## 3. Research Strategy

**A. Significance:** Autism spectrum disorder is a developmental disability associated with significant social, communication, and behavioral challenges. The prevalence of ASD has risen dramatically in the United States from 1 in 10,000 in 1972 to 1 in 59 children in 2014, with males diagnosed at nearly four times the rate of females.[8], [13] There is a current lack of consensus on whether increased awareness and recent changes in diagnostic practices[1] can fully explain this trend.[14] Nevertheless, with possibly over 1% of individuals affected worldwide,[15] ASD is a human condition with potentially serious negative impacts on individuals, families, and communities. Early detection can and does improve outcomes,[1] and is of paramount importance when designing interventions, and is aligned with the envisioned goal of this initiative, as supported by the National Advisory Mental Health Council (NAMHC) (https://www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2018/early-screening-for-autism-spectrum.shtml).

Even though ASD may be reliably diagnosed as early as the age of two,[8] children frequently remain undiagnosed until after the fourth birthday.[16] At this time, there are no laboratory tests for ASD, so a careful review of behavioral history and social interactions is necessary for a clinical diagnosis.[1], [17] Starting with being flagged by an initial screen based on standardized checklists presented to parents at the ages between 1.5 and 2 years, a confirmed ASD diagnosis is a multi-step process that very often spans 3 months to 1 year. Most of this time is spent waiting to see qualified providers who can carry out the evaluation necessary for a clinical diagnosis. This extended wait is stressful to families, and impacts patient outcomes by delaying entry into time-critical intervention programs. While lengthy evaluations,[2] cost of care,[3] lack of providers,[4] and lack of comfort in diagnosing ASD by primary care providers[4] are all responsible to varying degrees,[18] one obvious factor responsible is the number of false positives produced in the initial ASD-specific screening tools in use today. For example, the M-CHAT/F, the most widely used screen,[1], [6] produces about over 85 false positives out of every 100 people flagged for further diagnostic evaluation, contributing to extended queues.[18] The impact from an excessive number of false positives is exacerbated by the current limited access to care and sparse availability of resources except near urban academic centers.[18], [19]

The standardized questionnaires attempt to measure risk by direct observation of behavioral symptoms, as reported by untrained observers (parents). Hence the current screening tests are only as good as the ability of the questions to discern and disambiguate behavior in infants and toddlers on casual observation, and on the ability of parents and caregivers to correctly interpret and answer the items without bias. This has lead to possibility of under-diagnosis in diverse communities as reflected by the lower apparent prevalence among African-American and Hispanic children. Also, children with average or higher-than-average cognitive abilities seem to have been under-diagnosed as reported is large scale population studies.[1] Borderline cases are typically problematic to screen for due to the possibility of subjective interpretation that is built into questionnaire based risk assessment. Responses to checklists are clearly confounded by a host of socio-economic (SES) variables, potential interpretive biases, and cultural differences. The heterogeneity of presentation also causes issues, since a potential plurality of symptom classes makes it harder for clinicians to recognize borderline cases, or on-the-fly combine observed co-morbidities with scores from standardized screening tools.

In this study, we operationalize a documented aspect of ASD symptomology in that it has a wide range of co-morbidities occurring at much higher rates than in the general population.[1] The ACoR methodology we propose in this grant can address the aforementioned complicated challenges of ASD screening, by distilling incipient predictive patterns of elevated risk from past medical history of individual patients, gleaned by machine learning algorithms from large de-identified databases of retrospective patient records. Powered by novel stochastic learning algorithms, we reverse-engineer sparse noisy uncertain diagnostic code sequences into actionable signatures; in effect giving us a fundamentally new approach to ASD screening and risk evaluation.

***Potential Impact of ACoR In Science and Health:*** An automated diagnostic or screening capability that might be administered with little or no specialized training, requires no behavioral observations, and is functionally independent of the current tools has the potential for immediate transformative impact on patient care.

That such predictions of neuropsychiatric disorders might be possible from analyzing patterns in individual medical histories is suggested by the fact that parents of children with ASD often notice a diagnosable developmental problem before their child's first birthday; while vision and hearing problems are not uncommon in the first year, differences in social, communication, and fine motor skills have been reported to be evident from about 6 months of age.[20]–[22]

***Parallel Screening in Pediatric Primary Clinic:*** To achieve the specific goals outlined in the specific aims of this study, we will gather data from both the child and the primary caregiver of all patients in the participating UCM primary pediatric primary care clinic, and test the sensitivity and specificity of ACoR against the most commonly used screening tool M-CHAT/F, by an ADOS-2 based (near) gold-standard evaluation of patients flagged by either tool. Using these data and the inferred statistical dependency (or the lack thereof) properties between the screening tools, ACoR can tailor the selection of sensitivity/specificity trade-offs to the particular informant and to the age of the child, with the view to optimizing global characteristics such as maximizing the PPV or the sensitivity of the screening process. Additionally, the ACoR algorithm will identify categories of heterogeneity that will lead to mechanistic insights into ASD pathobiology.

Thus, the ACoR methodology is potentially able to screen instantaneously every child in primary care, for whom past medical history is available, with zero administrative and resource burden. Our results suggest that ACoR has superior predictive performance to existing screening tools, along with a host of other advantages that directly relate to the stated specific aims of this study.

***Significance of Specific Aim 1:*** A crucial question in this study is the quantification of the lead time of our positive predictions to the actual clinical diagnosis in individual patients: how much earlier can we trigger an intervention? In our preliminary studies, computing the time in weeks from the first time the relative risk crosses the $90\%$ of the threshold maximizing the $F1$-score, to the week of the actual diagnosis, we found the mean value is greater than $150$ weeks across independent data sets. This does not necessarily indicate that we are leading a possible clinical diagnosis by over $3$ years; a significant portion of this delay arises from families waiting in queue for diagnostic evaluations. Nevertheless, given than such delays are estimated to be within one year,[18] we are likely to produce valid red flags significantly earlier than current practice. Specific aim 1 is designed to answer this question on the precise lead time of ACoR over competing state-of-the-art screening tools. Importantly a corrected lead time in excess of 2 years would imply the potential of bringing down the median diagnostic age by the same amount, with particular impact on borderline cases and children with above average cognitive abilities, and "high-functioning" presentation ASD.

