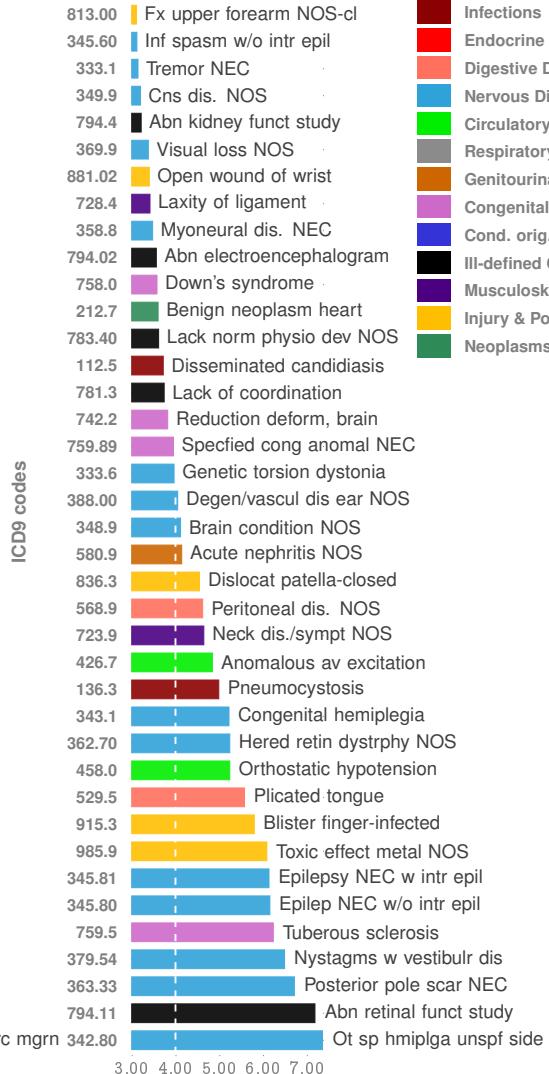
In addition to the category specific Markov models, we use a range of engineered features that reflect various aspects of the diagnostic histories, including the proportion of weeks in which a diagnostic code is generated, the maximum length of consecutive weeks with codes, the maximum length of weeks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

Since we need to infer the Markov models prior to the calculation of the likelihood defects, we need two training sets: one that is used to infer the models, and one that subsequently trains the final classifier with features derived from the inferred models along with other engineered features. Thus, the analysis proceeds by first carrying out a random 3-way split of the set of unique patients (in the Truven dataset) into *Markov model inference* (25%), *classifier training* (25%) and *test* (50%) sets. The approximate sample sizes of the three sets are as follows:  $\approx 700K$  for each of the training sets, and  $\approx 1.5M$  for the test set. The features used in our pipeline may be

### A. Male (3 YR)



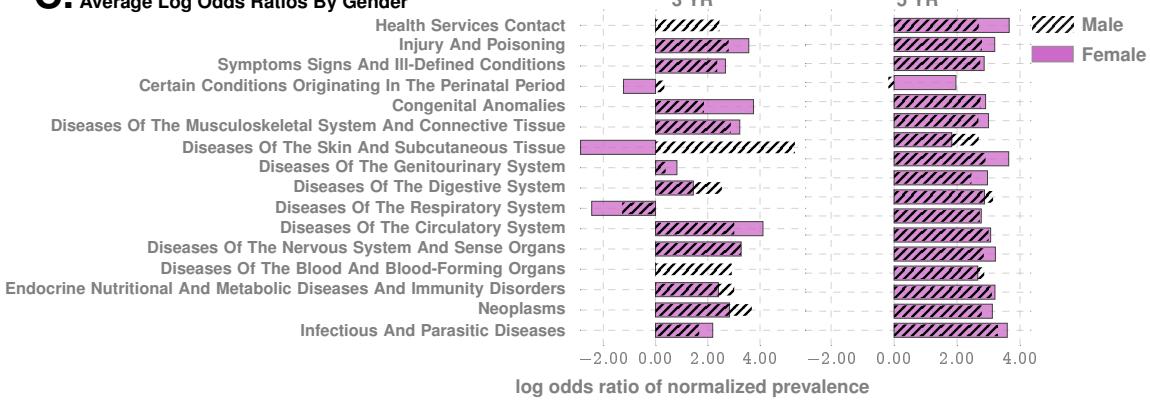
### B. Female (3 YR)



### ICD9 Class



### C. Average Log Odds Ratios By Gender



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

ranked in order of their relative importance (See Fig. 1B for the top 15 features), by estimating the loss in performance when dropped out of the analysis. We verified that different random splits do not adversely affect performance. The UCM dataset in its entirety is used as a test set, with no retraining of the pipeline.

Our pipeline maps medical histories to a raw indicator of risk. Ultimately, to make crisp predictions, we must

choose a decision threshold for this raw score. In this study, we base our analysis on maximizing the  $F_1$ -score, defined as the harmonic mean of sensitivity and specificity, to make a balanced trade-off between Type 1 and Type 2 errors (See Supplementary text, Section IV). The *relative risk* is then defined as the ratio of the raw risk to the decision threshold, and a value  $> 1$  predicts a future ASD diagnosis. Our two-step learning algorithm outperforms standard tools, and achieves stable performance across datasets strictly superior to documented M-CHAT/F.

The independence of our approach from questionnaire-based screening implies that we can further boost our performance by conditioning the sensitivity/specificity trade-offs on individual M-CHAT/F scores. In particular, we leverage the population stratification induced by M-CHAT/F to improve combined performance. Here a combination of ACoR with M-CHAT/F refers to the conditional tuning of the sensitivity/specificity for ACoR in each sub-population such that the overall performance is maximized. To describe this approach briefly, we assume that there are  $m$  sub-populations with the sensitivities, specificities achieved, and the prevalences in each sub-population are given by  $s_i, c_i$  and  $\rho_i$  respectively, with  $i \in \{1, \dots, m\}$ . Let  $\beta_i$  be the relative size of each sub-population. Then, we have (See Supplementary text, Section VIII-A):

$$s = \sum_{i=1}^m s_i \gamma_i \quad (5a)$$

$$c = \sum_{i=1}^m c_i \gamma'_i \quad (5b)$$

where we have denoted:

$$\gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma'_i = \beta_i \frac{1 - \rho_i}{1 - \rho} \quad (5c)$$

and  $s, c, \rho$  are the overall sensitivity, specificity, and prevalence. Knowing the values of  $\gamma_i, \gamma'_i$ , 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  $\leq 2$  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  $\geq 8$ , screening ASD positive. (See SI-Table III). The “follow-up” in the context of M-CHAT/F refers to the re-evaluation of responses by qualified personnel. We use published data from the CHOP study (3) on the relative sizes and the prevalence statistics in these sub-populations to compute the feasible conditional choices of our operating point to vastly supersede standalone M-CHAT/F performance. The CHOP study is the only large-scale study of M-CHAT/F we are aware of with sufficient follow-up after the age of four years to provide a reasonable degree of confidence in the sensitivity of M-CHAT/F.

Two limiting operating conditions are of particular interest in this optimization scheme, 1) where we maximize PPV under some minimum specificity and sensitivity (denoted as the High Precision or the HP operating point), and 2) where we maximize sensitivity under some minimum PPV and specificity (denoted as the High Recall or the HR operating point). Taking these minimum values of specificity, sensitivity, and PPV to be those reported for M-CHAT/F, we identify the feasible set of conditional choices in a four-dimensional decision space that would significantly outperform M-CHAT/F in universal screening. The results are shown in Fig. 3B.

We carried out a battery of tests to ensure that our results are not significantly impacted by class imbalance (since our control cohort is orders of magnitude larger) or systematic coding errors, *e.g.*, we verified that restricting the positive cohort to children with at least two distinct ASD diagnostic codes in their medical histories instead of one, has little impact on out-of-sample predictive performance (SI-Fig. 6B).

## RESULTS

We measure our performance using several standard metrics including the AUC, sensitivity, specificity and the PPV. For the prediction of the eventual ASD status, we achieve an out-of-sample AUC of 82.3% and 82.5% for males and females respectively at 125 weeks for the Truven dataset. In the UCM dataset, our performance is comparable: 83.1% and 81.3% for males and females respectively (Fig. 1 and 2). Our AUC is shown to improve approximately linearly with patient age: Fig. 2A illustrates that the AUC reaches 90% in the Truven dataset at the age of four. Importantly, note that Fig. 2B suggests that the risk progressions are somewhat monotonic. Additionally, the computed confidence bounds suggest that the odds of a child with low risk upto 100 or 150 weeks of age abruptly shifting to a high risk trajectory is low. Thus, “borderline” cases need to be surveiled longer, while robust decisions may be obtained in other cases much earlier.

Recall, that the UCM dataset is used purely for validation, and good performance on these independent datasets lends strong evidence for our claims. Furthermore, applicability in new datasets *without local re-training* makes it readily deployable in clinical settings.

Enumerating the top 15 predictive features (Fig. 1B), ranked according to their automatically inferred weights (the feature “importances”), we found that while infections and immunologic disorders are the most predictive, there is

a significant effect from each of the 17 disease categories. Thus, the co-morbid indicators are distributed across the disease spectrum, and no single disorder is uniquely implicated (See also Fig. 2F). Importantly, predictability is relatively agnostic to the number of local cases across US counties (Fig. 1C-D) which is important in light of the current uneven distribution of diagnostic resources (16, 39) 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  $\approx$  3 years, which corresponds to the median diagnosis age in the database. Thus, if a child is not diagnosed up to that age, then the risk falls, since the probability of a diagnosis in the population starts to go down after this age. While average discrimination is not useful for individual patients, these reveal important clues as to how the risk evolves over time. Additionally, while each new diagnostic code within the first year of life increases the risk burden by approximately 2% irrespective of sex (Fig. 2D), distinct categories modulate the risk differently, *e.g.*, for a single random patient illustrated in Fig. 2F infections and immunological disorders dominate early, while diseases of the nervous system and sensory organs, as well as ill-defined symptoms, dominate the latter period.

Given these results, it is important to ask how much earlier can we trigger an intervention? On average, the first time the relative risk (risk divided by the decision threshold set to maximize F1-score, see Methods) crosses the 90% threshold precedes diagnosis by  $\approx$  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 (16), we are still likely to produce valid red flags significantly earlier than the current practice.

Our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at least 14% (14.1-33.6%) when sensitivity and specificity are held at comparable values (approx. 38% and 95%) around the age of 26 months ( $\approx$  112 weeks, note 26 consecutive months is approximately 112-113 weeks). Fig. 3A and Table IIa show the out-of-sample PPV vs sensitivity curves for the two databases, stratified by sex, computed at 100, 112 and 100 weeks. A single illustrative operating point is also shown on the ROC curve in Fig. 1C, where at 150 weeks, we have a sensitivity of 51.8% and a PPV of 15.8% and 18.8% for males and females respectively, both at a specificity of 95%.

Beyond standalone performance, independence from standardized questionnaires implies that we stand to gain substantially from a combined operation. With the recently reported population stratification induced by M-CHAT/F scores (3) (SI-Table IV), we can compute a conditional choice of sensitivity for our tool, in each sub-population (M-CHAT/F score brackets: 0 – 2, 3 – 7 (negative assessment), 3 – 7 (positive assessment), and > 8), leading to a significant performance boost. With such conditional operation, we get a PPV close to or exceeding 30% at the high precision (HP) operating point across datasets (> 33% for Truven, > 28% for UCM), or a sensitivity close to or exceeding 50% for the high recall (HR) operating point (> 58% for Truven, > 50% for UCM), when we restrict specificities to above 95% (See Table IIb, Fig. 3B(i), and SI-Fig. 10 in the supplementary text). Comparing with standalone M-CHAT/F performance (Fig. 3B(ii)), we show that for any prevalence between 1.7% and 2.23%, we can double the PPV without losing sensitivity at > 98% specificity, or increase the sensitivity by  $\sim$  50% without sacrificing PPV and keeping specificity  $\geq$  94%.

## DISCUSSION

In this study, we operationalize a documented aspect of ASD symptomology in that it has a wide range of co-morbidities (18, 19, 40) occurring at above-average rates (12). Association of ASD with epilepsy (41), gastrointestinal disorders(42–47), mental health disorders (48), insomnia, decreased motor skills (49), allergies including eczema (42–47), immunologic (40, 50–56) and metabolic(46, 57, 58) disorders are widely reported. These studies, along with support from large-scale exome sequencing (59, 60), have linked the disorder to putative mechanisms of chronic neuroinflammation, implicating immune dysregulation and microglial activation (52, 55, 61–64) 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 (65).

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 (66). Additionally, no single gene can be considered “causal” for more than 1% of cases of idiopathic autism (67).

Despite these hurdles, laboratory tests and potential biomarkers for ASD have begun to emerge (7, 9, 10). These tools are still in their infancy, and have not demonstrated performance in the 18-24 month age group. In the absence of clinically useful biomarkers, current screening in pediatric primary care visits uses standardized questionnaires to categorize behavior. This is susceptible to potential interpretative biases arising from language barriers, as well as social and cultural differences, often leading to systematic under-diagnosis in diverse

populations (12). In this study we use time-stamped sequence of past disorders to elicit crucial information on the developing risk of an eventual diagnosis, and formulate the autism comorbid risk score. The ACoR is largely free from aforementioned biases (See performance comparison in out-of-sample populations stratified by race, gender and ethnicity in SI-Fig. 2, where we find no significant differences in predictive performances across racial and ethnic boundaries), and yet significantly outperforms the tools in current practice. Our ability to maintain high performance in diverse populations is an important feature of ACoR; it is not obvious *a priori* that ACoR resists inheriting or compounding the systemic biases that might exist in diagnostic procedures. Nevertheless, our investigations with the UCM dataset shows that we have no significant differences in predictive performance in African-American, white or multi-racial sub-populations. While the average performance amongst Hispanic children were somewhat lower, the differences were not significant.

Going beyond screening performance, this approach provides a new tool to uncover clues to ASD pathobiology, potentially linking the observed vulnerability to diverse immunological, endocrinological and neurological impairments to the possibility of allostatic stress load disrupting key regulators of CNS organization and synaptogenesis. Charting individual disorders in the co-morbidity burden further reveals novel associations in normalized prevalence — the odds of experiencing a specific disorder, particularly in the early years (age < 3 years), normalized over all unique disorders experienced in the specified time-frame. We focus on the true positives in the positive cohort and the true negatives in the control cohort to investigate patterns that correctly disambiguate ASD status. On these lines Fig. 4 and SI-Fig. 7 in the supplementary text outline two key observations: 1) *negative associations*: some diseases that are negatively associated with ASD with respect to normalized prevalence, *i.e.*, having those codes relatively over-represented in one's diagnostic history favors ending up in the control cohort, 2) *impact of sex*: there are sex-specific differences in the impact of specific disorders, and given a fixed level of impact, the number of codes that drive the outcomes is significantly more in males (Fig. 4A vs B).

Some of the disorders that show up in Fig. 4, panels A and B are surprising, *e.g.*, congenital hemiplegia or diplegia of the upper limbs indicative of either cerebral palsy (CP) or a spinal cord/brain injury, neither of which has a direct link to autism. However, this effect is easily explainable: since only about 7% of the children with cerebral palsy (CP) are estimated to have a co-occurring ASD (68, 69), and with the prevalence of CP significantly lower (1 in 352 vs 1 in 59 for autism), it follows that only a small number of children (approximately 1.17%) with autism have co-occurring CP. Thus, with significantly higher prevalence in children diagnosed with autism compared to the general population (1.7% vs 0.28%), CP codes show up with higher odds in the true positive set. Other patterns are harder to explain. For example, SI-Fig. 7A shows that the immunological, metabolic, and endocrine disorders are almost completely risk-increasing, and respiratory diseases (panel B) are largely risk-decreasing. On the other hand, infectious diseases have roughly equal representations in the risk-increasing and risk-decreasing classes (panel C). The risk-decreasing infectious diseases tend to be due to viral or fungal organisms, which might point to the use of antibiotics in bacterial infections, and the consequent dysbiosis of the gut microbiota (44, 58) as a risk factor.

Any predictive analysis of ASD must address if we can discriminate ASD from general developmental and behavioral disorders. The DSM-5 established a single category of ASD to replace the subtypes of autistic disorder, Asperger syndrome, and pervasive developmental disorders (12). This aligns with our use of diagnostic codes from ICD9 299.X as specification of an ASD diagnosis, and use standardized mapping to 299.X from ICD10 codes when we encounter them. For other psychiatric disorders, we get high discrimination reaching AUCs over 90% at 100 – 125 weeks of age (SI-Fig. 6A), which establishes that our pipeline is indeed largely specific to ASD.

Can our performance be matched by simply asking how often a child is sick? While the density of ICD codes in a child's medical history is higher for those eventually diagnosed with autism (See Table 1a), we found this to be a crude measure. While somewhat predictive, achieving  $AUC \approx 75\%$  in the Truven database at 150 weeks (See SI-Fig. 6, panel D in the supplementary text), code density by itself does not have stable performance across the two databases, with particularly poor performance in validation in the UCM database. Furthermore, adding code density as a feature shows no appreciable improvement in our pipeline.

We also investigated the impact of removing all psychiatric codes (ICD9 290 - 319, and corresponding ICD10) from the patient histories to eliminate the possibility that our performance is only reflective of prior psychiatric evaluation results. Training and validation on this modified data found no appreciable difference in performance (See SI-Fig. 9). Additionally, we found that including information on prescribed medications and medical procedures in addition to diagnostic codes did not improve results.

A key limitation of our approach is that automated pattern recognition strategies are not guaranteed to reveal true causal precursors of future diseases. **Thus, we must further scrutinize the patterns we find to be predictive. This is particularly true for the inferred negative associations of some disorders with the future odds of a ASD diagnosis; these might be indicative of collider bias (70) suggesting spurious associations arising from conditioning on the common effect of unrelated causes. Here the putative collider in our framework is the presence of the child in a clinic or a medical facility for a medical issue. In future, we will investigate the possibility of identifying more transparent and interpretable risk precursors via post-processing of our inferred models.**

Our approach is also impacted by the uncurated nature of medical history, which invariably includes coding

---

mistakes and other artifacts, *e.g.* potential for 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 (71). Discontinuities in patient medical histories from change in provider-networks can also introduce uncertainties in risk estimates, and socio-economic status of patients and differential access to care 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.

In future, 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.

One of our immediate future goals is to prospectively validate ACoR in a clinical setting. Ultimately, we hope to see widespread adoption of this new tool. A key hurdle to adoption is a general lack of trust in non-transparent decision algorithms, particularly in the light of the biases, mis-classifications, misinterpretations, and subjectivity in the training datasets (72). We hope that results from prospective trials, and the key advantages laid out in this study will help alleviate such concerns.

#### Data & Software Availability

All data and models needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The Truven and UCM datasets cannot be made available due to ethical and patient privacy considerations. Preliminary 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/>. To enable fast execution, some more compute intensive features are disabled in this version. Results from this software are for demonstration purposes only, and must not be interpreted as medical advice, or serve as replacement for such.

#### ACKNOWLEDGMENT

This work is funded in part by the Defense Advanced Research Projects Agency (DARPA) project number HR00111890043/P00004. The claims made in this study do not reflect the position or the policy of the US Government. The UCM dataset is provided by the Clinical Research Data Warehouse (CRDW) maintained by the Center for Research Informatics (CRI) at the University of Chicago. The Center for Research Informatics is funded by the Biological Sciences Division, the Institute for Translational Medicine/CTSA (NIH UL1 TR000430) at the University of Chicago. IRB exemption was granted due to de-identified subjects from University of Chicago, IRB Committee: BSD (Contact: Amy Horst ahorst@medicine.bsd.uchicago.edu). IRB#: Predictive Diagnoses IRB19-1040.

