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

Dmytro Onishchenko[a], Yi Huang[a], James van Horne[a], Peter J. Smith[d,g], Michael M. Msall[e,f], and Ishanu Chattopadhyay[a,b,c,1]

[a]Department of Medicine, University of Chicago; [b]Committee on Genetics, Genomics & Systems Biology, University of Chicago; [c]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago; [d]Department of Pediatrics, Section of Developmental and Behavioral Pediatrics, University of Chicago; [e]Department of Pediatrics, Section Chief of Developmental and Behavioral Pediatrics, University of Chicago; [f]Joseph P. Kennedy Research Center on Intellectual and Neurodevelopmental Disabilities, University of Chicago; [g]Executive Committee Chair, American Academy of Pediatrics' Section on Developmental and Behavioral Pediatrics

This manuscript was compiled on Friday 19th June, 2020

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

autism | electronic health record | stochastic learning | co-morbid risk

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

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

A screening tool that tracks the risk of an eventual ASD diagnosis, based on the information already being gathered during regular doctor's visits, and which may be implemented as a fully automated background process requiring no time commitment from providers has the potential to reduce avoid-

Significance Statement

In this study past medical history in the form of diagnostic codes generated during medical encounters is used to formulate a clinically useful risk estimator for autism screening. We can formulate such a predictor despite considerable heterogenity in ASD presentation that has made reliable biomarkers hard to identify, and has implicated over 1000 genes as contributing to ASD risk. Furthermore, the ability to carry out this evaluation without any blood tests, or questionaires, makes it possible to avoid uncertainties from language barriers and other socio-economic artifacts. In joint operation with existing screening tools we can boost the predictor performance significantly higher than the current state of art, potentially cutting down false positives by half without losing specificity.

DO implemented the algorithm, DO, YH, JH and IC worked on mathematical modeling, IC, PS and MM designed study parameters, IC wrote the paper

[2]To whom correspondence should be addressed. E-mail: ishanu@uchicago.edu

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

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

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

Table 3. Engineered Features (Total Count: 165)

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

‡ Disease categories are described in Table 2

able diagnostic delays at no additional burden of time, money and personnel resources. While still lacking the certainty of a diagnostic blood test, use of patterns emergent in the diagnostic history to estimate risk might help reduce the subjective component in questionnaire-based screening tools, resulting in 1) reduced effect of potential language and cultural barriers in diverse populations, and 2) possibly better identify children with milder symptoms (8). Furthermore, being functionally independent of the M-CHAT/F, we show that there is clear advantage to combining the outcomes of the two tools: we can take advantage of any population stratification induced by the M-CHAT/F scores to significantly boost combined screening performance (See Materials & Methods, and Supplementary text, section 9).

Results

We measure our performance using several standard metrics including the AUC, sensitivity, specificity and the PPV. For the prediction of the eventual ASD status, we achieve an out-of-sample AUC of 82.3% and 82.5% for males and females respectively at 125 weeks for the Truven dataset. In the UCM dataset, our performance is comparable: 83.1% and 81.3% for males and females respectively (Fig. 2 and 3). Our AUC is shown to improve approximately linearly with patient age: Fig. 3A illustrates that the AUC reaches 90% in the Truven dataset at the age of four. Importantly, we train our pipeline on 50% of the Truven dataset, and use held back data from Truven, and the entirety of the UCM dataset for validation: *No new training is done in the UCM dataset.* Good performance on these independent datasets lends strong evidence for our claims. Furthermore, applicability in new datasets *without local re-training* makes it readily deployable in clinical settings.

What are the inferred patterns that elevate risk? Enumerating the top 15 predictive features (Fig. 2B), ranked according to their automatically inferred weights (the feature "importances"), we found that while infections and immunologic disorders are the most predictive, there is significant effect

Table 2. Disease Categories (A few ICD9 codes shown from the complete set of $9,835$ unique ICD9 codes considered. See SI-Table 4 in Supplementary text for complete list)

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

† Categories inferred to be important for risk modulation are proportionately highlighted.
⋆ ICD10 codes when present were mapped back to closest ICD9 matches using published General Equivalence Mappings (17).

from all the 17 disease categories. Thus, the co-morbid indicators are distributed across the disease spectrum, and no single disorder is uniquely implicated (See also Fig. 3F). Importantly, predictability is relatively agnostic to the number of local cases across US counties (Fig. 2C-D) which is important in light of the current uneven distribution of diagnostic resources (12, 18) across states and regions.

Unlike individual predictions which only become relevant over 2 years, the average risk over the populations is clearly different from around the first birthday (Fig. 3B), with the risk for the positive cohort rapidly rising. Also, we see a saturation of the risk after ≈ 3 years, which corresponds to the median diagnosis age in the database (See Fig. 1B). Thus, if a child is not diagnosed up to that age, then the risk falls, since the probability of a diagnosis in the population starts to go down after this age. While average discrimination is not useful for individual patients, these reveal important clues as to how the risk evolves over time. Additionally, while each new diagnostic code within the first year of life increases the risk burden by approximately 2% irrespective of sex (Fig. 3D), distinct categories modulate the risk differently, *e.g.*, for a single random patient illustrated in Fig. 3F infections and immunological disorders dominate early, while diseases of the nervous system and sensory organs, as well as ill-defined symptoms, dominate the latter period.

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

Our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at least 14% (14.1-33.6%) when sensitivity and specificity are held at comparable values (approx. 38% and 95%) around the age of 26 months (≈ 112 weeks). Fig. 4A and Table 4 show the out-of-sample PPV vs sensitivity curves

Table 4. PPV Achieved at 100, 112 and 150 Weeks For Each Dataset and Gender (M-CHAT/F: sensitivity=38.8%, specificity=95%, PPV=14.6% between 16 and 26 months (≈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

Table 5. Personalized Operation Conditioned on M-CHAT/F Scores at 26 months

M-CHAT/F Outcome				global performance (Truven)			global performance (UCM)			prevalence*
0-2 NEG	3-7 NEG	3-7 POS	≥ 8 POS	speci-ficity	sensi-tivity	PPV	speci-ficity	sensi-tivity	PPV	
specificity choices										
0.2	0.54	0.83	0.98	0.95	0.585	0.209	0.95	0.505	0.186	0.022
0.21	0.53	0.83	0.98	0.95	0.586	0.208	0.95	0.506	0.184	0.022
0.42	0.87	0.98	0.99	0.98	0.433	0.331	0.98	0.347	0.284	0.022
0.48	0.87	0.97	0.99	0.98	0.432	0.331	0.98	0.355	0.289	0.022
0.38	0.54	0.94	0.98	0.95	0.736	0.203	0.95	0.628	0.178	0.017
0.3	0.55	0.94	0.98	0.95	0.737	0.203	0.95	0.633	0.179	0.017
0.58	0.96	0.98	0.99	0.98	0.492	0.302	0.98	0.373	0.247	0.017
0.59	0.96	0.98	0.99	0.98	0.491	0.303	0.98	0.372	0.248	0.017
0.46	0.92	0.97	0.99	0.977	0.534	0.291	0.977	0.448	0.256	0.017
0.48	0.92	0.97	0.99	0.978	0.533	0.292	0.978	0.448	0.257	0.017

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

for the two databases, stratified by sex, computed at 100, 112 and 100 weeks. A single illustrative operating point is also shown on the ROC curve in Fig. 2C, where at 150 weeks, we have a sensitivity of 51.8% and a PPV of 15.8% and 18.8% for males and females respectively, both at a specificity of 95%.

Beyond standalone performance, independence from standardized questionnaires implies that we stand to gain substantially from combined operation. With the recently reported population stratification induced by M-CHAT/F scores (3) (SI-Table 3), we can compute a conditional choice of sensitivity for our tool, in each sub-population (M-CHAT/F score brackets: $0-2$, $3-7$ (negative assessment), $3-7$ (positive assessment), and >8), leading to a significant performance boost. With such conditional operation, we get a PPV close to or exceeding 30% at the high precision (HP) operating point across datasets ($>33\%$ for Truven, $>28\%$ for UCM), or a sensitivity close to or exceeding 50% for the high recall (HR) operating point ($>58\%$ for Truven, $>50\%$ for UCM), when we restrict specificities to above 95% (See Table 5, Fig. 4B(i), and SI-Fig. 4 in the supplementary text). Comparing with standalone M-CHAT/F performance (Fig. 4B(ii)), we show that for any prevalence between 1.7% and 2.23%, we can *double the PPV* without losing sensitivity at $>98\%$ specificity, or increase the sensitivity by $\sim 50\%$ without sacrificing PPV and keeping specificity $\geqq 94\%$.

Discussion

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

Attempts at curating genetic biomarkers has also met with limited success. ASD genes exhibit extensive phenotypic variability, with identical variants associated with diverse individual outcomes not limited to ASD, including schizophrenia, intellectual disability, language impairment, epilepsy, neuropsychiatric disorders and, also typical development (45). Additionally, no single gene can be considered "causal" for

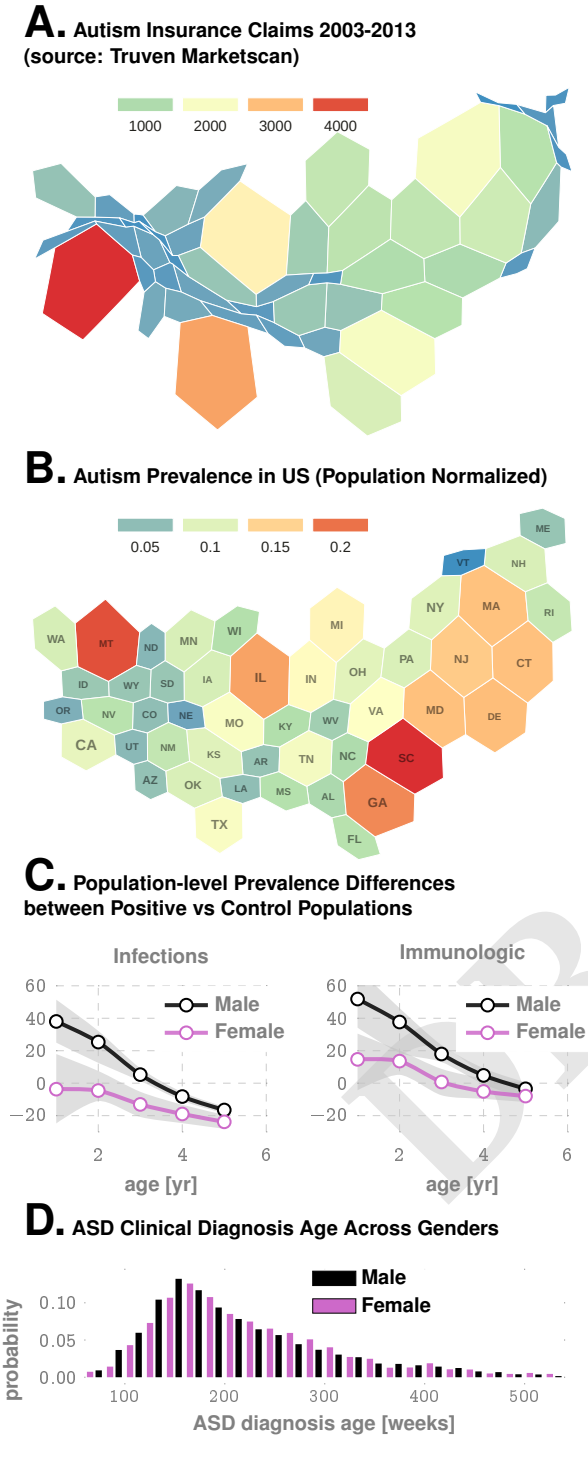


Fig. 1. ASD Occurrence Patterns. Panel A illustrates the differential representation of different disease categories in the positive and control cohorts, and panel B shows the distribution of the age of diagnosis for males and females in the Truven dataset. Panel C illustrates the spatial distribution of ASD insurance claims, and panel D shows the same data after population normalization, illustrating the relatively small demographic skew to ASD prevalence within the general population with access to medical insurance, which is consistent with the suggestion that prevalence variation might be linked to regional and socioeconomic disparities in access to services (16).

more than 1% of cases of idiopathic autism (46).

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

Going beyond screening performance, this approach provides a new tool to uncover clues to ASD pathobiology. Perhaps this vulnerability to diverse immunological, endocrinological and neurological impairments reflects how allostatic loads of medical stress get under the skin and disrupt key regulators of CNS organization and synaptogenesis. Charting individual disorders in the co-morbidity burden further reveals novel associations in normalized prevalence | the odds of experiencing a specific disorder, particularly in the early years (age < 3 years), normalized over all unique disorders experienced in the specified time-frame. We focus on the true positives in the positive cohort and the true negatives in the control cohort to investigate patterns that correctly disambiguate ASD status. On these lines Fig. 5 and SI-Fig. 1 in the supplementary text outline two key observations: 1) *negative associations:* some diseases that are negatively associated with ASD with respect to normalized prevalence, *i.e.*, having those codes relatively over-represented in one's diagnostic history favors ending up in the control cohort, 2) *impact of sex:* there are sex-specific differences in the impact of specific disorders, and given a fixed level of impact, the number of codes that drive the outcomes is significantly more in males (Fig. 5A vs B).

Some of the disorders that show up in Fig. 5, panels A and B are surprising, *e.g.*, congenital hemiplegia or diplegia of the upper limbs indicative of either cerebral palsy (CP) or a spinal cord/brain injury, neither of which has a direct link to autism. Since only about 7% of the children with cerebral palsy (CP) are estimated to have a co-occurring ASD (47, 48), and with the prevalence of CP significantly lower (1 in 352 vs 1 in 59 for autism), it follows that only a small number of children (approximately 1.17%) with autism have co-occurring CP. Thus, with significantly higher prevalence in children diagnosed with autism compared to the general population (1.7% vs 0.28%), CP codes show up with higher odds in the true positive set. Also, SI-Fig. 1A shows that the immunological, metabolic, and endocrine disorders are almost completely risk-increasing. In contrast, respiratory diseases (panel B) are largely risk-decreasing. On the other hand, infectious diseases have roughly equal representations in the risk-increasing and risk-decreasing classes (panel C). The risk-decreasing infectious diseases tend to be due to viral or fungal organisms, which might point to the use of antibiotics in bacterial infections, and the consequent dysbiosis of the gut microbiota (23, 37) as a risk factor.

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

We carried out a battery of tests to ensure that our results are not significantly impacted by class imbalance (since our

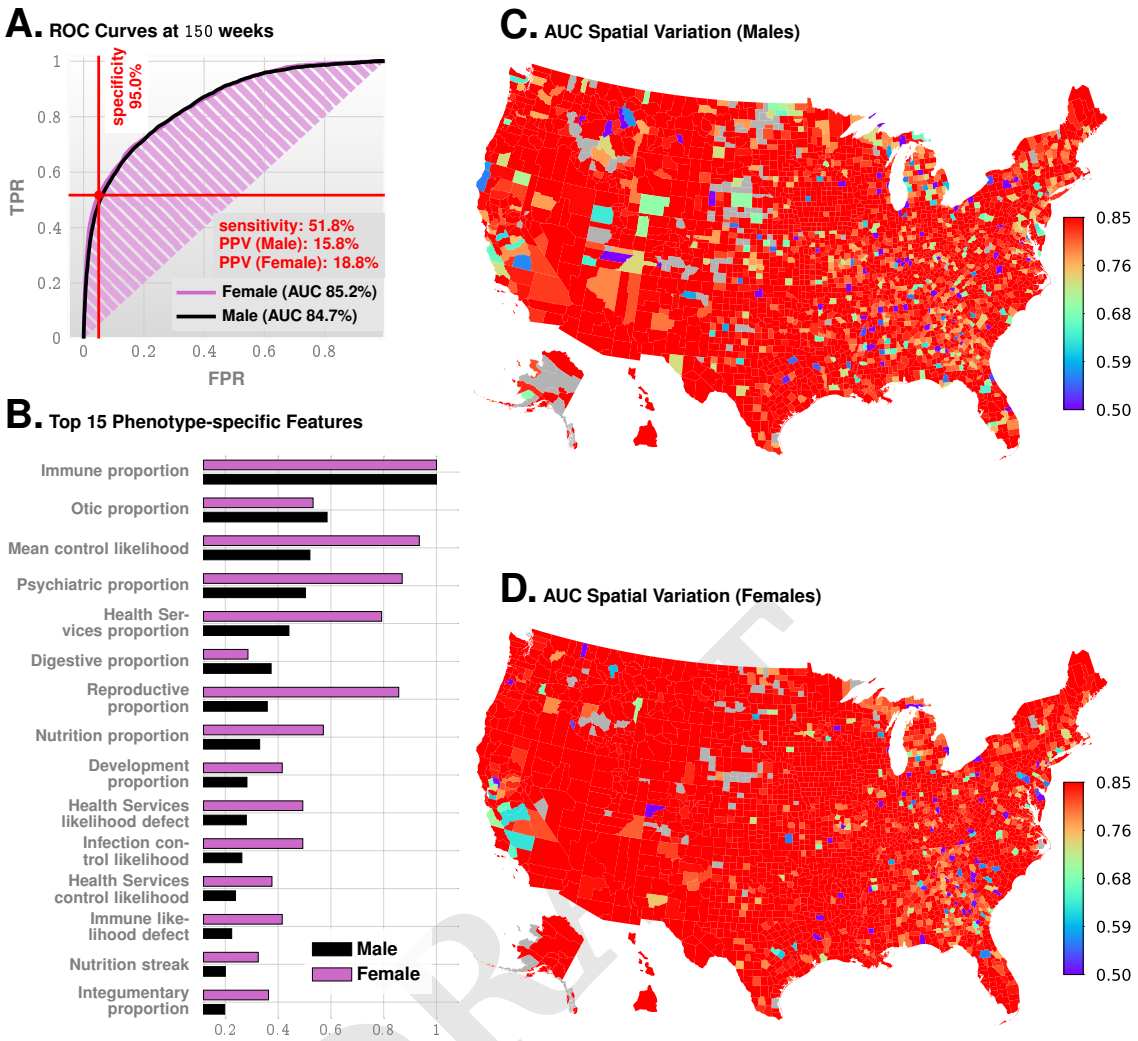


Fig. 2. Predictive Performance. Panel A shows the ROC curves for males and females. Panel B shows the feature importance inferred by our prediction pipeline. The detailed description of the features is given in Table 2. The most import feature is related to immunologic disorders, and we note that in addition to features related to individual disease categories, we also have the mean control likelihood (rank 3), which may be interpreted as the average likelihood of the diagnostic patterns correspond to the control category as opposed to the positive category. Panels C and D show the spatial variation in the achieved predictive performance at 150 weeks, measured by AUC, for males and females, respectively. Gray areas lack data on either positive or negative cases. These county-specific AUC plots show that the performance of the algorithm has relatively weak geospatial dependence, which is important in the light of current uneven distribution of diagnostic resources.

control cohort is orders of magnitude larger) or systematic coding errors (See Methods), *e.g.*, we verified that restricting the positive cohort to children with at least two distinct ASD diagnostic codes in their medical histories instead of one, has little impact on out-of-sample predictive performance (SI-Fig. 5B).

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

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

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

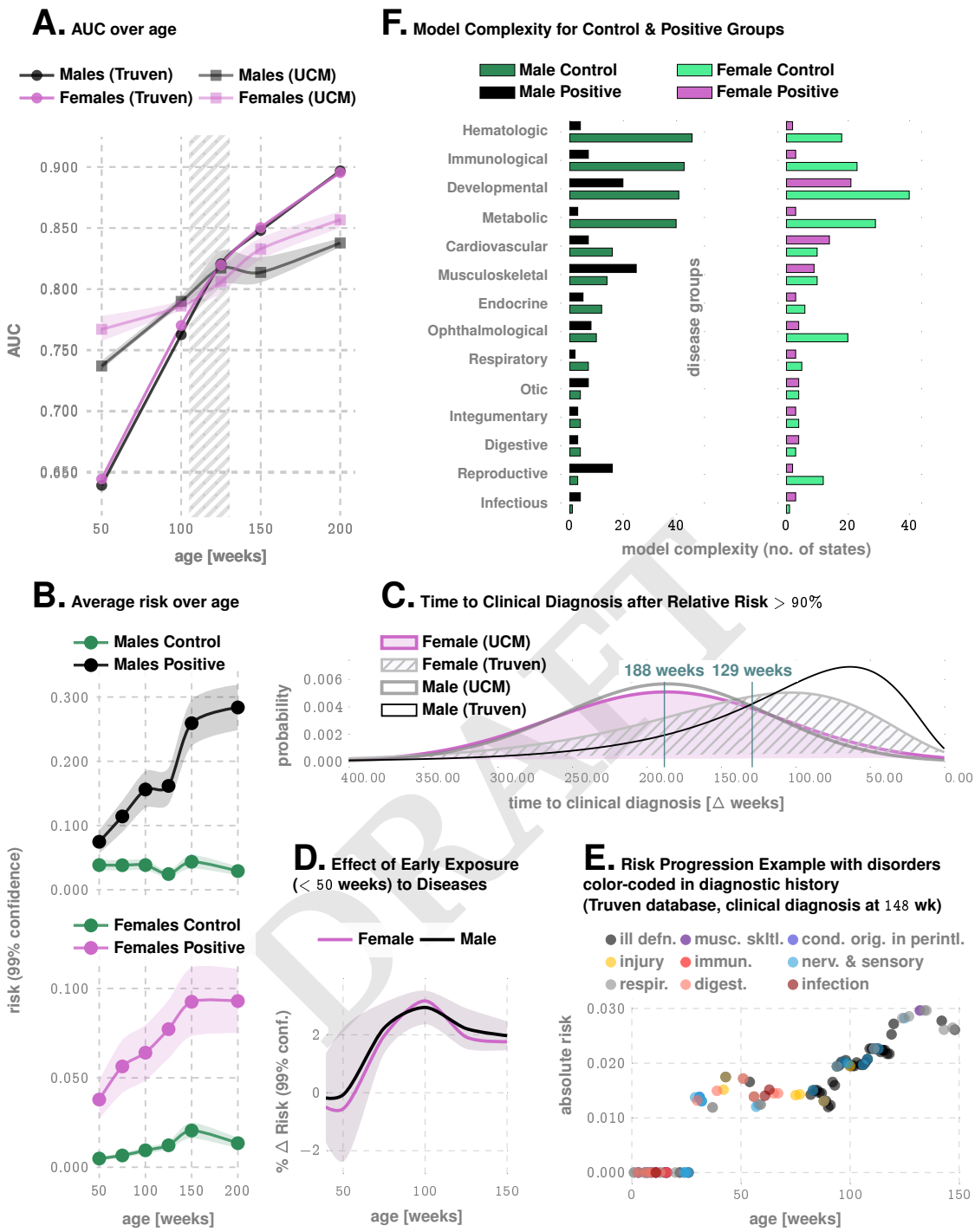


Fig. 3. More details on Predictive Performance and Variation of Inferred Risk. Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve $> 80\%$ AUC for either sex from shortly after 2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

of general comorbidity-aware risk predictions from electronic health record databases.