***Significance of Specific Aim 2:*** A key clinical contribution of this study is the formalization of subtle co-morbidity patterns as a reliable screening tool, and potentially improve waiting-times for diagnostic evaluations by significantly reducing the number of false positives encountered in initial screens in current practice. In our preliminary studies, we established that ACoR outperforms M-CHAT/F by considering the average reported performance of M-CHAT/F in a recent study[23] on a large cohort with near-universal screening carried out at the Children's Hospital of Philadelphia (CHOP). However, not having observed the ACoR and the M-CHAT/F scores jointly for individual patients, our preliminary studies lack objective assessment of statistical dependence between the two scores. While the very nature of the methodologies suggest functional independence, specific Aim 2 will investigate and establish this rigorously.

Functional independence from existing tools implies we can combine the scores; especially leveraging the population stratification induced by the M-CHAT/F scores as reported by the CHOP study to significantly boost combined screening performance. In particular, since patients in the lower M-CHAT/F score bracket have a smaller chance of an ASD diagnosis compared to the high risk upper brackets, we can tailor the sensitivity/specificity trade-offs in the ACoR to maximize either the global PPV or the global sensitivity without losing specificity. Our preliminary results suggest that the expected gains are substantial, with the possibility of doubling the PPV, or increasing the sensitivity by over $50\%$ while keeping the specificity above $95\%$. Specific aim 2 will investigate the e viability of the preliminary results in a pediatric primary care setting.

***Significance of Specific Aim 3:*** While still lacking the certainty of a diagnostic blood test, use of subtle patterns emergent in the diagnostic history to estimate risk might help reduce the subjective component in questionnaire-based screening tools, resulting in reduced effect of potential language and cultural barriers in diverse populations.[1] With a significant portion of the cohort expected to be African Americans and Hispanics in our primary care clinic, our comparative investigations will be able to explicitly answer these questions.

***Significance of Specific Aim 4:*** Despite unprecedented advances in charting the numerous genetic variations,[24] and established to be highly heritable,[25] the etiology of autism is still unclear.[26], [27] Despite tremendous recent progress, efforts to identify causal biomarkers for ASD have had limited success. Currently, over one hundred genes have been shown to contribute to autism risk,[24], [28], [29] and it is estimated that up to 1000 genes might be involved in ASD pathogenesis.[30] Still, genetic interactions and mechanisms

have accounted for a limited number of ASD cases,[31] potentially implicating environmental triggers that work alongside genetic predispositions. The plausible sources of risk are estimated to range from prenatal factors such as maternal infection and inflammation, diet, and household chemical exposures, to autoimmune conditions and localized inflammation of the central nervous system after birth.[26], [27], [32]–[37] The heterogeneity of ASD presentation admits the possibility of a plurality of etiologies with converging pathophysiological pathways, making the investigation of the etiology of future risk modulation extremely challenging. Specific Aim 4 will aim to unravel clues to mechanistic drivers by charting and categorizing the heterogeneous presentation by the nature of past diagnostic pattern in individual medical histories.

In our preliminary investigations standard machine learning tools failed to achieve clinically meaningful performance; the available data is too sparse for off-the-shelf deep learning frameworks[38] to make personalized predictions, and standalone classifiers or regressors fail to exploit the temporal dynamics embedded in the sparse diagnostic histories, requiring us to devise novel machine inference algorithms and feature engineering approaches to distill effective risk predictors. Compared to attempts at identifying biomarkers from differential expression of genes,[39]–[42] implicated in neuroinflammation and other comorbid disorders, our predictive performance in preliminary studies is substantially better, with no extra tests that require drawing blood, and is validated on positive cohort sizes at least two orders of magnitude, and control cohorts over three orders of magnitude larger.

### B. Innovation:

***Paradigm Shift in ASD Screening:*** Despite extensive documentation of co-morbidities, a risk estimator that makes reliable predictions for individuals — based purely on co-morbidity patterns — has never been reported to the best of our knowledge. The sparsity of available diagnostic codes corresponding to individual subjects, and the general absence of physiological disorders that would uniquely signal the eventual emergence of symptoms indicative of a clinical ASD diagnosis, combined with the heterogeneity of ASD presentation, make such an endeavor challenging. In this study we leverage our preliminary work on the formulation of a *first-of-its-kind* framework to make predictions based on models of statistically curated patterns of diagnostic code sequences automatically learned from sufficiently large databases of electronic health records (EHR), that achieves an out-of-sample AUC exceeding $80\%$ for either gender from just over 2 years of age. The machine learning tools that make this possible are also fundamentally novel, designed to address the specific issues in handling sparse, noisy categorical diagnostic sequences.

**C. Approach:** The technical approach of this study consists of the development and extension of the ACoR methodology from our preliminary studies, and application in a primary care setting with view to carrying out objective comparisons between M-CHAT/F and ACoR. Furthermore, we plan to assess our ability to boost performance from a conditional combination of the two scores.

***Preliminary Studies:*** We describe the formulation of the ACoR score, and the underlying principles that distill the estimated risk from EHR databases.

**Source of Electronic Patient Records in Preliminary Studies:** Of the two independent sources of clinical incidence data used in our preliminary 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[43] (referred to as the Truven dataset). For our analysis, we extracted histories of patients within the age of $0-9$ years, and excluded patients who do not satisfy the following criteria: 1) At least one code of any available phenotypes is present, 2) Lag between first and last available record for a patient should be at least 15 weeks. These exclusion criteria ensure that we are not considering patients who have too few observations to either train on, or predict outcomes from. For training, we analyzed over 4M children ($n = 4.4M$), with 30M diagnostic records (16,649,548 for males and 14,318,303 for females with 9,835 unique diagnostic codes).