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

#### AUTHOR CONTRIBUTIONS

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

#### SUPPLEMENTARY MATERIALS

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

#### REFERENCES

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

- [8] G. Li, O. Lee, H. Rabitz, *PLoS one* **13**, e0192867 (2018).
- [9] S. D. Hicks, *et al.*, *Frontiers in genetics* **9**, 534 (2018).
- [10] A. M. Smith, *et al.*, *Autism Research* **13**, 1270 (2020).
- [11] F. Volkmar, *et al.*, *Journal of the American Academy of Child & Adolescent Psychiatry* **53**, 237 (2014).
- [12] S. L. Hyman, S. E. Levy, S. M. Myers, *et al.*, *Pediatrics* **145** (2020).
- [13] L. G. Kalb, *et al.*, *Journal of Developmental & Behavioral Pediatrics* **33**, 685 (2012).
- [14] J. Bisgaier, D. Levinson, D. B. Cutts, K. V. Rhodes, *Archives of pediatrics & adolescent medicine* **165**, 673 (2011).
- [15] T. S. Fenikilé, K. Ellerbeck, M. K. Filippi, C. M. Daley, *Primary health care research & development* **16**, 356 (2015).
- [16] E. Gordon-Lipkin, J. Foster, G. Peacock, *Pediatric Clinics* **63**, 851 (2016).
- [17] D. L. Robins, *et al.*, *Pediatrics* **133**, 37 (2014).
- [18] C. Tye, A. K. Runcles, A. J. O. Whitehouse, G. A. Alvares, *Front Psychiatry* **9**, 751 (2018).
- [19] I. S. Kohane, *et al.*, *PLoS ONE* **7**, e33224 (2012).
- [20] K. K. Hyde, *et al.*, *Review Journal of Autism and Developmental Disorders* **6**, 128 (2019).
- [21] H. Abbas, F. Garberson, S. Liu-Mayo, E. Glover, D. P. Wall, *Scientific reports* **10**, 1 (2020).
- [22] M. Duda, J. Daniels, D. P. Wall, *Journal of autism and developmental disorders* **46**, 1953 (2016).
- [23] M. Duda, J. Kosmicki, D. Wall, *Translational psychiatry* **4**, e424 (2014).
- [24] V. A. Fusaro, *et al.*, *PLOS one* **9**, e93533 (2014).
- [25] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, T. F. DeLuca, *PLoS one* **7**, e43855 (2012).
- [26] D. P. Wall, J. Kosmicki, T. Deluca, E. Harstad, V. A. Fusaro, *Translational psychiatry* **2**, e100 (2012).
- [27] F. Doshi-Velez, Y. Ge, I. Kohane, *Pediatrics* **133**, e54 (2014).
- [28] L. Bishop-Fitzpatrick, *et al.*, *Autism Research* **11**, 1120 (2018).
- [29] T. Lingren, *et al.*, *PLoS one* **11**, e0159621 (2016).
- [30] L. Hansen, *Truven Health Analytics IBM Watson Health* (2017).
- [31] J. Baio, *et al.*, *MMWR Surveill Summ* **67**, 1 (2018).
- [32] I. Chattopadhyay, A. Ray, *International Journal of Control* **81**, 820 (2008).
- [33] I. Chattopadhyay, H. Lipson, *Philos Trans A* **371**, 20110543 (2013).
- [34] I. Chattopadhyay, H. Lipson, *Journal of The Royal Society Interface* **11** (2014).
- [35] Y. Huang, I. Chattopadhyay, *arXiv preprint arXiv:1909.12243* (2019).
- [36] T. M. Cover, J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).
- [37] S. Kullback, R. A. Leibler, *Ann. Math. Statist.* **22**, 79 (1951).
- [38] J. Doob, *Stochastic Processes*, Wiley Publications in Statistics (John Wiley & Sons, 1953).
- [39] L. A. Althouse, J. A. Stockman, *The Journal of pediatrics* **148**, 575 (2006).
- [40] O. Zerbo, *et al.*, *Brain Behav. Immun.* **46**, 232 (2015).
- [41] H. Won, W. Mah, E. Kim, *Front Mol Neurosci* **6**, 19 (2013).
- [42] G. Xu, *et al.*, *JAMA Netw Open* **1**, e180279 (2018).
- [43] J. B. Adams, *et al.*, *Nutr Metab (Lond)* **8**, 34 (2011).
- [44] A. Fattorusso, L. Di Genova, G. B. Dell'Isola, E. Mencaroni, S. Esposito, *Nutrients* **11** (2019).
- [45] R. Diaz Heijtz, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3047 (2011).
- [46] S. Rose, *et al.*, *PLoS ONE* **12**, e0186377 (2017).
- [47] E. M. Sajdel-Sulkowska, *et al.*, *Cerebellum* **18**, 255 (2019).
- [48] M. S. Kayser, J. Dalmau, *Curr Psychiatry Rev* **7**, 189 (2011).
- [49] O. I. Dadalko, B. G. Travers, *Front Integr Neurosci* **12**, 47 (2018).
- [50] Y. Yamashita, *et al.*, *Front Psychiatry* **10**, 152 (2019).
- [51] L. Shen, *et al.*, *Front Cell Neurosci* **13**, 105 (2019).
- [52] K. Ohja, *et al.*, *Neuromolecular Med.* **20**, 161 (2018).
- [53] D. Gadysz, A. Krzywdziska, K. K. Hozyasz, *Mol. Neurobiol.* **55**, 6387 (2018).
- [54] T. C. Theoharides, I. Tsilioni, A. B. Patel, R. Doyle, *Transl Psychiatry* **6**, e844 (2016).
- [55] A. M. Young, *et al.*, *Mol Autism* **7**, 9 (2016).
- [56] L. A. Croen, *et al.*, *Autism Res* **12**, 123 (2019).
- [57] T. Vargason, D. L. McGuinness, J. Hahn, *J Autism Dev Disord* **49**, 647 (2019).
- [58] M. Fiorentino, *et al.*, *Mol Autism* **7**, 49 (2016).
- [59] F. K. Satterstrom, *et al.*, *bioRxiv* (2019).
- [60] T. Gaugler, *et al.*, *Nat. Genet.* **46**, 881 (2014).
- [61] D. L. Vargas, C. Nascimbene, C. Krishnan, A. W. Zimmerman, C. A. Pardo, *Ann. Neurol.* **57**, 67 (2005).
- [62] H. Wei, *et al.*, *J Neuroinflammation* **8**, 52 (2011).
- [63] A. M. Young, E. Campbell, S. Lynch, J. Suckling, S. J. Powis, *Front Psychiatry* **2**, 27 (2011).
- [64] H. K. Hughes, E. Mills Ko, D. Rose, P. Ashwood, *Front Cell Neurosci* **12**, 405 (2018).
- [65] N. Pearce, *Journal of epidemiology and community health* **54**, 326 (2000).
- [66] J. D. Murdoch, M. W. State, *Curr. Opin. Genet. Dev.* **23**, 310 (2013).
- [67] V. W. Hu, *Future Neurol* **8**, 29 (2013).
- [68] Centers for Disease Control and Prevention, Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning (2020).

- 
- [69] D. Christensen, et al., *Developmental Medicine & Child Neurology* **56**, 59 (2014).
- [70] J. Berkson, *Biometrics Bulletin* **2**, 47 (1946).
- [71] E.-M. Rødgaard, K. Jensen, J.-N. Vergnes, I. Soulières, L. Mottron, *JAMA Psychiatry* **76**, 1124 (2019).
- [72] M. Cheng, S. Nazarian, P. Bogdan, *Frontiers in Artificial Intelligence* **3**, 54 (2020).
- [73] General equivalence mappings.
- [74] J. Baio (2014).
- [75] P. F. Bolton, J. Golding, A. Emond, C. D. Steer, *Journal of the American Academy of Child & Adolescent Psychiatry* **51**, 249 (2012).
- [76] J. Bondy, U. Murty, *Grad. Texts in Math* (2008).
- [77] L. Breiman, *Mach. Learn.* **45**, 5 (2001).
- [78] I. Chattopadhyay, A. Ray, *International Journal of Control* **81**, 820 (2008).
- [79] I. Chattopadhyay, H. Lipson, *Journal of The Royal Society Interface* **11**, 20140826 (2014).
- [80] C. Chlebowski, J. A. Green, M. L. Barton, D. Fein, *Journal of autism and developmental disorders* **40**, 787 (2010).
- [81] T. M. Cover, J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).
- [82] J. Doob, *Stochastic processes*, Wiley publications in statistics (Wiley, 1990).
- [83] A. N. Esler, et al., *Journal of Autism and Developmental Disorders* **45**, 2704 (2015).
- [84] T. Falkmer, K. Anderson, M. Falkmer, C. Horlin, *European child & adolescent psychiatry* **22**, 329 (2013).
- [85] J. H. Friedman, *Comput. Stat. Data Anal.* **38**, 367 (2002).
- [86] G. Hardy, *Éditions Jacques Gabay, Sceaux* (1992).
- [87] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **9**, 1735 (1997).
- [88] J. E. Hopcroft, *Introduction to automata theory, languages, and computation* (Pearson Education India, 2008).
- [89] V. G. Jarquin, L. D. Wiggins, L. A. Schieve, K. Van Naarden-Braun, *Journal of Developmental & Behavioral Pediatrics* **32**, 179 (2011).
- [90] C. P. Johnson, S. M. Myers, et al., *Pediatrics* **120**, 1183 (2007).
- [91] L. C. Kai, *Markov Chains: With Stationary Transition Probabilities* (Springer-Verlag, 1967).
- [92] J. M. Kleinman, et al., *Journal of autism and developmental disorders* **38**, 606 (2008).
- [93] A. Klenke, *Probability theory: a comprehensive course* (Springer Science & Business Media, 2013).
- [94] A. M. Kozlowski, J. L. Matson, M. Horovitz, J. A. Worley, D. Neal, *Developmental neurorehabilitation* **14**, 72 (2011).
- [95] C. Lord, et al., *Archives of general psychiatry* **63**, 694 (2006).
- [96] C. W. J. Granger, R. Joyeux, *Journal of Time Series Analysis* **1**, 15 (1980).
- [97] A. G. d. G. Matthews, J. Hensman, R. Turner, Z. Ghahramani, *Journal of Machine Learning Research* **51**, 231 (2016).
- [98] M. Penner, E. Anagnostou, W. J. Ungar, *Molecular autism* **9**, 16 (2018).
- [99] A. N. Trahtman, *Proc. of Prague stringology conference* (Citeseer, 2008), vol. 1, p. 12.
- [100] M. Vidyasagar, *Hidden markov processes: Theory and applications to biology*, vol. 44 (Princeton University Press, 2014).
- [101] L. Zwaigenbaum, et al., *Pediatrics* **136**, S60 (2015).

# Supplementary Information:

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

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

<sup>1</sup>Department of Medicine, University of Chicago, Chicago, IL USA

<sup>2</sup>Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA

<sup>3</sup>Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA

<sup>4</sup>Department of Pediatrics, Section of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL USA

<sup>5</sup>Department of Pediatrics, Section Chief of Developmental and Behavioral Pediatrics, University of Chicago, Chicago, IL USA

<sup>6</sup>Joseph P. Kennedy Research Center on Intellectual and Neurodevelopmental Disabilities, University of Chicago, Chicago, IL USA

<sup>7</sup>Executive Committee Chair, American Academy of Pediatrics' Section on Developmental and Behavioral Pediatrics

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

## CONTENTS

<b>I</b>	<b>Detailed Mathematical Approach</b>	9
I-A	Time-series Modeling of Diagnostic History . . . . .	9
I-B	Step 1: Partitioning The Human Disease Spectrum . . . . .	10
I-C	Step 2: Model Inference & The Sequence Likelihood Defect . . . . .	11
I-D	Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules . .	11
<b>II</b>	<b>Comparison With State of the Art Off-the-shelf ML Algorithms</b>	13
<b>III</b>	<b>Comparison With Pipeline Variations, Feature Subsets and Neural Network (NN) Post-processing</b>	13
III-A	Feature Subset Evaluations & Code Density As A Feature . . . . .	14
<b>IV</b>	<b>Threshold Selection on ROC Curve</b>	15
<b>V</b>	<b>Note on Receiver Operating Characteristics (ROC) and Precision-recall Curves</b>	18
<b>VI</b>	<b>Effect of Class Imbalance</b>	19
<b>VII</b>	<b>Note on ASD Clinical Diagnosis &amp; Uncertainty of EHR Record</b>	20
VII-A	Diagnostic Evaluations . . . . .	20
VII-B	Change In Diagnostic Criteria for ASD, Inclusion of PDD, Asperger, and Disambiguation From Unrelated Psychiatric Phenotypes . . . . .	20
VII-C	Performance Comparison With M-CHAT/F . . . . .	21

<b>VIII</b>	<b>Improving Wait-times For Diagnostic Evaluations by Reducing False Positives in Routine Screening</b>	21
	VIII-A 4D Decision Optimization Using M-CHAT/F Population Stratification To Boost PPV . . . . .	21
<b>IX</b>	<b>Generating PFSA Models From Set of Input Streams with Variable Input Lengths</b>	24
<b>X</b>	<b>Probabilistic Finite State Automata Inference</b>	24
	X-A Probabilistic Finite-State Automaton . . . . .	24
<b>XI</b>	<b>Sequence Likelihood Defect</b>	26
<b>XII</b>	<b>Pipeline Optimization</b>	28
	XII-A Input Data Format . . . . .	28
	XII-B Algorithms . . . . .	28
<b>XIII</b>	<b>Example Run with Released Application</b>	29
	XIII-A Prerequisites & Installation . . . . .	29
	XIII-B EHR data format . . . . .	29
	XIII-C Sample Python code risk estimation . . . . .	30
	XIII-D Sample Python script risk estimation . . . . .	31

#### LIST OF TABLES

I	Disease Categories With Detailed Set of ICD9 Codes Used . . . . .	3
II	Boosted Sensitivity, Specificity and PPV Achieved at <b>150 weeks</b> Conditioned on M-CHAT/F Scores . . . . .	23
III	Population Stratification Results on large M-CHAT/F Study(n=20,375) (3) . . . . .	23
IV	$\gamma, \gamma'$ Computed from Population Stratification Recorded In M-CHAT/F Study (3) ( $\rho = 0.0223$ ) . . . . .	23

**SI-Table I: Disease Categories With Detailed Set of ICD9 Codes Used**

Cat.	De-scrip-tion	Constituent ICD9 Codes
Hematologic	Diseases Of The Blood And Blood-Forming Organs	286.9 286.7 286.6 283.19 283.10 283.11 283.9 283 283.1 284.9 284.8 284.81 284.0 284.89 284.09 284 284.01 282.2 287.49 287.41 287.39 287.4 287.5 287.32 287.3 287.30 287.31 286.3 286.2 286.1 286.0 286.4 282.1 282.6 282.5 282.41 282.42 282.68 282.69 282.62 282.63 282.60 282.61 282.64 282 282.8 287.33 281.2 281.3 280.0 282.9 285.8 285.9 280.9 284.2 285.1 285.2 285.3 280.1 285.22 285.21 282.3 776.5 285 283.0 285.29 280.8 282.7 282.40 282.49 284.1 284.19 284.12 284.11 281.8 281.9 281.4 281.0 281.1 286.5 287 287.8 287.9 287.2 287.0 287.1 285 289.52 289.50 289.51 289.59 289.4 289.5 289.81 289.83 289.82 289.89 289 289.7 289.8 289.9
Psychiatric	Mental Disorders (Except ASD)	290 through 319 (except 299.x)
Metabolic	Metabolic Disorders (Distinct from respiratory, digestive and immunological conditions)	273.4 270 270.2 270.3 712.11 712.10 712.13 712.12 712.15 712.14 712.17 712.16 712.19 712.18 712.31 712.30 712.37 712.36 712.35 712.34 712.38 712.33 712.32 712.28 712.29 712.24 712.25 712.26 712.27 712.20 712.21 712.22 712.23 712.39 712.1 712.3 712.2 277.6 275.1 277.85 277.87 270.7 270.6 276.6 276.7 276.4 276.2 276.3 276.0 275.41 276.1 276.8 276.9 276.69 275.5 275.42 271.1 330.2 272.7 271 274.89 712.85 274.81 274.82 712.99 712.98 274.01 274.00 274.03 274.02 712.91 712.90 712.93 712.92 712.95 712.94 712.97 712.96 712.88 712.89 274.10 274.11 712.82 712.83 712.80 712.81 712.86 712.87 712.84 274.19 712.9 712.8 274.0 274.1 274.2 274.8 274.9 271.2 275.01 270.5 270.4 278.8 272.3 275.03 275.09 271.3 272.6 272.5 278.1 271.8 277.5 263.0 263.2 262 260 261 263 263.1 269.3 269.9 263.8 263.9 263 277.7 272.2 330.3 271.9 275.40 272.8 277.8 275.49 275.5 277.88 275.4 269.3 275.9 275.8 277.9 277.89 251.2 251.1 251.0 278.01 278.00 278.03 270.8 270.9 278 278.02 277.86 270.1 275.3 277.1 277.81 277.82 272 272.2 277.2 272.4 272.9 273.9 273.8 268.1 265.2 268.0 268.2 268.5 265.0 265.1 266.1 266.0 266.2 266.9 264.3 264.2 264.1 264.0 264.7 264.6 264.5 264.9 264.8 268.9 267 266 265 264 269.2 269.0 269.1 278.2 278.3 278.4
Cardiovascular	Diseases Of Arteries, Arterioles, And Capillaries	442.89 441.6 442.84 442.82 442.83 442.8 441.03 441.02 441.01 441.00 441 414.19 414.12 442 414.10 414.11 447.70 447.71 447.72 447.73 414.1 442.81 441.9 442.1 442.0 442.3 442.2 441.2 441.3 441.0 441.1 442.9 441.7 441.4 441.5 437.3 447.7 443.29 443.23 443.22 443.21 443.24 443.2 444.9 444.8 444.81 444.2 444.1 444.0 444.89 444 444.22 444.21 445.81 440.31 440.30 440.32 441.01 414.00 414.03 414.02 414.05 414.04 414.07 414.06 445.89 411.81 445.02 445.01 440.24 440.22 440.23 440.20 440.21 440 445 440.29 414.0 414.2 440.4 440.3 440.2 440.1 440.0 440.9 440.8 445.8 445.0 414.3 414.4 426.54 426.53 426.52 426.51 426.50 426.13 426.12 426.11 426.10 426.89 426.9 426.8 426.81 426.3 426.2 426.1 426.0 426.7 426.6 426.5 426.4 427.61 427.60 427.5 427.89 427.69 427.32 427.41 427.9 427.8 427.4 427.2 427.6 427.4 427.3 427.0 427.1 427.5 425.9 425.8 425.7 425.0 425.1 425.2 425.3 346.6 438.51 438.52 438.53 438.50 290.4 431.0 438.42 438.41 438.40 432.9 290.43 290.42 290.41 290.40 432.0 432.1 433.00 433.01 434.9 346.61 346.60 346.63 346.62 433.80 433.81 433.11 433.10 438.32 438.30 438.31 434.0 433.91 433.90 430.0 434.10 434.11 434.1 438.21 438.20 438.22 433.20 433.21 438.6 438.7 438.4 438.5 438.2 438.3 438.0 438.1 438.9 433.31 433.30 438.85 438.84 438.83 438.82 438.81 438.89 433.9 433.8 433.1 433.0 433.3 433.2 437 435.3 435.2 435.1 435.0 435.9 435.8 437.5 437.4 437.7 437.6 437.1 437.0 431 437.9 437.8 459.3 459.33 459.32 459.31 459.30 459.39 452 453 453.7 453.6 453.5 453.4 453.3 453.2 453.1 453.0 453.9 453.8 453.52 453.51 453.50 453.79 453.71 453.73 453.72 453.75 453.74 453.77 453.76 453.84 453.89 453.40 453.41 453.42 453.81 453.82 453.83 415.11 453.85 453.86 453.87 405.0 405.1 404.9 403.11 402.00 402.01 404.1 404.0 402.1 402.0 403.0 405.99 402.9 405.91 402.91 402.90 405.11 401.0 401.1 404.00 404.01 404.02 404.03 405.19 401.9 405.40 404.40 403.40 402.40 405.9 403.01 403.00 402.11 402.10 403.10 404.13 404.12 404.11 404.10 405.01 403.9 405.09 403.90 437.2 403.1 403.91 404.93 404.92 404.91 404.90 448 458 458.0 458.2 458.1 458.8 458.9 458.29 458.21 426.82 429.71 410.01 410.00 410.02 410.41 410.40 410.42 410.22 410.21 410.20 429.7 410.70 410.71 410.72 429.79 410.92 410.90 410.91 410.30 410.31 410.32 410.12 410.10 410.11 410.52 410.50 410.51 410.4 410.5 410.6 410.7 410.0 410.1 410.2 410.3 410.8 410.9 411.0 410.62 410.61 410.60 410 410.42 410.81 410.80 410.82 424.1 424.0 424.3 424.2 424.2 429.8 429.9 429.5 429.6 429.8 429.9 459.4 427.6.5 429.2 429.3 428.9 428.4 428.1 428.0 428.3 428.2 429.81 429.83 429.82 428.33 428.32 428.31 428.30 428 459.9 459.8 459.0 276.50 276.51 428.42 428.43 428.40 428.41 276.52 428.20 428.21 428.22 428.23 428.00 459.89 448.1 454 455 455.9 455.8 454.8 454.7 455.1 455.0 455.3 455.2 455.5 455.4 455.6 455.7 455.6 454.2 447 454.0 454.1 757.32 447.8 447.9 448.9 447.4 447.5 447.0 447.2 447.3 414.413.1 413.0 413.9 411.89 411.1 411 414.9 413 414.8 411.8 443.89 443.459.2 443.8 443.9 443.81 459.81 443.0 416.2 415.19 415.1 415.13 415.12 416.
Nutrition	Nutrition, metabolism, and development	7830,78321,7833,78340,78342,7837,7839