Data Sharing. Software implementation of the pipeline is available at: https://github.com/zeroknowledgediscovery/ehrzero, and installation in standard python environments may be

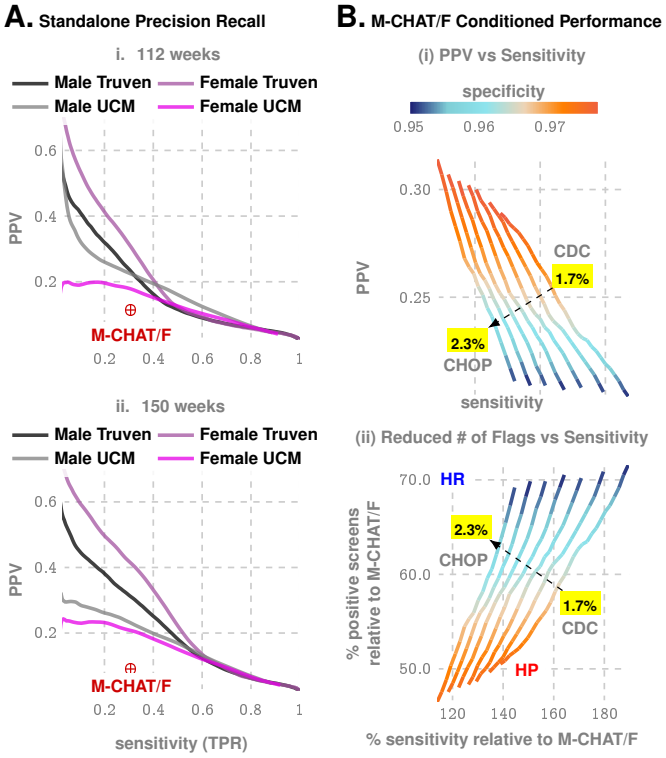


Fig. 4. Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs. Panel A shows the precision/recall curves, *i.e.*, the trade-off between PPV and sensitivity. Panel B shows how we can boost performance using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study (3). Panel C illustrates the boosted performance compared to M-CHAT/F alone, measured by the relative percentage increase in sensitivity, and percentage decrease in positive screens. Note that the population prevalence impacts this optimization, and hence we have a distinct curve for each prevalence value (1.7% is the CDC estimate, while 2.23% is reported by the CHOP study). The two extreme operating zones marked as High Precision (HP) and High Recall (HR): if we choose to operate in HR, then we do not reduce the number of positive screens by much, but maximize sensitivity, while by operating in HP, we do not increase sensitivity by much but double the PPV achieved in current practice. Note in all these zones, we maintain specificity above 95%, which is the current state of art, implying that by doubling the PPV, we can halve the number of positive screens currently reported, thus potentially sharply reducing the queues and wait-times.

done from https://pypi.org/project/ehrzero/. A sample of de-identified data from the UCM database is be shared as part of the public software package.

Materials and Methods

Source of Electronic Patient Records. We view the task of predicting ASD diagnosis as a binary classification problem: sequences of diagnostic codes are classified into positive and control categories, where "positive" refers to children eventually diagnosed with ASD, as indicated by the presence of a clinical diagnosis (ICD9 code 299.X) in their medical records. Of the two independent sources of clinical incidence data used in this study, the primary source used to train our predictive pipeline is the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database for the years 2003 to 2012 (50) (referred to as the Truven dataset). This US national database contains data contributed by over 150 insurance carriers, large self-insuring companies, and Medi and is a culmination of over 4.6 billion inpatient and outpatient service claims and almost six billion diagnosis codes. We extracted histories of patients within the age of $0-9$ years, and excluded patients for whom: 1) At least one code of any available phenotypes is present, 2) Lag between first and last available record for a patient should be at least 15 weeks. These exclusion criteria ensure that we are not considering patients who have too few observations to either train on. Additionally, during validation runs, we restricted the control set to patients observable in the databases to those whose last record is not before the first 150 weeks of life. Characteristics of excluded patients is shown in Table 1. We trained with over 30M diagnostic records (16,649,548 for males and 14,318,303 for females with 9,835 unique codes).

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

Our datasets are consistent with documented prevalence, with no significant geospatial prevalence variation (Fig. 1D) and reveal that infections and immunological disorders have differential representation in the positive and control groups (Fig. 1C). The median diagnosis age is just over 3 years in the claims database (Fig. 1B) versus 3 years 10 months to 4 years in US (51). Cohort details are given in Table 1 and discussed in Methods. Importantly, for the positive cohort, we only consider diagnostic history up to the first ASD code.

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

Time-series Modeling of Diagnostic History. Individual diagnostic histories can have long-term memory (52), implying that the order, frequency, and comorbid interactions between diseases are important for assessing the future risk of our target phenotype. We analyze patient-specific diagnostic code sequences by first representing the medical history of each patient as a set of stochastic categorical time-series | one each for a specific group of related disorders | followed by the inference of stochastic models for these individual data streams. These inferred generators are from a special class of Hidden Markov Models (HMMs), referred to as Probabilistic Finite State Automata (PFSA) (53). The inference algorithm we use is distinct from classical HMM learning, and has important advantages related to its ability to infer structure, and its sample complexity (See Supplementary text, Section 11). 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.

Step 1: Partitioning The Human Disease Spectrum. We begin by partitioning the human disease spectrum into 17 non-overlapping categories, as shown in Table 2. Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9) (See Table 2 in the main text and Table SI-4 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) (17) 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;

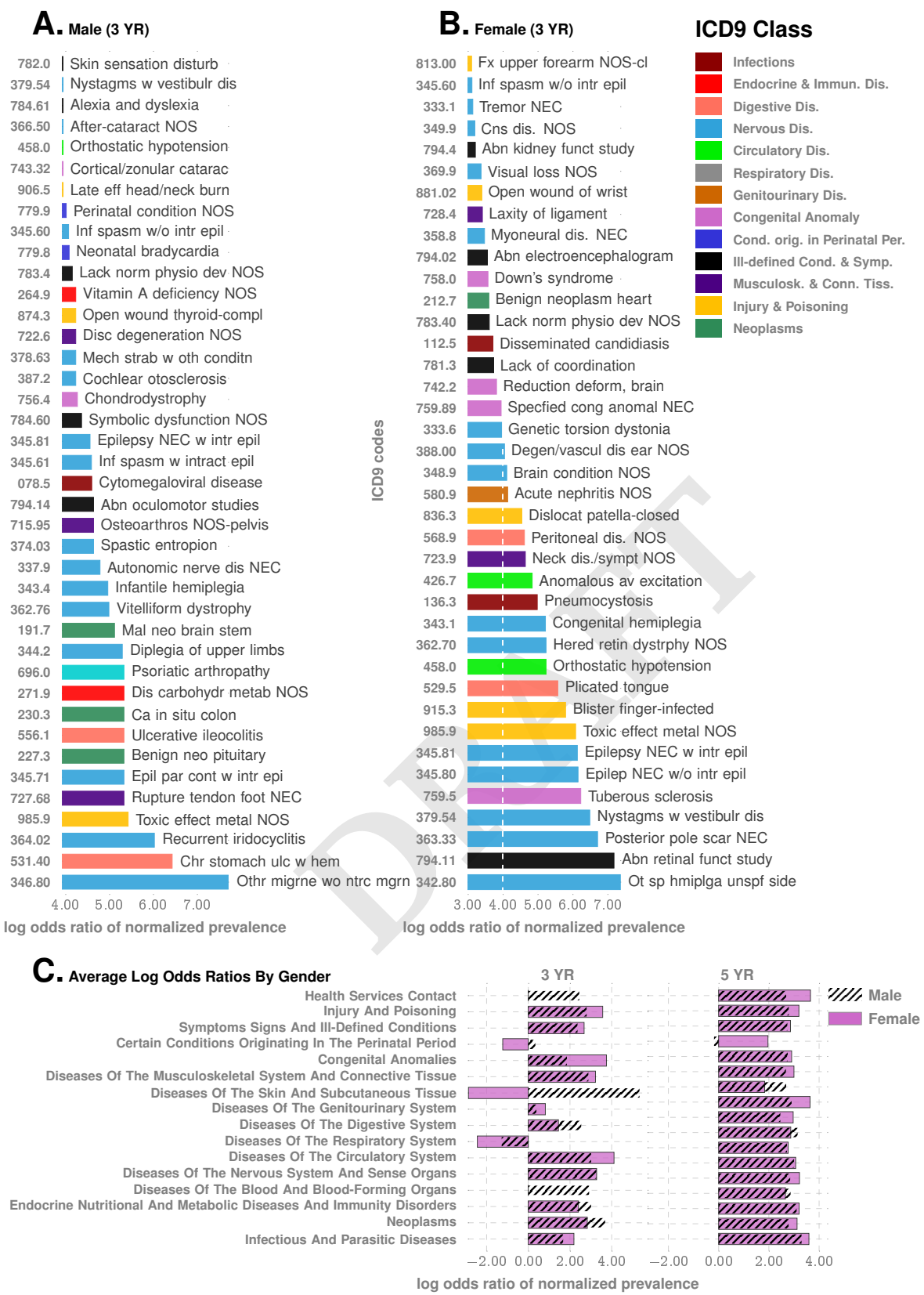


Fig. 5. Co-morbidity Patterns Panel A and B. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions. The dotted line on panel B shows the abscissa lower cut-off in Panel A, illustrating the lower prevalence of codes in females. Panel C illustrates log-odds ratios for ICD9 disease categories at different ages. Importantly, the negative associations disappear when we consider older children, consistent with the lack of such reports in the literature which lack studies on very young cohorts.

we want our algorithm to seek out the important patterns without any manual curation of the input data. The limitation of the set of phenotypes to 9835 unique codes arises from excluding patients from the database who have very few and rare codes that will skew the statistical estimates. As shown in Table 1, we exclude a very small number of patients, and who have very short diagnostic histories with a very small number of codes.

For each patient, the past medical history is a sequence $(t_1, x_1), \cdots, (t_m, x_m)$, where t_i are timestamps and x_i are ICD9 codes diagnosed at time t_i. We map individual patient history to a three-alphabet categorical time series z^k corresponding to the disease category k, as follows. For each week i, we have:

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

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.

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 (PFSAs). 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. 7 in the supplementary text provides some examples). Each of these inferred models is a PFSA; a directed graph with probability-weighted edges, and acts as an optimal generator of the stochastic process driving the sequential appearance of the three letters (as defined by Eq. Eq. (1)) corresponding to each sex, disease category, and cohort-type (See Section 11 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 sequnce being generated by the corresponding cohort model, we generalize the notion of Kullbeck-Leibler (KL) divergence (54, 55) between probability distributions to a divergence $\mathcal{D}_{\mathrm{KL}}(G||H)$ between ergodic stationary categorical stochastic processes (56) G, H as:

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

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 $x \mid$ which is actually generated by the hidden process $G \mid$ can be formally inferred from observations based on the following provable relationships (See Suppl. text Section 11, Theorem 6 and 7):

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

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

where $\mathcal{H}(\cdot)$ is the entropy rate of a process (54). Importantly, Eq. Eq. (4) shows that the computed likelihood has an additional non-negative contribution from the divergence term when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with ASD, then we expect that the disease-specific mapped series corresponding to her diagnostic history be modeled by the PFSA in the positive cohort. Denoting the PFSA corresponding to disease category j for positive and control cohorts as G_+^j, G_0^j respectively, we can compute the *sequence likelihood defect* (SLD, Δ^j) as:

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

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

Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules. The risk estimation pipeline operates on patient specific information limited to the sex and available diagnostic history from birth, and produces an estimate of the relative risk of ASD diagnosis at a specific age, with an associated confidence value. To learn the parameters and associated model structures of this pipeline, we transform the patient specific data to a set of engineered features, and the feature vectors realized on the positive and control sets are used to train a gradient-boosting classifier (57). The complete list of 165 features used is provided in Table 3.

We need two training sets: one to infer the models, and one to train the classifier with features derived from the inferred models. Thus, we do a random 3-way split of the set of unique patients into *feature-engineering* (25%), *training* (25%) and *test* (50%) sets. We use the feature-engineering set of ids first to infer our PFSA models *(unsupervised model inference in each category)*, which then allows us to train the gradient-boosting classifier using the training set and PFSA models *(classical supervised learning)*, and we finally execute out-of-sample validation on the test set. Fig. 2B shows the top 15 features ranked in order of their relative importance (relative loss in performance when dropped out of the analysis).

Calculating Relative Risk. Our pipeline maps medical histories to a raw indicator of risk. However, to make crisp predictions, we must choose a decision threshold for this raw score. Conceptually identical to the notion of Type 1 and Type 2 errors in classical statistical analyses, the choice of a threshold trades off false positives (Type 1 error) for false negatives (Type 2 error): choosing a small threshold results in predicting a larger fraction of future diagnoses correctly, *i.e.* have a high true positive rate (TPR), while simultaneously suffering from a higher false positive rate (FPR), and vice versa. Therefore, a choice of a specific decision threshold reflects a choice of the maximum FPR and minimum TPR, and is driven by the application at hand. In this study, we base our analysis on maximizing the F_1-score, defined as the harmonic mean of sensitivity and specificity, to make a balanced trade-off between the two kinds of errors (See Supplementary text, Section 5). The *relative risk* is then defined as the ratio of the raw risk to the decision threshold, and a value > 1 predicts a future ASD diagnosis.

Boosting Performance Via Leveraging Population Stratification Induced By Existing Tests. We leverage the population stratification induced by an existing independent screening test (M-CHAT/F) to improve combined performance. Here a combination refers to the conditional choice of the sensitivity/specificity trade-offs for our tool in each sub-population such that the overall performance is optimized with respect to whether we wish to maximize the PPV or the sensitivity at a specified minimum level of specificity. Assume that there are m sub-populations such that: the sensitivities, specificities achieved, and the prevalences in each sub-population are given by s_i, c_i and ρ_i respectively, with $i \in \{1, \cdots, m\}$. Let β_i be the relative size of each sub-population. Then, we have (See Supplementary text, Section 9.1):

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

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

where we have denoted:

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

and s, c, ρ are the overall sensitivity, specificity, and prevalence. Knowing the values of γ_i, γ_i', we can carry out an m-dimensional search to identify the feasible choices of s_i, c_i pairs for each i, such that some global constraint is satisfied, *e.g.* minimum values of

specificity, sensitivity, and PPV. We consider 4 sub-populations defined by M-CHAT/F score brackets (3), and if the screen result is considered a positive (high risk, indicating the need for a full diagnostic evaluation) or a negative, *i.e.*, low risk: 1) score ≤ 2 screening ASD negative, 2) score $[3-7]$ screening ASD negative on follow-up, 3) score $[3-7]$ and screening ASD positive on follow-up, and 4) score ≥ 8, screening ASD positive. (See SI-Table 2). The "follow-up" in the context of M-CHAT/F refers to the re-evaluation of responses by qualified personnel. We use published data on the relative sizes and the prevalence statistics in these sub-populations (3) to compute the feasible conditional choices of our operating point to strictly supersede M-CHAT/F performance. Two limiting operating conditions are of special interest here, where we maximize PPV under some minimum specificity and sensitivity (denoted as the High Precision or the HP operating point), and where we maximize sensitivity under some minimum PPV and specificity (denoted as the High Recall or the HR operating point). Taking these minimum values of specificity, sensitivity, and PPV to be those reported for M-CHAT/F, we identify the set feasible set of conditional choices in a four-dimensional decision space that would outperform M-CHAT/F in universal screening. The results are shown in Fig. 4B.

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

1. Data & statistics on autism spectrum disorder | cdc (2019).
2. L Gilotty, Early screening for autism spectrum (2019).
3. W Guthrie, et al., Accuracy of Autism Screening in a Large Pediatric Network. *Pediatrics* **144** (2019).
4. E Gordon-Lipkin, J Foster, G Peacock, Whittling Down the Wait Time: Exploring Models to Minimize the Delay from Initial Concern to Diagnosis and Treatment of Autism Spectrum Disorder. *Pediatr. Clin. North Am.* **63**, 851–859 (2016).
5. Data & statistics on autism spectrum disorder | cdc (2019).
6. LA Schieve, et al., Population attributable fractions for three perinatal risk factors for autism spectrum disorders, 2002 and 2008 autism and developmental disabilities monitoring network. *Ann Epidemiol* **24**, 260–266 (2014).
7. F Volkmar, et al., Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *J. Am. Acad. Child & Adolesc. Psychiatry* **53**, 237–257 (2014).
8. SL Hyman, SE Levy, SM Myers, , et al., Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics* **145** (2020).
9. LG Kalb, et al., Determinants of appointment absenteeism at an outpatient pediatric autism clinic. *J. Dev. & Behav. Pediatr.* **33**, 685–697 (2012).
10. J Bisgaier, D Levinson, DB Cutts, KV Rhodes, Access to autism evaluation appointments with developmental-behavioral and neurodevelopmental subspecialists. *Arch. pediatrics & adolescent medicine* **165**, 673–674 (2011).
11. TS Fenikilé, K Ellerbeck, MK Filippi, CM Daley, Barriers to autism screening in family medicine practice: a qualitative study. *Prim. health care research & development* **16**, 356–366 (2015).
12. E Gordon-Lipkin, J Foster, G Peacock, Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. *Pediatr. Clin.* **63**, 851–859 (2016).
13. DL Robins, et al., Validation of the modified checklist for autism in toddlers, revised with follow-up (m-chat-r/f). *Pediatrics* **133**, 37–45 (2014).
14. C Tye, AK Runicles, AJO Whitehouse, GA Alvares, Characterizing the Interplay Between Autism Spectrum Disorder and Comorbid Medical Conditions: An Integrative Review. *Front Psychiatry* **9**, 751 (2018).
15. IS Kohane, et al., The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS ONE* **7**, e33224 (2012).
16. VG Jarquin, LD Wiggins, LA Schieve, K Van Naarden-Braun, Racial disparities in community identification of autism spectrum disorders over time; metropolitan atlanta, georgia, 2000–2006. *J. Dev. & Behav. Pediatr.* **32**, 179–187 (2011).
17. General equivalence mappings (year?).
18. LA Althouse, JA Stockman, Pediatric workforce: A look at pediatric nephrology data from the american board of pediatrics. *The J. pediatrics* **148**, 575–576 (2006).
19. O Zerbo, et al., Immune mediated conditions in autism spectrum disorders. *Brain Behav. Immun.* **46**, 232–236 (2015).
20. H Won, W Mah, E Kim, Autism spectrum disorder causes, mechanisms, and treatments: focus on neuronal synapses. *Front Mol Neurosci* **6**, 19 (2013).
21. G Xu, et al., Association of Food Allergy and Other Allergic Conditions With Autism Spectrum Disorder in Children. *JAMA Netw Open* **1**, e180279 (2018).
22. JB Adams, et al., Nutritional and metabolic status of children with autism vs. neurotypical children, and the association with autism severity. *Nutr Metab (Lond)* **8**, 34 (2011).
23. A Fattorusso, L Di Genova, GB Dell'Isola, E Mencaroni, S Esposito, Autism Spectrum Disorders and the Gut Microbiota. *Nutrients* **11** (2019).
24. R Diaz Heijtz, et al., Normal gut microbiota modulates brain development and behavior. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 3047–3052 (2011).
25. S Rose, et al., Mitochondrial dysfunction in the gastrointestinal mucosa of children with autism: A blinded case-control study. *PLoS ONE* **12**, e0186377 (2017).
26. EM Sajdel-Sulkowska, et al., Common Genetic Variants Link the Abnormalities in the Gut-Brain Axis in Prematurity and Autism. *Cerebellum* **18**, 255–265 (2019).
27. MS Kayser, J Dalmau, Anti-NMDA Receptor Encephalitis in Psychiatry. *Curr Psychiatry Rev* **7**, 189–193 (2011).
28. OI Dadalko, BG Travers, Evidence for Brainstem Contributions to Autism Spectrum Disorders. *Front Integr Neurosci* **12**, 47 (2018).
29. Y Yamashita, et al., Anti-inflammatory Effect of Ghrelin in Lymphoblastoid Cell Lines From Children With Autism Spectrum Disorder. *Front Psychiatry* **10**, 152 (2019).
30. L Shen, et al., Proteomics Study of Peripheral Blood Mononuclear Cells (PBMCs) in Autistic Children. *Front Cell Neurosci* **13**, 105 (2019).
31. K Ohja, et al., Neuroimmunologic and Neurotrophic Interactions in Autism Spectrum Disorders: Relationship to Neuroinflammation. *Neuromolecular Med.* **20**, 161–173 (2018).
32. D Gadysz, A Krzywdziska, KK Hozyasz, Immune Abnormalities in Autism Spectrum Disorder-Could They Hold Promise for Causative Treatment? *Mol. Neurobiol.* **55**, 6387–6435 (2018).
33. TC Theoharides, I Tsilioni, AB Patel, R Doyle, Atopic diseases and inflammation of the brain in the pathogenesis of autism spectrum disorders. *Transl Psychiatry* **6**, e844 (2016).
34. AM Young, et al., From molecules to neural morphology: understanding neuroinflammation in autism spectrum condition. *Mol Autism* **7**, 9 (2016).
35. LA Croen, et al., Family history of immune conditions and autism spectrum and developmental disorders: Findings from the study to explore early development. *Autism Res* **12**, 123–135 (2019).
36. T Vargason, DL McGuinness, J Hahn, Gastrointestinal Symptoms and Oral Antibiotic Use in Children with Autism Spectrum Disorder: Retrospective Analysis of a Privately Insured U.S. Population. *J Autism Dev Disord* **49**, 647–659 (2019).
37. M Fiorentino, et al., Blood-brain barrier and intestinal epithelial barrier alterations in autism spectrum disorders. *Mol Autism* **7**, 49 (2016).
38. FK Satterstrom, et al., Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *bioRxiv* (2019).
39. T Gaugler, et al., Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
40. DL Vargas, C Nascimbene, C Krishnan, AW Zimmerman, CA Pardo, Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann. Neurol.* **57**, 67–81 (2005).
41. H Wei, et al., IL-6 is increased in the cerebellum of autistic brain and alters neural cell adhesion, migration and synaptic formation. *J Neuroinflammation* **8**, 52 (2011).
42. AM Young, E Campbell, S Lynch, J Suckling, SJ Powis, Aberrant NF-kappaB expression in autism spectrum condition: a mechanism for neuroinflammation. *Front Psychiatry* **2**, 27 (2011).
43. HK Hughes, E Mills Ko, D Rose, P Ashwood, Immune Dysfunction and Autoimmunity as Pathological Mechanisms in Autism Spectrum Disorders. *Front Cell Neurosci* **12**, 405 (2018).
44. N Pearce, The ecological fallacy strikes back. *J. epidemiology community health* **54**, 326–7 (2000).
45. JD Murdoch, MW State, Recent developments in the genetics of autism spectrum disorders. *Curr. Opin. Genet. Dev.* **23**, 310–315 (2013).
46. VW Hu, The expanding genomic landscape of autism: discovering the 'forest' beyond the 'trees'. *Futur. Neurol* **8**, 29–42 (2013).
47. Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning (2020).
48. D Christensen, et al., Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning–a utism and d evelopmental d isabilities m onitoring n etwork, usa, 2008. *Dev. Medicine & Child Neurol.* **56**, 59–65 (2014).
49. EM Rødgaard, K Jensen, JN Vergnes, I Soulières, L Mottron, Temporal Changes in Effect Sizes of Studies Comparing Individuals With and Without Autism: A Meta-analysis. *JAMA Psychiatry* **76**, 1124–1132 (2019).
50. L Hansen, The truven health marketscan databases for life sciences researchers. *Truven Heal. Ananlytics IBM Watson Heal.* (2017).
51. J Baio, et al., Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR Surveill Summ* **67**, 1–23 (2018).
52. CWJ Granger, R Joyeux, An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Analysis* **1**, 15–29 (year?).
53. I Chattopadhyay, H Lipson, Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).
54. TM Cover, JA Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. (Wiley-Interscience, New York, NY, USA), (2006).
55. S Kullback, RA Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
56. J Doob, *Stochastic Processes*, Wiley Publications in Statistics. (John Wiley & Sons), (1953).
57. JH Friedman, Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).

