3. Research Strategy

Abbreviations Used

ASD	Autism Spectrum Disorder
MCHAT-F	Modified Checklist for Autism in Toddlers with Followup
ADOS / ADOS-2	Autism Diagnostic Observation Schedule
EHR	Electronic Health Records
ACoR	Autism Comorbid Risk

3.1. 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.[12], [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,[13] 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] Our methodology can address the aforementioned challenges of ASD screening, by leveraging predictive signatures of elevated risk gleaned from past medical history of individual patients. Powered by a suite of novel stochastic learning algorithms trained and validated in our preliminary studies on very large patient databases, we reverse-engineer sparse noisy diagnostic code sequences into actionable signatures; in effect giving us a new approach to ASD screening.

Potential Impact of ACoR In Science and Health: A screening capability independent of existing tools, deployable as an automated module of a standard EHR system at the point of care, requiring no behavioral observations or new blood-work or laboratory tests has considerable potential to transform ASD care.

ASD presentation has significant heterogeneity, with no simple comorbidity consistently signaling future diagnosis; our algorithms distill robust actionable signatures under such stochastic scenarios. Thus, ACoR opens the possibility of a new screening modality for neuropsychiatric diseases beyond ASD.

Significance of Specific Aim 1: Diagnostic delays partially arise from families waiting in queue for diagnostic evaluations.[18] We expect ACoR to reduce false positives significantly, and hence reduce the number of children flagged for diagnostic evaluation, potentially reducing diagnostic delays. We cannot measure diagnostic delay directly since evaluations might be fast-tracked, but will use false positive rate as a proxy for wait-time.

Significance of Specific Aim 2: In our retrospective preliminary studies, ACoR supersedes M-CHAT/F performance[20] 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 assessment of statistical dependence between the two scores. While the methodologies suggest functional independence, specific Aim 2 will investigate this rigorously. Independence from existing tools implies we can combine the scores to significantly boost standalone screening performance.

Significance of Specific Aim 3: Use of comorbidity patterns 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, we will be able to explicitly investigate these questions.

Significance of Specific Aim 4: Despite advances in charting heritability,[21], [22] efforts to identify causal biomarkers have had limited success.[23], [24] While $100 - 1000$ genes might modulate ASD risk,[21], [25]–[27] genetics have accounted for a limited number of cases.[28] Suspected sources of environmental risk range from maternal infection and inflammation, diet, and household chemical exposures, to autoimmune conditions and localized perinatal inflammation of the central nervous system.[23], [24], [29]–[34] A plurality of etiologies with converging pathophysiological pathways is also plausible, and we aim to unravel clues to mechanistic drivers by categorizing the heterogeneous presentation via signatures buried in longitudinal co-morbidity patterns.

3.2. 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 our knowledge. The sparsity of diagnostic codes in individuals, the absence of physiological disorders that would consistently signal the eventual emergence of ASD symptoms, combined with the heterogeneity of ASD presentation, make such an endeavor challenging. This *first-of-its-kind* study proposes to estimate risk of a complex neuropsychiatric disease based on longitudinal patterns learned from large databases of sparse uncurated medical history. In our preliminary studies, we achieve an out-of-sample AUC exceeding 80% for either sex from just over 2 years of age.

The machine learning (ML) tools that make this possible are also fundamentally novel, designed to address the specific issues in handling sparse, noisy categorical diagnostic sequences.

Sophisticated analytics to identify children at high risk is a topic of substantial current interest, with independent progress being made by several groups.[35]–[41] Many of these approaches focus on analyzing questionnaires, with recent efforts demonstrating the use of standard automated pattern recognition in video clips of toddler behavior. However, the inclusion of older children and small cohort sizes in these studies is problematic. More importantly, a common thread in these attempts is the use of standard machine learning (ML) tools on currently well established modalities to try replicate physician behavior. In contrast, the ACoR innovation is developing a new modality of screening, and importantly, aiming to model the disease itself, not the physician response.

3.3. Approach: We describe our approach in the context of our preliminary retrospective results, outlining the ACoR methodology, towards prospective application in a primary care setting.