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

**Time-series Modeling of Diagnostic History:** Individual diagnostic histories can have long-term memory,[44] implying that the order, frequency, and comorbid interactions between diseases are potentially important for

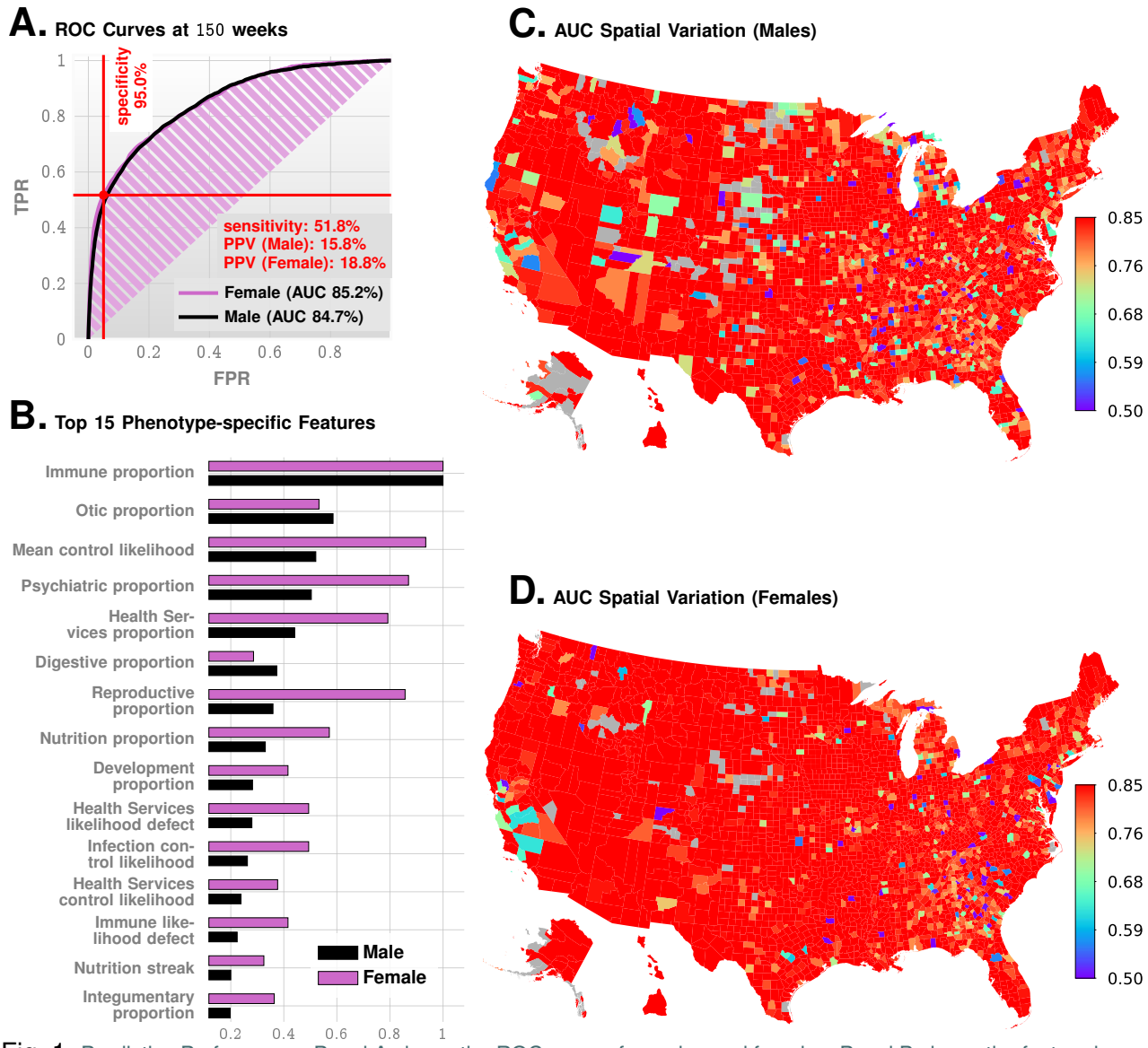


Fig. 1. Predictive Performance. Panel A shows the ROC curves for males and females. Panel B shows the feature importance inferred by our prediction pipeline. The most import feature is related to immunologic disorders, and we note that in addition to features related to individual disease categories, 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. Theses 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.

assessing the future risk of our target phenotype. Our approach to analyzing patient-specific diagnostic code sequences consists of representing the medical history of each patient as a set of stochastic categorical time-series — one each for a specific group of related disorders — followed by the inference of stochastic generators for these individual data streams. These inferred generators are from a special class of Hidden Markov Models (HMMs), referred to as Probabilistic Finite State Automata (PFSA).[45] The inference algorithm we use is distinct from classical HMM learning, and has important advantages related to the ability to infer structure, and sample complexity. We infer a separate class of models for the positive and control cohorts, and then the problem reduces to determining the probability that the short diagnostic history from a new patient arises from the positive as opposed to the control category of the inferred models. Importantly, the individual histories are typically short, often have large randomly varying gaps, and we have no guarantee that model-structural assumptions[46], [47] (linearity, additive noise structure, etc.) often used in the standard time-series analysis is applicable here. Also, the categorical observations are drawn from a large alphabet of possible diagnostic codes, which degrades statistical power.

**Step 1: Partitioning The Human Disease Spectrum:** To address these issues, we begin by partitioning the human disease spectrum into 17 non-overlapping categories, which remain fixed throughout the analysis.