† Categories inferred to be important for risk modulation are highlighted. Continued on next page

<sup>†</sup> Categories inferred to be important for risk modulation are highlighted. Continued on next page

<sup>†</sup> Categories inferred to be important for risk modulation are highlighted. Continued on next page

<sup>t</sup> Categories inferred to be important for risk modulation are highlighted. Continued on next page

	Integumentary	706.0 706.1 704.0 704.00 704.02 704.01 704.09 680.9 680.8 680.1 680.0 680.3 680.2 680.5 680.4 680.7 680.6 680 698.8 698 698.1 698.0 698.9 757.31 757.5 757.4 757.33 700 694.2 694.3 694.0 694.1 709.0 709.00 709.01 709.09 757.39 757.3 695.59 695 695.57 695.56 695.55 695.54 695.53 695.52 695.51 695.50 695.5 695.58 704 704.9 704.2 704.3 704.8 757 757.1 702.11 702.19 702.0 702.1 697.1 697.0 697.9 697 697.8 759.82 703.8 703.9 703.0 703.80 703 757.8 757.9 757.2 525.9 525.8 525.5 525.4 525.7 525.6 525.1 525.0 525.5 525.79 525.525.73 525.72 525.71 521.35 521.24 521.25 521.20 521.21 521.22 521.23 521.24 521.40 521.41 521.49 522.2 522.3 522.0 522.1 522.8 522.9 521.33 521.32 521.31 521.30 525.19 521 522 521.34 525.11 525.10 525.13 525.12 521.89 525.42 525.43 525.40 525.41 525.44 521.08 521.09 521.02 521.03 521.00 521.01 521.06 521.07 521.04 521.05 521.1 521.0 521.3 521.2 521.5 521.4 521.7 521.6 521.8 525.51 525.50 525.53 525.52 525.54 521.11 521.10 521.13 521.12 521.15 521.14 521.81 525.65 525.64 525.65 525.66 525.67 525.60 525.61 525.62 525.63 523.11 523.10 523.31 523.30 523.33 523.32 523.3 522.7 522.4 522.5 523.6 523.5 523.4 523.9 523.8 523 522.6 523.00 523.01 523.2 523.1 523.40 523.41 523.42 523.0 523.22 523.23 523.20 523.21 523.24 523.25 709.1 707.00 707.02 707.03 707.04 707.05 707.06 707.01 709.9 707.09 707 701 702 707.25 707.2 707.1 707.0 707.22 707.8 707.9 707.23 706.8 707.20 702.8 705.2 706.2 707.07 707.1 701.8 709.3 709.8 701.1 701.0 701.3 701.2 701.4 707.24 707.19 707.15 707.14 707.13 707.12 707.11 710 705.21 705.22 707.21 707.10 710.9 710.8 709 695.9 709.2 729 729.7 729.91 729.9 729.90 729.79 729.99 729.0 729.71 729.72 528.9 528.8 528.9 528.4 528.5 528.5 528.7 528.0 528.1 528.2 528.00 528.79 527.3 528.09 528.3 528 528.71 528.72 692.7 692.70 692.71 692.72 692.82 692.74 692.75 692.76 692.77 692.73 692.79 558.1 508.0 508.1 655.6 595.82 655.60 655.61 737.33 737.11 695.3 706 706.9 690 690.1 690.10 690.11 690.12 690.8 690.18 705.9 705.1 705.0 705.89 705 705.82 705.8 529.9 529.8 529.1 529.0 529.3 529.2 529.5 529.4 529.6 529
Respiratory	Dis-eases Of The Respiratory System (Distinct from Infectious)	516.31 516.30 516.33 516.32 516.35 516.34 516.37 516.36 516.9 516.8 516.3 516.2 516.1 516.0 516 277.0 277.00 277.01 277.02 277.03 277.09 470 478.0 491.9 491.8 515 491.1 491.0 491.2 515.0 491.22 491.21 494.0 492.0 492.8 494.1 494 496 491.20 491 492 478.32 478.33 478.30 478.31 478.34 478.75 478.4 478.79 478.3 478.74 478.5 478.6 478.7 478.70 478.71 471.8 471 471.9 471.1 471.0 519.09 519.00 519.02 495.2 495.5 495.4 495.7 495.6 495.1 495.0 495.3 518.4 495 495.9 495.8 517 503 519.8 507.0 517.8 504 505 502 517.3 500 501 518.84 518.82 518.83 518.81 514.0 519.11 518.89 519.0 519.4 518.8 518.5 519.9 518.1 518.0 518.3 518.2 501.0 512.1 512.0 517.2 512.8 519 518 510.0 510 512 510.9 514 500.0 327.24 327.25 327.26 327.2 327.20 327.21 327.22 327.23 327.27 327.29
Digestive	Dis-eases Of The Digestive System	540.0 540.1 540.9 541.0 542 543 540 541 543.9 543.0 562 562.1 562.11 562.03 562.02 562.01 562.00 562.10 562.0 562.12 562.13 530.3 530.13 530.11 530.0 530.6 530.5 530.4 456.21 530.9 530.8 456.2 456.20 456.0 456.1 530.89 530 530.84 530.85 530.81 530.82 530.83 530.80 530.79 530.60 530.50 530.4 530.01 564.09 560.1 564.01 564.00 564.02 564.81 560.30 564.89 560.31 560.32 579.9 536.8 564 537.0 537.1 537.2 537.3 536.9 537.5 537.6 564.5 537.8 537.9 560.2 560.3 536.1 536.0 536.3 537.81 560.8 560.9 537.89 537.3 536 564.8 564.9 537.4 564.7 564.0 538 532.51 532.50 533.21 533.20 569.41 534.11 534.10 531.70 531.71 534.91 534.90 532.60 532.61 533.50 533.51 533.46 534.61 531.61 531.60 531.6 531.7 531.4 531.5 531.2 531.3 531.0 531.1 531.9 534 533 532 531 531.50 531.51 533.41 533.40 532.71 532.70 534.71 534.70 530.20 530.21 534.9 531.41 531.40 532.00 532.01 534.40 534.41 532.11 532.10 531.30 531.31 534.51 534.50 539.86 533.70 533.71 536.9.82 533.10 533.11 532.1 532.0 532.3 532.2 531.21 531.20 532.7 532.6 532.9 532.91 532.90 534.20 534.21 533.61 533.60 533.01 533.0 531.90 531.91 531.10 531.11 534.31 534.30 532.20 532.21 533.90 533.91 533.30 533.31 532.40 532.41 534.7 534.6 534.5 534.4 534.3 534.2 534.1 534.0 533.4 533.5 533.6 533.7 533.0 533.1 533.2 533.3 533.9 531.01 531.00 530.2 534.00 534.01 532.31 532.30 532.5 532.4 751.3 568.0 578.7 569.9 569.8 558.9 579.1 569.3 558.569.87 579.2 560.81 569.84 569.85 558.42 569.83 578.9 558.41 579.8 578.7 578.0 569 569.81 537.83 537.82 530.7 537.84 558.4 569.89 568.569.8 568.81 568.82 568.89 251.5 251.4 577 251.9 251.8 577.8 577.9 579.4 251 577.0 577.1 577.2 569.4 569.1 565 566 569.2 569.42 569.43 566.0 569.44 565.1 565.0 564.6 569.49 527.9 527.8 527.7 527.6 527.5 527.4 527.2 527.1 527.0 527
Otic	Dis-eases Of The Ear And Mastoid Process	381.51 381.50 381.52 381.81 381.89 381.60 381.61 381.62 381.63 381 381.9 381.8 381.5 381.7 381.6 385.82 383.32 380.30 380.31 380.32 380.39 380.52 380.53 380.50 380.51 380.3 380.5 385.3 380.3 380.4 380.5 380.8 380.9 380.81 380.89 380 385.32 380.21 385.30 385.35 385.33 380.01 380.02 380.03 380.02 385.31 388.8 388.9 388.70 388.71 388.72 388.6 388.3 388.11 388.61 388.60 388 388.69 389.15 389.14 389.17 389.16 389.11 389.10 389.13 389.12 388.0 388.1 388.2 388.3 389.18 388.12 388.10 388.30 388.31 388.32 389.04 389.20 389.21 389.22 389.05 389.02 389.03 389.00 389.01 389.08 388.01 388.00 389.7 388.02 389.1 389.0 389.2 389.3 389.9 389.8 389.06 386.43 386.42 386.41 386.40 386.48 386.4 386.5 386.3 386.8 386.34 386.54 386.55 386.56 386.50 386.51 386.52 386.53 386.58 386.34 386.35 386.32 386.33 386.30 386.31 387.2 387.1 387.0 387.9 387.8 387 385.12 384.09 385.24 385.21 385.22 385.23 385.09 385.02 385.03 385.00 385.01 385.1 385.0 385.2 384.0 384.1 384.2 384.8 384.9 384 384.82 384.82 385.19 384.01 384.00 385.11 385.10 384.25 384.24 384.23 384.22 384.21 384.20 385.13 385.9 385.8 385.89 385
Endocrine	Dis-orders Of Thyroid and other En-docrine Glands	244 244.9 244.8 244.2 255.41 255.42 255.5 255.4 255.2 259.51 255.13 259.2 243 255 253.5 259.4 255.11 242.2 240.0 241 240 240.9 241.0 242.20 242.21 241.9 241.1 253.3 704.1 255.12 255.10 255.14 246.9 246.8 246.3 246.1 246 255 255.3 255.9 255.8 255.6 255 252.8 252.9 252.0 252.1 252.01 252.02 252 252.08 259.52 253.4 253.6 253.1 253.0 253.2 253 253.9 253.8 242.3 242.1 242.20 242.4 242.9 242.8 242.11 242.40 242.41 376.32 242 242.31 242.30 242.91 242.90 242.80 242.81 250.32 250.20 250.60 250.20 250.30 250.22 250.42 250.40 250.02 250.00 250.10 250.12 250.82 250.80 250.90 250.92 250.50 250.52 250.72 250.60 250.70 250.62 362.05 362.01 362.07 362.06 362.0 357.2 362.03 250.7 362.02 250.8 250.9 362.04 250.2 250.3 250 250.1 250.6 250.0 250.4 250.5 259 259.3 259.50 259.9 259.8 258.9 258.72
Vaccination etc.	Health Ser-vices etc.	V01-V91

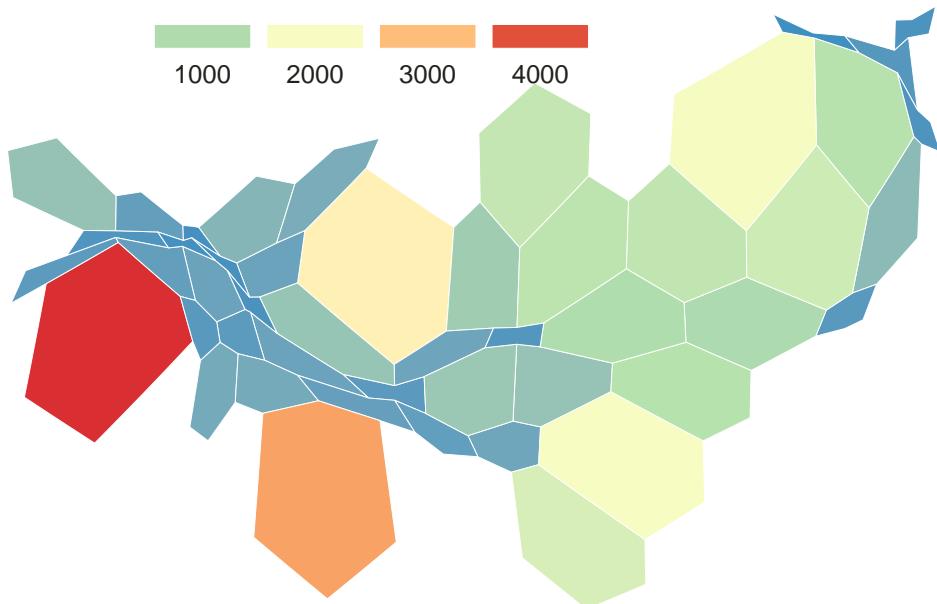
† Categories inferred to be important for risk modulation are highlighted. Continued on next page

Immunologic Diseases related to dysregulation of the Immune system	<p>580.81 580.89 580.4 580.0 580.8 580.9 461 461.8 461.9 461.0 461.1 461.2 461.3 477.9 477.0 477.2 477.4 477.8 493.22 493.20          493.90 493.91 493.12 493.11 493.10 493.00 493.01 493.02 493.92 493.9 493.8 493.81 493.82 493.2 493.1 493.0 493.21 493.21 493.20          708.5 708 708.1 708.2 708.3 708.8 708.9 708.0 692 693 691 692.81 692.83 692.84 692.89 373.32 691.8 693.9 692.0 692.1 692.2          692.3 692.4 693.8 692.6 692.8 692.9 693.0 279.41 258.1 733.01 711.29 711.28 711.25 711.2 711.27 711.26 711.21 711.20 711.23          711.22 136.1 711.24 579 582.4 582.0 582.1 582.8 582.9 582.8 582.9 473.9 473.8 473.1 473.0 473.3 473.2 475.0          474.2 474.0 474.1 474.8 474.9 474.10 474.11 474.12 474.475 474.02 474.01 474.00 279.06 279.04 288.01 288.02 595.2 595.9 595.1          597 597.80 595.3 595.81 597.89 597.0 595.0 597.8 710.3 710.4 279.11 530.12 530.1 530.10 530.19 379.09 360.04 360.03 360.02          360.01 360.00 379.01 379.01 379.03 379.02 379.05 379.04 379.07 379.06 363.20 363.21 363.22 379.92 379.93 379.91 364.2          364.3 364.0 364.1 364.10 364.01 364.00 364.05 364.11 379.0 372.31 372.33 370 373 372 360.0 360.1 373.33 373.31 376.12          376.11 376.10 372.14 372.13 372.12 372.11 372.10 370.64 370.60 370.61 370.32 373.02 373.00 373.11 373.12 373.13 372.1          370.5 370.6 370.9 370.8 363.1 363.0 373.01 363.2 363.0 363.01 363.05 363.07 363.04 363.05 363.08 370.59 370.50          370.52 370.55 370.54 364.02 376.1 364.21 364.23 364.22 364.24 363.11 363.10 363.13 363.12 363.15 363.14 370.62 370.63          360.11 360.12 364.04 360.14 373.8 373.9 373.2 373.3 373.0 373.1 277.31 728.71 729.4 477.1 558.3 693.1 692.5 558.2 535.01          535.00 535.61 535.60 535.20 535.20 535.41 535.40 535.2 535.0 535.1 535.6 535.7 535.4 535.5 535.10 535.11 535.50 535.51 535.50          535.70 535.71 288.2 242.0 242.00 242.01 289.6 279.05 555.0 555.1 555.2 564.1 555.9 279.01 279.02 721.1 720.81 720.0 720.1          721.4 720.8 720.9 721.41 721.42 720 721.91 720.89 446.1 710.0 373.34 695.4 457.9 457.8 457.2 457.1 457.0 457 289.1 289.2          289.3 683 683.0 341.1 341.0 340 341.9 358.0 358.0 358.01 279.13 716.48 716.49 716.42 716.43 716.40 716.41 716.46 716.47          716.44 716.45 716.52 716.88 716.8 716.38 716.89 026.1 716.86 716.87 716.84 716.85 716.37 716 711.83 003.23 716.58 716.80          716.51 716.50 716.53 716.81 716.55 716.54 716.57 716.56 716.5 716.4 716.36 716.6 716.1 716.0 711.8 711.9 711.6 711.7 711.4          711.5 716.9 711.3 711.1 711.72 711.73 711.70 711.71 711.76 711.77 711.74 711.75 716.59 711.78 711.79 056.71 716.64 716.65          716.66 716.67 716.60 716.61 716.62 716.63 716.68 716.8 719.3 036.82 711.61 711.60 711.63 711.62 711.65 711.64 711.67          711.66 711.69 711.68 716.29 716.82 711.58 711.59 711.54 711.55 711.56 711.57 711.51 711.50 711.51 711.52 711.53 716.39 711.38          711.39 711.36 711.37 711.34 711.35 711.32 711.33 711.30 711.31 716.06 716.07 716.04 716.05 716.02 716.03 716.00 716.01          716.35 716.08 716.09 716.34 716.33 716.32 711.49 711.48 716.31 711.43 711.42 711.41 711.40 711.47 711.46 711.45 711.44          716.99 715.99 715.22 735.2 715.09 715.21 715.20 715.23 715.08 715.25 715.24 715.27 715.26 715.28 715.26 715.25 715.26 715.27          715.28 715.29 715.30 715.31 715.32 715.33 715.34 715.35 715.36 715.37 715.38 715.39 715.30 715.31 715.36 715.37 715.38          715.39 715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49          715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40          715.41 715.42 715.43 715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41          715.42 715.43 715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42          715.43 715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43          715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44          715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45          715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46          715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47          715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47 715.48          715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45          715.46 715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46          715.47 715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47          715.48 715.49 715.40 715.41 715.46 715.47 715.48 715.49 715.40 715.41 715.42 715.43 715.44 715.45 715.46 715.47 715.48  </p>
--	---