Source of Electronic Patient Records in Preliminary Studies: Of the two independent sources of patient records used in our preliminary study, the primary source used to train our models is the Truven Health Analytics MarketScan® Commercial Claims and Encounters Database for the years 2003 to 2012[42] (referred to as the Truven dataset). We extracted histories of patients within the age of $0 - 5$ 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 with too few observations to either train on. 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).

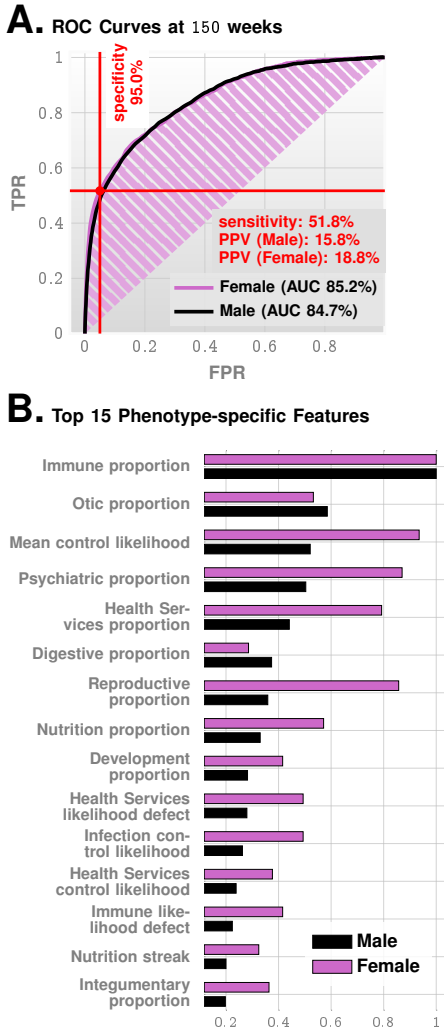

Fig. 1. Panel A: ROC curves. Panel B: feature importance inferred by our prediction pipeline. The most import feature is related to immunologic disorders, and we note that in addition to features related to individual disease categories, we also have the mean control likelihood (rank 3), which may be interpreted as the average likelihood of the diagnostic patterns in the control category vs the positive category.

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.

Predicting future ASD diagnosis is a binary classification problem: we classify time-stamped sequences of diagnostic codes into positive and control categories, where the "positive" category refers to patients eventually diagnosed with ASD (defined as people with one or more ICD9/10 codes corresponding to ASD in their medical history). For learning the differences in longitudinal patterns, we consider data from birth (or the earliest record) upto the time at which the prediction/screening is done. We do not pre-select any diagnostic code based on its suspected comorbidity with ASD.

Modeling & Prediction: The significant diversity of diagnostic codes, along with the sparsity of codes per patient (30-100 codes on average per patient per year of life, with 9,835 unique codes leads to very few consistent repeats for straightforward probability calculations) makes this a difficult learning problem. We proceed by partitioning the disease spectrum into 17 broad categories, *e.g.* infectious diseases, immunologic disorders, and endocrinal disorders. Some of these categories comprises a relatively large number of diagnostic codes aligning roughly with the ICD categories.[43] Each category yield a single time series over weeks (each week being identified as having a value '0' for no code corresponding to the diagnostic category, or '1' if some code is present, and '2' if a diagnostic code from any of the other categories is present). These time series are compressed into specialized Hidden Markov Models known as Probabilistic Finite Automata.[44], [45] These models are inferred separately for each phenotype, for each sex, and for the control and the positive cohorts, to identify distinctive average patterns emerging at the population level. Thus, we infer $17 \times 2 \times 2 = 68$ PFSA models in total in this study. Variation in these inferred models across positive and control groups quantify the divergence of comorbidity patterns with increasing risk.[46]

In addition, we use a range of engineered features that reflect various aspects of the patient-specific diagnostic histories, ultimately computing 701 features for each patient. These features are used to train a standard gradient boosting classifier[47] aiming to map individual patients to a raw risk score. 75% of our patients are randomly selected for training with the rest held-out as a validation set. We measure our performance using standard metrics including the Area Under the receiver-operating characteristic curve (AUC), sensitivity, specificity, and the Positive Predictive Value (PPV).