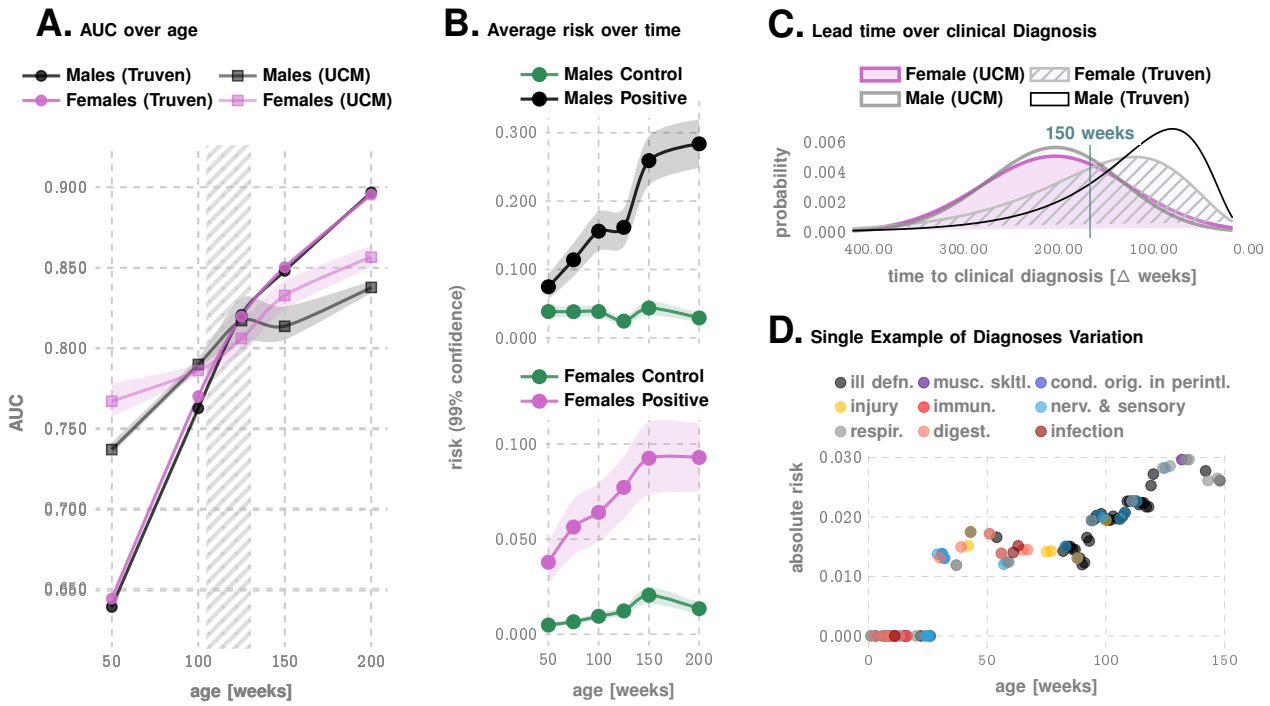


Fig. 2. More details on Predictive Performance and Variation of Inferred Risk. Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets. The shaded area outlines the 2 - 2.5 years of age, and shows that we achieve $> 80\%$ AUC for either gender from shortly after 2 years. Panel B illustrates how the average risk changes with time for the control and the positive cohorts. Panel C shows the distribution of the prediction horizon: the time to a clinical diagnosis after inferred relative risk crosses $90\%$. Panel d 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).

Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9). For this study, we considered $9,835$ distinct ICD9 codes (and their ICD10 General Equivalence Mappings (GEMS)[48] equivalents). We came across 6,089 distinct ICD-9 codes and 11,522 distinct ICD-10 codes in total in the two datasets we analyzed. Transforming the diagnostic histories to report only the broad categories reduces the number of distinct codes that the pipeline needs to handle, thus improving statistical power. The trade-offs for this increased power consist of 1) the loss of distinction between disorders in the same category, and 2) some inherent subjectivity in determining the constituent ICD9 codes that define each category, $e.g.$ an ear infection may be classified either an otic disease or an infectious one.

We do not pre-select the phenotypes; we want our algorithm to seek out the important patterns without any manual curation of the input data. The limitation of the set of phenotypes to $9835$ unique codes arises from excluding patients from the database who have very few and rare codes that will skew the statistical estimates. Next, we process raw diagnostic histories to generate data streams that report only the categories instead of the exact codes. For each patient, his or her past medical history is a sequence $(t_1, x_1), \cdots, (t_m, x_m)$, where $t_i$ are timestamps and $x_i$ are ICD9 codes diagnosed at time $t_i$. We map individual patient history to a three-alphabet time series $z^k$ corresponding to the disease category $k$, as follows. For each week $i$,

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

The time-series $z^k$ is terminated at a particular week if the patient is diagnosed with ASD the week after. In summary, each patient is now represented by $17$ mapped trinary series, which we use next to infer population-level PFSA models.

**Step 2: Model Inference & The Sequence Likelihood Defect:** The mapped series, stratified by gender, disease-category, and ASD diagnosis-status are considered to be independent realizations or sample paths from relatively invariant stochastic dynamical systems; and we explicitly model these systems as HMMs. We model the positive and the control cohorts for each gender, and in each disease category separately, ending up with a total of $68$ HMMs at the population level ($17$ categories, $2$ genders, $2$ cohort-types: positive and