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

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

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

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

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

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

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

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

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

Contents

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

SI-Table 1: Boosted Sensitivity, Specificity and PPV Achieved at **150 weeks** Conditioned on M-CHAT/F Scores

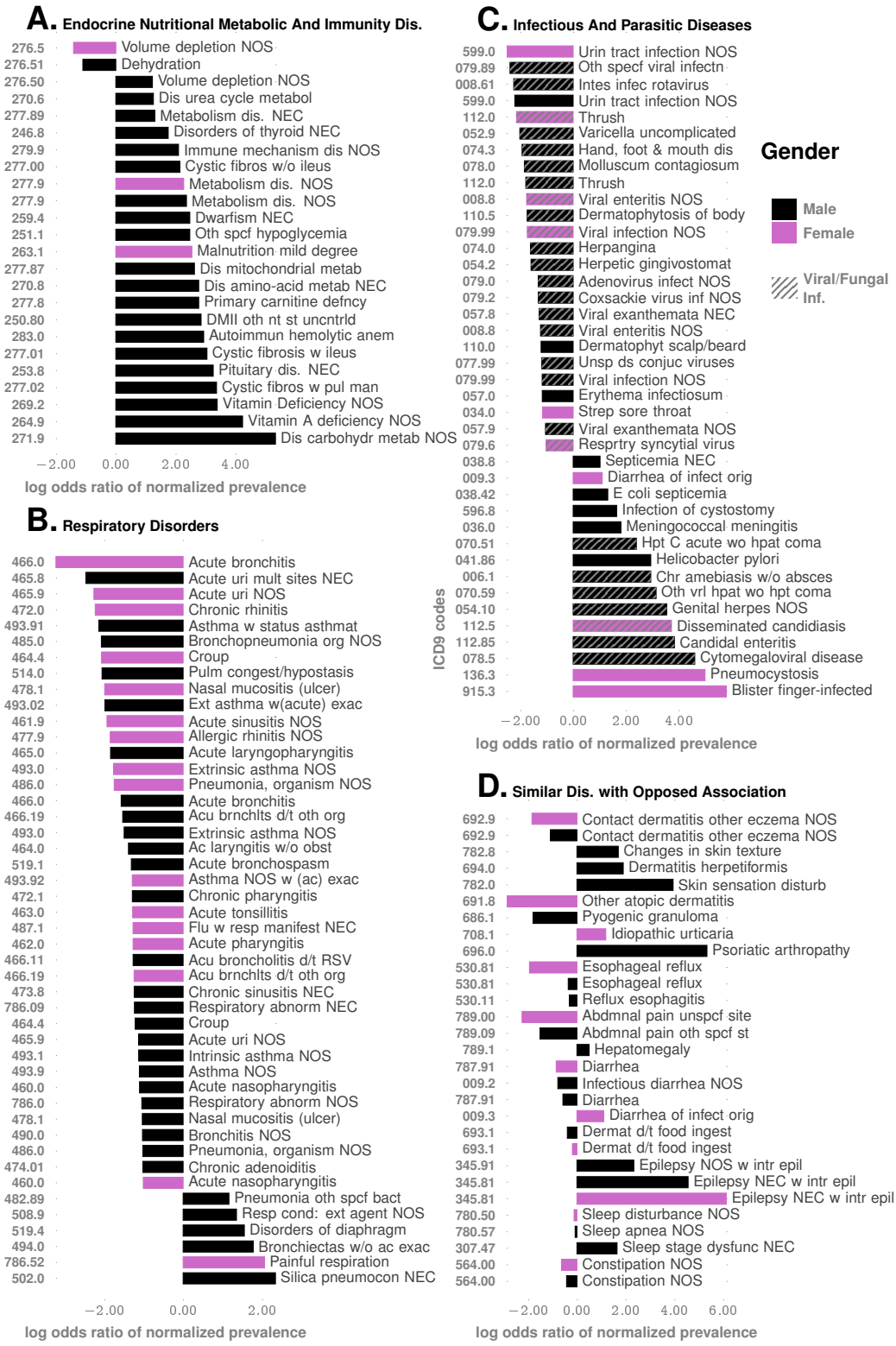
M-CHAT/F Outcome				global performance (Truven)			global performance (UCM)			prevalence
0-2 NEG	3-7 NEG	3-7 POS	≥ 8 POS	speci-ficity	sensi-tivity	PPV	speci-ficity	sensi-tivity	PPV	
specificity choices										
0.28	0.66	0.93	0.97	0.95	0.64	0.224	0.95	0.577	0.206	0.022
0.31	0.67	0.9	0.97	0.95	0.641	0.223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

SI-Table 2: Population Stratification Results on large M-CHAT/F Study(n=20,375)[1]

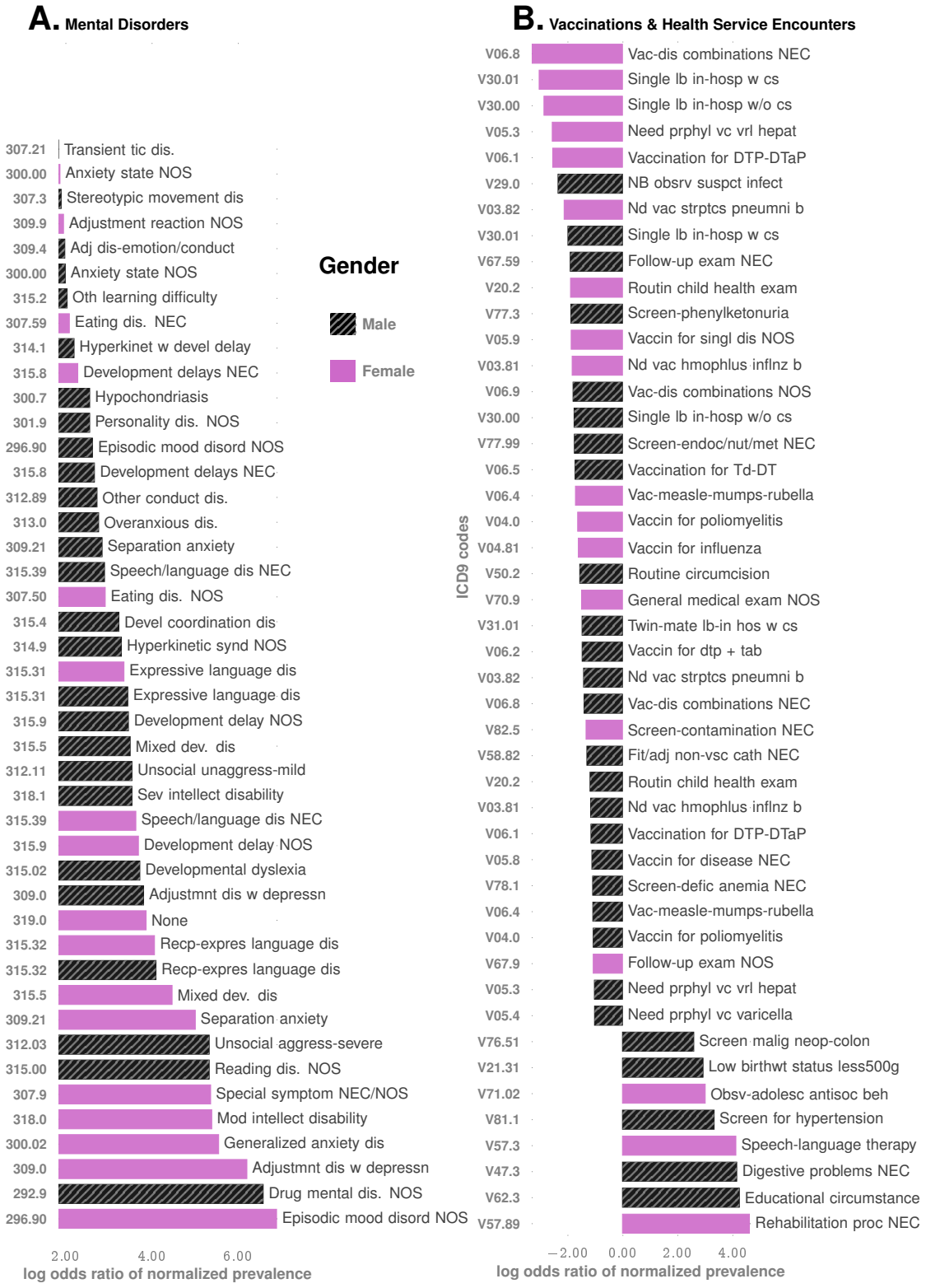
Id	Sub-population	Test Result	ASD positive	ASD Negative	Total %
A	M-CHAT/F $\geqq 8$	Positive	0.34%	0.64%	0.99%
B	M-CHAT/F $\in [3,7]$	Positive (after follow-up)	0.52%	4.39%	4.91%
C	M-CHAT/F $\in [3,7]$	Negative (after follow-up)	0.14%	3.1%	3.24%
D	M-CHAT/F $\in [0,2]$	Negative	1.22%	89.63%	90.86%
Total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

SI-Table 3: γ, γ' Computed from Population Stratification Recorded In M-CHAT/F Study[1] ($\rho = 0.0223$)

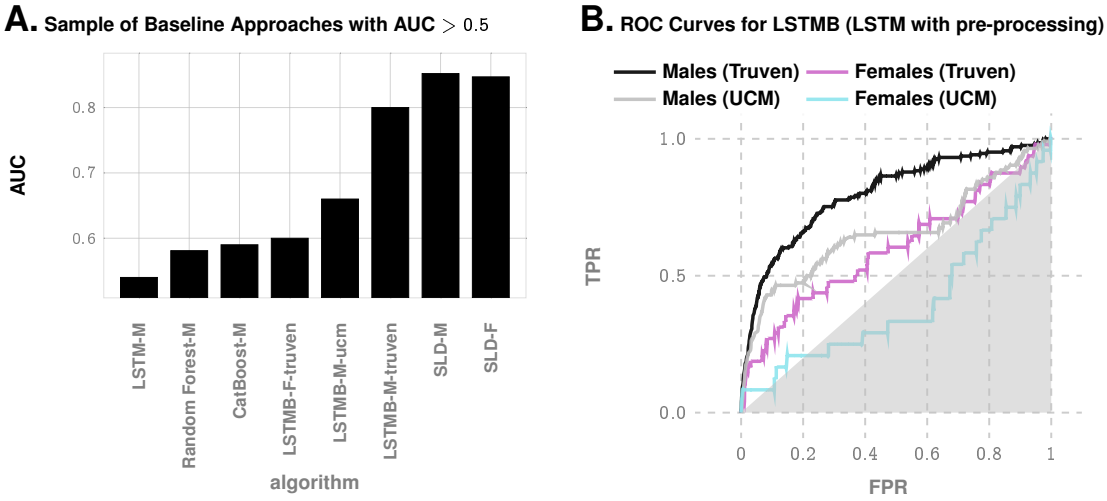
Id	Sub-population	Test Result	β_i	ρ_i	γ_i	γ_i'
A	M-CHAT/F $\geqq 8$	Positive	.0099	.3469	.1540	.0066
B	M-CHAT/F $\in [3,7]$	Positive (after follow-up)	.0491	.1059	.2331	.0449
C	M-CHAT/F $\in [3,7]$	Negative (after follow-up)	.0324	.0432	.0627	.0317
D	M-CHAT/F $\in [0,2]$	Negative	.9086	.0134	.5471	.9168



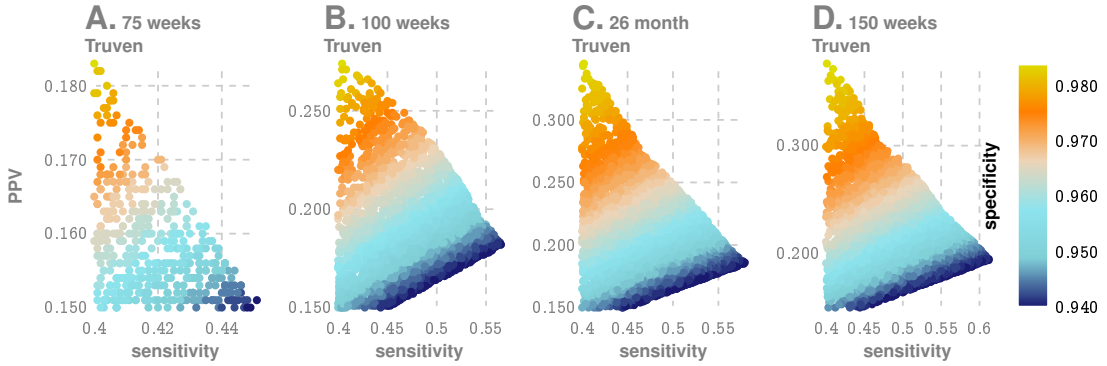
SI-Fig. 1: **Details of Co-morbidity Patterns (at age < 3 years)** for immunologic (panel A), respiratory (panel B), infections (panel C), and disorders with similar pathobiology manifesting opposing association with autism (panel D).



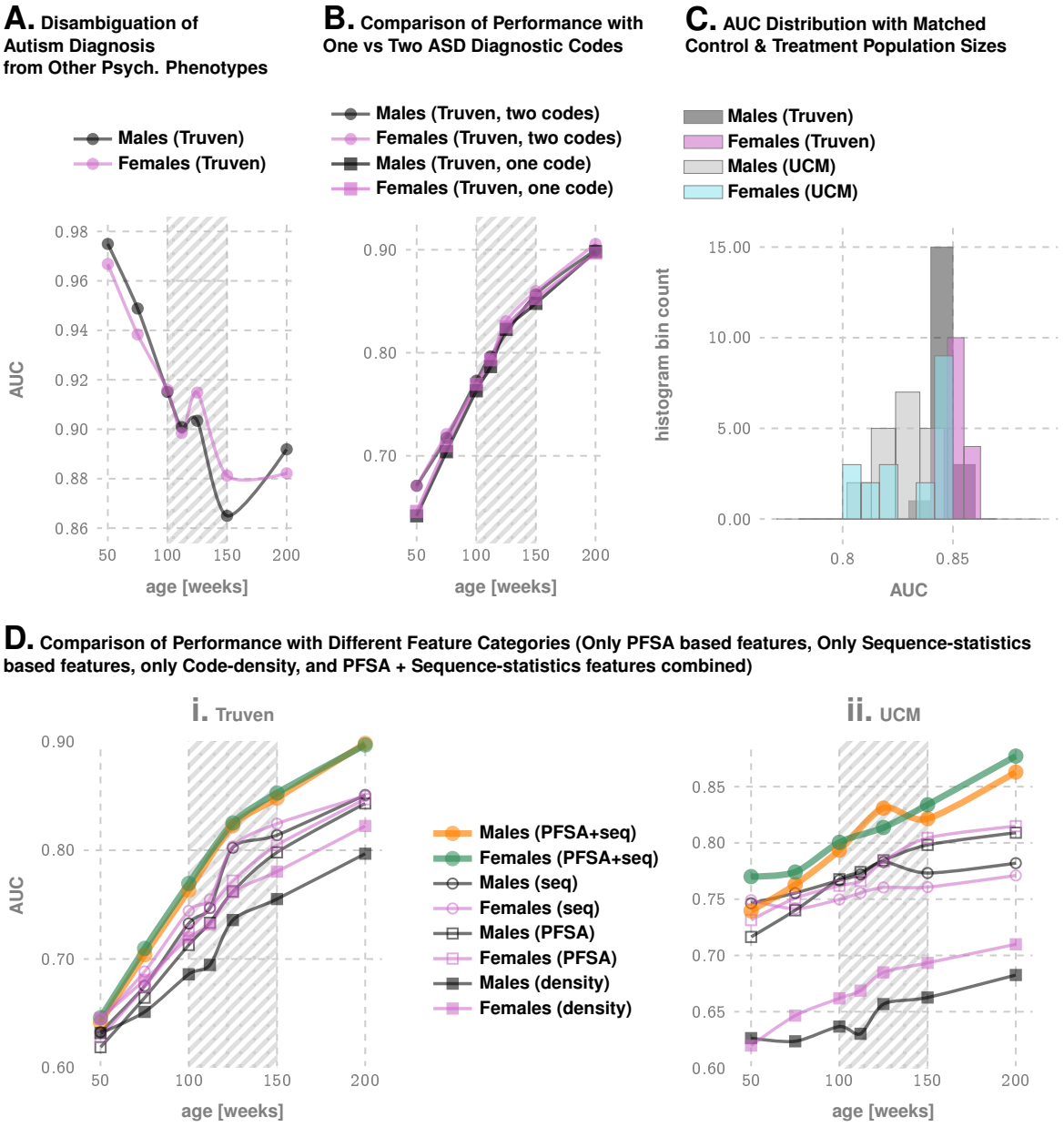
SI-Fig. 2: **Co-morbidity Patterns** for mental disorders, vaccinations and health-service encounters.



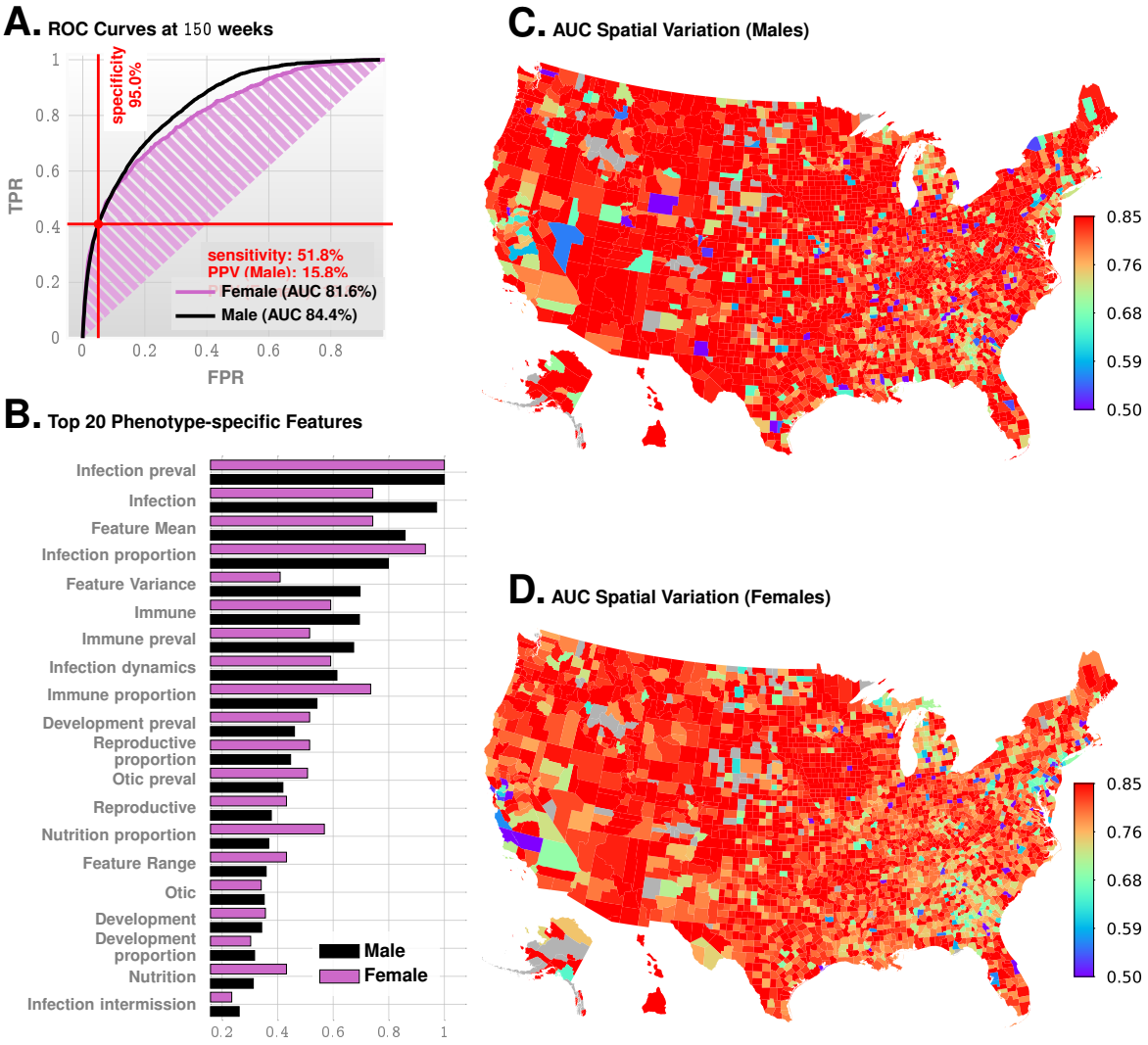
SI-Fig. 3: Performance of standard tools on correctly predicting eventual ASD diagnosis, computed at age 150 weeks of age. Long-short Term Memory (LSTM) networks are the state of the art variation of recurrent neural nets, and Random Forests and Gradient Boosting classifiers (CatBoost) are generally regarded as a representative state of the art classification algorithms. Sequence Likelihood Defect (SLD) is the approach developed in this study. LSTMB denotes LSTM with identical pre-processing as in our pipeline (instead of using raw diagnostic codes). We get much better performance with LSTMB with males in the Truven dataset, but the performance is sensitive to the sizes of the training set, and degrades for smaller samples available for females and in the UCM database, as shown in Panel B.