Calculation of the ACoR score offers insights into the relative importance of comorbidity categories, computed by estimating the mean change in the raw risk via random perturbation of a particular feature: this is the "feature importance" shown in Fig. 1c for top contributing categories, indicating that immunological, otic, digestive disorders and infections are important categories modulating the ACoR score. In our preliminary studies, we found excellent disambiguation from other intellectual disabilities.

Importantly, our features are based on data already available in the past medical records. We do not demand results from specific tests, or look for specific demographic, bio-molecular, physiological and other parameters; *we use what we get* in the diagnostic history of patients.

The standalone performance in preliminary studies is summarized in Figs. 1 and 2. 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

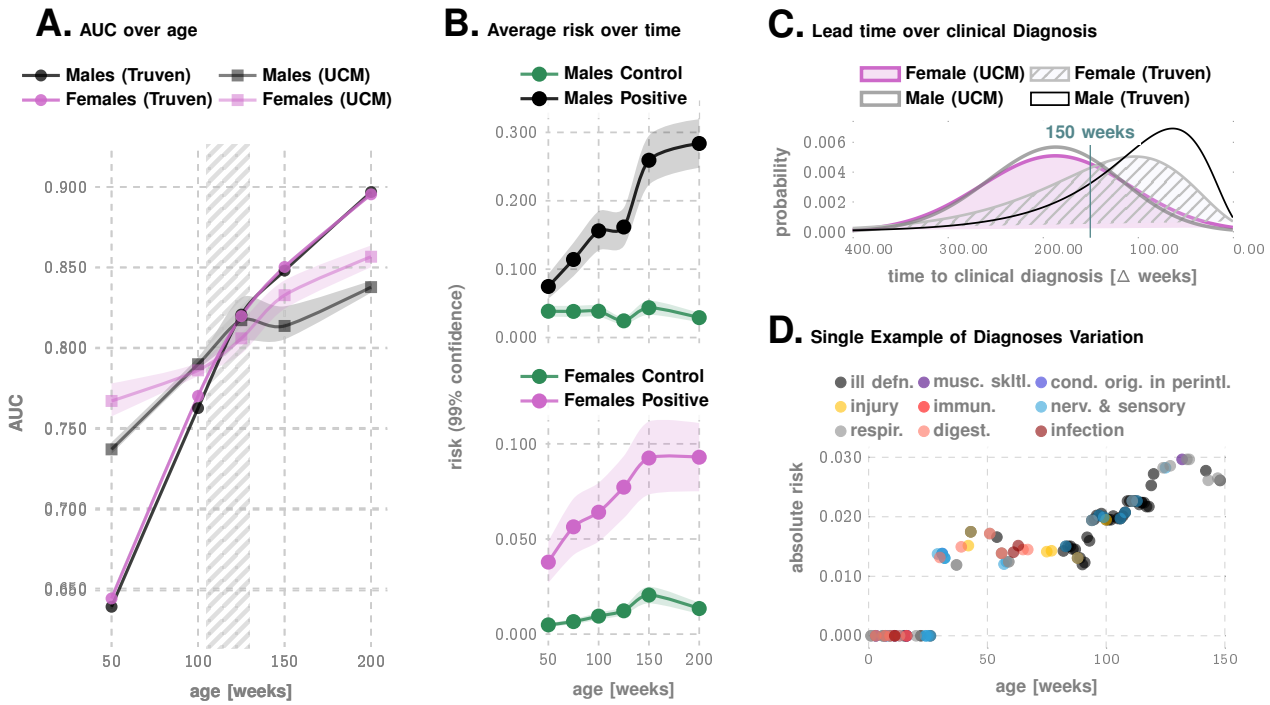

Fig. 2. Variation of Inferred Risk. Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets: we achieve $> 80\%$ AUC for either gender from shortly after 2 years. Panel B illustrates risk variation with time for the control and the positive cohorts. Panel C shows 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.

dataset, our performance is comparable: 83.1% and 81.3% for males and females respectively at 125 weeks of age. The good agreement of the out-of-sample performance on these independent datasets lends strong support for our claims. 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. Fig. 2A illustrates the variation of the AUC with increasing age of the subjects plotted with 99% confidence bounds, indicating the predictive performance increases with age. We find that while the AUC gradients are slightly different in the two datasets are comparable.