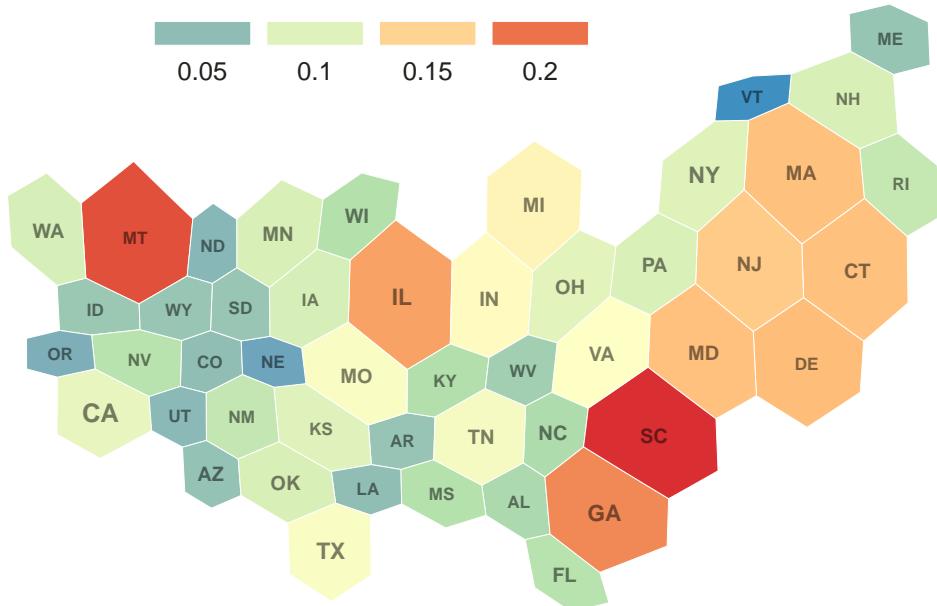
Musculoskeletal	Developmental
Con-genital muscu-loskeletal anomalies	756.52 756.53 733.03 733.02 733.0 733.00 733.09 737.43 737.42 737.41 737.40 754.2 737.21 737.20 737.22 737.29 737.4 737.2 737.3 737.0 737.1 737.8 737.9 737 737.10 737.34 737.32 737.30 737.31 737.19 737.39 737.12 359.7 359.71 359.89 359.79 359.81 359.8 359.9 359.6 359.24 359.4 359.5 359.3 736.9 736.8 735 736.89 736 756.83 756.55 732.9 732.8 732.7 732.6 732.5 732.4 732.3 732.2 732.1 732.0 732 729.31 721.00 722.2 729.30 723.6 729.39 722.70 722.71 722.72 722.73 729.3 721.90 722.92 722.93 722.90 722.91 722.30 722.31 722.32 724.9 724.7 722.10 722.11 722.52 722.39 724.8 722.51 722.4 722.5 722.6 722.7 722.8 722.0 722.1 723.7 722.3 723.9 723.8 722.8 722.9 721.3 721.2 721.0 721.7 721.6 721.5 724.71 721.9 721.8 724 722 723 724.70 721 723.5 724.79 722.81 722.80 722.83 722.82 359 359.1 550.02 550.03 550.00 550.01 553.29 551.55 550 550.1 550.0 550.29 551.0 550.01 553.8 553.9 553.2 552.2 552.1 552.0 553.2 553.3 553.0 553.1 550.9 552.20 552.21 551.8 551.9 550.1 550.0 550.29 551.0 550.01 552.02 551.3 550.11 550.10 550.13 550.12 551.21 552.00 553.21 551.29 553.20 551.01 551.00 551.03 551.02 551.20 550.91 550.90 550.93 550.92 551.1 550.11 550.13 550.12 551.21 552.00 553.21 551.29 553.20 551.01 551.00 551.03 551.02 551.20 550.91 524.0 524.1 524.19 524.6 524.11 525.2 524.05 524.51 526.52 524.10 524.09 524.04 524.02 524.69 524.01 524.00 524.03 524.64 524.63 524.62 524.61 524.60 526.89 524.07 526.81 524.06 526.69 526.61 526.63 526.62 525.20 525.21 525.22 525.23 525.24 525.25 525.26 756.56 726.64 726.69 726.65 728.79 726.63 726.62 726.61 726.60 728.19 735.0 728.10 728.11 728.12 728.13 729.89 726.90 729.8 726.30 726.31 728.8 728.9 728.7 729.1 728.5 728.2 728.3 728.1 726.70 726.71 726.72 726.19 726.32 726.79 726.10 726.33 726.0 726.1 726.2 726.3 726.4 726.5 726.6 726.7 726.8 726.9 727.1 728.4 728.89 728.88 728.87 728.85 728.84 728.83 728.81 726.39 726 728 359.22 359.21 717.89 733.13 717.81 717.83 717.82 717.85 717.84 736.75 734.0 736.76 736.71 736.70 736.73 736.72 736.79 738.5 738.4 738.7 738.6 738.1 738.0 738.3 738.2 738.9 738.8 717 717.8 736.41 736.42 717.0 717.1 717.2 717.3 717.4 717.5 717.6 717.7 717.8 717.9 717.10 718.41 717.82 718.43 717.84 718.45 717.86 718.47 718.48 718.49 733.8 718.4 735.4 735.5 733.98 735.1 735.3 733.94 733.95 733.96 733.97 735.8 735.9 730.3 733.93 730.37 730.36 730.35 730.34 730.33 730.32 730.31 730.30 730.31 730.39 730.38 726.12 726.11 726.11 726.12 736.22 733.82 736.20 736.21 738.11 738.10 738.12 718.39 718.13 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.1 718.02 736.03 736.09 733.15 717.49 733.16 717.41 717.40 717.43 717.42 729.81 729.82 733.10 736.81 736.3 736.2 736.1 733.11 736.0 736.7 736.6 736.5 717.9 736.4 733.12 755.14 754.31 754.30 754.31 754.33 754.32 754.33 755.30 755.31 755.38 755.39 756.79 756.73 756.72 755.52 755.53 755.55 755.59 755.36 755.37 755.34 755.35 755.32 755.33 755.30 755.31 755.38 755.39 756.79 756.73 756.72 756.71 756.70 755.25 755.24 755.27 755.26 755.21 755.20 755.23 755.22 755.65 755.29 755.28 755.64 756.82 756.81 754.51 754.50 754.81 754.82 755.10 755.11 754.12 755.13 754.62 754.60 754.61 754.69 754.3 754.7 754.6 754.5 754.4 728.6 756.3 754.71 754.70 756.7 756.6 756.6 756.9 754.79 755.02 755.01 755.00 755.40 754.41 754.42 754.43 754.44 755.6 755.4 755.5 755.2 755.3 755.0 755.1 755.8 755.9 755.69 754.59 755.61 755.60 755.63 755.62 754.53 754.52 755.67 755.66 719.10 719.11 719.12 718.65 719.14 719.15 719.80 719.17 719.18 719.19 719.18 718.3 718.2 718.1 719.13 718.17 718.7 718.6 718.5 718 718.9 719.88 719.16 718.93 718.92 718.91 718.90 718.97 718.96 718.95 718.94 718.93 718.13 718.12 718.11 718.10 718.17 718.15 718.14 718.19 718.18 718.75 718.74 718.77 718.76 718.71 718.70 719 719.08 719.07 719.06 719.05 719.04 719.03 719.02 719.01 719.00 719.98 719.99 718.0 719.90 719.91 719.92 719.93 719.94 719.95 719.96 719.97 719.78 719.79 719.70 719.76 719.77 719.75 719.6 719.7 719.4 719.5 719.2 719.0 719.1 718.9 718.9 719.8 719.9 719.87 719.86 719.85 719.84 719.83 719.82 719.81 719.80 719.89 719.88 718.57 719.60 719.63 719.69 719.68 719.62 718.59 718.58 719.61 718.56 718.55 718.54 719.65 719.64 719.67 719.66 718.53 718.51 718.50 718.28 718.29 718.22 718.23 718.20 718.21 718.26 718.27 718.24 718.25 718.25 718.39 718.38 718.31 718.30 718.33 718.32 718.35 718.34 718.37 718.36 719.54 719.55 719.56 719.57 719.50 719.51 719.52 719.53 719.58 719.59 718.52 718.73 719.09 718.72 719.43 719.42 719.41 719.40 719.47 719.46 719.45 719.44 719.49 719.48 719.25 719.24 719.27 719.26 719.26 719.21 719.20 719.23 719.22 718.79 719.29 719.28 718.78 718.88 718.89 718.84 718.85 718.86 718.87 718.86 718.87 718.80 718.81 718.82 718.83 718.04 718.05 718.07 718.00 718.01 718.02 718.03 718.08 718.09 733.2 733.3 733.7 733.4 733.5 733.6 733.9 731.1 731.2 731.3 731.8 733.90 733.91 733.92 726.91 731.71 756.50 726.73 733.44 733.29 733.49 733.21 733.20 733.45 733.22 733.33 733.42 733.41 733.40 736.51 359.29 756.59 731 756.54 359.23 724.03 724.02 724.0 724.00 724.09 724.01 723.0 727.67 727.42 727.43 727.42 727.40 727.41 727.68 727.69 727.64 727.65 727.66 727.49 727.60 727.61 727.62 727.63 727.09 727.06 727.04 727.05 727.02 727.03 727.00 727.01 727.82 727.83 727.81 727.89 727.9 727.8 727.5 727.4 727.6 727.0 727.3 727.2 727.2 727.51 727.50 727.59 727
Con-genital anomalies (Non-overlap. with musku-loskeletal)	755.55 743.45 743.12 743.11 743.10 743.06 743.00 743.03 743.1 743.44 743.41 743.42 743.43 743.22 743.2 743.22 743.2 743.20 743.21 758.4 745.1 745.0 745.3 745.2 745.5 747.5 746.09 747 745.4 745.8 746.02 746.01 746.00 745.9 745 747.89 745.7 745.6 747.82 747.83 745.60 745.61 745.69 747.21 746.6 746.7 746.4 746.5 746.2 746.3 746.0 746.1 747.9 747.8 746.8 746.9 746.87 746.86 746.85 746.84 746.83 746.82 746.81 746.89 745.19 747.11 745.11 745.10 745.12 756.4 744.4 743.63 743.61 743.66 744.1 743.64 743.65 743.69 744.8 744.9 744.0 745.1 745.0 748.1 748.0 744 744.49 744.2 744.4 744.3 744.5 743.6 744.29 744.47 744.42 744.89 744.09 744.81 744.83 744.82 744.05 744.04 744.01 744.00 744.03 744.02 744.02 744.41 744.43 744.24 744.23 744.22 744.22 744.21 744.46 743.35 742.51 756.2 756.19 742.53 756.11 756.10 756.13 756.12 756.15 756.14 758.31 315.4 315.5 315.8 315.3 315.9 758 758.2 752.4 752.1 752.0 752.3 752.2 752.4 752.43 752.40 752.41 752.46 752.47 752.44 752.45 749.25 752.36 752.35 752.34 752.33 752.32 752.31 752.39 752.19 752.11 752.10 759.83 751.61 751.7 751.6 751.5 751.0 751.6 751.5 751.4 751.3 751.2 750.2 50.0 520.0 520.1 520.2 756 752.7 758.7 756.16 752.5 752.6 752.51 752.52 752.69 752.64 752.65 752.61 752.62 752.63 740 740.0 740.2 741.9 740.1 741.02 741.0 741.00 741.01 741.01 741.02 741.91 741.90 741.93 741.92 741.03 756.17 758.5 758.39 758.81 758.78 743.49 743.49 743.47 743.9 743.8 743.5 743.3 743.59 743.53 743.53 743.52 743.51 743.57 743.56 743.55 743.54 743.39 743.34 743.37 743.36 743.31 743.30 743.33 743.32 750 751.4 751.5 751.2 751.0 751.1 750.4 750.9 750.8 750.7 750.6 750.24 751.9 750.3 751.8 750.5 753.0 753.1 753.2 753.3 753.4 753.5 753.6 753.7 753.8 753.9 753.10 753.11 753.12 753.13 753.14 753.15 753.16 753.17 753.18 753.19 752.81 753.12 753.10 753.11 753.15 753.17 752.89 753 752.75 753.23 753.22 758.99 753.0 753.0 753.0 753.0 753.0 753.0 753.0 753.0 753.0 742.59 742.2 742.3 742.1 742.4 742.5 742.9 742.5 742.7 742.4 742.7 742.3 742.4 742.7 742.4 742.5 742.6 742.7 742.8 742.9 742.8 742.7 520.9 750.9 750.25 750.26 750.27 750.25 750.22 750.23 750.25 750.21 752.42 752.43 752.44 752.45 752.46 752.47 752.48 752.49 752.48 752.47 749.03 749.09 749.01 754.39 750.16 750.15 750.13 750.12 750.11 750.10 750.15 750.19 749.14 749.11 749.10 749.13 749.11 749.12 749.11 524.82 524.89 749.20 749.21 749.22 749.23 749.24 749.25 524.34 524.35 524.36 524.37 524.30 524.31 524.32 524.33 750.2 750.1 750.0 524.56 524.57 524.54 524.55 524.52 524.53 524.54 524.50 749.0 524.59 749.1 749.0 749.2 524.23 524.22 524.21 524.20 524.27 524.26 524.25 524.24 524.29 524.28 758.1 759.81 259.0 759.81 759.81 759.81 759.81 759.81 759.81 759.81 759.81 759.81 748.5 748.6 758.33 759.81 759.81 759.81 313.9 313.9 313.9 309.7 307.0 307.6 313.8 313.8 313.9 759.9 759.9 759.9 759.0 759.1 759.7 759.4 747.40 747.41 747.64 747.29 747.62 747.63 747.60 747.61 747.22 747.49 747.20 747.81 747.69 747.3 747.2 747.1 747.0 747.6 747.4 747.42 747.10 758.32

<sup>†</sup> Categories inferred to be important for risk modulation are highlighted.

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



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

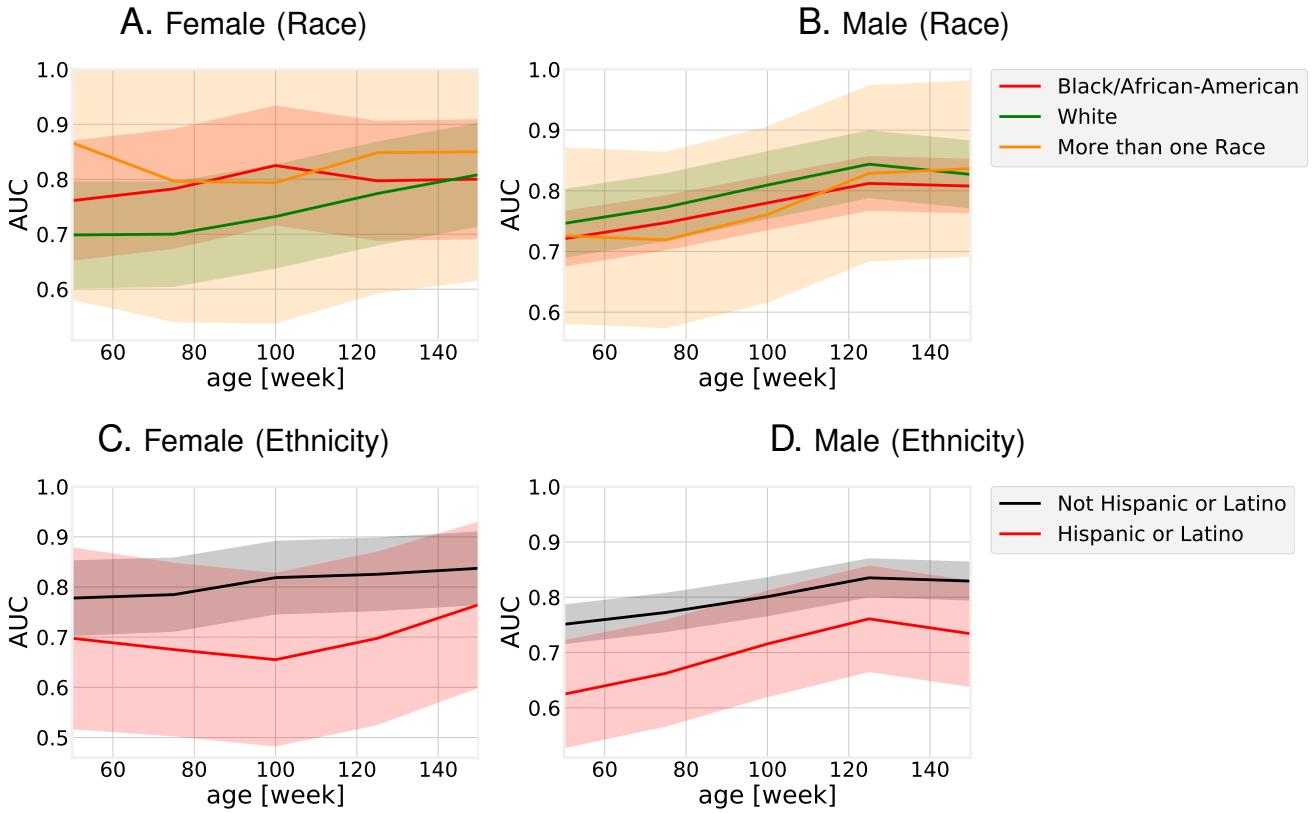


SI-Fig. 1. **ASD Occurrence Patterns** Panel A illustrates the spatial distribution of ASD insurance claims, and panel B shows the same data after population normalization, illustrating the relatively small demographic skew to ASD prevalence within the general population with access to medical insurance, which is consistent with the suggestion that prevalence variation might be linked to regional and socioeconomic disparities in access to services (89).

### I. DETAILED MATHEMATICAL APPROACH

#### A. Time-series Modeling of Diagnostic History

Individual diagnostic histories can have long-term memory (96), implying that the order, frequency, and comorbid interactions between diseases are important for assessing the future risk of our target phenotype. We analyze 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



**SI-Fig. 2. Effect of Race and Ethnicity on Predictive Performance with 95% Confidence Bounds from the UCM dataset.** Panels A and B show the variation of AUC achieved in out-of-sample data in three race-based population groups (White, African-American and multi-racial). We find no significant differences. Panels C and D show the performance in Hispanic vs non-Hispanic sub-populations. We find that children with Hispanic background have a lower AUC, but the differences are not significant. Other races/ethnicities were not considered due to lack of sufficient data.

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) (33). 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 X). 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.

### B. Step 1: Partitioning The Human Disease Spectrum

We begin by partitioning the human disease spectrum into 17 non-overlapping categories. Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9) (See Table SI-I 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) (73) 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 1a in the main text, 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 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.

### C. 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. 3 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. (1)) corresponding to each sex, disease category, and cohort-type (See Section X 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 (36, 37) between probability distributions to a divergence  $\mathcal{D}_{\text{KL}}(G||H)$  between ergodic stationary categorical stochastic processes (38)  $G, H$  as:

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

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. 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 X, Theorem 6 and 7):

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

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

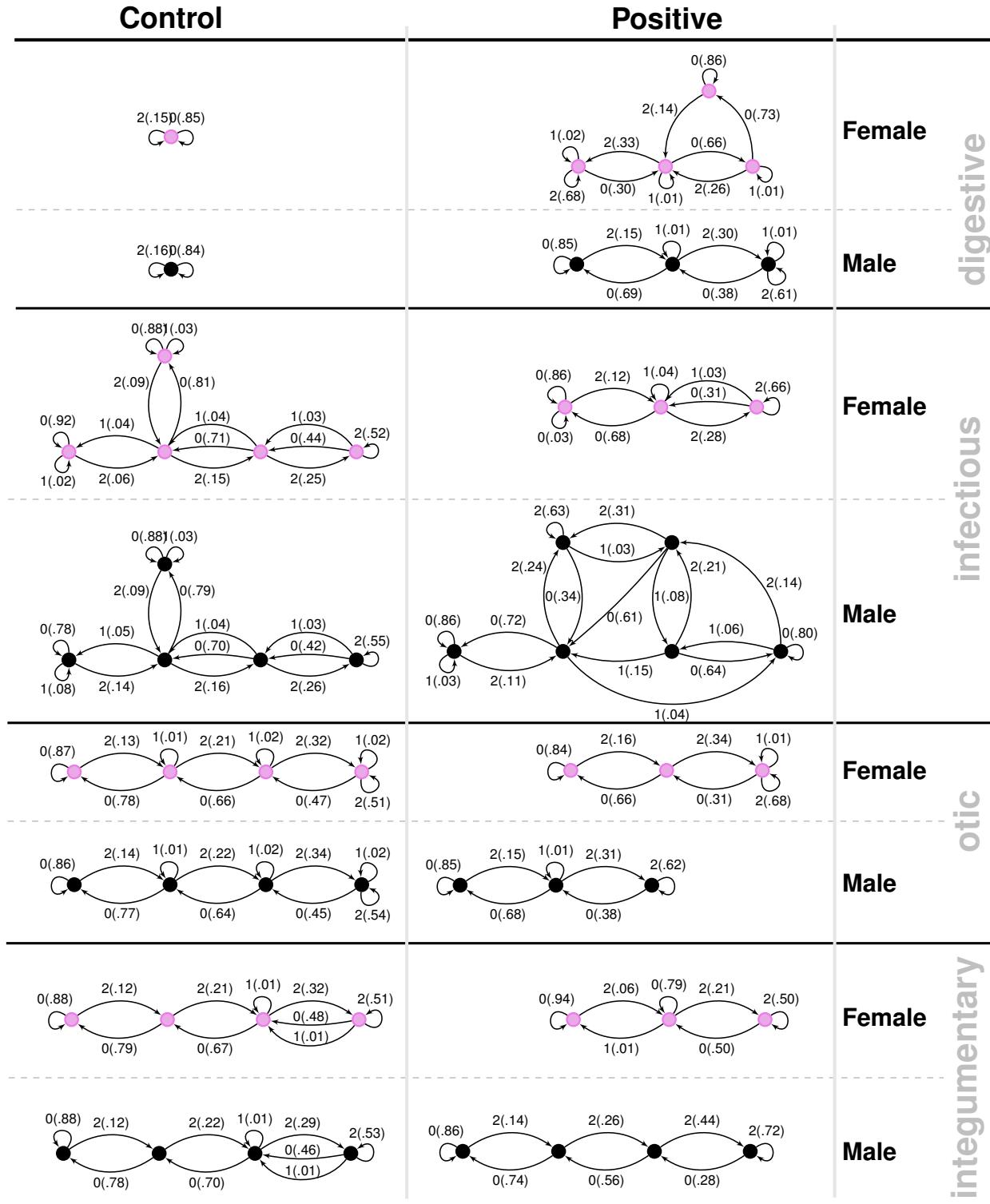
where  $\mathcal{H}(\cdot)$  is the entropy rate of a process (36). Importantly, Eq. (4) shows that the computed likelihood has an additional non-negative contribution from the divergence term when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with 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  $G_+^j, G_0^j$  respectively, we can compute the *sequence likelihood defect* (SLD,  $\Delta^j$ ) as:

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

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

### D. 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 (85). The complete list of 165 features used is provided in Tab. Ib in the main text.



SI-Fig. 3. 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 similar nature. In case of infectious diseases, the model gets more complex for males, and less complex for females, suggesting distinct sex-specific responses associated with high ASD risk.