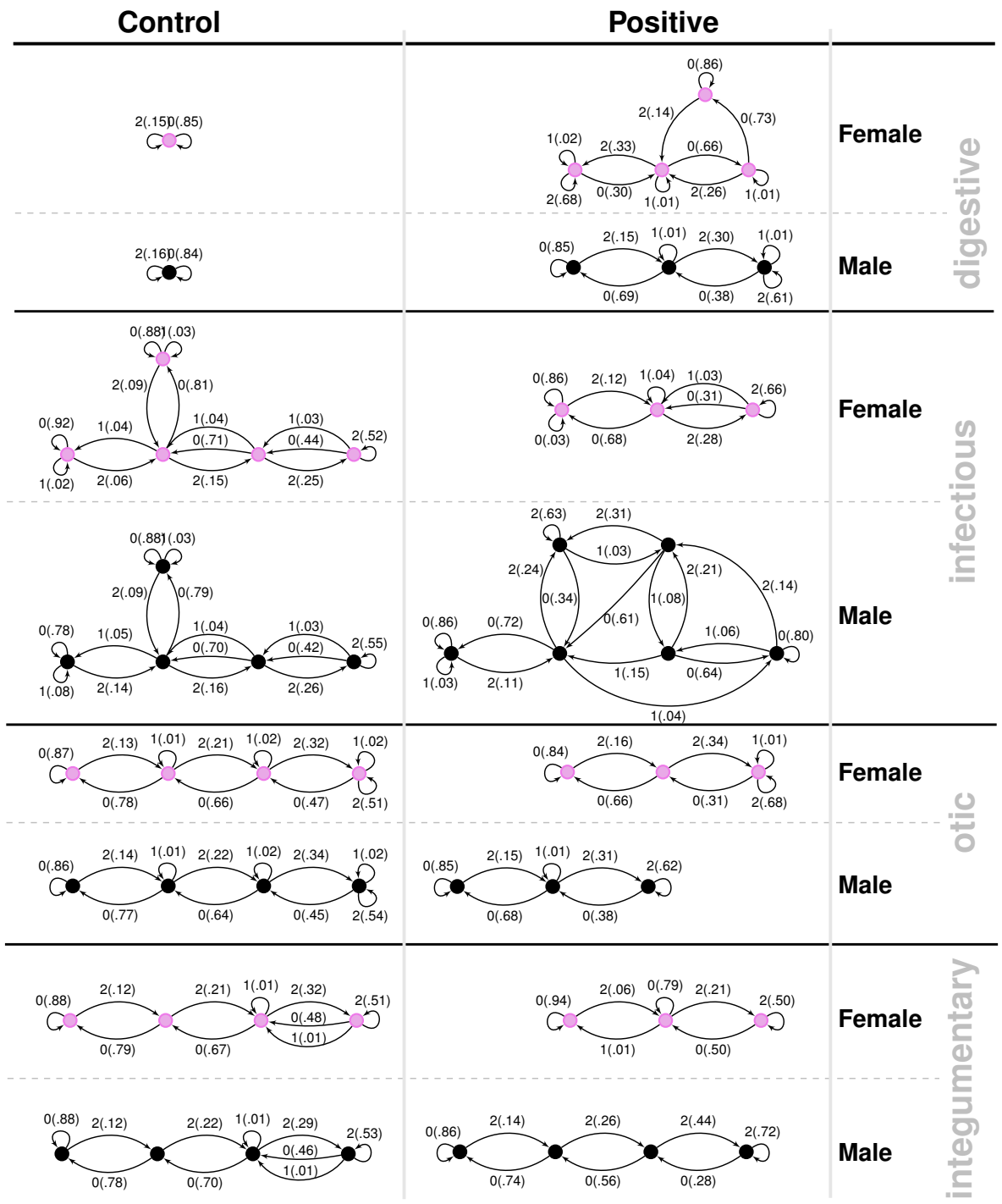
SI-Fig. 4: **4D Search To Take Advantage of Data on Population Stratification (Using Prevalence of 2.23% as reported by CHOP[1]).** While as a standalone tool our approach is comparable to M-CHAT/F at around the 26 month mark (and later), we can take advantage of the independence of the tests to devise 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.



SI-Fig. 5: **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.



SI-Fig. 6: 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. 2 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. 2A, the psychiatric codes all increase risk, and the vaccination codes (See Fig. 2B) 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.



SI-Fig. 7: Probabilistic Finite State Automata models generated for different disease categories for the control and positive cohorts. We note that in the first cases (digestive disorder), the models get more complex in the positive cohort, suggesting that the disorders become less random. However, in the categories of otic and integumentary disorders, the models become less complex suggesting increased independence from past events of similat nature. In case of infectious diseases, the model gets more complex for males, and less complex for females, suggesting distinct sex-specific responses associated with high ASD risk.

SI-Table 4: Disease Categories With Detailed Set of ICD9 Codes Used

Cat.†	Description	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)	273.4 270 270.2 270.3 712.11 712.10 712.13 712.12 712.15 712.14 712.17 712.16 712.19 712.18 712.31 712.30 712.37 712.36 712.35 712.34 712.38 712.33 712.32 712.28 712.29 712.24 712.25 712.26 712.27 712.20 712.21 712.22 712.23 712.39 712.1 712.3 712.2 277.6 275.1 277.85 277.87 270.7 270.6 276.6 276.7 276.4 276.2 276.3 276.0 275.41 276.1 276.8 276.9 276.69 275.5 275.42 271.1 330.2 272.7 271 274.89 712.85 274.81 274.82 712.99 712.98 274.01 274.00 274.03 274.02 712.91 712.90 712.93 712.92 712.95 712.94 712.97 712.96 712.88 712.89 274.10 274.11 712.82 712.83 712.80 712.81 712.86 712.87 712.84 274.19 712.9 712.8 274.0 274.1 274 274.8 274.9 271.2 275.01 270.5 270.4 278.8 272.3 275.03 275.09 271.3 272.6 272.5 278.1 271.8 277.5 263.0 263.2 262 260 261 263 263.1 269.8 269.9 263.8 263.9 269 277.7 272.2 330.3 271.9 275.40 272.8 277.8 275.49 275.2 277.88 275.4 269.3 275.9 275.8 277.9 277.89 251.2 251.1 251.0 278.01 278.00 278.03 270.8 270.9 278 278.0 278.02 277.86 270.1 275.3 277.1 277.81 277.82 272 272.1 277.2 272.4 272.9 273.9 273.8 268.1 265.2 268.0 268.2 268 265.0 265.1 266.1 266.0 266.2 266.9 264.3 264.2 264.1 264.0 264.7 264.6 264.5 264.4 264.9 264.8 268.9 267 266 265 264 269.2 269.0 269.1 278.2 278.3 278.4
Cardiovascular	Diseases Of Arteries, Arterioles, And Capillaries	442.89 441.6 442.84 442.82 442.83 442.8 441.03 441.02 441.01 441.00 441 414.19 414.12 442 414.10 414.11 447.70 447.71 447.72 447.73 414.1 442.81 441.9 442.1 442.0 442.3 442.2 441.2 441.3 441.0 441.1 442.9 441.7 441.4 441.5 437.3 447.7 443.29 443.23 443.22 443.21 443.24 443.2 444.9 444.8 444.81 444.2 444.1 444.0 444.89 444 444.22 444.21 445.81 440.31 440.30 440.32 414.01 414.00 414.03 414.02 414.05 414.04 414.07 414.06 445.89 411.81 445.02 445.01 440.24 440.22 440.23 440.20 440.21 440 445 440.29 414.0 414.2 440.4 440.3 440.2 440.1 440.0 440.9 440.8 445.8 445.0 414.3 414.4 426.54 426.53 426.52 426.51 426.50 426.13 426.12 426.11 426.10 426.89 426.9 426.8 426.81 426.3 426.2 426.1 426.0 426.7 426.6 426.5 426.4 427.61 427.60 427.5 427.89 427.69 427.32 427.41 427.9 427.81 427.8 427.42 427.6 427.4 427.31 427.2 427.3 427.0 427.1 425.8 425.9 425.4 425.7 425.0 425.1 425.2 425.3 346.6 438.51 438.52 438.53 438.50 290.4 431.0 438.42 438.41 438.40 432.9 290.43 290.42 290.41 290.40 432.0 432.1 433.00 433.01 434.9 346.61 346.60 346.63 346.62 433.80 433.81 433.11 433.10 438.32 438.30 438.31 434.0 433.91 433.90 430.0 434.10 434.11 434.1 438.21 438.20 438.22 433.20 433.21 438.6 438.7 438.4 438.5 438.2 438.3 438.0 438.1 438.8 438.9 438 436 434.90 434 435 432 433 430 434.91 434.01 434.00 438.10 438.11 438.12 438.13 438.14 438.19 433.31 433.30 438.85 438.84 438.83 438.82 438.81 438.89 433.9 433.8 433.1 433.0 433.3 433.2 437 435.3 435.2 435.1 435.0 435.9 435.8 437.5 437.4 437.7 437.6 437.1 437.0 431 437.9 437.8 459.3 459.33 459.32 459.31 459.30 459.39 452 453 453.7 453.6 453.5 453.4 453.3 453.2 453.1 453.0 453.9 453.8 453.52 453.51 453.50 453.79 453.71 453.73 453.72 453.75 453.74 453.77 453.76 453.84 453.89 453.40 453.41 453.42 453.81 453.82 453.83 415.11 453.85 453.86 453.87 405.0 405.1 404.9 403.11 402.00 402.01 404.1 404.0 402.1 402.0 403.0 405.99 402.9 405.91 402.91 402.90 405.11 401.0 401.1 404.00 404.01 404.02 404.03 405.19 401.9 405 404 403 402 401 405.9 403.01 403.00 402.11 402.10 403.10 404.13 404.12 404.11 404.10 405.01 403.9 405.09 403.90 437.2 403.1 403.91 404.93 404.92 404.91 404.90 448 458 458.0 458.2 458.1 458.8 458.9 458.29 458.21 426.82 429.71 410.01 410.00 410.02 410.41 410.40 410.42 410.22 410.21 410.20 429.7 410.70 410.71 410.72 429.79 410.92 410.90 410.91 410.30 410.31 410.32 410.12 410.10 410.11 410.52 410.50 410.51 410.4 410.5 410.6 410.7 410.0 410.1 410.2 410.3 410.8 410.9 411.0 410.62 410.61 410.60 410 412 410.81 410.80 410.82 424.1 424.0 424.3 424.2 429.89 429 429.1 429.5 429.6 429.8 429.9 459 276.5 429.2 429.3 428.9 428.4 428.1 428.0 428.3 428.2 429.81 429.83 429.82 428.33 428.32 428.31 428.30 428 459.9 459.8 459.0 276.50 276.51 428.42 428.43 428.40 428.41 276.52 428.20 428.21 428.22 428.23 428.00 459.89 448.1 454 455 455.9 455.8 454.8 454.9 455.1 455.0 455.3 455.2 455.5 455.4 455.7 455.6 454.2 447 454.0 454.1 757.32 447.8 447.9 448.9 447.4 447.5 447.0 447.2 447.3 414 413.1 413.0 413.9 411.89 411.1 411 414.9 413 414.8 411.8 443.89 443 459.2 443.8 443.9 443.81 459.81 443.0 416.2 415.19 415.1 415.13 415.12 416.
Endocrine	Disorders Of Thyroid and other Endocrine Glands	244 244.9 244.8 244.2 255.41 255.42 255.5 255.4 255.2 259.51 255.13 259.2 243 255 253.5 259.4 255.11 242.2 240.0 241 240 240.9 241.0 242.20 242.21 241.9 241.1 253.3 704.1 255.12 255.10 255.14 246.9 246.8 246.3 246.1 246.0 246 255.3 255.9 255.8 255.6 255 252.8 252.9 252.0 252.1 252.01 252.00 252.02 252 252.08 259.52 253.4 253.6 253.1 253.0 253.2 253 253.9 253.8 242.3 242.1 242.10 242.4 242.9 242.8 242.11 242.40 242.41 376.32 242 242.31 242.30 242.91 242.90 242.80 242.81 250.32 250.20 250.30 250.22 250.42 250.40 250.02 250.00 250.10 250.12 250.82 250.80 250.90 250.92 250.50 250.52 250.72 250.60 250.70 250.62 362.05 362.01 362.07 362.06 362.0 357.2 362.03 250.7 362.02 250.8 250.9 362.04 250.2 250.3 250 250.1 250.6 250.0 250.4 250.5 259 259.3 259.50 259.9 259.8 258.8 258.92

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

Reproductive	Diseases Of The Genitourinary System	611.79 611.71 611.72 611.89 611.81 611.83 676.64 611 676.61 676.60 676.63 676.62 611.7 611.6 611.5 611.4 611.3 611.2 611.1 611.0 611.9 611.8 676.6 611.82 757.6 778.7 622 621 654.43 616.81 626.7 654.0 616.89 654.32 654.33 654.30 654.31 629 654.34 654.41 256.8 256.9 256.4 610.0 610.1 256.0 256.1 618.2 618.3 618.0 618.1 618.6 618.7 618.4 618.5 618.8 618.9 628.1 629.89 628.0 616.51 665.20 665.2 654.70 622.7 654.60 616.50 625.79 622.9 622.8 622.5 622.4 625.70 625.71 622.0 622.2 624.1 624.0 654.83 654.82 624.9 624.8 654.84 654.03 654.02 654.01 654.00 654.04 623 654.3 618.89 620 654.6 654.7 625 624 618.83 618.82 618.81 628 618.84 618.03 618.02 618.01 618.00 628.3 628.2 618.05 618.04 618.09 256 654.81 628.9 628.8 654.80 654.74 654.72 654.73 620.9 620.8 620.5 620.4 620.3 620.2 620.1 620.0 623.1 623.2 623.3 623.5 623.8 623.9 654.71 625.6 625.7 625.5 625.0 625.1 625.8 625.9 615.9 615.1 615.0 624.09 617.9 617.8 665.22 654.61 665.24 654.63 654.62 617.1 617.0 617.3 617.2 617.5 617.4 617.6 654.4 654.42 629.8 629.9 629.0 629.1 654.64 621.0 621.1 621.6 621.7 621.8 621.9 618 616 617 614 615 628.4 616.10 616.11 654.44 614.8 614.9 614.6 614.7 614.4 614.5 614.2 614.3 614.0 614.1 616.8 616.9 616.4 616.5 654.8 654.40 616.0 616.1 616.2 616.3 601.4 601.9 607 606 603 601.3 608.24 608.20 608.21 608.22 608.23 605 257 604 607.2 608.89 607.3 608.82 608.83 608.81 608.86 608.87 608.84 608.85 608.9 608.8 607.0 607.1 608 604.90 608.0 606.8 608.1 604.91 607.8 607.9 604.99 608.4 606.1 606.0 607.89 604.9 606.9 607.82 607.83 302.72 607.81 604.0 607.85 607.84 603.0 603.1 608.3 603.8 603.9 257.9 257.8 608.2 257.0 257.2 601.8 643.83 671.54 643.81 643.80 671.51 671.50 671.53 671.52 664.51 664.50 664.54 642.44 642.40 642.41 642.42 642.43 674.34 674.30 674.32 772.13 772.12 772.11 772.10 772.14 676.03 770.10 676.01 676.00 770.15 770.14 770.17 770.16 652.41 652.40 652.43 662.30 653.03 653.00 653.01 675.04 675.02 675.03 675.00 675.01 651.31 651.30 651.33 633.91 633.90 641.30 641.31 648.64 648.60 648.61 648.62 648.63 675.23 665.83 665.82 665.81 665.80 665.84 642.7 642.6 642.5 642.4 668.24 655.31 655.30 655.33 679.02 679.03 679.00 679.01 679.04 640.8 670.24 659.53 670.22 659.51 659.50 669.1 762.8 762.9 762.0 762.1 762.2 762.3 762.4 762.5 762.6 762.7 764.9 764.2 764.0 764.1 662.01 766.0 766.1 766.2 651.80 649.31 649.33 764.29 764.28 764.25 764.24 764.27 764.26 764.21 764.20 764.23 764.22 642.53 642.52 642.51 642.50 642.54 666.32 768.2 768.3 768.0 768.1 768.6 768.7 768.4 768.5 768.9 674.22 674.20 674.24 673.22 673.23 673.20 673.21 673.24 651.00 651.01 651.03 653.11 653.10 653.13 675.14 675.11 675.10 675.13 675.12 634.12 652.71 634.10 634.11 648.73 648.72 648.71 648.70 648.74 665.74 665.72 665.70 665.71 657.03 657.00 657.01 664.6 664.5 664.4 664.3 664.2 664.1 664.0 664.9 664.8 666.1 666.0 666.3 770.18 655.00 655.01 655.03 670.10 679.11 679.10 679.13 679.12 679.14 659.23 659.20 659.21 674.8 674.9 676.02 674.0 674.1 674.2 674.3 674.4 674.5 770.12 672.0 659.9 659.8 670.8 659.3 659.2 659.1 659.0 670.0 659.6 659.5 659.4 655.9 655.8 655.7 655.4 655.3 655.2 655.1 655.0 657.0 651.3 651.2 651.1 651.0 651.7 651.6 651.5 651.4 649.0 649.1 651.9 649.3 649.4 649.5 649.6 649.7 662.33 662.31 764.18 764.19 669.34 643.1 669.32 640.83 669.30 648.14 764.10 764.11 764.12 643.9 764.14 764.15 764.16 764.17 641.0 641.1 641.2 641.3 641.8 641.9 648.10 647.6 647.4 647.5 647.2 647.3 647.0 647.1 648.12 647.8 647.9 673.31 673.30 673.33 673.32 673.34 774.31 774.30 774.39 651.13 647.43 651.11 651.10 647.42 647.41 675.20 675.21 675.22 647.40 675.24 771.0 771.1 647.44 652.63 652.61 652.60 634.01 634.00 634.02 670 648.84 648.82 648.83 648.80 648.81 653.20 653.21 653.23 665.64 655.10 665.61 665.60 669.90 655.13 655.11 669.91 669.92 669.93 659.31 659.30 659.33 647.34 647.32 647.33 647.30 647.31 671.02 671.03 671.00 671.01 671.04 642.63 642.60 648.11 642.61 664.80 664.81 664.84 669.24 669.21 669.20 669.23 669.22 771.89 668.84 648.13 668.80 668.81 668.82 668.83 639.5 639.4 639.6 639.1 639.0 639.3 639.2 639.9 639.8 764.07 764.06 764.05 764.04 764.03 764.02 764.01 764.00 673.04 673.00 673.01 673.02 673.03 642.94 642.93 642.92 642.91 642.90 649.53 649.50 649.51 644.1 644.0 644.2 619.0 648.5 634.31 648.7 648.6 648.1 648.0 619.9 619.8 648.9 652.93 652.90 652.91 648.94 648.91 648.90 648.93 648.92 779.84 653.33 653.31 653.30 660.80 660.81 660.83 665.50 665.51 665.54 658.43 658.41 658.40 645.13 645.10 645.11 659.00 659.01 659.03 647.24 647.21 647.20 647.23 647.22 646.84 646.80 646.81 646.82 646.83 671.11 671.10 671.13 671.12 671.14 646.64 646.62 646.63 646.60 646.61 664.94 664.91 664.90 669.10 669.11 669.12 669.13 669.14 651.41 668.93 668.92 668.91 668.90 668.94 673.13 643.20 673.11 673.10 673.14 674.92 763.82 674.90 674.94 763.81 649.44 649.41 649.40 649.43 649.42 640.91 640.90 640.93 652.81 634.22 634.21 634.20 661.33 661.30 661.31 669.2 656.93 656.90 656.91 660.93 660.91 660.90 654.94 654.90 654.91 654.92 654.93 760.8 760.9 760.2 760.3 760.0 760.1 760.6 760.7 760.4 760.5 622.3 655.71 655.70 655.73 624.5 624.4 665.44 665.41 665.40 645.01 645.00 645.03 659.13 771.81 659.11 659.10 647.10 647.11 647.12 647.13 647.14 771.83 771.82 646.93 620.6 646.91 646.90 766.22 766.21 671.24 641.13 641.10 641.11 671.20 671.21 671.22 671.23 646.71 646.70 646.73 658.90 658.91 658.93 040.41 669.04 669.03 669.02 669.01 669.00 670.84 670.80 670.82 643.10 643.11 643.13 676.94 676.90 676.91 676.92 676.93 674.80 674.82 674.84 649.21 670.20 649.20 765.22 765.23 765.21 765.26 765.27 765.24 765.25 765.28 649.70 649.71 649.73 633.20 633.21 676.04 634.52 634.50 634.51 661.21 661.20 661.23 656.81 656.80 656.83 678.0 678.1 660.63 660.60 660.61 676.8 676.9 676.2 676.3 676.0 676.1 676.4 676.5 764.09 764.08 655.40 655.41 655.43 665.30 665.31 665.34 630 631 632 633 634 639 647.04 647.03 647.02 647.01 647.00 641.01 641.00 641.03 671.30 646.40 646.41 646.42 646.43 646.44 765.1 765.0 658.83 658.81 658.80 767.9 767.8 767.1 767.0 767.3 767.2 767.5 767.4 767.7 767.6 669.70 669.71 773.4 773.5 773.2 773.3 773.0 773.1 654.9 643.03 643.01 643.00 676.83 676.82 676.81 676.80 676.84 663.13 663.10 663.11 634.41 634.40 634.42 661.10 661.11 661.13 649.64 649.63 649.62 649.61 649.60 656.53 656.50 656.51 660.71 660.70 660.73 631.0* 667.12 667.10 667.14 645.23 645.21 645.20 634.30 648.4 643.2 634.32 654.50 654.51 654.52 654.53 654.54 648.2 649.23 646.54 646.53 646.52 646.51 646.50 673.8 673.3 673.2 673.1 673.0 648.8 671.9 671.8 619.1 671.1 671.0 671.3 671.2 671.5 671.4 664.24 664.20 664.21 669.61 669.60 667.0 667.1 661.9 661.4 661.2 661.3 661.0 661.1 663.8 663.9 641.33 663.4 663.5 663.6 663.0 663.1 663.2 663.3 642.9 668.22 668.23 668.20 668.21 642.3 642.2 642.1 642.0 669.6 669.7 669.4 669.5 770.86 669.3 669.0 770.85 770.88 640.0 669.8 669.9 646.3 646.2 646.1 646.0 646.7 646.6 646.5 646.4 646.9 646.8 652.4 652.5 652.6 652.7 652.0 652.1 652.2 652.3 652.8 652.9 765.08 765.09 765.04 765.05 765.06 765.07 765.00 765.01 765.02 765.03 633.00 633.01 663.01 663.00 663.03 634.70 634.71 634.72 649.12 649.13 649.10 649.11 649.14 671.31 656.41 656.40 656.43 660.40 660.41 660.43 661.03 661.01 661.00 667.00 667.02 667.04 659.70 665.14 665.10 665.11 665.12 619 647.61 647.60 647.63 647.62 647.64 646.22 646.23 646.20 646.21 646.24 655.80 655.81 655.83 767.11 664.34 664.31 664.30 767.19 675.1 641.23 641.21 641.20 669.50 669.51 764.90 764.91 764.92 764.93 764.94 764.95 764.96 764.97 764.98 764.99 642.24 642.22 642.23 642.20 642.21 765.19 765.18 765.13 765.12 765.11 765.10 765.17 765.16 765.15 765.14 633.11 633.10 675.84 675.82 675.83 675.80 675.81 663.30 663.31 663.33 634.62 634.61 634.60 643.90 674.54 674.52 674.53 674.50 674.51 653.83 653.80 653.81 649.01 649.00 649.03 649.02 649.04 656.70 656.71 656.73 774 771 772 648.04 648.02 648.03 648.00 648.01 619.2 632.0 660.53 660.51 660.50 665.03 665.01 665.00 647.54 647.50 647.51 647.52 647.53 667 666 665 664 663 662 661 660 768.72 768.73 768.70 768.71 644.03 644.00 666.04 666.02 666.00 646.31 646.30 646.33 655.93 655.91 655.90 664.00 664.01 664.04 652.70 652.73 671.82 671.83 671.80 671.81 671.84 669.43 669.42 669.41 669.40 669.44 673.84 630.0 673.80 673.81 673.82 673.83 676.50 676.51 676.52 676.53 676.54 642.34 642.31 642.30 642.33 642.32 668.04 668.00 668.01 668.02 668.03 663.23 663.21 663.20 675.94 675.91 675.90 675.93 675.92 640.80 634.92 634.90 634.91 674.44 674.40 674.42 652.13 652.10 652.11 653.91 653.90 653.93 649.34 649.30 651.81 649.32 651.83 658.8 658.9 656.63 656.61 656.60 651.63 651.60 651.61 768 769 762 763 761 766 767 764 765 761.7 761.6 761.5 761.4 761.3 761.2 761.1 761.0 761.9 761.8 763.5 763.4 763.7 763.6 763.1 763.0 763.3 763.2 763.9 763.8 660.23 660.20 660.21 771.4 771.5 771.6 771.7 678 679 771.2 771.3 674 675 676 677 771.8 671 672 673 659.93 656.2 656.3 678.13 678.10 678.11 654.21 654.20 654.23 646.00 646.01 646.03 644.10 644.13 666.14 666.10 666.12 664.14 664.11 664.10 659.80 659.81 659.83 671.91 671.90 671.93 671.92 671.94 640.03 640.00 640.01 642.00 642.01 642.02 642.03 642.04 668.13 668.12 668.11 668.10 668.14 663.53 663.50 663.51 634.81 634.80 634.82 676.44 676.43 676.42 676.41 676.40 652.01 652.00 652.03 651.93 649.22 651.91 651.90 649.24 653.43 653.40 653.41 656.13 656.10 656.11 651.71 651.70 651.73 669 668 679.1 679.0 648.20 648.21 648.22 648.23 648.24 675.0 675.2 675.9 675.8 660.31 660.30 660.33 654.14 665.8 665.9 654.10 654.11 654.12 654.13 665.3 665.0 665.1 665.6 665.7 665.4 665.5 678.01 678.00 678.03 646.13 646.12 646.11 646.10 646.14 666.20 666.22 666.24 658.33 658.30 658.31 666.2 641 640 643 642 645 644 647 646 649 648 658.4 658.2 658.3 658.0 658.1 654.2 654.1 654.5 664.60 664.61 664.64 656.0 659.91 659.90 656.4 656.5 656.6 656.7 656.8 656.9 650.0 644.20 644.21 774.6 642.14 642.12 642.11 774.5 774.4 642.10 774.1 774.0 774.3 774.2 772.3 772.2 772.1 772.0 642.13 772.6 772.5 772.4 772.9 772.8 663.41 663.40 663.43 760.78 760.79 760.76 760.77 760.74 760.70 676.32 676.33 676.30 676.31 676.34 652.30 662.03 662.00 652.33 653.51 653.50 653.53 656.01 656.00 656.03 651.40 643.21 651.43 673.12 643.23 661.43 661.41 661.40 660.00 660.01 660.03 659.7 670.1 670.2 670.3 666.34 658.21 658.20 658.23 666.30 656 657 654 655 652 653 650 651 658 659 670.14 670.12 659.63 659.60 659.61 623.4 660.9 660.8 623.6 660.3 660.2 660.1 660.0 660.7 660.6 660.5 660.4 662.1 662.0 662.3 662.2 669.94 634.4 634.5 634.6 634.7 634.0 634.1 634.2 634.3 634.8 634.9 642.62 668.2 668.1 668.0 642.64 668.9 668.8 620.7 649.2 651.8 652.31 760.64 760.61 760.63 760.62 653.1 653.0 653.3 653.2 653.5 653.4 653.7 653.6 653.9 653.8 674.12 674.10 674.14 676.21 676.20 676.23 676.22 676.24 662.11 662.10 662.13 763.89 763.84 652.23 763.83 652.21 652.20 645.0 645.1 645.2 653.60 653.61 653.63 643.0 651.53 651.51 651.50 663.93 663.90 663.91 640.81 643.8 764.13 661.90 661.91 661.93 648.42 648.43 648.40 648.41 648.44 656.30 656.31 656.33 660.13 660.11 660.10 647.90 647.91 647.92 647.93 647.94 658.10 658.11 658.13 652.80 652.83 670.04 770.11 659.71 670.00 659.73 670.02 641.93 641.90 641.91 770.13 671.44 671.42 643.91 671.40 643.93 669.84 669.83 669.82 669.81 669.80 664.44 664.40 664.41 642.71 642.70 642.73 642.72 642.74 672.04 672.02 672.00 663.63 663.61 663.60 674.01 674.00 674.03 674.02 674.04 676.14 676.10 676.11 676.12 676.13 652.53 652.50 652.51 662.20 662.21 662.23 653.73 653.71 653.70 651.23 651.20 651.21 633.80 633.81 663.81 663.80 663.83 670.30 648.51 648.50 648.53 648.52 648.54 656.23 656.21 656.20 665.94 665.90 665.91 665.92 665.93 640.9 671.33 621.5 633.9 633.8 633.2 633.1 633.0 658.03 658.01 658.00 655.23 655.20 655.21 647.84 647.83 647.82 647.81 647.80 659.40 659.41 670.32 659.43 670.34 641.81 641.80 641.83 627.4 627 627.1 627.2 627.3 627.8 627.9 626.1 626.0 626.3 626.2 627.0 626 346.4 256.39 626.9 626.8 626.6 2