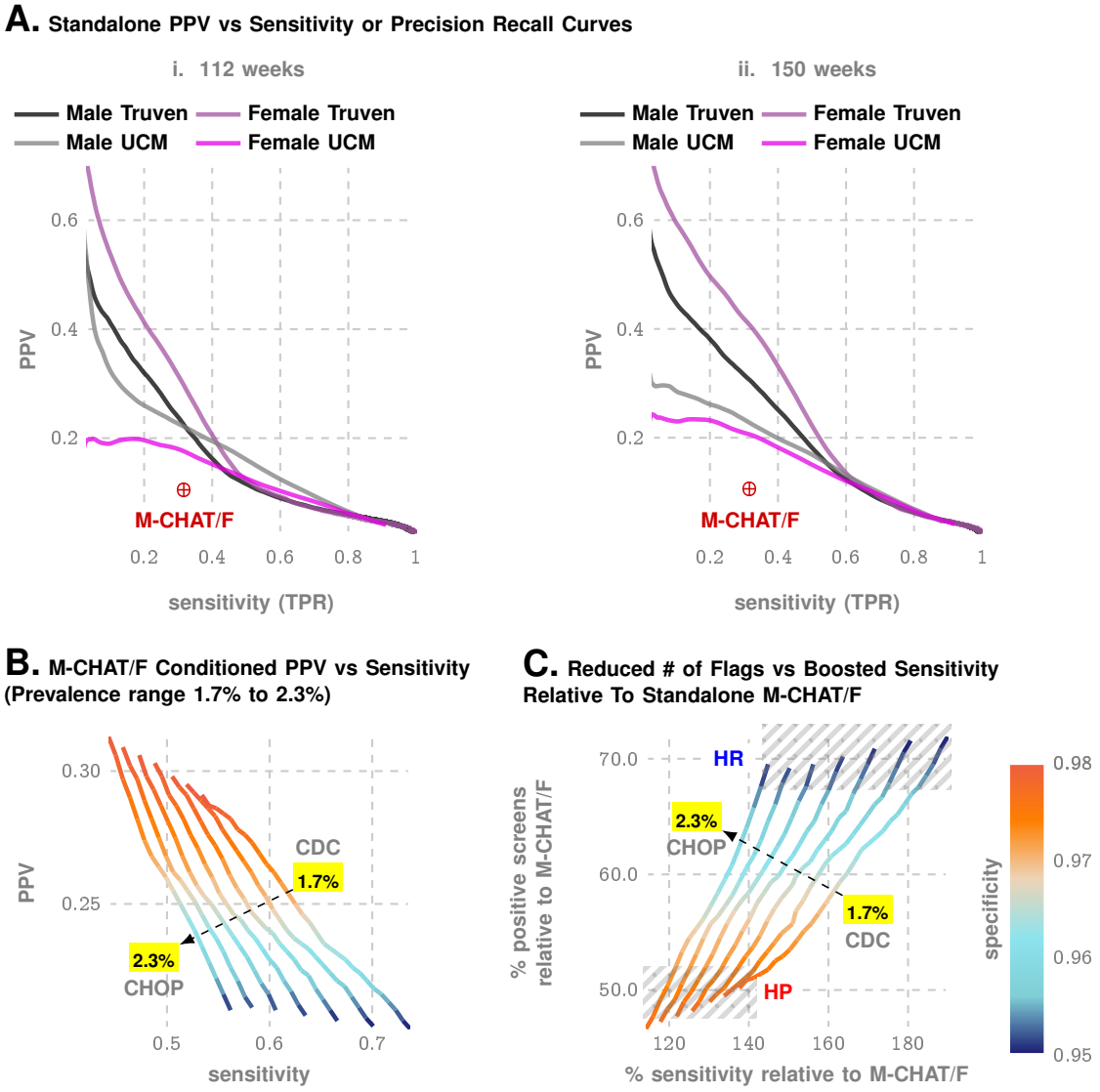


Fig. 3. **Metrics relevant to clinical practice: PPV vs Sensitivity trade-offs.** Panel A shows the precision/recall curves, *i.e.*, the trade-off between PPV and sensitivity. Panel B shows how we can boost performance using population stratification from the distribution of M-CHAT/F scores in the population, as reported by the CHOP study.[23] Panel C illustrates the boosted performance compared to M-CHAT/F alone, mesured by the relative percentage increase in sensitivity, and percentage decrease in postive 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 seinsitivity 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.

control). Each of these inferred models is a PFSA; a directed graph with probability-weighted edges, and acts as an optimal generator of the stochastic process driving the sequential appearance of the three letters (as defined by Eq. (1)) corresponding to each gender, disease category, and cohort-type. The modeling objective here is to exploit the relative differences in these probabilistic models to reliably infer the cohort-type of a new patient from their individual sequence of past diagnostic codes. To that effect, we generalized the well-known notion of Kullbeck-Leibler (KL) divergence[49], [50] between probability distributions to a divergence $\mathcal{D}_{\mathrm{KL}}(G||H)$ between ergodic stationary categorical stochastic processes[51] $G, H$ as:

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

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

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

The cohort-type for an observed sequence $x$ — which is actually generated by the hidden process $G$ — can be formally inferred from observations based on the following provable relationships:

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

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

where $\mathcal{H}(\cdot)$ is the entropy rate of a process.[49] Importantly, Eq. (4) shows that the computed likelihood has an additional non-negative contribution from the divergence term, when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with ASD, then we expect that the disease-specific mapped series corresponding to her diagnostic history be modeled by the PFSA in the positive cohort. Denoting the PFSA corresponding to disease category $j$ for positive and control cohorts as $G_+^j, G_0^j$ respectively, we can compute the *sequence likelihood defect* (SLD, $\Delta^j$) as:

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

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

**Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules:** Ultimately, the risk estimation pipeline operates on patient specific information limited to the gender and available diagnostic history from birth, and produces an estimate of the relative risk of ASD diagnosis at a specific age, with an associated confidence value. To learn the parameters and associated model structures of this pipeline, we transform the patient specific data to a set of engineered features, and the feature vectors realized on the positive and control sets are then used to train a gradient-boosting classifier.[52] Of the set of engineered features, the most important are the disease-category-specific SLD described above. For example, if $SLD > 0$ for a specific patient for every disease category, then he or she is likely to have an ASD diagnosis eventually. However, not all disease categories are equally important for this decision; parametric tuning of the classifier allows us to infer the optimal combination weights, as well as compute the relative risk with associated confidence. In addition to category-specific SLDs, we use a range of other derived quantities as features, including the mean and variance of the defects computed over all disease categories, the occurrence frequency of the different disease groups, etc. The top $15$ features used in our pipeline may be ranked in order of their relative importance (See Fig. 1B), by estimating the loss in performance when dropped out of the analysis, and the importance of infections and immunologic disorders are clearly evident (See Fig. 1B).

**Calculating Relative Risk:** Our pipeline maps medical histories to a score, which is interpreted as a raw indicator of risk — higher this value, higher the probability of a future diagnosis. However, to make crisp predictions, we must choose a decision threshold for this raw score. Conceptually identical to the notion of Type 1 and Type 2 errors in classical statistical analyses, the choice of a threshold trades off false positives (Type 1 error) for false negatives (Type 2 error): choosing a small threshold results in predicting a larger fraction of future diagnoses correctly, *i.e.* have a high true positive rate (TPR), while simultaneously suffering from a higher false positive rate (FPR), and vice versa. We choose thresholds for the standalone ACoR method by 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. The *relative risk* is then defined as the ratio of the raw pipeline score to the chosen decision threshold. While the raw score does not give us actionable information, the relative risk being close to or greater than 1.0 for a specific child signals the need for intervention.