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 PFSA models (*classical supervised learning*), and we finally execute out-of-sample validation on the test set. Fig. 1B in the main text shows the top 15 features ranked in order of their relative importance (relative loss in performance when dropped out of the analysis).

## II. COMPARISON WITH STATE OF THE ART OFF-THE-SHELF ML ALGORITHMS

SeaGreen4

Model: "sequential\_4"

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 150)	180600
dense_5 (Dense)	(None, 32)	4832
dense_6 (Dense)	(None, 1)	33
<hr/>		
Total params: 185,465		
Trainable params: 185,465		
Non-trainable params: 0		

SI-Fig. 4. The simplest LSTM investigated as a baseline. More complex models have significantly larger number of trainable parameters (1 to 10 Million). In contrast our pipeline has 13,744 trainable parameters.

Off the shelf algorithms with little or no pre-processing, *i.e.*, using the diagnostic codes themselves are time-stamped categorical features failed to produce clinically relevant performance (See SI-Fig. 5). Classifiers such as random forests (77), and gradient boosters (85) might be penalized due to their inability to take into account long-range temporal information. Since the number of diagnostic codes available per patient is small, recurrent neural network implementations such as LSTM (87) might be suffering from the data sparsity in training. It is possible that the performance of the competing approaches might be improved with extensive tuning or clever feature-engineering.

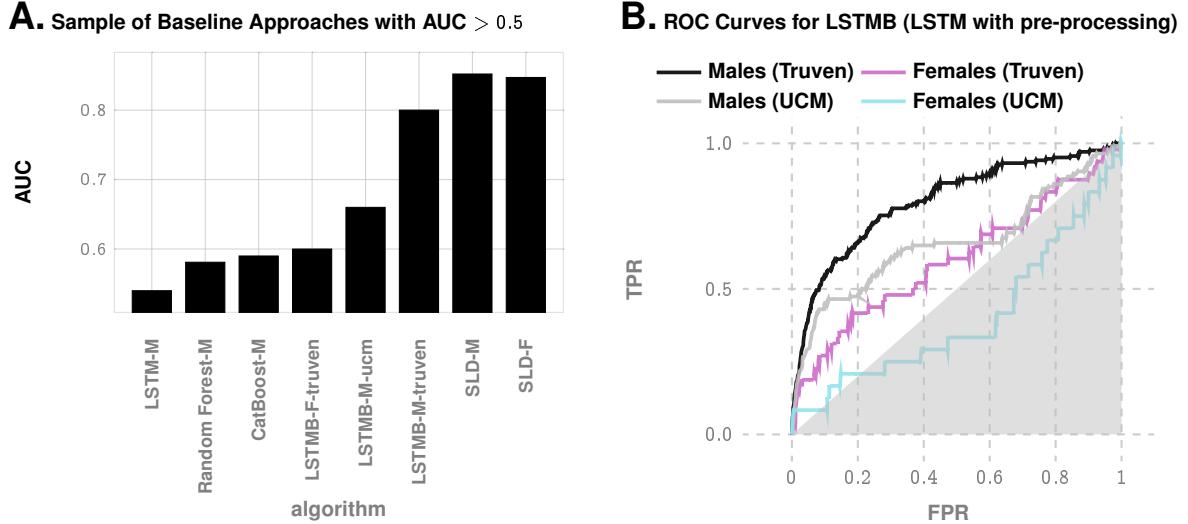
## III. COMPARISON WITH PIPELINE VARIATIONS, FEATURE SUBSETS AND NEURAL NETWORK (NN) POST-PROCESSING

In addition to the naive baseline approaches, we also evaluated the performance achievable with LSTMs (denoted as LSTMb in SI-Fig. 5) that use identical pre-processing as our pipeline, *i.e.*, representation of diagnostic histories as trinary sequences in 18 categories for each patient, and achieved  $\sim 80\%$  AUC at 150 weeks for males in the Truven database (compared to  $> 85\%$  for our approach). However, the performances drop significantly when the number of positive samples is reduced, yielding an AUC of 66% on the UCM dataset for males, 60% for females on the Truven dataset, and a worse-than-random 40% on the UCM dataset respectively (See Fig. 5).

Much better results were obtained when we compared our optimized pipeline to pipelines that use only a subset of our features: namely, the ones that use only features derived from sequence statistics and exclude the ones derived from learning PFSAs (recall that PFSAs are special HMMs we learn using our novel algorithms) from the disease categories as described in Methods in the main text, or using only the PFSA-based SLD features, or using simply the density of diagnostic codes (See Fig. 6, panel D). In all these cases, our pipeline has a clearly demonstrable advantage (See Fig. 6, panel D) that is stable across databases, under reductions in sample sizes, and in balanced re-sampling experiments (See Fig. 6, panel C).

While it is difficult to explain the exact source of a modeling framework's performance, we can point to the following advantages that our approach has over existing techniques:

- 1) **Purely Classification Algorithms With No Pre-processing Do not Do well.** Pure classifiers such as random forests, gradient boosters, etc. are not time series modeling frameworks, and might not capture stochastic temporal patterns well. While features are not certainly assumed to be independent in these algorithms, it is problematic to learn patterns that do not appear at fixed time points in the diagnostic history.
- 2) **Lower Sample Complexity Compared to Deep Learning Frameworks.** Compared to LSTMs and RNNs, we are able to capture stochastic behavior with more compact models, which results in better sample complexity. In other words, if we have less data, our models do better, because we estimate fewer parameters.
- 3) **Designed Bottom-up for Learning Stochastic Processes.** It is easily demonstrated that LSTMs and RNNs, while good models of complicated time series in many cases, do not work well for data that are generated by stochastic processes, *i.e.* are sample paths of a hidden process.



SI-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. LSTM-B denotes LSTM with identical pre-processing as in our pipeline (instead of using raw diagnostic codes). We get much better performance with LSTM-B 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.

Thus, there are two related issues: NN models generally have a relatively large number of nominal free parameters, requiring a substantial amount of data to train. In our case, we indeed have a large number of control examples, but a relatively smaller number of positive examples (children who get diagnosed with autism eventually). The amount of training data, even with our large patient database, is not enough to train anything but the simplest of the NN architectures, and even then the number of parameters for such NN models is substantially larger compared to our architecture. This is shown in our example (See SI-Fig. 4), where one of the LSTM models we investigate (the simplest one we tested) has 185,465 parameters, whereas our complete end to end pipeline has  $7025+6719=13,744$  trainable parameters (an order of magnitude less, where the first contribution is from inferred PFSA models, the second from the ensemble of decision trees in the final gradient boosting model). Additionally, the number of parameters for deep learning models rapidly increase to millions or tens of millions as the models get more complicated.

Also note that our framework is adaptive (See examples of PFSA models inferred in SI-Fig. 3), and the network structures of the inferred PFSA models is inferred from data, and not fixed a priori; it produces more states where it needs, thus adapting the model resolution to more complex contexts, and obtaining significantly more compact yet effective representations. We must however add the caveat that with sufficient experience it might be possible to devise a specific NN architecture with similar performance to our framework; however, in our approach such artful insight is unnecessary.

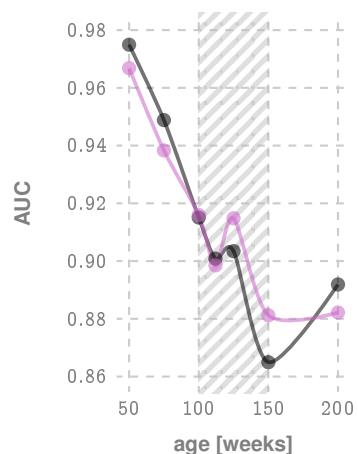
We also found that NN models generalize poorly across databases for the problem at hand; the performance is relatively better in the Truven dataset (See SI-Fig. 5), but worse in an independent database.

#### A. Feature Subset Evaluations & Code Density As A Feature

With regards to Fig. 6, panel D, we note that the PFSA based features by themselves are comparable to those engineered manually from sequence statistics (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 main text Table Ib in the main text for details), but the combined runs produce significantly superior results. Also, it is interesting to note that simply using the density of diagnostic codes in a child's history is quite predictive of future ASD diagnosis, with the AUC from using just the density of codes as a feature rising to over 75% in the Truven database at 150 weeks. However, it does not have stable predictive performance across databases, and is also the least performing predictor. We did not include code density in our combined feature set, since it has no effect once the rest of the features are combined.

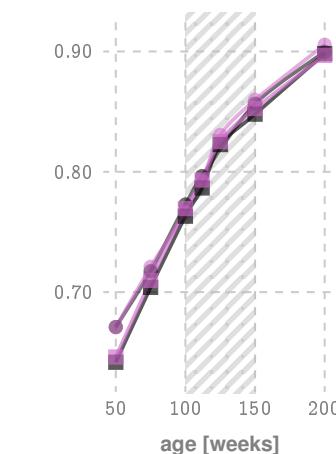
### A. Disambiguation of Autism Diagnosis from Other Psych. Phenotypes

—●— Males (Truven)  
—●— Females (Truven)



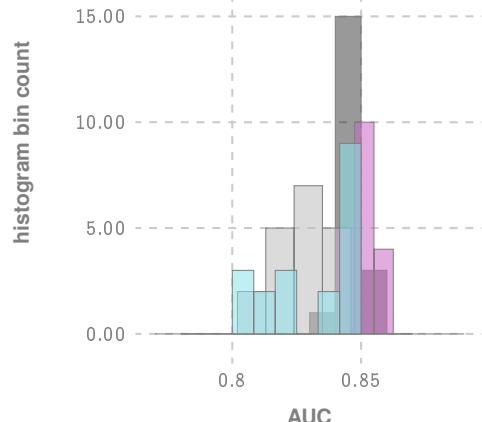
### B. Comparison of Performance with One vs Two ASD Diagnostic Codes

—●— Males (Truven, two codes)  
—●— Females (Truven, two codes)  
—■— Males (Truven, one code)  
—■— Females (Truven, one code)

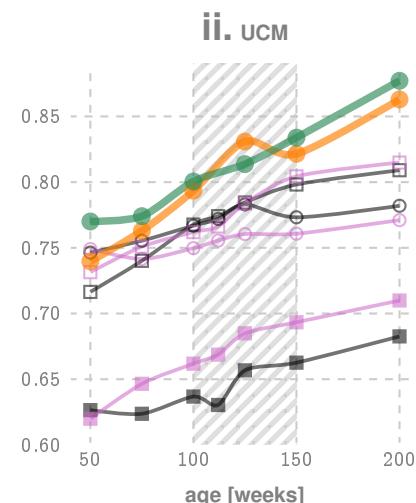
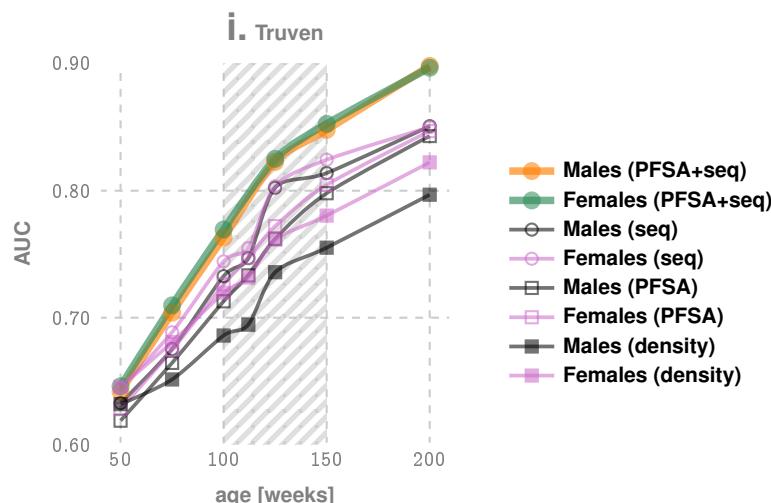


### C. AUC Distribution with Matched Control & Treatment Population Sizes

—●— Males (Truven)  
—●— Females (Truven)  
—■— Males (UCM)  
—■— Females (UCM)



### D. Comparison of Performance with Different Feature Categories (Only PFSA based features, Only Sequence-statistics based features, only Code-density, and PFSA + Sequence-statistics features combined)

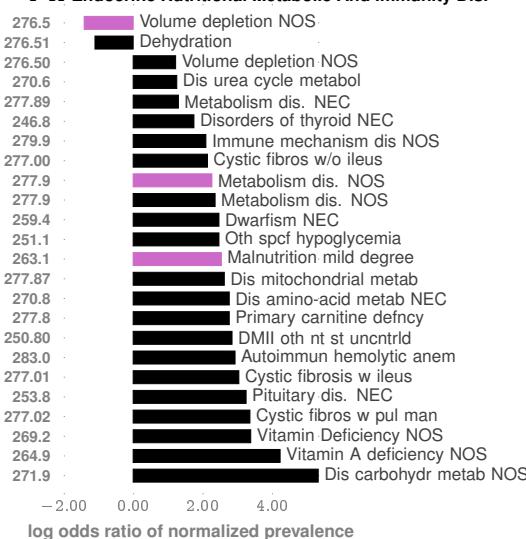


SI-Fig. 6. Evaluations of Feature Subsets, Class Imbalance, Code Density, Coding Uncertainty, & Disambiguation from Other Psychiatric Phenotypes. Panel A illustrates that the pipeline performance where the control group is restricted to children to have at least one psychiatric phenotype other than ASD. 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.

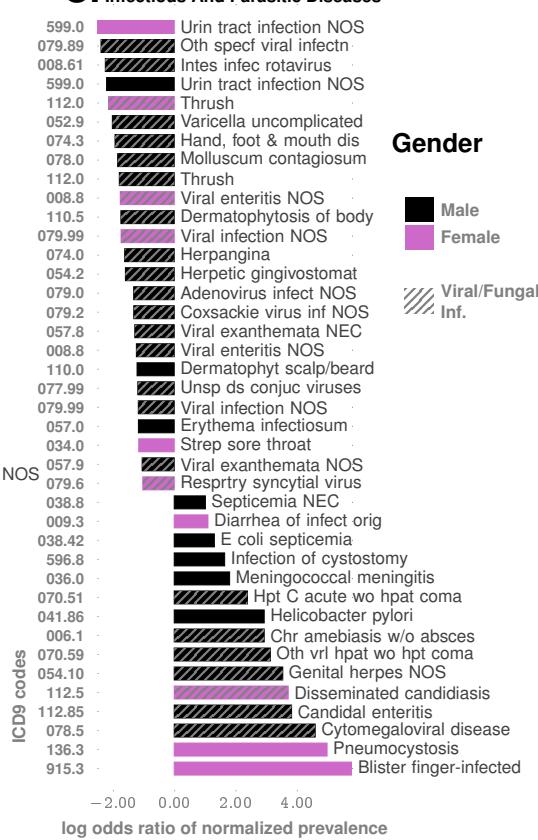
#### IV. THRESHOLD SELECTION ON ROC CURVE

Once the ROC curve has been computed, we must choose a decision threshold to trade-off true positive rate and false positive rate. In situations where the number of negatives vastly outnumber the number of positives (which is the case in our problem), it is better to base this trade-off on a measure that is independent of the

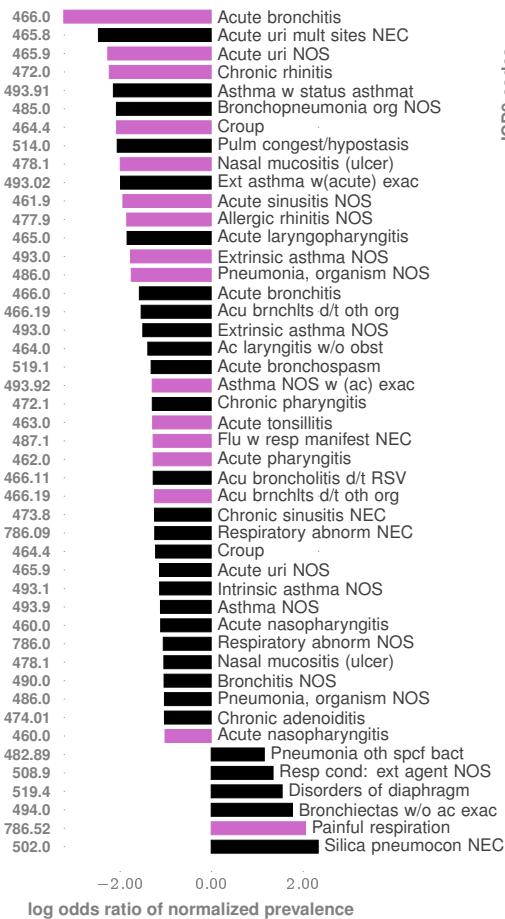
### A. Endocrine Nutritional Metabolic And Immunity Dis.



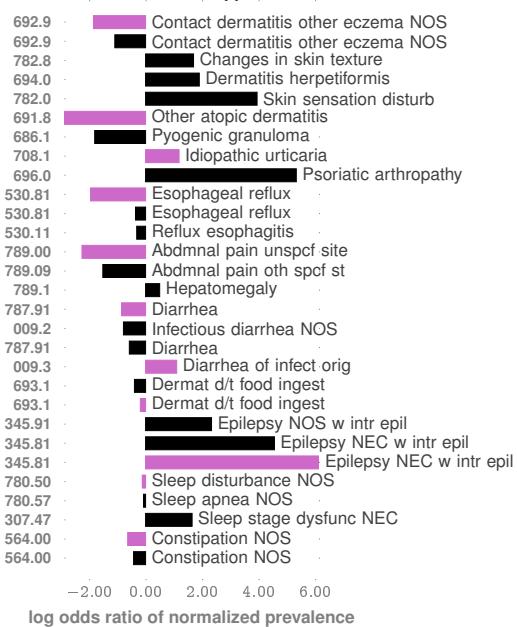
### C. Infectious And Parasitic Diseases



### B. Respiratory Disorders



### D. Similar Dis. with Opposed Association



SI-Fig. 7. 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).