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

Infectious	Diseases Caused By Pathogens	487.8 488.12 488.1 488.0 488.01 488.02 487.0 487.1 488.19 488.09 488 487 488.11 464.4 466.19 466 466.11 466.1 466.0 490 464.5 465.8 464.3 465.9 032.0 032.1 032.2 032.3 460 463 465.0 465 464.31 464.51 464.50 034.0 464.30 085.5 085.2 084.6 085.0 086.4 085.1 088.89 088 086 085 084 088.82 088.81 084.9 084.8 085.9 084.3 084.2 085.4 084.0 084.7 085.3 084.5 084.4 086.9 088.9 088.8 086.1 086.0 086.3 086.2 086.5 084.1 088.0 730.86 730.87 730.84 730.85 730.82 730.83 730.80 730.81 730.88 730.89 730.9 730.8 730 730.95 730.94 730.97 730.96 730.91 730.90 730.93 730.92 730.99 730.98 048.0 324.9 326 324.0 324.1 424.9 398.0 036.42 429.0 423.0 423.1 112.81 036.41 115.13 422.0 115.03 115.04 422.9 074.2 036.40 036.43 423.9 420.0 421 423.3 423 422 420.9 423.8 115.93 115.94 420 421.0 423.2 391.1 420.99 421.1 420.91 420.90 391.2 421.9 391.0 032.82 074.21 074.20 074.23 074.22 130.3 115.14 424.99 036.4 422.93 422.92 422.91 422.90 424.90 393 422.99 031.1 039.0 035 681.01 681.00 681.02 020.1 682.9 682.8 021.0 682.3 682.2 682.1 682.0 032.85 682.6 682.5 682.4 022.0 681.9 681.0 681.1 682.7 681.10 681.11 686 681 682 472 472.0 472.2 472.1 567.31 728.86 728.0 710.5 383.0 382.02 382.01 382.00 382.9 383.31 383.30 383.33 381.20 383.3 383.1 055.2 381.06 381.04 381.05 381.02 381.03 381.00 381.01 382.2 382.3 382.0 382.1 382.4 383.8 380.12 380.13 380.10 380.11 380.16 383.2 380.14 380.15 380.1 380.2 381.4 381.1 381.0 381.3 381.2 383.00 383.01 383.02 383.22 383.20 383.21 380.23 380.22 382 383 381.19 383.89 381.10 383.9 383.81 381.29 323.01 323.63 072.2 323.81 323.82 323.61 066.2 341.2 323.9 055.0 323.5 323.4 139.0 323.6 323.1 323.0 064 323.51 323.52 058.29 323.72 056.01 058.21 323 323.2 341.22 063 062 323.8 058.2 054.3 049.8 049.9 323.41 323.71 046.2 323.7 063.8 052.0 323.42 063.9 323.02 036.1 323.62 130.0 063.2 063.0 063.1 341.20 341.21 062.9 062.8 062.5 062.4 062.3 062.2 062.1 062.0 041.43 041.42 041.4 041.49 041.41 076 076.0 053.29 372.06 077 376.04 376.00 376.01 376.02 376.03 375.41 372.01 077.2 375.42 379.60 372.05 372.20 372.21 139.1 053.22 053.20 053.21 054.49 054.44 115.02 054.41 054.40 054.43 054.42 376.13 055.71 372.30 021.3 032.81 372.22 379.63 376.0 077.99 077.98 115.92 076.1 077.9 077.8 372.02 364.03 076.9 054.4 077.0 077.3 379.62 360.19 379.61 379.6 053.2 372.15 360.13 372.39 077.4 372.00 372.3 372.2 372.0 375.30 375.31 375.32 375.33 130.2 130.1 115.12 372.03 077.1 372.04 373.6 373.4 373.5 112.3 112.2 112.1 112.0 112.5 112.9 112.8 115 114 117 116 111 110 112 112.84 112.85 116.0 118 112.82 115.09 110.1 111.3 115.1 115.00 111.0 110.5 111.1 115.90 115.99 115.19 115.9 518.6 117.2 117.3 117.0 117.1 117.6 117.7 117.4 117.5 111.8 117.8 117.9 116.2 116.1 111.9 115.0 110.0 110.3 110.2 114.9 110.4 111.2 110.6 110.9 110.8 114.1 115.10 114.3 112.89 005.81 005.89 001.0 001.1 001.9 003.0 003.2 003.8 003.9 008.6 008.5 008.4 008.3 008.2 008.1 008.0 008.8 003.20 003.24 003.29 008.44 008.45 008.46 008.47 008.41 008.42 008.43 008.49 006.0 006.3 006.2 006.5 006.4 006.6 006.9 006.8 009.2 009.3 009.0 009.1 007.5 007.2 007.3 007.9 004.3 004.2 004.1 004.0 004.9 004.8 008.00 008.01 008.02 008.03 008.04 008.09 002.9 021.1 002.1 002.0 002.3 002.2 022.2 005.4 005.2 005.3 005.0 005.1 005.8 005.9 003 002 001 007 006 005 004 009 008 008.69 008.62 008.63 008.61 008.66 008.67 008.64 008.65 041.5 041.7 041.9 037 040.3 039 033.1 033.0 041.8 040.2 032 033 040.1 040.0 033.9 033.8 041.2 040.4 027.2 027.0 027.1 041.89 026.9 036 027.8 027.9 041.6 026.0 032.89 020.0 020.9 020.8 021.8 021.9 032.8 032.9 041.84 023.2 023.3 023.0 023.1 022.9 022.8 041.83 041.82 023.8 023.9 025 024 027 026 021 020 023 022 036.89 041 040 036.8 036.9 039.9 039.8 041.3 039.4 039.3 039.2 036.3 041.81 040.8 040.81 040.82 041.85 036.81 040.42 040.89 072.7 072.0 072.3 072.8 072.9 078.3 056.8 056.9 056.0 056.7 056.09 056.00 074.3 074.0 074.1 074.8 050.9 050.0 050.1 050.2 079.52 079.51 056.79 079.59 066.9 066.8 055.79 079.81 079.82 079.83 079.50 079.88 079.89 078 059.09 059.00 059.01 051.02 079.3 051.01 079.2 059.12 060.9 079.99 079.98 060.1 060.0 066.3 066.1 066.0 059 066.4 055 056 057 050 051 059.19 079.9 079.8 059.11 059.10 079.1 071 079.6 079.5 073.8 061 060 065 066 059.20 059.21 059.22 079 066.40 066.41 074 075 072 073 073.7 057.0 073.9 066.49 057.9 057.8 055.7 078.89 078.88 055.9 055.8 078.81 072.71 072.72 078.6 078.7 078.4 078.5 078.2 072.79 078.0 078.82 075.0 065.4 065.0 065.1 065.2 065.3 065.8 065.9 079.0 051.9 078.8 051.1 051.0 051.2 059.1 059.0 059.2 059.9 059.8 066.42 042.9 043.9 044.0 042.1 042.0 042.2 043.0 043.1 044.9 043.3 079.53 043.2 044 043 042 41.86 058 054 058.81 058.82 054.8 054.9 054.0 054.1 054.2 054.6 054.7 054.12 054.13 054.10 054.11 054.19 054.79 058.12 058.10 058.11 054.71 054.73 054.74 058.1 058.89 058.8 053.19 053.13 053.7 053.11 053.10 053.14 053.9 053.8 053.79 053.12 053.1 053.71 053 133.0 134.1 134.0 134.9 134.8 133 132 132.9 134 133.8 133.9 132.1 132.0 132.3 132.2 590 590.9 590.8 590.11 590.2 590.1 590.10 590.00 590.0 590.3 590.80 590.81 590.01 047.8 047.9 320.0 320.1 072.1 320.3 047.0 047.1 320.8 320.9 321.3 321.2 321.1 321.0 322.0 322.1 322.2 321.4 322.9 112.83 115.01 321.8 322 320 321 115.91 049.0 049.1 003.21 047 320.82 320.81 053.0 320.7 054.72 115.11 036.0 320.89 100.81 320.2 114.2 030.1 031.2 031.9 031.8 030.8 030.9 030.2 030.3 030.0 031 134.2 139.8 082.49 136 082.40 082.41 139 083 082 081 080 083.8 083.9 136.21 083.1 083.2 136.29 082.4 081.9 082.1 082.0 082.3 082.2 081.2 081.0 081.1 082.9 082.8 136.0 136.2 136.5 136.4 136.9 136.8 070.6 070 573.2 070.43 070.42 070.4 070.5 070.52 070.53 573.1 070.49 573.3 070.59 070.9 128.8 128.9 122.9 128.0 128.1 127 123.2 120.3 127.7 127.6 122.2 125.1 122.0 122.1 122.6 122.7 122.4 120.1 127.5 123.6 122.8 120.8 123.3 120.2 123.1 123.0 120.0 125.0 125.3 125.2 121.9 121.8 125.7 125.6 121.5 121.4 127.4 121.6 121.1 121.0 121.3 121.2 127.3 127.2 127.1 127.0 126.2 126.3 126.0 126.1 123.8 127.9 127.8 125.5 126.8 125.9 120 121 122 123 124 125 126 120.9 128 129 126.9 123.9 122.5 123.4 125.4 122.3 123.5 567.9 567.8 567.81 567.82 567.3 567.2 567.1 567.0 567.89 567 567.22 567.23 567.21 567.39 567.38 032.83 567.29 569.5 073.0 031.0 481 115.05 136.3 484.5 483.1 482.82 480.9 480.8 055.1 482.83 480.3 480.2 480.1 480.0 485.0 481.0 482.9 482.8 482.81 483.8 482.1 482.0 482.3 482.2 483.0 482.4 020.3 020.5 020.4 021.2 480 517.1 484.7 484.6 022.1 484.3 513.0 484.1 482.49 482.42 482.40 484.8 003.22 482.89 114.4 083.0 486 485 484 483 482 482.84 052.1 513.1 112.4 130.4 513 039.1 115.15 482.39 115.95 482.41 114.5 482.32 482.31 482.30 114.0 049 138.0 730.77 730.76 730.75 730.7 138 045.0 730.74 730.79 730.78 730.73 730.72 730.71 730.70 045.10 045.11 045.12 045.13 045.90 045.91 045.92 045.93 324 046.7 045.9 046.0 046.3 048 045.2 046 045 045.1 046.9 046.8 045.21 045.20 045.23 045.22 045.03 045.02 045.01 045.00 046.72 046.71 046.79 046.11 046.1 046.19 131.03 131.02 131.0 131.00 131.8 130.9 130.8 130.7 130.5 131.9 131.01 007.4 006.1 131 130 007.0 007.1 007.8 131.09 098 099 093.24 093.22 093.23 093.20 093.21 090 091 092 093 094 095 096 097 098.81 090.9 098.83 098.82 098.85 093.2 093.1 093.0 090.0 090.1 090.2 090.3 090.4 090.5 090.6 090.7 091.5 091.4 091.7 091.6 091.1 091.0 091.3 091.2 091.9 091.8 094.8 094.9 094.0 094.1 094.2 094.3 098.0 098.1 098.2 098.3 098.4 098.5 098.6 091.89 098.8 091.82 091.81 099.40 099.41 099.49 098.34 098.35 098.36 098.37 098.30 098.31 098.32 098.33 098.39 098.84 098.86 099.51 098.89 099.53 099.52 099.55 099.54 099.56 099.59 099.8 092.9 092.0 090.49 093.9 093.8 090.41 090.40 090.42 091.69 094.85 094.84 094.87 094.86 094.81 094.83 094.82 094.89 093.89 098.52 098.53 098.50 098.51 093.81 093.82 098.59 098.41 098.40 098.43 098.42 098.49 099.50 091.51 091.50 091.52 097.9 097.1 097.0 095.9 095.8 095.1 095.0 095.3 095.2 095.5 095.4 095.7 095.6 099.9 091.61 091.62 099.5 099.4 099.1 099.0 099.2 098.7 098.19 098.12 098.13 098.10 098.11 098.16 098.17 098.14 098.15 038.10 038.11 038.12 038.19 038 003.1 020.2 022.3 449 054.5 038.49 038.44 038.43 038.42 038.41 038.40 038.2 038.3 038.0 038.1 038.4 038.8 038.9 036.2 686.1 685.0 685.1 686.9 686.8 684.0 685 684 102.8 102.9 102.0 102.1 102.2 102.3 102.4 102.5 102.6 102.7 104.8 104.9 100 104.0 087 100.8 100.9 103.1 103.0 103.3 103.2 102 103 100.0 101 103.9 104 087.9 100.89 087.0 087.1 041.10 041.11 041.12 041.19 041.1 041.05 041.04 041.03 041.02 041.01 041.00 041.09 390.0 392 391 390 392.9 034 041.0 034.1 016.6 016.7 016.4 010 011 012 013 014 015 016 017 011.15 011.14 011.16 011.11 011.10 011.13 011.12 016.1 012.21 012.20 012.23 012.22 012.25 012.24 017.95 012.26 018 014.85 014.84 014.86 014.81 014.80 014.83 014.82 017.96 016.10 016.11 013.15 013.14 016.14 013.12 016.16 013.10 017.53 017.03 015.1 015.0 015.2 015.5 017.70 015.7 015.6 015.9 015.8 018.90 018.91 018.96 018.94 018.95 011.51 011.50 011.53 011.52 011.55 011.54 011.56 015.55 015.54 015.56 015.51 015.50 015.53 015.52 013.26 013.24 013.25 013.22 013.23 013.20 013.21 016.54 016.55 016.56 016.50 016.51 016.52 016.53 013.53 013.52 013.51 013.50 013.56 013.55 013.54 015.82 015.83 015.80 015.81 015.86 015.84 015.85 017.31 017.30 017.33 017.32 017.35 017.34 017.36 017.93 137 017.92 017.91 017.90 015.11 015.10 015.13 015.12 015.15 015.14 015.16 017.94 016.5 011.20 011.21 011.22 011.23 011.24 011.25 011.26 010.12 010.13 010.10 010.11 010.16 010.14 010.15 016.3 012.14 012.15 012.16 012.10 012.11 012.12 012.13 016.90 016.91 016.92 016.93 016.94 016.95 016.96 013.62 013.63 013.60 013.61 013.66 013.64 013.65 016.06 016.05 016.04 016.03 016.02 016.01 016.00 017.80 017.40 017.41 017.42 017.43 017.44 017.45 017.46 018.81 018.80 018.83 018.82 018.85 018.84 018.86 011.64 011.65 011.66 011.60 011.61 011.62 011.63 015.95 015.60 015.61 015.62 015.63 015.64 015.65 015.66 018.02 013.35 013.34 013.36 013.31 013.30 013.33 013.32 011.86 016.42 011.84 011.85 011.82 011.83 011.80 011.81 015.91 015.90 015.93 015.92 016.2 015.94 016.0 015.96 016.9 013.80 013.81 013.82 013.83 013.84 013.85 013.86 017.04 017.05 017.06 017.00 017.01 017.02 013.16 016.12 016.36 016.34 016.35 016.32 016.33 016.30 016.31 013.13 016.15 010.92 010.93 010.90 010.91 010.96 013.11 010.94 010.95 015.24 015.25 015.26 015.20 015.21 015.22 015.23 012.2 012.3 012.0 012.1 012.8 010.0 010.1 010.8 010.9 011.36 011.35 011.34 011.33 011.32 011.31 011.30 017.84 017.85 017.86 010.01 010.00 010.03 010.02 010.05 010.04 010.06 017.82 017.83 012.03 012.02 012.01 012.00 012.06 012.05 012.04 137.3 016.61 016.60 016.63 016.65 016.64 016.66 018.01 018.00 018.03 017.0 018.05 018.04 018.06 016.72 016.73 016.70 016.71 016.76 016.74 016.75 018.8 018.9 018.0 014.0 017.56 017.55 017.54 014.8 017.52 017.51 017.50 011.73 011.72 011.71 011.70 011.76 011.75 011.74 017.75 017.74 015.76 015.75 015.74 015.73 015.72 015.71 015.70 017.76 017.71 017.73 017.72 018.92 018.93 011.06 011.04 011.05 011.02 011.03 011.00 011.01 013.00 013.01 013.02 013.03 013.04 013.05 013.06 012.32 012.33 012.30 012.31 012.36 017.81 012.34 012.35 011.95 011.94 011.96 011.91 011.90 011.93 011.92 013.96 013.95 013.94 013.93 013.92 013.91 013.90 017.13 017.12 017.11 017.10 017.16 017.15 017.14 016.25 016.24 016.26 016.21 016.20 016.23 016.22 010.81 010.80 010.83 010.82 010.85 010.84 010.86 013.5 014.05 014.04 014.06 014.01 014.00 014.03 014.02 017.66 017.64 017.65 017.62 017.63 017.60 017.61 011.42 011.43 011.40 011.41 011.46 016.13 011.44 011.45 012.83 012.82 012.81 012.80 012.86 012.85 012.84 016.43 016.41 016.40 016.46 016.45 016.44 137.4 137.0 137.1 137.2 013.6 013.44 013.45 013.46 016.62 013.40 013.41 013.42 013.43 017.3 017.2 017.1 013.4 017.7 017.6 017.5 017.4 013.3 017.9 017.8 013.1 013.0 017.22 017.23 017.20 017.21 017.26 017.24 017.25 013.9 013.8 015.02 015.03 015.00 015.01 015.06 013.2 015.04 015.05 011.9 011.8 011.5 011.4 011.7 011.6 011.1 011.0 011.3 011.2 599.0 598 598.0 598.00 598.01 052.7 052.2 052.8 052 052.9 070.0 070.1 070.21 070.20 070.23 070.22 070.31 070.2 070.3 070.32 070.30 070.33 070.70 070.71 070.54 070.7 070.41 070.51 070.44 078.19 078.10 078.11 078.12 079.4 078.1