**Standalone Predictive Performance:** The standalone performance of our risk estimator is summarized in Fig. 1 and Fig. 2. The task of predicting if a patient has a future ASD diagnosis, at an earlier date compared to the clinical diagnosis, may be viewed as a binary classification problem. In our case, we achieve an out-of-sample AUC of $82.3\%$ for males and $82.5\%$ for females at $125$ weeks of age for the Truven dataset. In the UCM dataset, our performance is comparable: $83.1\%$ and $81.3\%$ for males and females respectively at $125$ weeks

TABLE 1

Standalone ACoR PPV Achieved

(For Comparison M-CHAT/F Performance: sensitivity=38.8%, specificity=95%, PPV=14.6% between within 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 |

of age. The good agreement of the out-of-sample performance on these independent datasets lends strong evidence for the claims made in this study. The specificity, sensitivity, PPV trade-offs are shown in Table 1.

We enumerate the top 15 predictive features in Fig. 1B. We also computed the county-specific performance of the risk pipeline for the Truven dataset, and we got nearly uniform performance across the country for both genders, with the exception of few isolated counties lacking patients in the appropriate age groups (See Fig. 1, panels C and D) which prevented us from estimating AUC for those counties. Thus the performance of the algorithm is relatively agnostic to the number of local diagnoses, which is import in light of the fact that crucial diagnostic resources currently have a very uneven distribution (only 7% of developmental pediatricians practice in rural areas, and some states in US do not even have a developmental pediatrician[18], [19]).

Fig. 2A illustrates the variation of the AUC with increasing age of the subjects plotted with 99% confidence bounds: the increase is very nearly linear, with a change of gradient near the 150 week mark. We suspect that the median diagnosis age in the databases ($\approx$ 150 weeks) manifests this inflection. The curves for the smaller UCM dataset are less smooth, probably due to more uncertainty. We find that while the AUC gradients are different in the two datasets, they tend to match up in later ages. The differences in the early ages are possibly due to differences in patient statistics: a larger number of patients in Truven at the earlier ages with a relatively smaller number of observations on average.

We plot the absolute or raw risk over time for males and females for the out-of-sample control and positive cohorts in Fig. 2B. We see that while the risk for the control cohort remains more or less stable, that for the positive cohort rapidly increases. Notably, in these risk plots, averaged over the population, we see disambiguation early, right from 50 weeks. Also, we see a saturation of the risk after 150 weeks, which corresponds to the median diagnosis age in the database (approx. 150 weeks). 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.

**Calculating PPV, Sensitivity & Specificity Trade-offs & M-CHAT/F Comparison:** The sensitivity vs PPV plots, also known as the precision-recall curves (See Fig. 3A) are constructed in a similar fashion as the ROC curves by varying the decision threshold. These curves allow direct comparison with the state of the art screening tests,*e.g.*, M-CHAT/F, in a manner that is most relevant to clinical practitioners. Guthrie *et al.*[23] from Children's Hospital of Philadelphia (CHOP) has recently demonstrated that when applied as a nearly universal screening tool, M-CHAT/F has a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%, implying that out of every 100 children who in fact ave ASD, the M-CHAT/F flags about 39, and out of every 100 children it flags, about 85 are false positives. The PPV is affected by the prevalence of the disease. This work is the only large-scale study of M-CHAT/F (n=20,375) we are aware of with sufficient follow-up after the age of four years to provide a reasonable degree of confidence in the reported performance values.

Comparing the performance metrics achieved at different age groups across data sets and genders for our

TABLE 2

Population Stratification Results on large M-CHAT/F Study(n=20,375) reproduced from Guthrie *et al.*[23]

| Id | Sub-population | Test Result | ASD pos. | ASD Neg. | Total % |
|---|---|---|---|---|---|
| A | M-CHAT/F $\geqq 8$ | Positive | 0.34% | 0.64% | 0.99% |
| B | M-CHAT/F $\in [3,7]$ | Positive (follow-up) | 0.52% | 4.39% | 4.91% |
| C | M-CHAT/F $\in [3,7]$ | Negative (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% |

pipeline (See Table 1), we conclude that our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at $14\%$ (14.1-33.6%) when sensitivity and specificity are held at comparable values around the age of 26 months ($\approx$ 112 weeks). We cannot compare at other operating points due to a lack of M-CHAT/F performance characterization anywhere else.

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

$$s = \sum_{i=1}^{m} s_i \gamma_i, \text{ and } c = \sum_{i=1}^{m} c_i \gamma_i' \text{ where we have denoted: } \gamma_i = \beta_i \frac{\rho_i}{\rho}, \text{ and } \gamma_i' = \beta_i \frac{1-\rho_i}{1-\rho} \tag{6a}$$

and $s, c, \rho$ are the overall sensitivity, specificity, and prevalence. Knowing the values of $\gamma_i, \gamma_i'$, we can carry out an $m$-dimensional search to identify the feasible choices of $s_i, c_i$ pairs for each $i$, such that some global constraint is satisfied, $e.g.$ minimum values of specificity, sensitivity, and PPV. We consider $4$ sub-populations defined by M-CHAT/F score brackets,[23] and if the screen result is considered a positive (high risk, indicating the need for a full diagnostic evaluation) or a negative, $i.e.$, low risk: 1) score $\leq 2$ screening ASD negative, 2) score $[3-7]$ screening ASD negative on follow-up, 3) score $[3-7]$ and screening ASD positive on follow-up, and 4) score $\geq 8$, screening ASD positive. (See 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[23] 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.