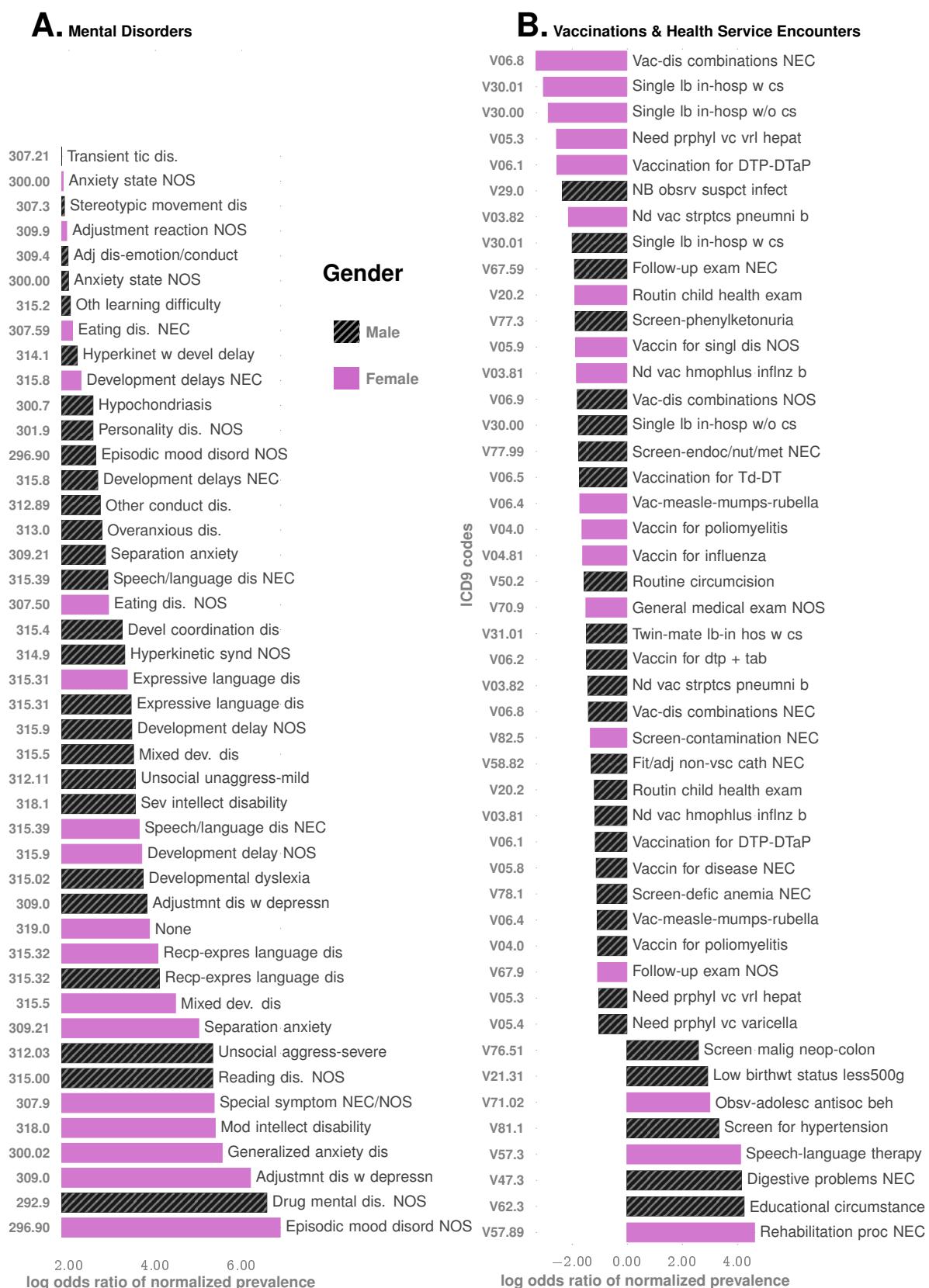
number of true negatives. The two popular measures considered in the literature are accuracy and the F1-score:

$$\text{accuracy} = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \quad (6)$$

$$F1 = \frac{2t_p}{2t_p + f_p + f_n} \quad (7)$$

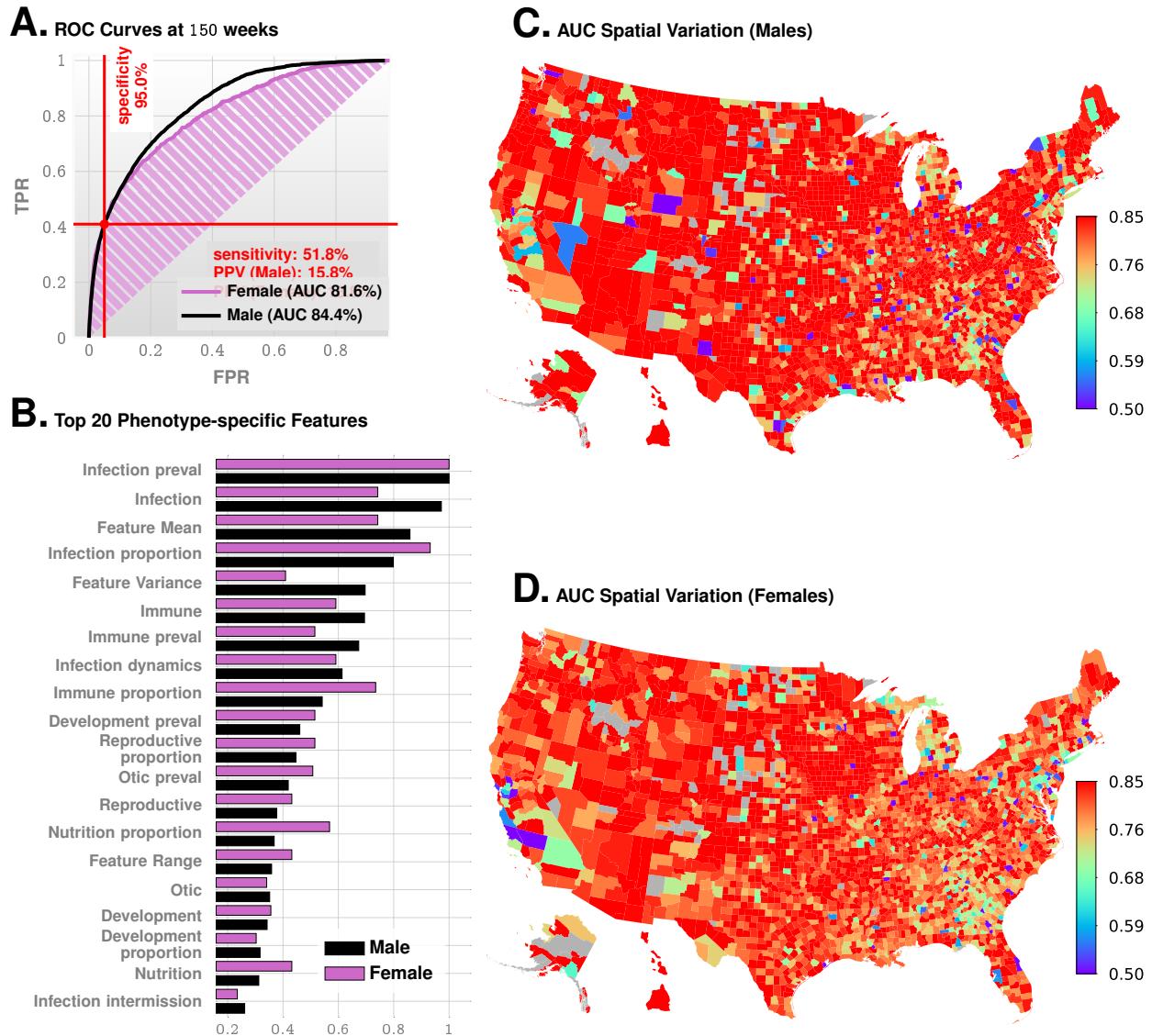
The F1-score is the same as accuracy where the number of true negatives is the same as the number of true positives, thus partially correcting for the class imbalance.

The selection of the threshold may also be dictated by the current practice of ensuring high specificities in screening tests. Thus, the most relevant clinically operating point is probably the one corresponding to 95%



SI-Fig. 8. **Co-morbidity Patterns** for mental disorders, vaccinations and health-service encounters.

specificity, which is highlighted in Fig. 1C in the main text.



SI-Fig. 9. 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 Fig. 8A, the psychiatric codes all increase risk, and the vaccination codes (See Fig. 8B) 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.

## V. NOTE ON RECEIVER OPERATING CHARACTERISTICS (ROC) AND PRECISION-RECALL CURVES

The ROC curve is a plot between the False Positive rate (TPR) and the True Positive Rate (TPR), and the area under the ROC curve (AUC) is often used as a measure of classifier performance. For the same of completeness, we introduce the relevant definitions:

In the following  $P$  denotes the total number of positive samples (number of patients who are eventually diagnosed), and  $N$  denotes the total number of negative samples (number of patients in the control group).

**Definition 1.** *True positive rate, true negative rate, false positive rate, positive predictive value (PPV), and*

**prevalence** ( $\rho$ ) are defined as:

$$TPR = \frac{t_p}{P} = \frac{t_p}{t_p + f_n} \quad (8)$$

$$TNR = \frac{t_n}{N} = \frac{t_n}{t_n + f_p} \quad (9)$$

$$FPR = 1 - TNR \quad (10)$$

$$PPV = \frac{t_p}{t_p + f_p} \quad (11)$$

$$\rho = \frac{P}{N + P} \quad (12)$$

where as before  $t_p, t_n, f_p, f_n$  are true positives, true negatives, false positives, and false negatives respectively.

Note that TPR is also referred to as **recall** or **sensitivity**, and PPV is also referred to as **precision**. True negative rate is also known as **specificity**.

A **precision-recall curve**, or a PPV-sensitivity curve is a plot between PPV and TPR.

Denoting sensitivity by  $s$ , and specificity by  $c$ , it follows that:

$$PPV = \frac{t_p/P}{t_p/P + (f_p/N)(N/P)} = \frac{TPR}{TPR + ((N - t_n)/N)(N/P)} \quad (13)$$

$$\Rightarrow PPV = \frac{s}{s + (1 - c)(\frac{1}{\rho} - 1)} \quad (14)$$

Thus, we note that for a fixed specificity and sensitivity, the PPV depends on prevalence. Indeed, it is clear from the above argument that PPV decreases with decreasing prevalence, and vice versa, if specificity and sensitivity are held constant. Also, if prevalence is limited to 2%, and specificity is held at 95%, then the maximum PPV is limited to:

$$PPV = s/(s + 2.45) \leq 1/3.45 \sim 29\% \quad (15)$$

This shows that for ASD screening, we can hope for a maximum PPV of ~29% at 95% specificity, if the prevalence is stable at around 2%.

Compare this with the PPV of 15.8% (M) and 18.8% (F) that we achieve at 51.8% sensitivity, where the specificity is held at 95% in Fig. 1C in the main text. Note here that M-Chat/F with follow-up has a PPV of 14.6% as reported by the recent CHOP study (3).

## VI. EFFECT OF CLASS IMBALANCE

ROC curves are generally assumed to be robust to class imbalance. Note that if we assume that patient outcomes are independent (which is well-justified in the case of a non-communicable condition, particularly in large databases), then  $t_p$  should scale linearly with the total number of positives  $P$ , implying:

$$TPR = \frac{t_p}{P} = \frac{t'_p}{P'} \quad (16)$$

implying that with different sizes of the set of positive samples (or negative samples), the ROC curve remains unchanged. In particular, note that even if the prevalence is very small (say 0.01%), we cannot cheat to boost the AUC by labeling all predictions as negative, or stating that risk is always zero: in that case, our  $P$  is very small, but our  $t_p = 0$  strictly, implying that our  $TPR = 0$ , thus leading to a zero AUC. We can cheat to boost the accuracy (See the previous section), but not the AUC.

Note that while relative class sizes or imbalance does not affect the ROC (under the assumption that true positives and true negatives scale with the number of positives and negatives), very small absolute sample sizes might still result in poor performance of the model.

We do have significant class imbalance in our datasets. This arises naturally from the low prevalence rate of ASD (small in the sense of comparison of sizes of the control and the positive cohorts). Thus, we validated if the performance of our predictive pipeline remains unchanged by replacing the full control cohort with a random sample of size equal to that of the positive cohort. The results, shown in Fig. 6C, illustrate that class imbalance has no appreciable effect on our pipeline, as far as the AUC metric is considered.

The precision-recall curves do get affected by class imbalance, or the prevalence, as shown by Eq (14). However, in diagnostic analysis, they are important since we are generally less interested in the number of true negatives; the ratio of false positives to the total number of positive recommendations by the algorithm is much more relevant, *i.e.*, the PPV or the precision.

We have used this to our advantage. Note that since the PPV is affected by prevalence, a stratification of the total population with different prevalence in each sub-population suggests the possibility of a conditional choice of the operating point, thus boosting the overall PPV. We describe this approach in the sequel, in Section VIII-A. First, we establish that our pipeline does not suffer from some important pitfalls arising in the work-flows associated with ASD diagnosis, and how the diagnostic codes in Electronic Health Records (EHR) are generated.

## VII. NOTE ON ASD CLINICAL DIAGNOSIS & UNCERTAINTY OF EHR RECORD

With no precise laboratory test for ASD, most families experience the following sequence of events (12, 16, 98): 1) routine screening at 18 and 24 months of age identifies high risk, and is followed by 2) a diagnostic evaluation. The American Academy of Pediatrics (AAP) recommends screening all children for symptoms of ASD at 18 and 24 months of age in their primary care visits (90, 101). However, results of a screening test are not diagnostic (*and hence do not produce an EHR diagnostic code*); they help the primary care provider identify children who are at risk for a diagnosis of ASD and require additional evaluation. The M-CHAT/F is the most studied and widely used tool for screening toddlers for ASD (12, 17).

Unfortunately, children with milder symptoms are harder to screen for. The AAP warns that children with milder symptoms and/or average or above-average intelligence may not be identified with symptoms until school age, when differences in social language or personal rigidities affect function (12).

### A. Diagnostic Evaluations

Once a child is determined to be at risk for a diagnosis of ASD, either by screening or surveillance, a timely referral is needed for clinical diagnostic evaluation (98), which will, on positive identification, assign a clinical diagnosis, and produce an EHR record.

The history of symptoms of ASD presentation in individual patients may be elucidated by questionnaires such as the Social Communication Questionnaire (SCQ), or Social Responsiveness Scale (SRS), or the Autism Diagnostic Interview-Revised (ADI-R) (12). These questionnaires alone are insufficient for making a clinical diagnosis, but provide a structured approach to elicit symptoms. Validated observation tools used to provide structured data to confirm a clinical diagnosis include the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) (83) and the Childhood Autism Rating Scale, Second Edition (CARS-2) (80). Current guidance from the American Academy of Pediatrics (12) notes that no single observation tool is universally appropriate, and that such tools are meant to support the application of the diagnostic criteria informed by history and other data.

At present, the Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS) are considered the “gold standard” tools to enable the diagnosis of ASD (84). The true “gold standard” classification and diagnosis of autism is historically taken to be a multi-disciplinary team (MDT) clinical assessment, including use of the ADOS and ADI-R, as well as other assessments with consensus clinical judgment (84). The MDT clinical diagnosis correct classification rate for ASD is approximately 80.8%. Thus, any individual tool that correctly classifies ASD at a rate of 80 % or over could be considered to be just as accurate as the “gold standard” (84). With ADOS-2 and associated tools verifiably reaching this classification rate, the current APA guidance suggests that individual general pediatricians might hand out initial diagnoses if they are familiar with the relevant DSM diagnostic criteria. This simultaneously raises the prevalence, and the possibility that some diagnostic codes pertaining to ASD in medical history databases could be arising from less restrictive work-flows, and thus might carry more uncertainty.

In our study, we checked if restricting the treatment cohort to children with at least two distinct ASD diagnostic codes in their medical histories instead of one (which significantly reduces the possibility of erroneous coding) changes the performance of the algorithm. The results shown in Fig. 6B illustrate that we have very little change in out-of-sample predictive performance, thus alleviating this concern.

### B. Change In Diagnostic Criteria for ASD, Inclusion of PDD, Asperger, and Disambiguation From Unrelated Psychiatric Phenotypes

The DSM-5 established a single category of ASD to replace the subtypes of autistic disorder, Asperger syndrome, and pervasive developmental disorder not otherwise specified in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR) (12). This justifies our use of diagnostic codes from ICD9 299.X as specification of an ASD diagnosis, and use of GEMS mapping to 299.X for ICD10 codes when we encounter them. Future renditions of our pipeline will use purely ICD-10 specification, which does not change the algorithm, but merely how we input data into it.

It is interesting to note that we would be actually unable to discriminate between those phenotypes effectively for high predictability even if we wanted: in our initial efforts, we found it is very difficult to design a high performing pipeline that recognizes these sub-types separately.

The question then arises as to how well we can discriminate between ASD and other unrelated psychiatric phenotypes. Does our pipeline pick up on any psychiatric conditions, or is it specific to ASD? We directly evaluated this, by restricting the test control cohort to patients with at least one psychiatric code other than ASD. We get very high discrimination reaching AUCs over 90% at 100-125 weeks of age, which establishes that our pipeline is indeed largely specific to ASD.

### C. Performance Comparison With M-CHAT/F

The M-CHAT/F is the most studied and widely used tool for screening toddlers for ASD (12, 17).

Guthrie *et al.* (3) from the Children's Hospital of Philadelphia (CHOP) demonstrate that when applied as a universal screening tool, M-CHAT/F has a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%. 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 sensitivity of M-CHAT/F.

Comparing the performance metrics achieved at different age groups across data sets and sexes for our pipeline (See main text Table IIa in the main text), 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 ( $\approx 112$  weeks). We cannot compare at other operating points due to a lack of M-CHAT/F performance characterization anywhere else.

Apart from standalone performance, our proposed approach has several key advantages: it is clearly immune to parental educational level, and language barriers. Since access to insurance and medical records do get impacted by socio-economic variables, there is the possibility of some indirect impact from the demographic makeup of the training datasets. But overall, diagnostic histories are free from biases that have historically plagued questionnaire-based screens (12). Additionally, while M-CHAT/F is relatively easy and quick to administer, the issue of time and resource commitment cannot be ignored (12). These factors conspire to produce reduced coverage, which in turn casts doubt upon the necessity of universal screening programs despite clear guidance on the contrary from the AAP (3).

Additionally, being functionally independent of the M-CHAT/F, we can take advantage of any population stratification induced by the M-CHAT/F results to significantly boost combined screening performance.

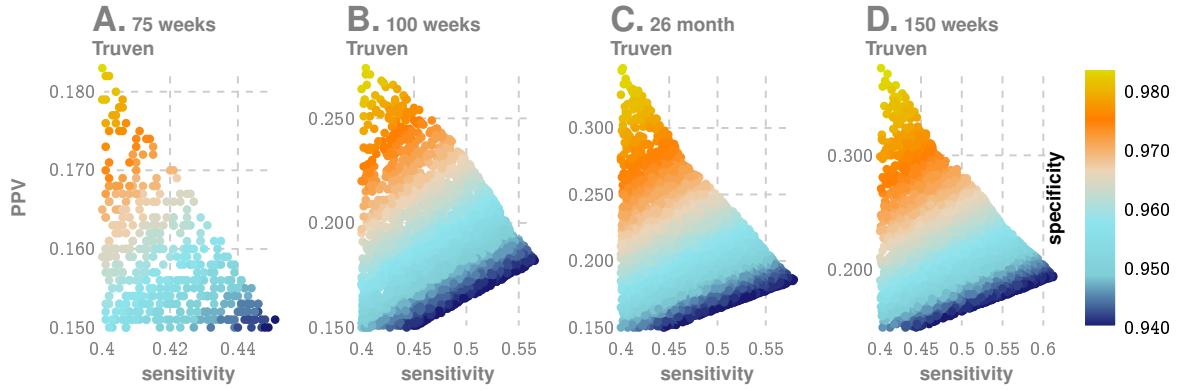
## VIII. IMPROVING WAIT-TIMES FOR DIAGNOSTIC EVALUATIONS BY REDUCING FALSE POSITIVES IN ROUTINE SCREENING

While children with ASD can be diagnosed as toddlers (92, 95) (developmental concerns may show up before the first birthday (75, 94)), the mean age of diagnosis is over 4 years (74). Since a clinical diagnosis of ASD requires the multi-step process described in the previous section, this delay mainly arises from extended wait-times and queues, which ultimately delays entry into early intervention (EI) programs. While time-consuming evaluations (13), cost of care (14), lack of providers (15), lack of comfort in diagnosing by primary care providers (15), and other challenges, are all responsible to varying degrees that culminate in these delays (16), one rather obvious source is the limited PPV of screening tests that are available today. With the PPV of M-CHAT/F being around 14.6%, over 85 out of 100 people flagged for diagnostic evaluation are false positives, leading to wait times that currently range from 3 months to 1 year. To make matters worse, access to care and resources are sparse except near urban centers. For example, only 7% of developmental pediatricians practice in rural areas, and some states do not even have a developmental pediatrician (16, 39).

A key contribution of this work is to be able to significantly reduce the number of false positives without sacrificing specificity, and thus significantly improving wait-times and patient outcomes.

### A. 4D Decision Optimization Using M-CHAT/F Population Stratification To Boost PPV

Assume that there are  $m$  sub-populations such that: the total number of positives and negatives, and the prevalences in each sub-population are given by  $P_i, N_i$  and  $\rho_i$  respectively, with  $i \in \{1, \dots, m\}$ . Let  $\beta_i$  be



**SI-Fig. 10. 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.

the relative size of the sub-populations. Thus, we have:

$$P = \sum_i P_i \quad (17)$$

$$N = \sum_i N_i \quad (18)$$

$$\beta_i = \frac{N_i + P_i}{N + P} \quad (19)$$

$$\rho_i = \frac{P_i}{N_i + P_i} = \frac{P_i}{\beta_i(N + P)} \quad (20)$$

Therefore, denoting the sensitivity and specificity of the sub-populations as  $s_i$  and  $c_i$  respectively, we have:

$$s = t_p/P = \frac{\sum_i t_p|_i}{P} = \frac{\sum_i (t_p|_i/P_i) \times (\beta_i \rho_i (P + N))}{P} = \sum_i s_i \beta_i \frac{\rho_i}{\rho} \quad (21)$$

Thus, we end up with:

$$s = \sum_{i=1}^m s_i \gamma_i \quad (22a)$$

$$c = \sum_{i=1}^m c_i \gamma'_i \quad (22b)$$

$$PPV = \frac{s}{s + (1 - c)(\frac{1}{\rho} - 1)} \quad (22c)$$

where we have denoted:

$$\gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma'_i = \beta_i \frac{1 - \rho_i}{1 - \rho} \quad (22d)$$

Now, using Table III, we can compute the values for  $\gamma_i, \gamma'_i$ , as shown below.