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

Integumentary	Diseases Of Skin And Subcutaneous Tissue	706.0 706.1 704.0 704.00 704.02 704.01 704.09 680.9 680.8 680.1 680.0 680.3 680.2 680.5 680.4 680.7 680.6 680 698.8 698 698.1 698.0 698.9 757.31 757.5 757.4 757.33 700 694.2 694.3 694.0 694.1 709.0 709.00 709.01 709.09 757.39 757.3 695.59 695 695.57 695.56 695.55 695.54 695.53 695.52 695.51 695.50 695.5 695.58 704 704.9 704.2 704.3 704.8 757 757.1 702.11 702.19 702.0 702.1 697.1 697.0 697.9 697 697.8 759.82 703.8 703.9 703.0 703.80 703 757.8 757.9 757.2 525.9 525.8 525.5 525.4 525.7 525.6 525.1 525.0 525.3 525.79 525 525.73 525.72 525.71 521.35 521.24 521.25 521.20 521.21 521.22 521.23 521.42 521.40 521.41 521.49 522.2 522.3 522.0 522.1 522.8 522.9 521.33 521.32 521.31 521.30 525.19 521 522 521.34 525.11 525.10 525.13 525.12 521.89 525.42 525.43 525.40 525.41 525.44 521.08 521.09 521.02 521.03 521.00 521.01 521.06 521.07 521.04 521.05 521.1 521.0 521.3 521.2 521.5 521.4 521.7 521.6 521.9 521.8 525.51 525.50 525.53 525.52 525.54 521.11 521.10 521.13 521.12 521.15 521.14 521.81 525.69 525.64 525.65 525.66 525.67 525.60 525.61 525.62 525.63 523.11 523.10 523.31 523.30 523.33 523.32 523.3 522.7 522.4 522.5 523.6 523.5 523.4 523.9 523.8 523 522.6 523.00 523.01 523.2 523.1 523.40 523.41 523.42 523.0 523.22 523.23 523.20 523.21 523.24 523.25 709.1 707.00 707.02 707.03 707.04 707.05 707.06 707.01 709.9 707.09 707 701 702 707.25 707.2 707.1 707.0 707.22 707.8 707.9 707.23 706.8 707.20 702.8 705.2 706.2 707.07 701.9 701.8 709.3 709.8 701.1 701.0 701.3 701.2 701.4 707.24 707.19 707.15 707.14 707.13 707.12 707.11 710 705.21 705.22 707.21 707.10 710.9 710.8 709 695.9 709.2 729 729.7 729.91 729.92 729.9 729.90 729.79 729.99 729.0 729.71 729.73 729.72 528.8 528.9 528.4 528.5 528.6 528.7 528.0 528.1 528.2 528.00 528.79 527.3 528.09 528.3 528 528.71 528.72 692.7 692.70 692.71 692.72 692.82 692.74 692.75 692.76 692.77 692.73 692.79 558.1 508.0 508.1 655.6 595.82 655.63 655.60 655.61 737.33 737.11 695.3 706 706.9 690 690.1 690.10 690.11 690.12 690.8 690.18 705.9 705.1 705.0 705.89 705 705.82 705.8 529.9 529.8 529.1 529.0 529.3 529.2 529.5 529.4 529.6 529
Respiratory	Diseases 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	Diseases Of The Digestive System	540.0 540.1 540.9 541.0 542 543 540 541 543.9 543.0 562 562.1 562.11 562.03 562.02 562.01 562.00 562.10 562.0 562.12 562.13 530.3 530.13 530.11 530.0 530.6 530.5 530.4 456.21 530.9 530.8 456.2 456.20 456.0 456.1 530.89 530 530.84 530.85 530.81 530.82 530.83 560.89 579 560.30 560.0 564.09 560.1 564.01 564.00 564.02 564.81 560.39 564.89 560.31 560.32 579.9 536.8 564 537.0 537.1 537.2 537.3 536.9 537.5 537.6 564.5 537.8 537.9 560.2 560.3 536.1 536.0 536.3 537.81 560.8 560.9 537.89 537 536 564.8 564.9 537.4 564.7 564.0 538 532.51 532.50 533.21 533.20 569.41 534.11 534.10 531.70 531.71 534.91 534.90 532.60 532.61 533.50 533.51 534.60 534.61 531.61 531.60 531.6 531.7 531.4 531.5 531.2 531.3 531.0 531.1 531.9 534 533 532 531 531.50 531.51 533.41 533.40 532.71 532.70 534.71 534.70 530.20 530.21 534.9 531.41 531.40 532.00 532.01 534.40 534.41 532.11 532.10 531.30 531.31 534.51 534.50 569.86 533.70 533.71 569.82 533.10 533.11 532.1 532.0 532.3 532.2 531.21 531.20 532.7 532.6 532.9 532.91 532.90 534.20 534.21 533.61 533.60 533.01 533.00 531.90 531.91 531.10 531.11 534.31 534.30 532.20 532.21 533.90 533.91 533.30 533.31 532.40 532.41 534.7 534.6 534.5 534.4 534.3 534.2 534.1 534.0 533.4 533.5 533.6 533.7 533.0 533.1 533.2 533.3 533.9 531.01 531.00 530.2 534.00 534.01 532.31 532.30 532.5 532.4 751.3 568.0 578 569.9 569.8 558.9 579.1 569.3 558 569.87 579.2 560.81 569.84 569.85 558.42 569.83 578.9 558.41 579.8 578.1 578.0 569 569.81 537.83 537.82 530.7 537.84 558.4 569.89 568 568.8 568.9 568.81 568.82 568.89 251.5 251.4 577 251.9 251.8 577.8 577.9 579.4 251 577.0 577.1 577.2 569.4 569.1 565 566 569.2 569.42 569.43 566.0 569.44 565.1 565.0 564.6 569.49 527.9 527.8 527.7 527.6 527.5 527.4 527.2 527.1 527.0 527

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

Immunologic	Diseases related to dys-regulation of the Immune system	580.81 580.89 580.4 580.0 580.8 580.9 461 461.8 461.9 461.0 461.1 461.2 461.3 477.9 477.0 477.2 477 477.8 493.22 493.20 493.90 493.91 493.12 493.11 493.10 493.00 493.01 493.02 493.92 493.9 493.8 493.81 493.82 493.2 493.1 493.0 493.21 493 708.4 708.5 708 708.1 708.2 708.3 708.8 708.9 708.0 692 693 691 692.81 692.83 692.84 692.89 373.32 691.8 693.9 692.0 692.1 692.2 692.3 692.4 693.8 692.6 692.8 692.9 693.0 279.41 258.1 733.01 711.29 711.28 711.25 711.2 711.27 711.26 711.21 711.20 711.23 711.22 136.1 711.24 579 582.4 582.0 582.1 582.2 582.81 582.8 582.9 582.89 473 473.9 473.8 473.1 473.0 473.3 473.2 475.0 474.2 474.0 474.1 474.8 474.9 474.10 474.11 474.12 474 475 474.02 474.01 474.00 279.06 279.04 288.01 288.02 595.2 595 595.1 597 597.80 595.3 595.81 597.89 597.0 595.0 597.8 710.3 710.4 279.11 530.12 530.1 530.10 530.19 379.09 360.04 360.03 360.02 360.01 360.00 379.01 379.00 379.03 379.02 379.05 379.04 379.07 379.06 363.20 363.21 363.22 379.92 379.93 379.91 364.2 364.3 364.0 364.1 364.10 364.01 364.00 364.05 364.11 379.0 372.31 372.33 370 373 372 360.0 360.1 373.33 373.31 376.12 376.11 376.10 372.14 372.13 372.12 372.11 372.10 370.64 370.60 370.61 370.32 373.02 373.00 373.11 373.12 373.13 372.1 370.5 370.6 370.9 370.8 363.1 363.0 373.01 363.2 363.03 363.00 363.01 363.06 363.07 363.04 363.05 363.08 370.59 370.50 370.52 370.55 370.54 364.02 376.1 364.21 364.23 364.22 364.24 363.11 363.10 363.13 363.12 363.15 363.14 370.62 370.63 360.11 360.12 364.04 360.14 373.8 373.9 373.2 373.3 373.0 373.1 277.31 728.71 729.4 477.1 558.3 693.1 692.5 558.2 535.01 535.00 535.61 535.60 535.21 535.20 535.41 535.40 535.2 535.0 535.1 535.6 535.7 535.4 535.5 535.10 535.11 535.50 535.51 535 535.70 535.71 288.2 242.0 242.00 242.01 289.6 279.05 555.0 555.1 555.2 564.1 555.9 279.01 279.02 721.1 720.81 720.0 720.1 721.4 720.8 720.9 721.41 721.42 720 721.91 720.89 446.1 710.0 373.34 695.4 457.9 457.8 457.2 457.1 457.0 457 289.1 289.2 289.3 683 683.0 341.1 341.0 340 341.9 358.0 358.00 358.01 279.13 716.48 716.49 716.42 716.43 716.40 716.41 716.46 716.47 716.44 716.45 716.52 716.88 716.38 716.89 026.1 716.86 716.87 716.84 716.85 716.37 716 711 716.83 003.23 716.58 716.80 716.51 716.50 716.53 716.81 716.55 716.54 716.57 716.56 716.5 716.4 716.36 716.6 716.1 716.0 711.8 711.9 711.6 711.7 711.4 711.5 716.9 711.3 711.1 711.72 711.73 711.70 711.71 711.76 711.77 711.74 711.75 716.59 711.78 711.79 056.71 716.64 716.65 716.66 716.67 716.60 716.61 716.62 716.63 716.68 716.8 719.3 036.82 711.61 711.60 711.63 711.62 711.65 711.64 711.67 711.66 711.69 711.68 716.29 716.82 711.58 711.59 711.54 711.55 711.56 711.57 711.50 711.51 711.52 711.53 716.39 711.38 711.39 711.36 711.37 711.34 711.35 711.32 711.33 711.30 711.31 716.06 716.07 716.04 716.05 716.02 716.03 716.00 716.01 716.35 716.08 716.09 716.34 716.33 716.32 711.49 711.48 716.31 711.43 711.42 711.41 711.40 711.47 711.46 711.45 711.44 716.99 716.98 716.95 716.94 716.97 716.96 716.91 716.3 716.93 716.92 716.15 716.14 716.17 716.16 716.11 716.10 716.13 716.12 716.19 716.18 716.20 716.21 716.2 716.22 716.23 716.24 716.25 719.36 719.37 719.34 719.35 719.32 716.26 719.30 719.31 719.33 716.27 719.38 719.39 711.90 711.91 711.92 711.93 711.94 711.95 711.96 711.97 711.98 711.99 716.30 716.90 711.89 711.88 711.87 711.86 711.85 711.84 711.83 711.82 711.81 711.80 716.28 279.03 279.00 279.09 695.13 695.12 695.11 695.10 695.15 695.14 686.09 695.19 694 694.61 694.60 695.1 694.5 686.01 695.89 698.2 698.3 695.81 698.4 705.83 705.81 686.00 686.0 710.1 694.6 695.2 694.4 695.0 706.3 695.8 694.8 694.9 715.94 715.95 715.96 715.97 715.90 715.91 715.92 715.93 715.98 715.99 715.22 735.2 715.09 715.21 715.20 715.23 715.08 715.25 715.24 715.27 715.26 715.28 715.00 715.06 715.05 715.04 715.18 715.80 715 715.89 715.16 715.8 715.9 715.2 715.3 715.0 715.1 715.32 715.33 715.30 715.31 715.36 715.37 715.34 715.35 715.14 715.15 715.38 715.17 715.10 715.11 715.12 715.13 733.6 730.1 730.0 730.2 730.19 730.18 730.15 730.14 730.17 730.16 730.11 730.10 730.13 730.12 730.20 730.21 730.27 730.22 730.23 730.24 730.25 730.28 730.29 730.26 730.08 730.09 730.06 730.07 730.04 730.05 730.02 730.03 730.00 730.01 451 459.10 451.11 443.1 451.19 443.10 459.19 451.0 451.2 451.9 451.8 459.13 459.12 325 451.89 459.1 451.84 451.82 451.83 451.81 459.11 451.1 511.8 511.9 511.0 511.1 511.89 511 273 288.1 696.4 696.5 696 696.1 696.2 696.3 696.8 696.0 711.18 711.19 711.10 711.11 711.12 711.13 711.14 711.15 099.3 711.17 711.16 394.2 394.1 394.0 395.2 395.0 395.1 394.9 398.9 395.9 397 396 395 394 398 424 424.91 398.91 398.90 398.99 391.8 391.9 392.0 396.1 396.0 396.3 396.2 397.0 397.1 396.9 396.8 397.9 714.31 714.30 714.33 714.32 714 725.0 714.89 714.4 725 714.3 714.2 714.1 714.0 714.81 714.9 714.8 135 720.2 713 713.8 713.4 713.5 713.6 713.7 713.0 713.1 713.2 713.3 249.01 249.00 249.81 249.80 249.8 249.9 249.4 249.5 249.6 249.7 249.0 249.1 249.2 249.3 249.50 249.51 249.70 249.71 249.30 249.31 249.10 249.11 249.90 249.91 249 249.41 249.40 249.21 249.20 249.61 249.60 711.0 711.07 711.06 711.05 711.04 711.03 711.02 711.01 711.00 711.09 711.08 279.2 710.2 254.9 254.8 254.0 254 254.1 245.0 246.2 245 245.1 245.2 245.3 245.4 245.8 245.9 250.21 250.33 250.23 250.31 250.43 250.41 250.03 250.01 250.11 250.13 250.83 250.81 250.91 250.93 250.51 250.53 250.61 250.73 250.63 250.71 556 556.6 556.5 556.4 556.3 556.2 556.1 556.0 556.9 556.8 277.39 277.30 279.49 288.3 571.49 571.42 571.41 571.4 571.40 279.19 279.10 279.3 728.82 701.5 279.9 279.8 709.4 289.53 288.9 288.8 288.5 288.4 288.6 288.0 288.51 288.50 288.59 288.69 288.64 288.65 288.66 288.60 288.61 288.62 288.63 288.09 288.04 288.03 288.00 478.21 478.20 478.22 478.25 478.24 478.26 478.29 464.20 464.21 464.00 464.01 519.19 478.19 464 478.11 462 519.1 519.2 519.3 464.11 464.10 464.1 464.0 464.2 478.8 478.9 476 478.2 476.0 478 478.1 476.1 446.29 447.6 446.20 446 446.21 446.0 446.3 446.2 446.5 446.4 446.7 446.6 279.12
Otic	Diseases Of The Ear And Mastoid Process	381.51 381.50 381.52 381.81 381.89 381.60 381.61 381.62 381.63 381 381.9 381.8 381.5 381.7 381.6 385.82 383.32 380.30 380.31 380.32 380.39 380.52 380.53 380.50 380.51 380.0 385.3 380.3 380.4 380.5 380.8 380.9 380.81 380.89 380 385.32 380.21 385.30 385.35 385.33 380.01 380.00 380.03 380.02 385.31 388.8 388.9 388.70 388.71 388.72 388.6 388.7 388.11 388.61 388.60 388 388.69 389.15 389.14 389.17 389.16 389.11 389.10 389.13 389.12 388.0 388.1 388.2 388.3 389.18 388.12 388.10 388.30 388.31 388.32 389.04 389.20 389.21 389.22 389.05 389.02 389.03 389.00 389.01 389.08 388.01 388.00 389.7 388.02 389.1 389.0 389.2 389 389.9 389.8 389.06 386.43 386.42 386.41 386.40 386.48 386.4 386.5 386.3 386.8 386.54 386.55 386.56 386.50 386.51 386.52 386.53 386.58 386.34 386.35 386.32 386.33 386.30 386.31 387.2 387.1 387.0 387.9 387.8 387 385.12 384.09 385.24 385.21 385.22 385.23 385.09 385.02 385.03 385.00 385.01 385.1 385.0 385.2 384.0 384.1 384.2 384.8 384.9 384 384.81 384.82 385.19 384.01 384.00 385.11 385.10 384.25 384.24 384.23 384.22 384.21 384.20 385.13 385.9 385.8 385.89 385

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

Ophthalmologic	Disorders Of The Eye And Adnexa	362.8 362.9 362.4 362.6 362.1 362.2 362.3 362.18 362.16 362.17 362.14 362.15 362.12 362.13 362.10 362.11 363.33 363.32 363.31 363.30 363.35 363.34 362.85 362.84 362.81 362.83 362.82 362.89 362.30 362.31 362.32 362.33 362.34 362.35 362.36 362.37 362 361 362.27 362.26 362.25 362.24 362.23 362.22 362.21 362.20 362.29 361.31 361.30 361.33 361.32 362.55 361.11 361.04 361.05 361.06 361.07 361.00 361.01 361.02 361.03 362.41 362.40 362.43 362.42 361.81 361.89 361.19 361.13 361.12 363.3 361.10 361.14 362.63 362.62 362.61 362.60 362.66 362.65 362.64 361.9 361.8 361.3 361.2 361.1 361.0 369.13 369.12 369.11 369.10 369.17 369.16 369.15 369.14 369.18 369.3 369.2 369.1 369.0 369.7 369.6 369.4 369.9 369.8 369.66 369.67 369.64 369.65 369.62 369.63 369.60 369.61 369.68 369.69 368.12 367.51 367.53 367.52 369.75 369.74 369.76 369.71 369.70 369.73 369.72 368.03 368.02 368.01 368.00 368.6 368.4 368.2 368.3 368.0 368.1 368.8 368.9 368 369 368.16 368.69 368.60 368.63 368.62 367.81 368.55 368.59 369.22 369.23 369.20 369.21 369.24 369.25 368.43 368.42 368.41 368.40 368.47 368.46 368.45 368.44 360.21 367.20 367.21 367.22 367.5 367.4 367.1 367.0 367.3 367.2 369.04 369.05 369.06 369.07 369.00 369.01 369.02 369.03 369.08 368.34 368.32 368.33 368.30 368.31 368.10 368.11 367.32 367.31 368.13 368.14 368.15 366.52 366.53 366.50 366.51 366.01 366.00 366.03 366.02 366.04 366 366.22 366.21 366.20 366.23 366.45 366.44 366.46 366.41 366.43 366.42 366.09 366.8 366.9 366.0 366.1 366.2 366.3 366.4 366.5 366.12 366.13 366.10 366.11 366.16 366.17 366.14 366.15 366.34 366.18 366.19 366.30 366.31 366.32 366.33 363.9 363.8 363.5 363.4 363.7 363.6 363.61 363.62 363.63 363.42 363.43 363.40 363.41 363 363.72 363.71 363.70 363.55 363.54 363.57 363.56 363.51 363.50 363.53 363.52 378.72 372.44 372.45 372.40 372.41 372.42 372.43 372.81 372.62 372.63 372.61 372.64 372.53 372.52 372.51 372.50 372.56 372.55 372.54 372.34 372.71 372.73 372.72 372.75 372.74 372.7 372.6 372.5 372.4 372.9 372.8 372.89 370.34 371.49 371.48 370.33 371.45 371.44 371.46 371.41 371.40 371.43 371.42 371.89 370.31 371.81 371.82 370.2 371.8 371.9 370.4 371.0 371.1 371.2 371.3 371.4 371.6 371.7 371.12 371.13 371.10 371.11 371.16 371.14 371.15 371.30 371.31 371.32 371.33 370.06 370.05 370.3 370.07 371.70 371.71 371.72 371.73 370.04 371.23 370.23 370.02 371.04 370.20 370.21 370.22 370.49 370.24 370.03 370.00 370.01 370.40 370.44 370.0 371.01 371.00 371.03 371.02 371.05 371.62 371.61 371.60 371 371.22 371.21 371.20 370.35 371.24 364.59 364.74 364.77 364.76 364.51 364.52 364.75 364.72 364.55 364.70 364.71 364.53 364.54 364.73 364.8 364.9 364.56 364 364.57 364.6 364.7 364.4 364.5 364.42 364.41 364.82 364.81 364.64 364.89 364.61 364.60 364.63 364.62 374.55 374.54 374.56 374.51 374.50 374.52 374.89 374.9 374.8 374.2 374.5 374.4 374.20 374.21 374.22 374.23 374.46 374.44 374.45 374.43 374.41 374.87 374.82 374.83 374.81 374 374.84 374.85 365.52 365.51 365.20 365.12 365.59 365.31 365.15 365.32 365.71 365.70 365.73 365.72 365.13 365.74 365.11 365.10 365.7 365.6 365.5 365.4 365.3 365.2 365.1 365.0 365.9 365.8 365 365.44 365.89 365.41 365.42 365.43 365.81 365.82 365.83 365.14 365.00 365.01 365.02 365.03 365.04 365.05 365.06 365.21 365.62 365.63 365.60 365.61 365.64 365.65 365.24 365.22 365.23 360.89 360.43 360.41 360.40 374.53 360.81 360.44 360.29 360 360.8 360.9 360.24 360.4 360.2 360.20 360.23 368.54 368.51 368.52 368.53 371.58 371.56 371.57 371.54 371.55 371.52 371.53 371.50 371.51 368.61 360.42 360.3 360.34 360.32 360.33 360.30 360.31 375.43 375.22 375.21 375.20 375.61 375.00 375.03 375.02 375.81 375.89 375.8 375.9 375.4 375.5 375.6 375.0 375.1 375.2 375.3 375.52 375.53 375.51 375.56 375.57 375.54 375.55 375.16 375.14 375.12 375.13 375.11 375.69 375 362.5 362.52 362.53 362.50 362.51 362.56 362.57 362.54 374.11 374.10 374.13 374.12 374.14 379.90 379.99 375.01 379.34 379.12 379.31 379.32 374.0 379.39 379.19 374.1 379.13 379.11 379.16 379.14 379.15 374.02 374.03 374.00 374.01 374.04 374.05 379.8 379.9 379.1 375.15 379 379.3 379.33 377.54 377.5 377.51 377.52 377.53 377.04 377.03 377.02 377.01 377.2 377.24 377.21 377.23 377.22 377.49 377.4 377.12 377.43 377.42 377.41 377.13 377.3 377.15 377.16 377.14 377.39 377.10 377.11 377.34 377.1 377.32 377.33 377.30 377.31 376.22 376.21 376.81 376.82 376.40 376.41 376.42 376.43 376.44 376.89 376.46 376.47 376.50 376.45 376.9 376.8 376.3 376.2 376.5 376.4 376.31 376.30 376.33 376.35 376.52 376.51 376.36 376.34 376 377.0 377.00 362.74 362.75 362.76 362.77 362.73 367 367.9 367.8 367.89 362.70 362.71 362.72 377.61 377.72 377.63 377.62 377.73 377.75 377.71 377.9 377.6 377.7 379.21 379.23 379.22 379.2 379.26 379.25 379.24 379.29
Musculoskeletal	Congenital musculoskeletal anomalies	756.52 756.53 733.03 733.02 733.0 733.00 733.09 737.43 737.42 737.41 737.40 754.2 737.21 737.20 737.22 737.29 737.4 737.2 737.3 737.0 737.1 737.8 737.9 737 737.10 737.34 737.32 737.30 737.31 737.19 737.39 737.12 359.7 359.71 359.89 359.79 359.81 359.8 359.9 359.6 359.24 359.4 359.5 359.3 736.9 736.8 735 736.89 736 756.83 756.55 732.9 732.8 732.7 732.6 732.5 732.4 732.3 732.2 732.1 732.0 732 729.31 721.00 722.2 729.30 723.6 729.39 722.70 722.71 722.72 722.73 729.3 721.90 722.92 722.93 722.90 722.91 722.30 722.31 722.32 724.9 724.6 724.7 722.10 722.11 722.52 722.39 724.8 722.51 722.4 722.5 722.6 722.7 722.0 722.1 723.7 722.3 723.9 723.8 722.8 722.9 721.3 721.2 721.0 721.7 721.6 721.5 724.71 721.9 721.8 724 722 723 724.70 721 723.5 724.79 722.81 722.80 722.83 722.82 359 359.1 550.02 550.03 550.00 550.01 553.29 551 550 553 552 551.2 552.9 552.8 553.8 553.9 552.3 552.2 552.1 552.0 553.2 553.3 553.0 553.1 550.9 552.20 552.21 551.8 551.9 550.1 550.0 552.29 551.0 552.01 552.02 551.3 550.11 550.10 550.13 550.12 551.21 552.00 553.21 551.29 553.20 551.01 551.00 551.03 551.02 551.20 550.91 550.90 550.93 550.92 551.1 553.03 553.02 553.01 553.00 552.03 526.2 526.3 526.0 526.1 526.6 526.4 526.5 524.12 526.8 526.9 524.0 524.1 524.19 524.6 524.11 525.2 524.05 524.51 526 524.10 524.09 524.04 524.02 524.69 524.01 524.00 524.03 524.64 524.63 524.62 524.61 524.60 526.89 524.07 526.81 524.06 526.69 526.61 526.63 526.62 525.20 525.21 525.22 525.23 525.24 525.25 525.26 756.56 726.64 726.69 726.65 728.79 726.63 726.62 726.61 726.60 728.19 735.0 728.10 728.11 728.12 728.13 729.89 726.90 729.8 726.30 726.31 728.8 728.9 728.7 729.1 728.5 728.2 728.3 728.1 726.70 726.71 726.72 726.19 726.32 726.79 726.10 726.33 726.0 726.1 726.2 726.3 726.4 726.5 726.6 726.7 726.8 726.9 727.1 728.4 728.89 728.88 728.87 728.85 728.84 728.83 728.81 726.39 726 728 359.22 359.21 717.89 733.13 717.81 717.83 717.82 717.85 717.84 736.75 734.0 736.76 736.71 736.70 736.73 736.72 736.79 738.5 738.4 738.7 738.6 738.1 738.0 738.3 738.2 738.9 738.8 717 717.8 736.41 736.42 717.0 717.1 717.2 717.3 717.4 717.5 717.6 717.7 736.74 718.40 718.41 718.42 718.43 718.44 718.45 718.46 718.47 718.48 718.49 733.8 718.4 735.4 735.5 733.98 735.1 735.3 733.94 733.95 733.96 733.97 735.8 735.9 730.3 733.93 730.37 730.36 730.35 730.34 730.33 730.32 730.31 730.30 733.19 730.39 730.38 726.12 726.11 349.39 736.29 736.22 733.82 736.20 736.21 738.11 738.10 738.12 738.19 733.1 349.3 736.39 736.31 736.30 736.32 733.81 734 738 733.14 736.04 736.05 736.06 736.07 736.00 736.01 736.02 736.03 736.09 733.15 717.49 733.16 717.41 717.40 717.43 717.42 729.81 729.82 733.10 736.81 736.3 736.2 736.1 733.11 736.0 736.7 736.6 736.5 717.9 736.4 733.12 755.14 754.31 754.30 754.33 754.32 754.35 755.54 755.56 755.57 755.50 755.51 755.52 755.53 755.58 755.59 755.36 755.37 755.34 755.35 755.32 755.33 755.30 755.31 755.38 755.39 756.79 756.73 756.72 756.71 756.70 755.25 755.24 755.27 755.26 755.21 755.20 755.23 755.22 755.65 755.29 755.28 755.64 756.82 756.81 754.51 754.50 754.81 754.82 755.10 755.11 755.12 755.13 754.62 754.60 754.61 754.69 754.3 754.7 754.6 754.5 754.4 728.6 756.3 754.71 754.70 756.7 756.6 756.9 754.79 755.02 755.01 755.00 754.40 754.41 754.42 754.43 754.44 755.6 755.4 755.5 755.2 755.3 755.0 755.1 755.8 755.9 755.69 754.59 755.61 755.60 755.63 755.62 754.53 754.52 755.67 755.66 719.10 719.11 719.12 718.65 719.14 719.15 718.60 719.17 719.18 719.19 718.3 718.2 718.1 719.13 718.7 718.6 718.5 718 718.8 718.99 718.98 719.16 718.93 718.92 718.91 718.90 718.97 718.96 718.95 718.94 718.13 718.12 718.11 718.10 718.17 718.15 718.14 718.19 718.18 718.75 718.74 718.77 718.76 718.71 718.70 719 719.08 719.07 719.06 719.05 719.04 719.03 719.02 719.01 719.00 719.98 719.99 718.0 719.90 719.91 719.92 719.93 719.94 719.95 719.96 719.97 719.78 719.79 719.70 719.76 719.77 719.75 719.6 719.7 719.4 719.5 719.2 719.0 719.1 718.9 719.8 719.9 719.87 719.86 719.85 719.84 719.83 719.82 719.81 719.80 719.89 719.88 718.57 719.60 719.63 719.69 719.68 719.62 718.59 718.58 719.61 718.56 718.55 718.54 719.65 719.64 719.67 719.66 718.53 718.51 718.50 718.28 718.29 718.22 718.23 718.20 718.21 718.26 718.27 718.24 718.25 718.39 718.38 718.31 718.30 718.33 718.32 718.35 718.34 718.37 718.36 719.54 719.55 719.56 719.57 719.50 719.51 719.52 719.53 719.58 719.59 718.52 718.73 719.09 718.72 719.43 719.42 719.41 719.40 719.47 719.46 719.45 719.44 719.49 719.48 719.25 719.24 719.27 719.26 719.21 719.20 719.23 719.22 718.79 719.29 719.28 718.78 718.88 718.89 718.84 718.85 718.86 718.87 718.80 718.81 718.82 718.83 718.04 718.05 718.07 718.00 718.01 718.02 718.03 718.08 718.09 733.2 733.3 733.7 733.4 733.5 733.9 731.1 731.2 731.3 731.8 733.90 733.91 733.92 726.91 731 756.50 726.73 733.44 733.29 733.49 733.21 733.20 733.45 733.22 733.43 733.42 733.41 733.40 756.51 359.29 756.59 731 756.54 359.23 724.03 724.02 724.0 724.00 724.09 724.01 723.0 727.67 727.42 727.43 727.40 727.41 727.68 727.69 727.64 727.65 727.66 727.49 727.60 727.61 727.62 727.63 727.09 727.06 727.04 727.05 727.02 727.03 727.00 727.01 727.82 727.83 727.81 727.89 727.9 727.8 727.5 727.4 727.6 727.0 727.3 727.2 727.51 727.50 727.59 727