TABLE 1

Standalone ACoR performance (M-CHAT/F: sensitivity=38.8%,specificity=95%, PPV=14.6%)

week	spec.	sens.	PPV	sex	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

We plot the raw risk over time for males and females for the out-of-sample control and positive cohorts in Fig. 2B. Notably, averaged over the population, the risks differ from 50 weeks showing that early disambiguation is possible. Guthrie *et al.*[20] has demonstrated that as a nearly universal screening tool (n=20,375) M-CHAT/F has a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%, which suggests (See Table 1) that our approach produces a 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).

TABLE 2

Boosted Sensitivity, specificity and PPV at 26 months with ACoR Conditioned on M-CHAT/F Scores

M-CHAT/F Outcome				perf. (Truven)			perf. (UCM)		
0-2 NEG	3-7 NEG	3-7 POS	> 8 POS	speci-ficity	sensi-tivity	PPV	speci-ficity	sensi-tivity	PPV
specificity choices									
0.48	0.87	0.97	0.99	0.98	0.432	0.331	0.98	0.355	0.289
0.38	0.54	0.94	0.98	0.95	0.736	0.203	0.95	0.628	0.178

Ultimately, depending on Aim 2, we would attempt to combine ACoR amd M-CHAT/F via a conditional choice of sensitivity/specificity trade-offs. In our preliminary studies, this boosts overall performance significantly, with a PPV $\approx 30\%$ across datasets, or a sensitivity close to or exceeding 50%, when we restrict specificities to above 95% (See Table 2).

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

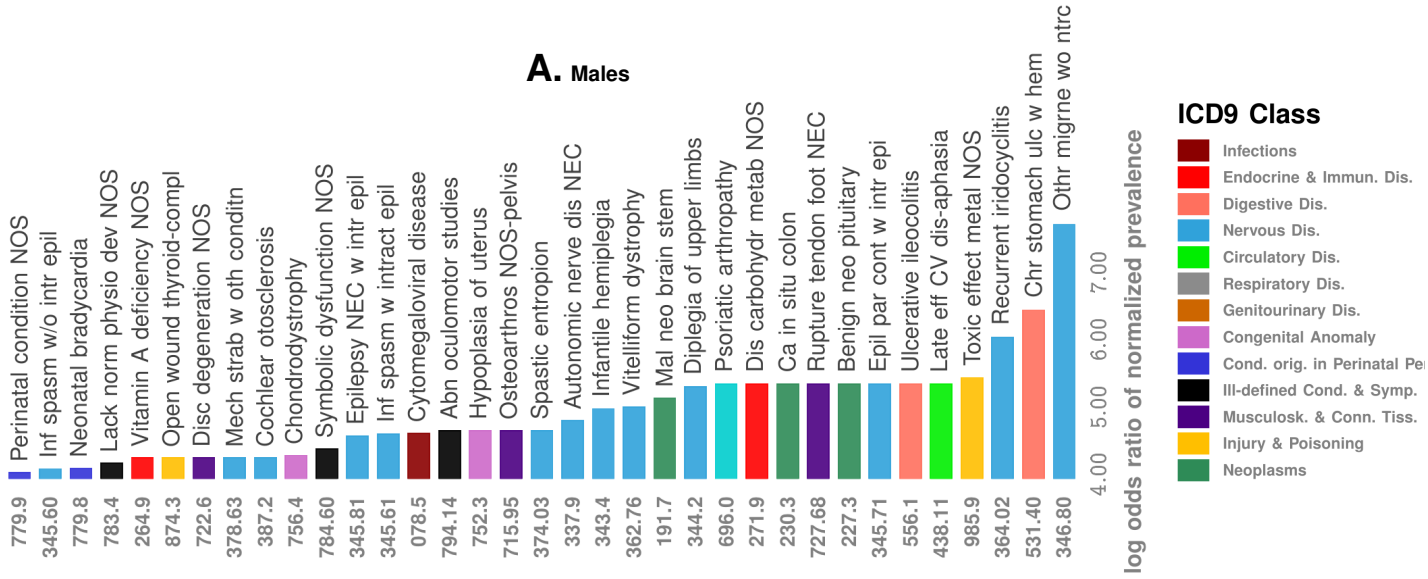

Fig. 3. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions in males. The color coding shows the disease categories of the co-morbidities.