Using the prevalence and stratification parameters calculated from the CHOP study (See main text Table IV (3)), we can compute a conditional choice of sensitivity and specificity for our tool, in each sub-population to ultimately yield an overall performance significantly superior to M-CHAT/F. We carry out a four-dimensional search at the age the CHOP population stratification is reported (26 months or 112 weeks approximately) to identify the feasible region with  $PPV > 14.6\%$ , or sensitivity  $> 38.8\%$  while keeping specificity  $> 94.9\%$  where each of these dimensions represent the independent choice of sensitivity in the corresponding sub-population. For each set of 4 choices, the corresponding specificities are read-off from our computed ROC curve, and then the overall sensitivity, specificity and PPV are calculated using Eq. (22). The results are shown in Fig. 10, where we include the computations at 75 weeks, 125 weeks, and 150 weeks, with the same population stratification (although understandably the stratification will deviate from the values obtained at 26 months for those other ages).

An important assumption here is that the two tests are independent. Since M-CHAT/F is based on the detection of behavioral signals of developmental delay associated with autism via questionnaires completed by the primary care-givers, while our pipeline is based on physical comorbidities, independence is reasonable. Hence, we can

SI-Table II  
BOOSTED SENSITIVITY, SPECIFICITY AND PPV ACHIEVED AT 150 WEEKS 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	$\geq 8$ POS	specificity	sensitivity	PPV	specificity	sensitivity	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

SI-Table III  
POPULATION STRATIFICATION RESULTS ON LARGE M-CHAT/F STUDY(N=20,375) (3)

Id	Sub-population	Test Result	ASD positive	ASD Negative	Total %
A	M-CHAT/F $\geq 8$	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%

SI-Table IV  
 $\gamma, \gamma'$  COMPUTED FROM POPULATION STRATIFICATION RECORDED IN M-CHAT/F STUDY (3) ( $\rho = 0.0223$ )

Id	Sub-population	Test Result	$\beta_i$	$\rho_i$	$\gamma_i$	$\gamma'_i$
A	M-CHAT/F $\geq 8$	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

simulate the application of the pipeline to each sub-population, and compute the overall performance quantities using a pre-computed ROC curve. Here we use the curve corresponding to the age in weeks, but average the male and female ROC curves, which are close as shown in Fig. 1 in the main text. The male-female averaging is necessary since the results from the CHOP study does not report sex stratified data.

We show the feasible region obtained by this computation in Fig. 10 of this document, and in main text Fig. 3 of the main text. Particularly, note that we get a PPV close to or higher than 30% at the high precision (HP) operating point, or a sensitivity above 55% for the high recall (HR) operating point, when we restrict specificities to above 95%.

It is important to note that Eq. (22) and hence the results are dependent on the population prevalence  $\rho$ . We report the dependence of the solution to the 4D optimization for population prevalence between 1.7% (CDC estimate (12)), and 2.23% (CHOP estimate (3)). In particular, it is illuminating to compare these results directly with M-CHAT/F performance, as shown in Fig. 3, panels B and C in the main text. In panel C, we show that for any stable population prevalence between 1.7% and 2.24%, we can achieve nearly double the PPV without

losing sensitivity, or increase the sensitivity by about 50% without sacrificing PPV, while holding not letting the specificity to drop below 94%.

## IX. GENERATING PFSA MODELS FROM SET OF INPUT STREAMS WITH VARIABLE INPUT LENGTHS

Our PFSA reconstruction algorithm (33) is distinct from standard HMM learning. We do not need to pre-specify structures, or the number of states in the algorithm, and all model parameters are inferred directly from data. Additionally, we can operate either with 1) a single input stream, or 2) a set of input streams of possibly varying lengths which are assumed to be different and independent sample paths from the unknown stochastic generator we are trying to infer. At an intuitive level, we use the input data to infer the length of histories one must remember to estimate the current state, and predict futures for the process being modeled. Thus, we do not step through the symbol streams with a pre-specified model structure, and avoid the need to have equal-length inputs. More details of the algorithm are provided in the next section.

The ability to model a set of input streams of varying lengths is particularly important, since medical histories of different patients are typically of different lengths.

## X. PROBABILISTIC FINITE STATE AUTOMATA INFERENCE

### A. Probabilistic Finite-State Automaton

Let  $\Sigma$  be a finite alphabet of symbols with size  $|\Sigma|$ . The set of sequences of length  $d$  over  $\Sigma$  is denoted by  $\Sigma^d$ . The set of finite but unbounded sequences over  $\Sigma$  is denoted by  $\Sigma^*$ , the Kleene star operation (88), i.e.  $\Sigma^* = \bigcup_{d=0}^{\infty} \Sigma^d$ . We use lower case Greek, for example  $\sigma$  or  $\tau$ , for symbols in  $\Sigma$ , and lower case Latin, for example  $x$  or  $y$ , for sequences of symbols, i.e.  $x = \sigma_1 \sigma_2 \dots \sigma_n$ . We use  $|x|$  to denote the length of  $x$ . The empty sequence is denoted by  $\lambda$ .

We denote the set of strictly infinite sequences over  $\Sigma$  by  $\Sigma^\omega$ , and the set of strictly infinite sequences having  $x$  as prefix by  $x\Sigma^\omega$ . Let  $\mathcal{S} = \{x\Sigma^\omega : x \in \Sigma^*\} \cup \{\emptyset\}$ , we can verify that  $\mathcal{S}$  is a semiring (93) over  $\Sigma^\omega$ . We use  $\mathcal{F}$  to denote the sigma algebra generated by  $\mathcal{S}$ .

**Definition 2** (Stochastic Process over  $\Sigma$ ). *A stochastic process over a finite alphabet  $\Sigma$  is a collection of  $\Sigma$ -valued random variables  $\{X_t\}_{t \in \mathbb{N}}$  indexed by positive integers (82).*

We are specifically interested in processes in which the  $X_i$ s are not necessarily independently distributed.

**Definition 3** (Sequence-Induced Measure and Derivative). *For a process  $\mathcal{P}$ , let  $Pr_{\mathcal{P}}(x)$  or simply  $Pr(x)$  denote the probability  $\mathcal{P}$  producing a sample path prefixed by  $x$ . The measure  $\mu_x$  induced by a sequence  $x \in \Sigma^*$  is the extension (93) to  $\mathcal{F}$  of the premeasure defined on the semiring  $\mathcal{S}$  given by*

$$\forall x, y \in \Sigma^*, \mu_x(y\Sigma^\omega) \triangleq \frac{Pr(xy)}{Pr(x)}, \text{ if } Pr(x) > 0 \quad (23)$$

For any  $d \in \mathbb{N}$ , the  $d$ -th order derivative of a sequence  $x$ , written as  $\phi_x^d$ , is defined to be the marginal distribution of  $\mu_x$  on  $\Sigma^d$ , with the entry indexed by  $y$  denoted by  $\phi_x^d(y)$ . The first-order derivative is called the symbolic derivative and is denoted by  $\phi_x$  for short.

**Definition 4** (Probabilistic Nerode Equivalence and Causal States (78)). *For any pair of sequences  $x, y \in \Sigma^*$ ,  $x$  is equivalent to  $y$ , written as  $x \sim y$ , if and only if either  $Pr(x) = Pr(y) = 0$ , or  $\mu_x = \mu_y$ . The equivalence class of a sequence  $x$  is denoted by  $[x]$  and is called a causal state (79). The cardinality of the set of causal states is called the probabilistic Nerode index, or the Nerode index for simplicity.*

We can see from the definition that causal states captures how the history of a process influences its future. Since the probabilistic Nerode equivalence is right invariant, it gives rise naturally to a automaton structure introduced below.

**Definition 5** (Probabilistic Finite-State Automaton (PFSA)). *A PFSA  $G$  is defined by a quadruple  $(Q, \Sigma, \delta, \tilde{\pi})$ , where  $Q$  is a finite set,  $\Sigma$  is a finite alphabet,  $\delta : Q \times \Sigma \rightarrow \Sigma$  is called the transition map, and  $\tilde{\pi} : Q \rightarrow \mathbf{P}_\Sigma$ , where  $\mathbf{P}_\Sigma$  is the space of probability distributions over  $\Sigma$ , is called the transition probability. The entry of  $\tilde{\pi}(q)$  indexed by  $\sigma$  is denoted by  $\tilde{\pi}(q, \sigma)$ .*

**Definition 6** (Transition and Observation Matrices). *The transition matrix  $\Pi$  is the  $|Q| \times |Q|$  matrix with the entry indexed by  $q, q'$ , written as  $\pi_{q,q'}$ , satisfying*

$$\pi_{q,q'} \triangleq \sum_{\{\sigma \in \Sigma | \delta(q, \sigma) = q'\}} \tilde{\pi}(q, \sigma) \quad (24)$$

and the observation matrix  $\tilde{\Pi}$  is a  $|Q| \times |\Sigma|$  matrix with the entry indexed by  $q, \sigma$  equaling  $\tilde{\pi}(q, \sigma)$ .

We note that both  $\Pi$  and  $\tilde{\Pi}$  are stochastic, i.e. non-negative with rows summing up to 1.

**Definition 7** (Extension of  $\delta$  and  $\tilde{\pi}$  to  $\Sigma^*$ ). For any  $x = \sigma_1 \dots \sigma_k$ ,  $\delta(q, x)$  is defined recursively by

$$\delta(q, x) \triangleq \delta(\delta(q, \sigma_1 \dots \sigma_{k-1}), \sigma_k) \quad (25)$$

with  $\delta(q, \lambda) = q$ , and  $\tilde{\pi}(q, x)$  is defined recursively by

$$\tilde{\pi}(q, x) \triangleq \prod_{i=1}^k \tilde{\pi}(\delta(q, \sigma_1 \dots \sigma_{i-1}), \sigma_i) \quad (26)$$

with  $\tilde{\pi}(q, \lambda) = 1$ .

**Definition 8** (Strongly Connected PFSA). We say a PFSA is strongly connected if the underlying directed graph is strongly connected (76). More precisely, a PFSA  $G = (Q, \Sigma, \delta, \tilde{\pi})$  is strongly connected if for any pair of distinct states  $q$  and  $q' \in Q$ , there is an  $x \in \Sigma^*$  such that  $\delta(q, x) = q'$ .

We assume all PFSA in the discussions in the sequel are strongly connected if not specified otherwise. For strongly connected PFSA  $G$ , there is a unique probability distribution over  $Q$  that satisfies  $\mathbf{v}^T \Pi = \mathbf{v}^T$ . This is the **stationary distribution** (91, 100) of  $G$  and is denoted as  $\wp_G$ , or  $\wp$  if  $G$  is understood.

**Definition 9** ( $\Gamma$ -Expression). We can encode the information contained in  $\delta$  and  $\tilde{\pi}$  by a set of  $|Q| \times |Q|$  matrices  $\Gamma = \{\Gamma_\sigma | \sigma \in \Sigma\}$ , where

$$\Gamma_\sigma|_{q, q'} \triangleq \begin{cases} \tilde{\pi}(q, \sigma) & \text{if } \delta(q, \sigma) = q', \\ 0 & \text{if otherwise.} \end{cases} \quad (27)$$

$\Gamma_\sigma$  is called **event-specific transition matrix**, with the event being that  $\sigma$  is current the output.  $\Gamma_\sigma$  can also be extended to arbitrary  $x \in \Sigma^*$  by defining  $\Gamma_x = \prod_{i=1}^k \Gamma_{\sigma_i}$  with  $\Gamma_\lambda = I$ .

**Definition 10** (Sequence-Induced Distribution on States). For a PFSA  $G = (Q, \Sigma, \delta, \tilde{\pi})$  and a distribution  $\wp_0$  on  $Q$ , the **distribution on  $Q$  induced by a sequence  $x$**  is given by  $\wp_{G, \wp_0}^T(x) = [\wp_0^T \Gamma_x]$  with  $\wp_{G, \wp_0}(\lambda) = \wp_0$ . The entry indexed by  $q \in Q$  of the vector  $\wp_{G, \wp_0}(x)$  is written as  $\wp_{G, \wp_0}(x, q)$ . When  $\wp_0 = \wp_G$ , the stationary distribution of  $G$ , we write  $\wp_{G, \wp_0}(x)$  as  $\wp_G(x)$ , or simply as  $\wp(x)$ , if  $G$  is understood.

**Definition 11** (Stochastic Process Generated by a PFSA). Let  $G = (Q, \Sigma, \delta, \tilde{\pi})$  be a PFSA and let  $\wp_0$  be a distribution on  $Q$ , the  $\Sigma$ -valued stochastic process  $\{X_t\}_{t \in \Sigma}$  generated by  $G$  and  $\wp_0$  satisfies that  $X_1$  follows the distribution  $\wp_0$  and  $X_{t+1}$  follows the distribution  $\wp_{G, \wp_0}(X_1 \dots X_t)$  for  $t \in \mathbb{N}$ .

For the rest of this paper, we will assume  $\wp_0 = \wp_G$  if not specified otherwise. We can show that, when initialized with  $\wp_G$ , the process generated by a PFSA  $G$  is stationary and ergodic. We also note the, for the process generate by  $G$ , we have  $\phi_x = \wp_G(x)^T \tilde{\Pi}$ . Since  $\wp_G(\lambda) = \wp_G$ , the symbolic derivative of the empty sequence  $\phi_\lambda$  is the stationary distribution on the symbols.

**Definition 12** (Synchronizable PFSA and Synchronizing Sequence). A **synchronizing sequence** is a finite sequence that sends an arbitrary state of the PFSA to a fixed state (99). To be more precise, let  $G = (Q, \Sigma, \delta, \tilde{\pi})$  be a PFSA, we say a sequence  $x \in \Sigma^*$  is a synchronizing sequence to a state  $q \in Q$  if  $\delta(q', x) = q$  for all  $q' \in Q$ . A PFSA is **synchronizable** if it has at least one synchronizing sequence. Given a sample path generated by a PFSA, we say the PFSA is **synchronized** if a synchronizing sequence transpires in the sample path.

**Definition 13** (Equivalence and Irreducibility). Two PFSA  $G$  and  $H$  are **equivalent** if they generate the same stochastic process. A PFSA  $G$  is said to be **irreducible**, if there is not another PFSA with smaller state set that is equivalent to  $G$ .

**Definition 14.** Consider a PFSA  $G$  over state set  $Q$ . For a give  $\epsilon > 0$ , we say a sequence  $x$  is a  $\epsilon$ -synchronizing sequence to a state  $q \in Q$  if

$$\|\wp_G(x) - \mathbf{e}_q\|_\infty \leq \epsilon. \quad (28)$$

While there exists PFSA that is not synchronizable, we can show that an irreducible PFSA always has an  $\epsilon$ -synchronizing sequence for some state  $q$  for arbitrarily small  $\epsilon > 0$ . Moreover, we can show that as length increases, sequences produced by PFSA become uniformly  $\epsilon$ -synchronizing. These two are the underpinning properties for the inference algorithm of PFSA (See Alg. 1), because they imply that  $\phi_x$  can be used to approximate  $\tilde{\pi}(q)$  if  $x$  are properly prefixed and long enough.

**Definition 15** (Joint  $\epsilon$ -Synchronizing Sequence). Let  $G$  and  $H$  be two PFSA over state sets  $Q_G$  and  $Q_H$ , respectively. For a fixed  $\epsilon$ , a sequence  $x$  is said to be **jointly  $\epsilon$ -synchronizing** to  $(q, r) \in Q_G \times Q_H$  if  $x$  is

$\varepsilon$ -synchronizing to  $q$  and to  $r$  simultaneously. We define

$$\Sigma_{\varepsilon,(q,r)}^d \triangleq \{x \in \Sigma^d : x \text{ jointly } \varepsilon\text{-synchronizing to } (q, r)\} \quad (29)$$

**Definition 16** (Joint Pair of States). Let  $G$  and  $H$  be two PFSA over state sets  $Q_G$  and  $Q_H$ , respectively. Define

$$p_G(q, r) \triangleq \lim_{d \rightarrow \infty} p_G(\Sigma_{\varepsilon,(q,r)}^d) \quad (30)$$

A pair of states  $(q, r) \in Q_G \times Q_H$  is called a  **$G$ -joint pair** of states if  $p_G(q, r) > 0$ . We also define

$$Q_c \triangleq \{(q, r) \in Q_G \times Q_H : (q, r) \text{ is a } G\text{-joint pair}\} \quad (31)$$

The inference algorithm for PFSA is called **GenESeSS** for Generator Extraction Using Self-similar Semantics. With an input sequence  $x$  and a hyperparameter  $\varepsilon$ , **GenESeSS** outputs a PFSA in the following three steps: 1) approximate an almost synchronizing sequence; 2) identify the transition structure of the PFSA; 3) calculate the transition probabilities of the PFSA. See Alg. 1 for detail.

---

**Algorithm 1: GenESeSS**


---

**Data:** A sequence  $x$  over alphabet  $\Sigma$ ,  $0 < \varepsilon < 1$   
**Result:** State set  $Q$ , transition map  $\delta$ , and transition probability  $\tilde{\pi}$

```

/* Step One: Approximate ε-synchronizing sequence */
```

1 Let  $L = \lceil \log_{|\Sigma|} 1/\varepsilon \rceil$ ;

2 Calculate the **derivative heap**  $D_\varepsilon^x$  equaling  $\{\hat{\phi}_y^x : y \text{ is a sub-sequence of } x \text{ with } |y| \leq L\}$ ;

3 Let  $\mathcal{C}$  be the convex hull of  $D_\varepsilon^x$ ;

4 Select  $x_0$  with  $\hat{\phi}_{x_0}^x$  being a vertex of  $\mathcal{C}$  and has the highest frequency in  $x$ ;

```

/* Step Two: Identify transition structure */
```

5 Initialize  $Q = \{q_0\}$ ;

6 Associate to  $q_0$  the **sequence identifier**  $x_{q_0}^{\text{id}} = x_0$  and the probability vector  $d_{q_0} = \hat{\phi}_{x_0}^x$ ;

7 Let  $\tilde{Q}$  be the set of states that are just added and initialize it to be  $Q$ ;

8 **while**  $\tilde{Q} \neq \emptyset$  **do**

9   Let  $Q_{\text{new}} = \emptyset$  be the set of new states;

10   **for**  $(q, \sigma) \in \tilde{Q} \times \Sigma$  **do**

11     Let  $x = x_q^{\text{id}}$  and  $d = \hat{\phi}_{x\sigma}^x$ ;

12     **if**  $\|d - d_{q'}\|_\infty < \varepsilon$  **for some**  $q' \in Q$  **then**

13       Let  $\delta(q, \sigma) = q'$ ;

14     **else**

15       Let  $Q_{\text{new}} = Q_{\text{new}} \cup \{q_{\text{new}}\}$  and  $Q = Q \cup \{q_{\text{new}}\}$ ;

16       Associate to  $q_{\text{new}}$  the sequence identifier  $x_{q_{\text{new}}}^{\text{id}} = x\sigma$  and the probability vector  $d_{q_{\text{new}}} = d$ ;

17       Let  $\delta(q, \sigma) = q_{\text{new}}$ ;

18     Let  $\tilde{Q} = Q_{\text{new}}$ ;

19 Take a strongly connected subgraph of the labeled directed graph defined by  $Q$  and  $\delta$ , and denote the vertex set of the subgraph again by  $Q$ ;

```

/* Step Three: Identify transition probability */
```

20 Initialize counter  $N[q, \sigma]$  for each pair  $(q, \sigma) \in Q \times \Sigma$ ;

21 Choose a random starting state  $q \in Q$ ;

22 **for**  $\sigma \in x$  **do**

23   Let  $N[q, \sigma] = N[q, \sigma] + 1$ ;

24   Let  $q = \delta(q, \sigma)$ ;

25 Let  $\tilde{\pi}(q) = \llbracket (N[q, \sigma])_{\sigma \in \Sigma} \rrbracket$ ;

26 **return**  $Q, \delta, \tilde{\pi}$ ;

---

## XI. SEQUENCE LIKELIHOOD DEFECT

**Definition 17** (Entropy Rate and KL Divergence). By entropy rate of a PFSA, we mean the entropy rate of the stochastic process generated by the PFSA(81). Similarly, by KL divergence of two PFSA, we mean the KL divergence between the two processes generated by them (97). More precisely, we have

$$\mathcal{H}(G) = - \lim_{d \rightarrow \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p(x) \log p(x) \quad (32)$$

and the KL divergence

$$\mathcal{D}_{KL}(G \parallel H) = \lim_{d \rightarrow \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} \quad (33)$$

whenever the limits exist.

**Theorem 1** (Closed-form Formula for Entropy Rate and KL Divergence). *The entropy rate of a PFSA  $G = (\Sigma, Q, \delta, \tilde{\pi})$  is given by*

$$\mathcal{H}(G) = \sum_{q \in Q} \wp_G(q) \cdot h(\tilde{\pi}(q)) \quad (34)$$

where  $h(\mathbf{v})$  is the based-2 entropy of the probability vector  $\mathbf{v}$ .