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

Developmental	Congenital anomalies (Non-overlap. with musculoskeletal)	755.55 743.45 743.12 743.11 743.10 743.06 743.00 743.03 743.1 743.44 743.41 743.42 743.43 743.22 743.2 743.20 743.21 758.4 745.1 745.0 745.3 745.2 745.5 747.5 746.09 747 745.4 745.8 746.02 746.01 746.00 745.9 745 747.89 745.7 745.6 747.82 747.83 745.60 745.61 745.69 747.21 746.6 746.7 746.4 746.5 746.2 746.3 746.0 746.1 747.9 747.8 746.8 746.9 746.87 746.86 746.85 746.84 746.83 746.82 746.81 746.89 745.19 747.11 745.11 745.10 745.12 756.4 744.4 743.63 743.61 743.66 744.1 743.64 743.65 743.69 744.8 744.9 744.0 754.1 754.0 748.1 748.0 744 744.49 744.2 744.84 744.3 744.5 743.6 744.29 744.47 744.42 744.89 744.09 744.81 744.83 744.82 744.05 744.04 744.01 744.00 744.03 744.02 744.41 744.43 744.24 744.23 744.22 744.21 744.46 743.35 742.51 756.2 756.19 742.53 756.11 756.10 756.13 756.12 756.15 756.14 758.31 315.4 315.5 315.8 315 315.9 758 758.2 752.4 752.1 752.0 752.3 752.2 752.42 752.43 752.40 752.41 752.46 752.47 752.44 752.45 752.49 752.36 752.35 752.34 752.33 752.32 752.31 752.39 752.19 752.11 752.10 759.83 751.61 751.7 751.6 751.60 751.69 743.62 520.4 520.5 520.0 520.1 520.2 756 752.7 758.7 756.16 752.5 752.6 752.51 752.52 752.69 752.64 752.65 752.61 752.62 752.63 740 740.0 740.2 741.9 740.1 741.02 741.0 741.00 741.01 741 742.0 741.91 741.90 741.93 741.92 741.03 756.17 758.5 758.39 758.81 758.9 743.48 743.49 743.46 743.47 743.9 743.8 743.5 743.3 743.59 743.58 743.53 743.52 743.51 743.57 743.56 743.55 743.54 743.39 743.34 743.37 743.36 743.31 743.30 743.33 743.32 750 751.4 751.5 751.2 751.0 751.1 750.4 750.9 750.8 750.7 750.6 750.24 751.9 750.3 751.8 750.5 753.0 753.1 753.2 753.3 753.4 753.5 753.6 753.7 753.8 753.9 752.9 752.8 593.72 593.73 593.70 593.71 593.0 753.19 593.7 752.81 753.12 753.10 753.11 753.17 752.89 753 752 753.21 589 753.23 753.22 589.9 753.20 589.1 589.0 753.15 753.29 742 742.59 742.2 742.3 742.1 742.4 742.5 742.9 524.70 524.71 524.72 524.73 524.74 524.75 524.76 524.79 520.6 520.7 520.3 520.8 520.9 750.29 750.26 750.27 750.25 750.22 750.23 750.21 524.2 524.3 524.4 524.5 524.7 524.8 524.9 524 520 749.04 749.02 749.03 749.00 749.01 524.39 750.16 750.15 750.13 750.12 750.11 750.10 750.19 749.14 749.11 749.10 749.13 749.12 524.81 524.82 524.89 749.20 749.21 749.22 749.23 749.24 749.25 524.34 524.35 524.36 524.37 524.30 524.31 524.32 524.33 750.2 750.1 750.0 524.56 524.57 524.54 524.55 524.52 524.53 524.50 749 524.59 749.1 749.0 749.2 524.23 524.22 524.21 524.20 524.27 524.26 524.25 524.24 524.29 524.28 758.1 759.81 259.1 259.0 748.61 748.60 748 748.69 748.8 748.9 748.2 748.3 748.4 748.5 748.6 758.33 759.89 758.6 313.89 313.9 313 309.21 307.7 307.6 313.8 313.23 313.2 759.9 759.2 759.3 759.0 759.1 759.7 759.4 747.40 747.41 747.64 747.29 747.62 747.63 747.60 747.61 747.22 747.49 747.20 747.81 747.69 747.3 747.2 747.1 747.0 747.6 747.4 747.42 747.10 758.32
Nutrition	Nutrition, metabolism, and development	7830,78321,7833,78340,78342,7837,7839
Health Status & Services Contact.	Vaccination, Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services etc.	V01-V91

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

Algorithm 1: ICD-9 Encoding

input : Dataset, TargetDiseaseGroup, DiseaseGroups
output: Encoding

```
Encoding ← new Dictionary();
for diseaseGroup ∈ DiseaseGroups do
    Encoding[diseaseGroup][patientID] ← new List();
    Encoding[diseaseGroup][gender] ← new List();
    Encoding[diseaseGroup][record] ← new List();
    Encoding[diseaseGroup][target] ← new List();
    for record ∈ Dataset do
        //encode Dataset into a weekly trinary sequence;
        weeklyEncoding ← new List();
        for weeklyDiseaseRecord ∈ record do
            //no code recorded for the observed week;
            if weeklyDiseaseRecord.code == NIL then
                append "0" to weeklyEncoding;
            if weeklyDiseaseRecord.code ∈ diseaseGroup.codes then
                append "1" to weeklyEncoding;
            if weeklyDiseaseRecord.code ∉ diseaseGroup.codes then
                append "2" to weeklyEncoding;
        target ← 1 if any weeklyDiseaseRecord.code of record ∈ TargetDiseaseGroup;
        if target == 1 then
            cut weeklyEncoding up to (but not including) first occurrence of TargetDiseaseGroup member;
        append record.patientID to Encoding[diseaseGroup][patientID];
        append record.gender to Encoding[diseaseGroup][gender];
        append weeklyEncoding to Encoding[diseaseGroup][record];
        append target to Encoding[diseaseGroup][target];
return Encoding;
```

1. Pipeline Optimization

1.1. 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[2] (See Section 11), 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.

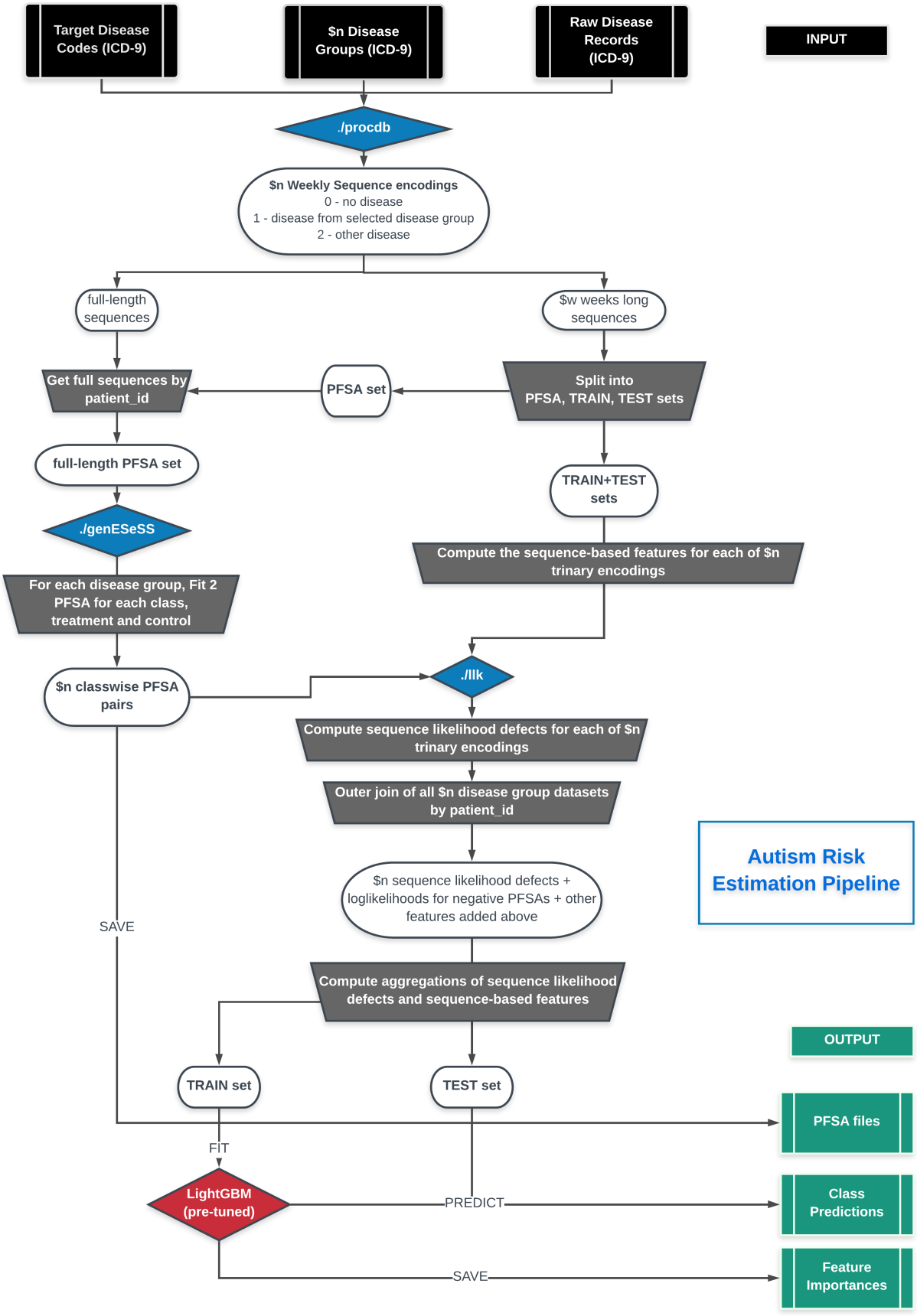
1.2. Algorithms

The key data processing approach is outlined in Algorithm 1. The remaining steps of the approach are sketched in Algorithm 2. Fig. 8 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 2: Prediction Pipeline Training

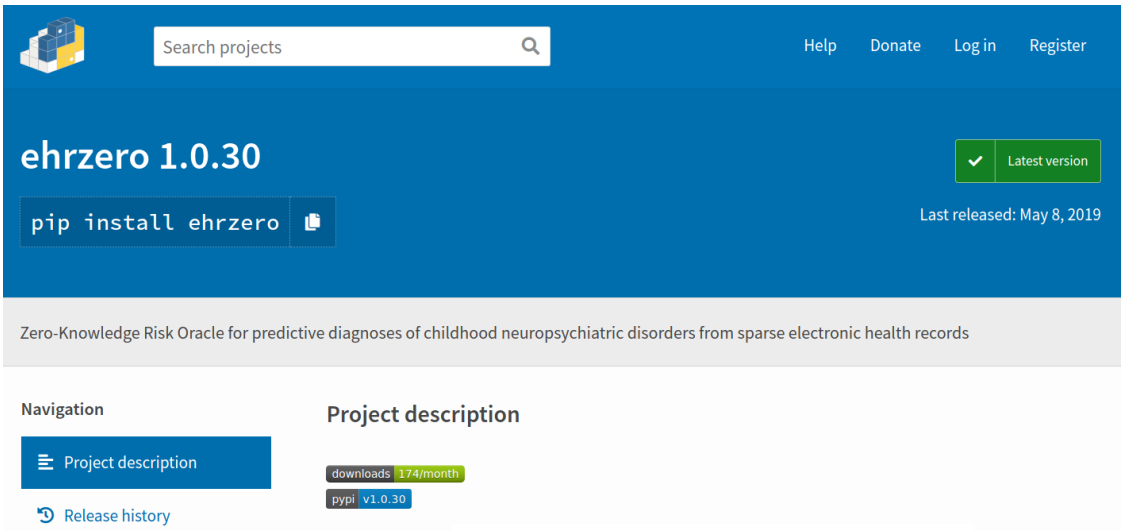
input : Encoding, DiseaseGroups, SequenceFeatures, hyperparameters
output: Predictions, FeatureImportances

DiseaseDatasets ← new `Dictionary()`;
for $Dataset, DiseaseGroup \in zip(Encoding, DiseaseGroups)$ **do**
    PFSAset, LLset ← `TrainTestSplit(`*Dataset, w.r.t = "target"*`)`;
    df ← new `Dataframe()`;
    df[patientID] ← LLset[patientID];
    df[target] ← LLset[target];
    //Generate 2 PFSAs for each class;
    PositivePFSAset ← PFSAset[PFSAset.target == 1];
    NegativePFSAset ← PFSAset[PFSAset.target == 0];
    PosPFSA ← `genESeSS(`*PositivePFSAset*`)`;
    NegPFSA ← `genESeSS(`*NegativePFSAset*`)`;
    //For each record, compute loglikelihoods of being generated by either of 2 PFSAs generated above;
    PosLLK ← `llk(`*LLset, PosPFSA*`)`;
    NegLLK ← `llk(`*LLset, NegPFSA*`)`;
    //Compute sequence likelihood defect;
    df[DiseaseGroup] ← `pairwise(`*PosLLK - NegLLK*`)`;
    df[DiseaseGroup + '_abs_neg'] ← NegLLK;
    for $SequenceFeature \in SequenceFeatures$ **do**
        df[DiseaseGroup + '_' + SequenceFeature] ← [`ComputeSequenceFeature(`*SequenceFeature, seq*`)` for each seq ∈ LLset['record']];
    DiseaseDatasets[DiseaseGroup] ← df;
Dataset ← `outerjoin(`*DiseaseDatasets.values, on = 'patientID'*`)`;
Aggregate all features in Dataset where feature_name ∈ DiseaseGroups (mean, std. deviation, range);
Aggregate all features in Dataset where feature name minus '_abs_neg' ∈ DiseaseGroups (mean, std. deviation, range);
Aggregate all sequence features in Dataset (mean, std. deviation, range, max);
TrainSet, TestSet ← `TrainTestSplit(`*Dataset, w.r.t = "target"*`)`;
LGBM ← new `LightGBM(`*hyperparameters*`)`;
LGBM.fit(TrainSet);
Predictions ← LGBM.predict(TestSet);
return *Predictions, LGBM.feature_importances*;



SI-Fig. 8: 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 ket algorithms here are `genESeSS`[2] and the llk which does the sequence likelihood computation described in Section 12

2. Example Run with Released Application



SI-Fig. 9: 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
```

[3]:

	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. 10: Python code prediction example

2.1. 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
```