This ultimately yields an overall performance significantly superior to M-CHAT/F alone. We carry out a four dimensional search at the age of 26 months ($\approx 112$ weeks) to identify the feasible region with PPV $> 14.6\%$ and sensitivity $> 38.8\%$ simultaneously while keeping specificity $> 94\%$. These four dimensions reflect the independent choice of sensitivities in the corresponding sub-populations. For each set of choices, the associated specificities are read-off from our fixed pre-computed ROC curve corresponding to $112$ weeks, and then the overall sensitivity, specificity and PPV are calculated using standard relationships.

We get a PPV close to or exceeding $30\%$ at the high precision (HP) operating point across datasets ($> 33\%$ fro Truven, $> 28\%$ for UCM), or a sensitivity close to or exceeding $50\%$ for the high recall (HR) operating point ($> 58\%$ for Truven, $> 50\%$ for UCM), when we restrict specificities to above $95\%$ (See Table 3).

It is important to compare these results directly with M-CHAT/F performance, as shown in Fig. 3, panels C. In panel C, we show that for any stable population prevalence between $1.7\%$ and $2.23\%$, the conditional operation can achieve double the PPV relative to M-CHAT/F alone without losing sensitivity at $> 98\%$ specificity,

TABLE 3

Boosted Sensitivity, specificity and PPV Achieved at 26 months Personalized Operation Conditioned on M-CHAT/F Scores

| M-CHAT/F Outcome | | | | global performance (Truven) | | | global performance (UCM) | | | prevalence* |
|---|---|---|---|---|---|---|---|---|---|---|
| 0-2 NEG | 3-7 NEG | 3-7 POS | $\geq 8$ 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 $2.23\%$. Results depend on the prevalence estimate.

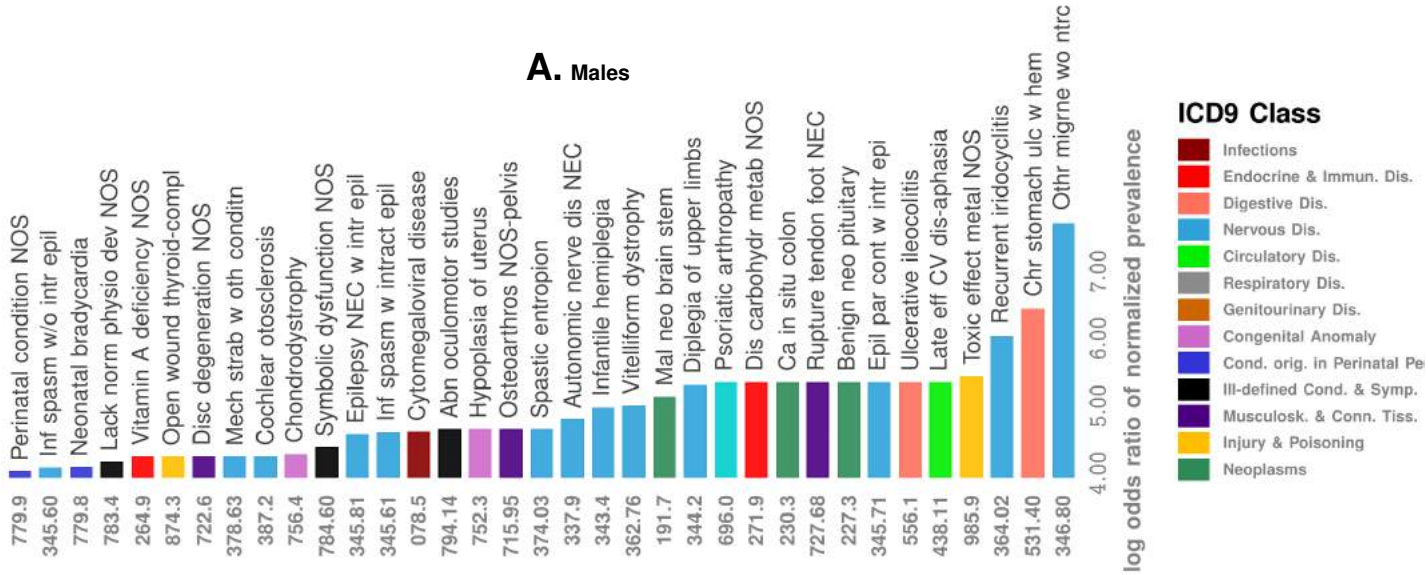


Fig. 4. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions in males. The color coding shows the disease categories of the co-morbidities.

or increase the sensitivity by $\sim 50\%$ without sacrificing PPV and not letting the specificity to drop below $94\%$.

Importantly, designing the rules for conditional operation only requires average population characteristics, *i.e.*, an estimate of ASD prevalence in the sub-populations defined by the relevant brackets of M-CHAT/F scores, and the prevalence of these score brackets in the general population. In particular, M-CHAT/F scores of individual patients are unnecessary for designing the rules themselves, or evaluating the overall expected performance in the population, provided the stratification statistics (Table 2) remains invariant. However, in the proposed study, we will be able to compute explicit dependency relationships between M-CHAT?F and ACoR scores.

**Inferred Co-morbidity Patterns & Normalized Prevalence Comparison:** The predictive ability of our pipeline arises from the difference in patterns of co-morbid disorders between the positive and the control cohorts: the diagnostic history of individual patients is not random and hides key signatures to future neuropsychiatric outcomes. As an illustrative example, a single random patient from the Truven database is illustrated in Fig. 2D.

Color-coding the diagnoses according to the broad ICD9 disease categories reveals that for this specific individual, infections and immunological disorders are experienced early to a much higher degree compared to other diseases, and diseases of the nervous system and sensory organs, as well as ill-defined symptoms dominate the latter period. This suggests the necessity of a deeper interrogation of the structure of co-morbid patterns, which we carried out in our preliminary investigations, as described next.

While the ASD co-morbidity burden is reported to be high for nearly the entire spectrum of physiological disorders, in our preliminary we find novel association patterns 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. Additionally, we only focus on the true positives in the positive cohort and the true negatives in the control cohort. This allows us to investigate patterns that correctly disambiguate future ASD status, $i.e.$, strongly favor one outcome over the other at the individual level (as opposed to population-level prevalence rates), as shown in Fig. 4 for males.