Consider two PFSA  $G = (Q_G, \Sigma, \delta_G, \tilde{\pi}_G)$  and  $H = (Q_H, \Sigma, \delta_H, \tilde{\pi}_H)$  with  $\mu_G$  being absolutely continuous with respect to  $\mu_H$ . Let  $Q_c$  be the set of  $G$ -joint pairs of states, we have

$$\mathcal{D}_{KL}(G \parallel H) = \sum_{(q,r) \in Q_c} p_G(q, r) \mathcal{D}_{KL}(\tilde{\pi}_G(q) \parallel \tilde{\pi}_H(r)) \quad (35)$$

**Definition 18** (Log-likelihood). *Let  $x \in \Sigma^d$ , the log-likelihood (81) of a PFSA  $G$  generating  $x$  is given by*

$$L(x, G) = -\frac{1}{d} \log p_G(x) \quad (36)$$

The calculation of log-likelihood is detailed in Alg. 2.

---

#### Algorithm 2: Log-likelihood

---

**Data:** A PFSA  $G = (\Sigma, Q, \delta, \tilde{\pi})$  and a sequence  $x$  over alphabet  $\Sigma$   
**Result:** Log-likelihood  $L(x, G)$  of  $G$  generating  $x$

- 1 Calculate the state transition matrix  $\Pi$  and observation  $\tilde{\Pi}$ ;
- 2 Calculate the stationary distribution over states  $\wp_G$  of  $G$  from  $\Pi$ ;
- 3 Calculate the stationary distribution of alphabet  $\phi_\lambda^T = \wp_G^T \tilde{\Pi}$ ;
- 4 Initialize  $\mathbf{p}$  by  $\wp_G$  and  $\mathbf{q}$  by  $\phi_\lambda$ ;
- 5 Let  $L = 0$ ;
- 6 **for**  $i$  from 1 to  $|x|$  **do**
- 7   Let  $\sigma$  be the  $i$ -th entry of  $x$ ;
- 8   Let  $L = L - \log \mathbf{q}|_\sigma$ ;
- 9   Let  $\mathbf{p}^T = [\mathbf{p}^T \Gamma_\sigma]$  where  $\Gamma_\sigma$  is defined in 9;
- 10   Let  $\mathbf{q}^T = \mathbf{p}^T \tilde{\Pi}$ ;
- 11 **return**  $L/|x|$ ;

---

**Theorem 2** (Convergence of log-likelihood). *Let  $G$  and  $H$  be two reduced PFSA, and let  $x \in \Sigma^d$  be a sequence generated by  $G$ . Then we have*

$$L(x, H) \rightarrow \mathcal{H}(G) + \mathcal{D}_{KL}(G \parallel H) \quad (37)$$

in probability as  $d \rightarrow \infty$ .

*Proof.* We first notice that

$$\sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} = \sum_{x \in \Sigma^{d-1}} \sum_{\sigma \in \Sigma} p_G(x) \wp_G(x) \tilde{\Pi}_G \Big|_{\sigma} \log \frac{p_G(x) \wp_G(x) \tilde{\Pi}_G \Big|_{\sigma}}{p_H(x) \wp_H(x) \tilde{\Pi}_H \Big|_{\sigma}} \quad (38)$$

$$= \sum_{x \in \Sigma^{d-1}} p_G(x) \log \frac{p_G(x)}{p_H(x)} + \underbrace{\sum_{x \in \Sigma^{d-1}} p_G(x) \sum_{\sigma \in \Sigma} \wp_G(x) \tilde{\Pi}_G \Big|_{\sigma} \log \frac{\wp_G(x) \tilde{\Pi}_G \Big|_{\sigma}}{\wp_H(x) \tilde{\Pi}_H \Big|_{\sigma}}}_{D_d} \quad (39)$$

By induction, we have  $\mathcal{D}_{KL}(G \parallel H) = \lim_{d \rightarrow \infty} \frac{1}{d} \sum_{i=1}^d D_i$ , and hence by Cesàro summation theorem (86), we have  $\mathcal{D}_{KL}(G \parallel H) = \lim_{d \rightarrow \infty} D_d$ . Let  $x = \sigma_1 \sigma_2 \dots \sigma_n$  be a sequence generated by  $G$ . Let  $x^{[i-1]}$  is the truncation of  $x$  at the  $(i-1)$ -th symbols, we have

$$-\frac{1}{n} \sum_{i=1}^n \log \wp_H(x^{[i-1]}) \tilde{\Pi}_H \Big|_{\sigma_i} = \underbrace{\frac{1}{n} \sum_{i=1}^n \log \frac{\wp_G(x^{[i-1]}) \tilde{\Pi}_G \Big|_{\sigma_i}}{\wp_H(x^{[i-1]}) \tilde{\Pi}_H \Big|_{\sigma_i}}}_{A_{x,n}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \log \wp_G(x^{[i-1]}) \tilde{\Pi}_G \Big|_{\sigma_i}}_{B_{x,n}} \quad (40)$$

**Algorithm 3:** ICD-9 Encoding

---

```

input : Dataset, TargetDiseaseGroup, DiseaseGroups
output: Encoding

1 Encoding ← new Dictionary();
2 for diseaseGroup ∈ DiseaseGroups do
3   Encoding[diseaseGroup][patientID] ← new List();
4   Encoding[diseaseGroup][gender] ← new List();
5   Encoding[diseaseGroup][record] ← new List();
6   Encoding[diseaseGroup][target] ← new List();
7   for record ∈ Dataset do
8     //encode Dataset into a weekly trinary sequence;
9     weeklyEncoding ← new List();
10    for weeklyDiseaseRecord ∈ record do
11      //no code recorded for the observed week;
12      if weeklyDiseaseRecord.code == NIL then
13        append "0" to weeklyEncoding;
14      if weeklyDiseaseRecord.code ∈ diseaseGroup.codes then
15        append "1" to weeklyEncoding;
16      if weeklyDiseaseRecord.code ∉ diseaseGroup.codes then
17        append "2" to weeklyEncoding;
18    target ← 1 if any weeklyDiseaseRecord.code of record ∈ TargetDiseaseGroup;
19    if target == 1 then
20      cut weeklyEncoding up to (but not including) first occurrence of TargetDiseaseGroup member;
21      append record.patientID to Encoding[diseaseGroup][patientID];
22      append record.gender to Encoding[diseaseGroup][gender];
23      append weeklyEncoding to Encoding[diseaseGroup][record];
24      append target to Encoding[diseaseGroup][target];
25 return Encoding;

```

---

Since the stochastic process  $G$  generates is ergodic, we have

$$\lim_{n \rightarrow \infty} A_{x,n} = \lim_{d \rightarrow \infty} D_d = \mathcal{D}_{\text{KL}}(G \parallel H) \quad (41)$$

and  $\lim_{n \rightarrow \infty} B_{x,n} = \mathcal{H}(G)$ .  $\square$

## XII. PIPELINE OPTIMIZATION

### A. Input Data Format

To encode the ICD-9 codes, 17 Disease Groups of codes are used to transform the raw health records into a format suitable for PFSA. As described in *Algorithm 1*, for each patient, the list of ICD-9 codes is encoded into a weekly array of three-symbol alphabet digits with respect to selected disease group, for each week: "0" - no disease "1" - disease from the selected group, "2" - other disease.

Once the trinary encodings are ready, the PFSA pairs are fit for each of the disease groups, on positive (treatment) and negative (control) sets using `genESESS` algorithm (33) (See Section X), as described in *Algorithm 2*. The PFSA pairs are then used to obtain the loglikelihood scores of belonging to a PFSA modeling the positive and the control cohorts accordingly for each of the encodings of a patient record. As a result, we yield the difference between positive and control loglikelihoods for each disease group of each patient. The positive value of difference means that with respect to a given disease group, a certain patient is more likely to be a positive one. Conversely, the negative value of difference signifies that a patient is more likely to be from the control group. These features, as well as their aggregations and the aggregations of the ternary encoding arrays, are used as the features for the final LightGBM gradient boosting classifier.

### B. Algorithms

The key data processing approach is outlined in Algorithm 3. The remaining steps of the approach are sketched in Algorithm 4. Fig. 11 shows the overall schema, including the breakdown of a database into a test set, and two training sets: one for training the HMM models, and one for training the boosting classifier.

**Algorithm 4:** Prediction Pipeline Training

---

```

input : Encoding, DiseaseGroups, SequenceFeatures, hyperparameters
output: Predictions, FeatureImportances

1 DiseaseDatasets ← new Dictionary();
2 for Dataset, DiseaseGroup ∈ zip(Encoding, DiseaseGroups) do
3   PFSAsset, LLset ← TrainTestSplit(Dataset, w.r.t = "target");
4   df ← new Dataframe();
5   df[patientID] ← LLset[patientID];
6   df[target] ← LLset[target];
7   //Generate 2 PFSAs for each class;
8   PositivePFSAsset ← PFSAsset[PFSAsset.target == 1];
9   NegativePFSAsset ← PFSAsset[PFSAsset.target == 0];
10  PosPFSA ← genESeSS(PositivePFSAsset);
11  NegPFSA ← genESeSS(NegativePFSAsset);
12  //For each record, compute loglikelihoods of being generated by either of 2 PFSAs generated above;
13  PosLLK ← llk(LLset, PosPFSA);
14  NegLLK ← llk(LLset, NegPFSA);
15  //Compute sequence likelihood defect;
16  df[DiseaseGroup] ← pairwise(PosLLK - NegLLK);
17  df[DiseaseGroup + '_abs_neg'] ← NegLLK;
18  for SequenceFeature ∈ SequenceFeatures do
19    df[DiseaseGroup + '_' + SequenceFeature] ← [ComputeSequenceFeature(SequenceFeature,
      seq) for each seq ∈ LLset['record']];
20  DiseaseDatasets[DiseaseGroup] ← df;
21 Dataset ← outerjoin(DiseaseDatasets.values, on = 'patientID');
22 Aggregate all features in Dataset where feature_name ∈ DiseaseGroups (mean, std. deviation, range);
23 Aggregate all features in Dataset where feature name minus '_abs_neg' ∈ DiseaseGroups (mean, std.
  deviation, range);
24 Aggregate all sequence features in Dataset (mean, std. deviation, range, max);
25 TrainSet, TestSet ← TrainTestSplit(Dataset, w.r.t = "target");
26 LGBM ← new LightGBM(hyperparameters);
27 LGBM.fit(TrainSet);
28 Predictions ← LGBM.predict(TestSet);
29 return Predictions, LGBM.feature_importances;

```

---

## XIII. EXAMPLE RUN WITH RELEASED APPLICATION

*A. Prerequisites & Installation*

The minimum prerequisites for running ehrzero are the following:

- 1) A x64 system running any flavor of Linux.
- 2) A working python 3.x installation
- 3) scikit-learn, version = 0.20.0

Installation:

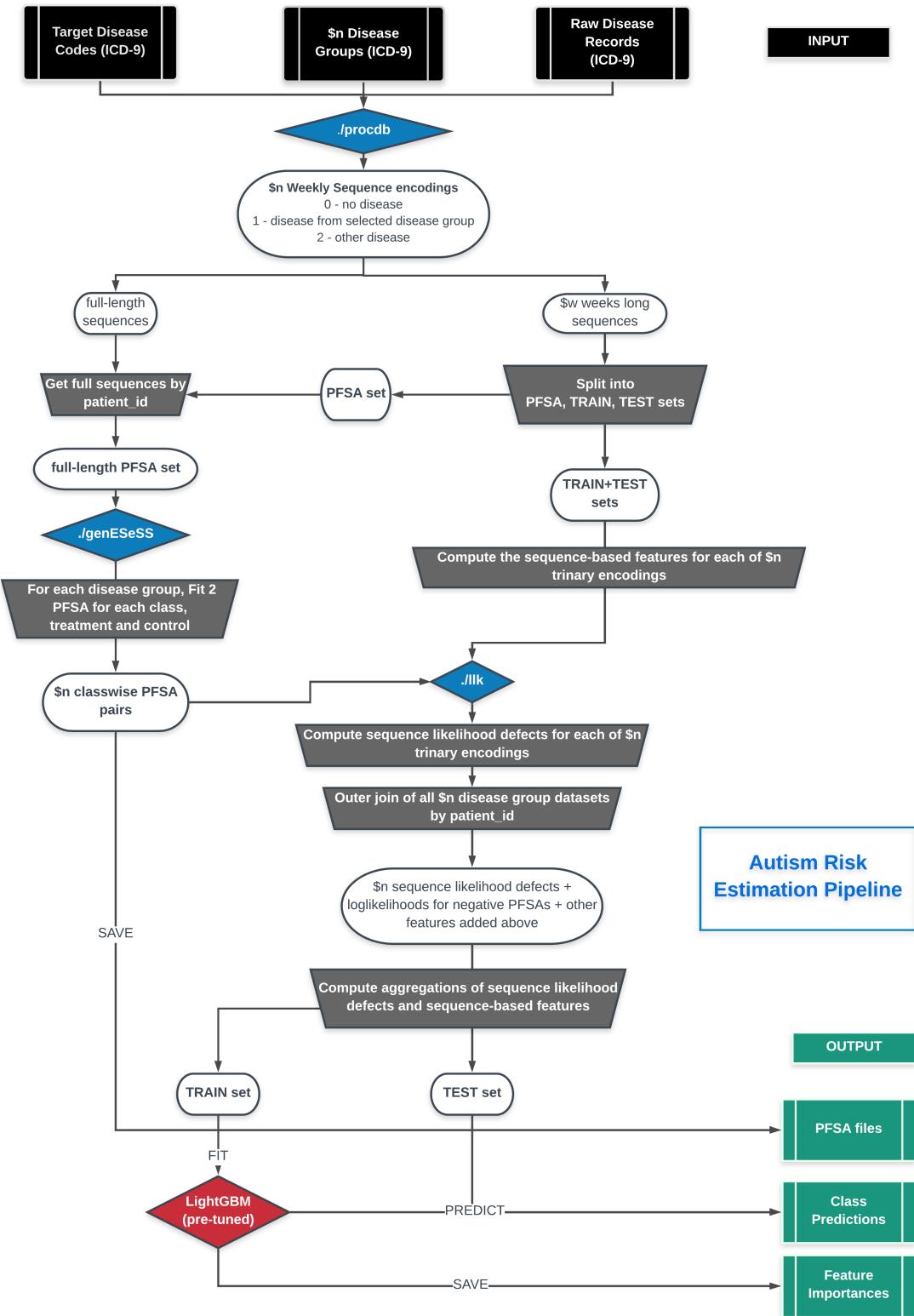
```
pip3 install ehrzero --user
```

*B. EHR data format*

Diagnostic data stored in text file, one line per patient as follows: patient id, gender, and list of space-separated, comma-delineated diagnosis records, all separated by spaces. Each diagnosis record consists of the week since the start of the observation, followed by a comma, and the ICD-9 code of the diagnosis.

Example of a patient line:

```
Lorax,M 5,277.03 10,611.79 18,057.8 58,157.8 78,057.8 108,057.8 128,057.8 148,057.8
```

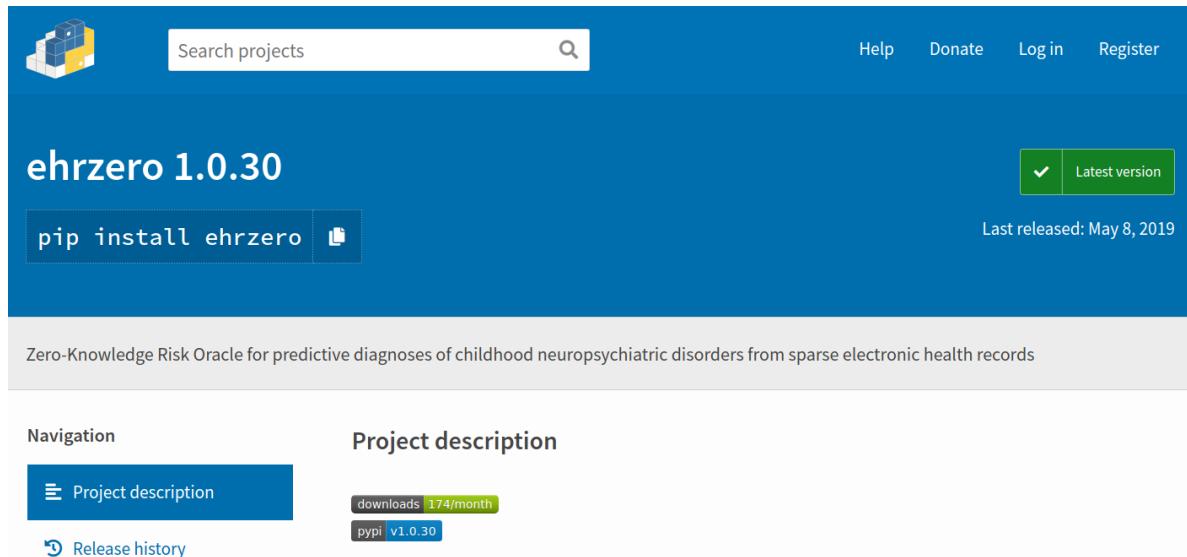


SI-Fig. 11. Pipeline schema: How the data set is split into test sets and two training sets: one for inferring HMM models, and one for training the boosting classifier. The two key algorithms here are genESeSS (33) and the llk which does the sequence likelihood computation described in Section XI

### C. Sample Python code risk estimation

Once the patient diagnostic data is in the required format, for function `predict_with_confidence` we specify the filepath of the data and the list of the cutoffs for the first weeks since the start of observations for the data we want to analyze. We also specify the separator and delimiter for the patients within file (space and comma are default values, but can be changed for user convenience).

The `predict_with_confidence` function returns the predicted risk of autism for every patient in the input file with all the specified numbers of first weeks to consider.



SI-Fig. 12. Screen capture of the page on pypi.org hosting the released application Link: <http://pypi.org/ehrzero>

```
[1]: from ehrzero import ehrzero as ehr
      import warnings
      warnings.filterwarnings("ignore")

[2]: source = 'test_free.dat'
      outfile = 'out.dat'
      first_weeks = [200, 100] # number of first weeks of the observations to consider
      risks = ehr.predict_with_confidence(source,
                                          outfile,
                                          separator = ',',
                                          delimiter = ' ',
                                          n_first_weeks = first_weeks)

[3]: risks
```

	patient_id	week	risk	relative_risk	confidence
0	AAbby	200	0.000174	0.028200	0.920977
0	AAbby	100	0.000135	0.021954	0.954741
1	ALorax	200	0.000101	0.016416	0.989583
1	ALorax	100	0.000099	0.016071	0.986710

SI-Fig. 13. Python code prediction example

#### D. Sample Python script risk estimation

The script version is similar to the one mentioned before.

Once ehrzero package is installed, locate its directory and go to `textt../ehrzero/example`. Select one of the `".dx"` or `".dat"` files in `/ehrzero/example/tests` as input and run the following command as an example:

```
python zero.py -data tests/ZEROexample.dat -outfile predictions.csv -nweeks 100 200
300 -Verbose 1
```

```
(base) [onishchenko@midway2-login1 example]$ cat tests/ZERO_example.dat
M:44 380.10:101 381.81:111 084.6
M:9 380.10:104 381.81:11 084.6
M:99 380.10:104 381.81:11 084.6:98 380.11
M:9 380.10:104 381.71:11 084.6
M:9 390.11:4 390.11:4 381.71:11 084.6
(base) [onishchenko@midway2-login1 example]$ python zero.py -data tests/ZERO_example.dat -outfile predictions.csv -n_weeks 100 200 300 -Verbose 1
patient_id week risk relative_risk confidence
A000000001 100 0.001703 0.098002 68.54
A000000001 200 0.001834 0.105517 72.43
A000000001 300 0.001834 0.105517 72.43
A000000002 100 0.002030 0.116816 60.94
A000000002 200 0.001426 0.082033 80.01
A000000002 300 0.001426 0.082033 80.01
A000000003 100 0.001703 0.098002 68.54
A000000003 200 0.001706 0.098170 74.53
A000000003 300 0.001853 0.106620 72.18
A000000004 100 0.000512 0.029449 94.86
A000000004 200 0.001450 0.083436 79.53
A000000004 300 0.001450 0.083436 79.53
A000000005 100 0.000658 0.037868 91.73
A000000005 200 0.000658 0.037868 95.21
A000000005 300 0.000658 0.037868 95.21
(base) [onishchenko@midway2-login1 example]$
```

SI-Fig. 14. Python script prediction example