2.2. 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
```

2.3. 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.

2.4. 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 texttt../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 0.106625 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. 11: Python script prediction example

3. Comparison With State of the Art Off-the-shelf ML Algorithms

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 Fig. 3). Classifiers such as random forests[3], and gradient boosters[4] 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[5] 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.

4. Comparison With Pipeline Variations, Feature Subsets and Neural Net Post-processing

In addition to the naive baseline approaches, we also evaluated the performance achievable with LSTMs (denoted as LSTMB in Fig. 3) that use identical preprocessing 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. 3).

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. 5, panel D). In all these cases we analyzed, our pipeline has a clearly demonstrable advantage (See Fig. 5, panel D) that is stable across databases, under reductions in sample sizes, and in balanced resampling experiments (See Fig. 5, panel C).

While it is difficult to explain the exact source of a modeling framework's performance, and even more difficult to explain non-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.
4. **We May Have Missed Some Clever Transformation.** It is possible that extensive tuning or feature selection with LSTMs, RNNs or CNNs or some combination thereof, can replicate our performance, or even do better. There will always be that possibility, notwithstanding how much effort we put in to evaluate competing techniques. The authors welcome *future work in this direction that surpasses our performance reported here; this is only going to help the patients which is what matters.*

4.1. Feature Subset Evaluations & Code Density As A Feature

With regards to Fig. 5, 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 3 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.

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

$$\text{accuracy} = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \tag{1}$$

$$\text{F1} = \frac{2t_p}{2t_p + f_p + f_n} \tag{2}$$

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% specificity, which is highlighted in Fig. 2C in the main text.

6. Note on Reciever 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** (ρ) are defined as:*

$$\mathit{TPR} = \frac{t_p}{P} = \frac{t_p}{t_p + f_n} \tag{3}$$

$$\mathit{TNR} = \frac{t_n}{N} = \frac{t_n}{t_n + f_p} \tag{4}$$

$$\mathit{FPR} = 1 - \mathit{TNR} \tag{5}$$

$$\mathit{PPV} = \frac{t_p}{t_p + f_p} \tag{6}$$

$$\rho = \frac{P}{N + P} \tag{7}$$

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 specifciyty by c, it follows that:

$$\text{PPV} = \frac{t_p/P}{t_p/P + (f_p/N)(N/P)} = \frac{\text{TPR}}{\text{TPR} + ((N - t_n)/N)(N/P)} \tag{8}$$

$$\Rightarrow \text{PPV} = \frac{s}{s + (1 - c)(\frac{1}{\rho} - 1)} \tag{9}$$

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:

$$\text{PPV} = s/(s + 2.45) \leqq 1/3.45 \sim 29\% \tag{10}$$

This shows that for ASD screening, we can hope for a maximum PPV of $\sim$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. 2C 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[1].

7. 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:

$$\text{TPR} = \frac{t_p}{P} = \frac{t_p'}{P'} \tag{11}$$

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. 5C, 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 (9). 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 9.1. First, we establish that our pipeline does not suffer from some important pitfalls arising in the workflows associated with ASD diagnosis, and how the diagnostic codes in Electronic Health Records (EHR) are generated.

8. Note on ASD Clinical Diagnosis & Uncertainty of EHR Record

With no precise laboratory test for ASD, most families experience the following sequence of events[6–8]: 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[9,10]. 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[8,11].

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[8].

8.1. 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[7], 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)[8]. 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)[12] and the Childhood Autism Rating Scale, Second Edition (CARS-2)[13]. Current guidance from the American Academy of Pediatrics[8] 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[14]. 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[14]. 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"[14]. With ADOS-2 and associated tools verfiably 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 workflows, 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. 5B illustrate that we have very little change in out-of-sample predictive performance, thus alleviating this concern.

8.2. 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)[8]. 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.

8.3. Performance Comparison With M-CHAT/F

The M-CHAT/F is the most studied and widely used tool for screening toddlers for ASD[8,11].

Guthrie *et al.*[1] 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 4 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 (≈ 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[8]. Additionally, while M-CHAT/F is relatively easy and quick to administer, the issue of time and resource commitment cannot be ignored[8]. 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[1].

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.

9. Improving Wait-times For Diagnostic Evaluations by Reducing False Positives in Routine Screening

While children with ASD can be diagnosed as toddlers[15,16] (developmental concerns may show up before the first birthday[17,18]), the mean age of diagnosis is over 4 years[19]. 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[20], cost of care[21], lack of providers[22], lack of comfort in diagnosing by primary care providers[22], and other challenges, are all responsible to varying degrees that culminate in these delays[6], 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[6,23].

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.

9.1. 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 ρ_i respectively, with $i \in \{1, \cdots, m\}$. Let β_i be the relative size of the sub-populations. Thus, we have:

$$P = \sum_i P_i \tag{12}$$

$$N = \sum_i N_i \tag{13}$$

$$\beta_i = \frac{N_i + P_i}{N + P} \tag{14}$$

$$\rho_i = \frac{P_i}{N_i + P_i} = \frac{P_i}{\beta_i (N + P)} \tag{15}$$

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

$$s = t_p/P = \frac{\sum_i t_p|_i}{P} = \frac{\sum_i (t_P|_i/P_i) \times (\beta_i \rho_i (P + N))}{P} = \sum_i s_i \beta_i \frac{\rho_i}{\rho} \tag{16}$$

Thus, we end up with:

$$s = \sum_{i=1}^{m} s_i \gamma_i \tag{17a}$$

$$c = \sum_{i=1}^{m} c_i \gamma_i' \tag{17b}$$

$$PPV = \frac{s}{s + (1 - c)(\frac{1}{\rho} - 1)} \tag{17c}$$

where we have denoted:

$$\gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma_i' = \beta_i \frac{1 - \rho_i}{1 - \rho} \tag{17d}$$

Now, using Table 2, we can compute the values for γ_i, γ_i', as shown below.

Using the prevalence and stratification parameters calculated from the CHOP study (See main text Table 3[1]), 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. (17). The results are shown in Fig. 4, 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 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. 2 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. 4 of this document, and in main text Fig. 4 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. (17) and hence the results are dependent on the population prevalence ρ. We report the dependence of the solution to the 4D optimization for population prevalence between 1.7% (CDC estimate[8]), and 2.23% (CHOP estimate[1]). In particular, it is illuminating to compare these results directly with M-CHAT/F performance, as shown in Fig. 4, 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%.

10. Generating PFSA Models From Set of Input Streams with Variable Input Lengths

Our PFSA reconstruction algorithm[2] 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.

11. Probabilsitic Finite State Automata Inference

11.1. Probabilistic Finite-State Automaton

Let Σ be a finite alphabet of symbols with size $|\Sigma|$. The set of sequences of length d over Σ is denoted by Σ^d. The set of finite but unbounded sequences over Σ is denoted by $\Sigma^\star$, the Kleene star operation[24], *i.e.* $\Sigma^\star = \bigcup_{d=0}^{\infty} \Sigma^d$. We use lower case Greek, for example σ or τ, for symbols in Σ, and lower case Latin, for example x or y, for sequences of symbols, *i.e.* $x = \sigma_1\sigma_2\ldots\sigma_n$. We use $|x|$ to denote the length of x. The empty sequence is denoted by λ.

We denote the set of strictly infinite sequences over Σ by Σ^ω, and the set of strictly infinite sequences having x as prefix by $x\Sigma^\omega$. Let $\mathcal{S} = \{x\Sigma^\omega : x \in \Sigma^\star\} \cup \{\emptyset\}$, we can verify that $\mathcal{S}$ is a semiring[25] over Σ^ω. We use $\mathcal{F}$ to denote the sigma algebra generated by $\mathcal{S}$.

Definition 2 (Stochastic Process over Σ)**.** *A stochastic process over a finite alphabet Σ is a collection of Σ-valued random variables $\{X_t\}_{t\in\mathbb{N}}$ indexed by positive integers*[26]*.*

We are specifically interested in processes in which the X_is are not necessarily independently distributed.

Definition 3 (Sequence-Induced Measure and Derivative)**.** *For a process $\mathscr{P}$, let $Pr_{\mathcal{P}}(x)$ or simply $Pr(x)$ denote the probability $\mathscr{P}$ producing a sample path prefixed by x. The* ***measure*** *μ_x* ***induced by a sequence*** *$x \in \Sigma^\star$ is the extension*[25] *to $\mathcal{F}$ of the premeasure defined on the semiring $\mathcal{S}$ given by*

$$\forall x, y \in \Sigma^\star, \mu_x\left(y\Sigma^\omega\right) \triangleq \frac{Pr\left(xy\right)}{Pr(x)},\ \textit{if } Pr(x) > 0 \tag{18}$$

*For any $d \in \mathbb{N}$, the d-**th order derivative** of a sequence x, written as ϕ_x^d, is defined to be the marginal distribution of μ_x on Σ^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 ϕ_x for short.*

Definition 4 (Probabilistic Nerode Equivalence and Causal States[27])**.** *For any pair of sequences $x, y \in \Sigma^\star$, 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***[28]*. 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 $(Q, \Sigma, \delta, \widetilde{\pi})$, where Q is a finite set, Σ is a finite alphabet, $\delta : Q \times \Sigma \to \Sigma$ is called the transition map, and $\widetilde{\pi} : Q \to \mathbf{P}_\Sigma$, where $\mathbf{P}_\Sigma$ is the space of probability distributions over Σ, is called the transition probability. The entry of $\widetilde{\pi}(q)$ indexed by σ is denoted by $\widetilde{\pi}(q, \sigma)$.*

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

$$\pi_{q,q'} \triangleq \sum_{\left\{\sigma\in\Sigma|\delta(q,\sigma)=q'\right\}} \widetilde{\pi}(q, \sigma) \tag{19}$$

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

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

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

$$\delta(q, x) \triangleq \delta\left(\delta\left(q, \sigma_1 \dots \sigma_{k-1}\right), \sigma_k\right) \tag{20}$$

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

$$\widetilde{\pi}(q, x) \triangleq \prod_{i=1}^{k} \widetilde{\pi}\left(\delta\left(q, \sigma_1 \dots \sigma_{i-1}\right), \sigma_i\right) \tag{21}$$

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

Definition 8 (Strongly Connected PFSA)**.** *We say a PFSA is strongly connected if the underlying directed graph is strongly connected*[29]*. More precisely, a PFSA $G = (Q, \Sigma, \delta, \widetilde{\pi})$ is strongly connected if for any pair of distinct states q and $q' \in Q$, there is an $x \in \Sigma^\star$ 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**[30,31] of G and is denoted as $\wp_G$, or $\wp$ if G is understood.

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

$$\Gamma_\sigma\big|_{q,q'} \triangleq \begin{cases} \widetilde{\pi}(q, \sigma) & \textit{if } \delta(q, \sigma) = q', \\ 0 & \textit{if otherwise.} \end{cases} \tag{22}$$

Γ_σ is called ***event-specific transition matrix****, with the event being that σ is current the output. Γ_σ can also be extended to arbitrary $x \in \Sigma^\star$ 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 $G = (Q, \Sigma, \delta, \widetilde{\pi})$ and a distribution $\wp_0$ on Q, the **distribution on Q induced by a sequence** x is given by $\wp^T_{G,\wp_0}(x) = [\![\wp_0^T \Gamma_x]\!]$ 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 $G = (Q, \Sigma, \delta, \widetilde{\pi})$ be a PFSA and let $\wp_0$ be a distribution on Q, the Σ-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 $\wp_{G,\wp_0}(X_1 \cdots X_t)$ 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 $\phi_x = \wp_G(x)^T \widetilde{\Pi}$. Since $\wp_G(\lambda) = \wp_G$, the symbolic derivative of the empty sequence ϕ_λ 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[32]. To be more precise, let $G = (Q, \Sigma, \delta, \widetilde{\pi})$ be a PFSA, we say a sequence $x \in \Sigma^\star$ 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 $\varepsilon > 0$, we say a sequence x is a ε-synchronizing sequence to a state $q \in Q$ if*

$$\|\wp_G(x) - \mathbf{e}_q\|_\infty \leq \varepsilon. \tag{23}$$

While there exists PFSA that is not synchronizable, we can show that an irreducible PFSA always has an ε-synchronizing sequence for some state q for arbitrarily small $\varepsilon > 0$. Moreover, we can show that as length increases, sequences produced by PFSA become uniformly ε-synchronizing. These two are the underpinning properties for the inference algorithm of PFSA (See Alg. 3), because they imply that ϕ_x can be used to approximate $\widetilde{\pi}(q)$ if x are properly prefixed and long enough.

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

$$\Sigma^d_{\varepsilon,(q,r)} \triangleq \left\{x \in \Sigma^d : \ x \text{ jointly } \varepsilon\text{-synchronizing to } (q, r)\right\} \tag{24}$$

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

$$p_G(q, r) \triangleq \lim_{d\to\infty} p_G\left(\Sigma^d_{\varepsilon,(q,r)}\right) \tag{25}$$

*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*

$$Q_c \triangleq \{(q, r) \in Q_G \times Q_H : \ (q, r) \text{ is a } G\text{-joint pair}\} \tag{26}$$

The inference algorithm for PFSA is called **GenESeSS** for Generator Extraction Using Self-similar Semantics. With an input sequence x and a hyperparameter ε, **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. 3 for detail.

12. 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[33]. Similarly, by KL divergence of two PFSA, we mean the KL divergence between the two processes generated by them[34]. More precisely, we have*

$$\mathcal{H}(G) = -\lim_{d\to\infty} \frac{1}{d} \sum_{x\in\Sigma^d} p(x) \log p(x) \tag{27}$$

Algorithm 3: `GenESeSS`

Data: A sequence x over alphabet Σ, $0 < \varepsilon < 1$
Result: State set Q, transition map δ, and transition probability $\widetilde{\pi}$
/* **Step One: Approximate ε-synchronizing sequence** */
Let $L = \left\lceil \log_{|\Sigma|} 1/\varepsilon \right\rceil$;
Calculate the **derivative heap** $\mathcal{D}_\varepsilon^x$ equaling $\left\{ \hat{\phi}_y^x \ : \ y \text{ is a sub-sequence of } x \text{ with } |y| \leq L \right\}$;
Let $\mathcal{C}$ be the convex hull of D_ε^x;
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** */
Initialize $Q = \{q_0\}$;
Associate to q_0 the **sequence identifier** $x_{q_0}^{\mathrm{id}} = x_0$ and the probability vector $d_{q_0} = \hat{\phi}_{x_0}^x$;
Let $\widetilde{Q}$ be the set of states that are just added and initialize it to be Q;
while $\widetilde{Q} \neq \emptyset$ **do**
    Let $Q_{\mathrm{new}} = \emptyset$ be the set of new states;
    for $(q, \sigma) \in \widetilde{Q} \times \Sigma$ **do**
        Let $x = x_q^{\mathrm{id}}$ and $d = \hat{\phi}_{x\sigma}^x$;
        if $\|d - d_{q'}\|_\infty < \varepsilon$ *for some* $q' \in Q$ **then**
            Let $\delta(q, \sigma) = q'$;
        else
            Let $Q_{\mathrm{new}} = Q_{\mathrm{new}} \cup \{q_{\mathrm{new}}\}$ and $Q = Q \cup \{q_{\mathrm{new}}\}$;
            Associate to q_{new} the sequence identifier $x_{q_{\mathrm{new}}}^{\mathrm{id}} = x\sigma$ and the probability vector $d_{q_{\mathrm{new}}} = d$;
            Let $\delta(q, \sigma) = q_{\mathrm{new}}$;
    Let $\widetilde{Q} = Q_{\mathrm{new}}$;
Take a strongly connected subgraph of the labeled directed graph defined by Q and δ, and denote the vertex set of the subgraph again by Q;
/* **Step Three: Identify transition probability** */
Initialize counter $N[q, \sigma]$ for each pair $(q, \sigma) \in Q \times \Sigma$;
Choose a random starting state $q \in Q$;
for $\sigma \in x$ **do**
    Let $N[q, \sigma] = N[q, \sigma] + 1$;
    Let $q = \delta(q, \sigma)$;
Let $\widetilde{\pi}(q) = [\![(N[q, \sigma])_{\sigma \in \Sigma}]\!]$;
return Q, δ, $\widetilde{\pi}$;

and the KL divergence

$$\mathcal{D}_{KL}(G \,\|\, H) = \lim_{d \to \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} \tag{28}$$

whenever the limits exist.

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

$$\mathcal{H}(G) = \sum_{q \in Q} \wp_G(q) \cdot h(\widetilde{\pi}(q)) \tag{29}$$

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

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

$$\mathcal{D}_{KL}(G \,\|\, H) = \sum_{(q,r) \in Q_c} p_G(q, r) \, D_{KL}(\widetilde{\pi}_G(q) \,\|\, \widetilde{\pi}_H(r)) \tag{30}$$

Definition 18 (Log-likelihood)**.** *Let* $x \in \Sigma^d$*, the log-likelihood*[33] *of a PFSA* G *generating* x *is given by*

$$L(x, G) = -\frac{1}{d} \log p_G(x) \tag{31}$$

The calculation of log-likelihood is detailed in Alg. 4.

Algorithm 4: Log-likelihood

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

Calculate the state transition matrix Π and observation $\tilde{\Pi}$;
Calculate the stationary distribution over states $\wp_G$ of G from Π;
Calculate the stationary distribution of alphabet $\phi_\lambda^T = \wp_G^T \tilde{\Pi}$;
Initialize $\mathbf{p}$ by $\wp_G$ and $\mathbf{q}$ by ϕ_λ;
Let $L = 0$;
for i *from* 1 *to* $|x|$ **do**
    Let σ be the i-th entry of x;
    Let $L = L - \log \mathbf{q}|_\sigma$;
    Let $\mathbf{p}^T = [\![\mathbf{p}^T \Gamma_\sigma]\!]$ where Γ_σ is defined in 9;
    Let $\mathbf{q}^T = \mathbf{p}^T \tilde{\Pi}$;
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*

$$L(x, H) \to \mathcal{H}(G) + \mathcal{D}_{KL}(G \,\|\, H) \tag{32}$$

in probability as $d \to \infty$.

Proof. We first notice that

$$\sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} = \sum_{x \in \Sigma^{d-1}} \sum_{\sigma \in \Sigma} p_G(x) \wp_G(x)\, \tilde{\Pi}_G\Big|_\sigma \log \frac{p_G(x) \wp_G(x)\, \tilde{\Pi}_G\Big|_\sigma}{p_H(x) \wp_H(x)\, \tilde{\Pi}_H\Big|_\sigma} \tag{33}$$

$$= \sum_{x \in \Sigma^{d-1}} p_G(x) \log \frac{p_G(x)}{p_H(x)} + \underbrace{\sum_{x \in \Sigma^{d-1}} p_G(x) \sum_{\sigma \in \Sigma} \wp_G(x)\, \tilde{\Pi}_G\Big|_\sigma \log \frac{\wp_G(x)\, \tilde{\Pi}_G\Big|_\sigma}{\wp_H(x)\, \tilde{\Pi}_H\Big|_\sigma}}_{D_d} \tag{34}$$

By induction, we have $\mathcal{D}_{\mathrm{KL}}(G \,\|\, H) = \lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^d D_i$, and hence by Cesàro summation theorem[35], we have $\mathcal{D}_{\mathrm{KL}}(G \,\|\, H) = \lim_{d\to\infty} D_d$. Let $x = \sigma_1\sigma_2 \ldots \sigma_n$ be a sequence generated by G. Let $x^{[i-1]}$ is the truncation of x at the $(i-1)$-th symbols, we have

$$-\frac{1}{n} \sum_{i=1}^n \log \wp_H\left(x^{[i-1]}\right) \tilde{\Pi}_H\Big|_{\sigma_i} = \underbrace{\frac{1}{n} \sum_{i=1}^n \log \frac{\wp_G\left(x^{[i-1]}\right) \tilde{\Pi}_G\Big|_{\sigma_i}}{\wp_H\left(x^{[i-1]}\right) \tilde{\Pi}_H\Big|_{\sigma_i}}}_{A_{x,n}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \log \wp_G\left(x^{[i-1]}\right) \tilde{\Pi}_G\Big|_{\sigma_i}}_{B_{x,n}} \tag{35}$$

Since the stochastic process G generates is ergodic, we have

$$\lim_{n\to\infty} A_{x,n} = \lim_{d\to\infty} D_d = \mathcal{D}_{\mathrm{KL}}(G \,\|\, H) \tag{36}$$

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

[1] Guthrie, W. *et al.* Accuracy of Autism Screening in a Large Pediatric Network. *Pediatrics* **144** (2019).
[2] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).
[3] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
[4] Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
[5] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
[6] Gordon-Lipkin, E., Foster, J. & Peacock, G. Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. *Pediatric Clinics* **63**, 851–859 (2016).
[7] Penner, M., Anagnostou, E. & Ungar, W. J. Practice patterns and determinants of wait time for autism spectrum disorder diagnosis in canada. *Molecular autism* **9**, 16 (2018).
[8] Hyman, S. L., Levy, S. E., Myers, S. M. *et al.* Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics* **145** (2020).
[9] Johnson, C. P., Myers, S. M. *et al.* Identification and evaluation of children with autism spectrum disorders. *Pediatrics* **120**, 1183–1215 (2007).

[10] Zwaigenbaum, L. *et al.* Early intervention for children with autism spectrum disorder under 3 years of age: recommendations for practice and research. *Pediatrics* **136**, S60–S81 (2015).

[11] Robins, D. L. *et al.* Validation of the modified checklist for autism in toddlers, revised with follow-up (m-chat-r/f). *Pediatrics* **133**, 37–45 (2014).

[12] Esler, A. N. *et al.* The autism diagnostic observation schedule, toddler module: standardized severity scores. *Journal of Autism and Developmental Disorders* **45**, 2704–2720 (2015).

[13] Chlebowski, C., Green, J. A., Barton, M. L. & Fein, D. Using the childhood autism rating scale to diagnose autism spectrum disorders. *Journal of autism and developmental disorders* **40**, 787–799 (2010).

[14] Falkmer, T., Anderson, K., Falkmer, M. & Horlin, C. Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European child & adolescent psychiatry* **22**, 329–340 (2013).

[15] Lord, C. *et al.* Autism from 2 to 9 years of age. *Archives of general psychiatry* **63**, 694–701 (2006).

[16] Kleinman, J. M. *et al.* Diagnostic stability in very young children with autism spectrum disorders. *Journal of autism and developmental disorders* **38**, 606–615 (2008).

[17] Bolton, P. F., Golding, J., Emond, A. & Steer, C. D. Autism spectrum disorder and autistic traits in the avon longitudinal study of parents and children: precursors and early signs. *Journal of the American Academy of Child & Adolescent Psychiatry* **51**, 249–260 (2012).

[18] Kozlowski, A. M., Matson, J. L., Horovitz, M., Worley, J. A. & Neal, D. Parents' first concerns of their child's development in toddlers with autism spectrum disorders. *Developmental neurorehabilitation* **14**, 72–78 (2011).

[19] Baio, J. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010 (2014).

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

[21] Bisgaier, J., Levinson, D., Cutts, D. B. & Rhodes, K. V. Access to autism evaluation appointments with developmental-behavioral and neurodevelopmental subspecialists. *Archives of pediatrics & adolescent medicine* **165**, 673–674 (2011).

[22] Fenikilé, T. S., Ellerbeck, K., Filippi, M. K. & Daley, C. M. Barriers to autism screening in family medicine practice: a qualitative study. *Primary health care research & development* **16**, 356–366 (2015).

[23] Althouse, L. A. & Stockman, J. A. Pediatric workforce: A look at pediatric nephrology data from the american board of pediatrics. *The Journal of pediatrics* **148**, 575–576 (2006).

[24] Hopcroft, J. E. *Introduction to automata theory, languages, and computation* (Pearson Education India, 2008).

[25] Klenke, A. *Probability theory: a comprehensive course* (Springer Science & Business Media, 2013).

[26] Doob, J. *Stochastic processes*. Wiley publications in statistics (Wiley, 1990). URL https://books.google.com/books?id=7Bu8jgEACAAJ.

[27] Chattopadhyay, I. & Ray, A. Structural transformations of probabilistic finite state machines. *International Journal of Control* **81**, 820–835 (2008).

[28] Chattopadhyay, I. & Lipson, H. Data smashing: uncovering lurking order in data. *Journal of The Royal Society Interface* **11**, 20140826 (2014).

[29] Bondy, J. & Murty, U. Graph theory (2008). *Grad. Texts in Math* (2008).

[30] Vidyasagar, M. *Hidden markov processes: Theory and applications to biology*, vol. 44 (Princeton University Press, 2014).

[31] Kai, L. C. *Markov Chains: With Stationary Transition Probabilities* (Springer-Verlag, 1967).

[32] Trahtman, A. N. The road coloring and Černý conjecture. In *Proc. of Prague stringology conference*, vol. 1, 12 (Citeseer, 2008).

[33] Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

[34] Matthews, A. G. d. G., Hensman, J., Turner, R. & Ghahramani, Z. On sparse variational methods and the kullback-leibler divergence between stochastic processes. *Journal of Machine Learning Research* **51**, 231–239 (2016).

[35] Hardy, G. Divergent series, with a preface by je littlewood and a note by ls bosanquet, reprint of the revised (1963) edition. *Éditions Jacques Gabay, Sceaux* (1992).