Additionally, we found in our preliminary studies indications of : 1) *negative associations:* there are diseases that are negatively associated with ASD diagnosis with respect to normalized prevalence, $i.e.$, having those codes over-represented relative to other codes in one's diagnostic history favors ending up in the control cohort, 2) *gendered impact:* there are gender-specific differences in the impact of specific disorders which will be further investigated in this study.

**Effect of Change In Diagnostic Criteria: Inclusion of PDD & Asperger's Syndrome:** 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).[1] This aligns with our use of diagnostic codes from ICD9 299.X as specification of an ASD diagnosis, and use GEMS mapping to 299.X from ICD10 codes when we encounter them. Importantly, we found that it is difficult to design a high performing pipeline that recognizes these ASD sub-types separately, even if we so wanted.

**Disambiguation From Unrelated Psychiatric Phenotypes:** 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 evaluated this question, by restricting the control cohort in validation 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.

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

We verified that class imbalance is not inappropriately enhancing our performance, by replacing the control cohort with a random sample of size equal to that of the positive cohort in out-of-sample tests.

We also found that the density of diagnostic codes in a child's medical history by itself is somewhat predictive of a future ASD diagnosis, but not at clinically significant levels.

***Research Design:*** The key steps in our research design are as follows (See Fig. 5):

□ 1. The pediatric clinic team (Dr. Mitchell and Dr. MSall) will enable adminstering of M-CHAT/F to incoming children in appropriate age groups (16-30 months)

□ 2. The University of Chicago Research Informatics Support team led by John Moses (Letter of Support included) will work with the PI and his team to integrate ACoR within the EPIC system deployed in the hospital to enable background processes to kick in automatically to compute the ACoR score corresponding to individual consenting patients

□ 3. On being flagged by either tool as high risk, the patients will be scheduled for ADOS-2 evaluation overseen

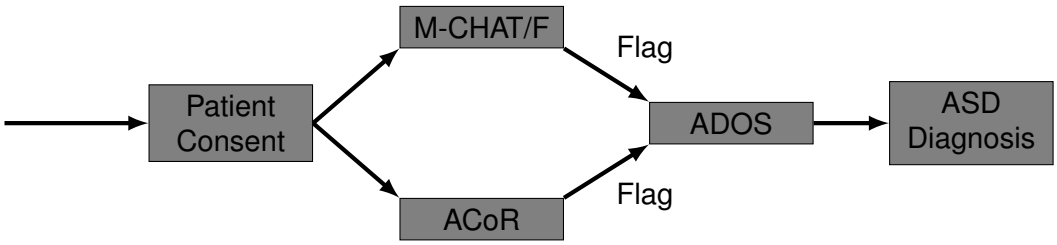


Fig. 5. Patient flow logic in this study. Consenting families with children in appropriate age group (16 to 30 months) are subjected to both M-CHAT/F and ACoR screens, and a flag in either screen is then scheduled for an ADOS evaluation for ascertaining ASD status.

by Dr. Smith and his team

□ 4. The evaluated scores will be post-processed and analyzed by the PI and his computational team to answer questions laid out in the specific aims.

**Cohort Selection:** We plan to establish the result as a universal procedure, that is applicable irrespective of population characteristics. Nevertheless, it is clearly conceivable that co-morbidities have demographic dependencies. Unfortunately, the key training dataset (Truven) does not have demographic information. Hence, the ACoR algorithm at this point does not take such variables into consideration. With a diverse patient population at the University of Chicago, we plan to *not* pre-select population characteristics, but analyze the performance of the proposed tool on the different demographic and ethnic groups in the post-processing.

Our power calculations suggest that we need to achieve a total cohort size (over $4$ years) of $n \geq 28.5K$, leading to $\approx 1600$ potential flags, which translates to $\approx 400$ ADOS evaluations per year, to achieve $> 95\%$ confidence bound.

**Risk To Patients:** The design of the study guarantees that patients suffer no negative impact from the added ACoR screen. Indeed, patients who are flagged are to be immediately scheduled for ADOS evaluation which eliminates their wait-times. For some borderline cases, which would have been missed by M-CHAT/F, might get flagged by ACoR, and be scheduled for ADOS, which they would not have had to do with just M-CHAT/F. But this is a positive outcome. The possibility that ACoR might have some false positives that are different from that of M-CHAT/F and hence schedule some children for ADOS, who do not have autism is a possibility, and might cause some stress in parents and families. The potential societal benefit gained in lieu of this discomfort is the validation of the expected performance boost for ASD screening at the population level PPV by up to 100%, or the sensitivity by 50%.

**Procedures:** Eligible patients at the Department of Pediatrics, University of Chicago (patients who present for a well-child visit or any other non-emergency reason) will be asked for consent for carrying out the ACoR screen in the background. The algorithm will be already integrated with EPIC, such that indicating this consent on the screen will automatically trigger pulling up the patient medical history if available, carrying out the calculations, and storing the results to be analyzed in post-processing. If there is a flag either in M-CHAT/F or in ACoR, the attending pediatrician will inform parents of a potential elevated risk of ASD diagnosis, and offer to schedule for an ADOS-2 evaluation. The ADOS-2 evaluation for the ACoR flags will be at no cost to the patient.

All study procedures and consent forms will be approved by the University of Chicago Institutional Review Board. For all assessments, basic demographic information, recruitment site, medications and diagnoses assigned by the current clinical treatment team, will be obtained from the parent/caregiver and medical record.

**Data Management:** Data collection forms for demographic and clinical history data, database design and data management procedures will be designed, created and conducted at the University of Chicago under the direction of Dr. Smith. Demographic and clinical history data will be collected and entered into an HIPAA compliant secure database. Data will be entered within one day of collection and will pass through rigorous quality control checks for accuracy and completeness before and after data entry. Monthly reports will be generated to monitor data timeliness, completeness, and accuracy as well as subject flow through the study. Data sets stripped of patient identifiers will be sent electronically to Dr. Chattopadhyay as needed for analysis.