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. 3 for males.

Disambiguation From Unrelated Psychiatric Phenotypes: In our retrospective analyses, 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.

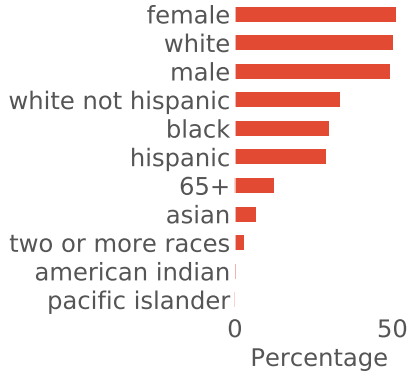

Fig. 4. Expected demographic makeup in proposed study

Sanity Checks: Uncertainty in EHR Records: 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 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: ACoR methodology can screen instantaneously every child in primary care, for whom past medical history is available, with zero administrative and resource burden. To achieve the specific aims, we will gather data from both the child and the primary caregiver in the participating primary care clinic. The key

steps are as follows (See Fig. 5):

□ 1. Pediatric clinic team will administer M-CHAT/F to incoming children with 16-26 months.

□ 2. The PI's team will compute individual ACoR with consent (steps 1 and 2 in Fig. 5).

□ 3. On being flagged by M-CHAT/F as high risk, or if the ACoR score indicates high risk and M-CHAT/F is borderline, the patients will be scheduled for ADOS-2 evaluation overseen by Dr. Smith and his team (Step 3 in in Fig. 5).

□ 4. The evaluation scores will be analyzed by PI and his team.

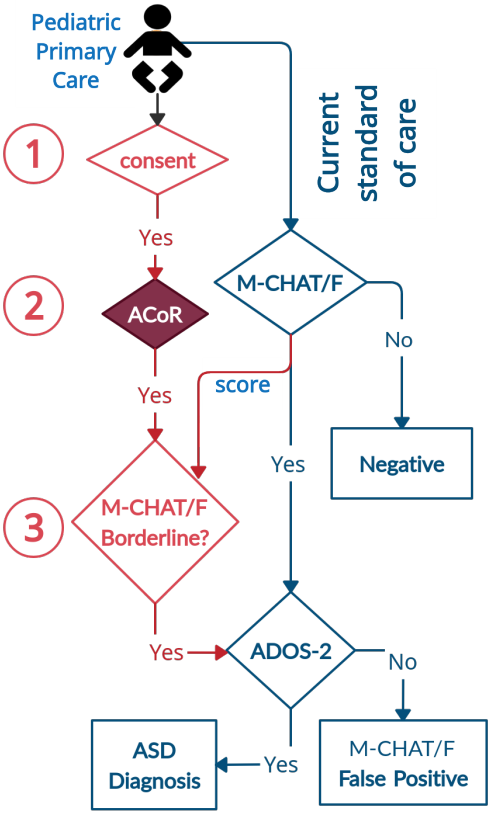

Fig. 5. Patient processing logic

The limited scope of this project implies that we need to be careful about the number of ADOS-2 referral generated due to ACoR, particularly since ADOS-2 evaluations involve significant resource and cost.

Cohort Selection: Our expected cohort is diverse (See Fig. 4). Participants (16-26 months) will be approximately 5000 children per year (producing approximately 300 ADOS-2 referrals from M-CHAT/F at no cost to the project) who will be evaluated via both the MCHAT-F screening during wellness visits at the 1 year, 1.5 year and 2 year mark, and the ACoR algorithm applied to their diagnostic history on file. Additional inclusion criteria: Child has diagnostic history on record with at least 5 diagnostic codes, and the first code is at least from 15 weeks in the past. Additional exclusion criteria: Diagnostic history only consists of health service contact codes.

Beyond the evaluation of ≈ 300 children as a part of standard clinical workflow, we will evaluate $100 - 120$ children at no cost to the patient family, to evaluate the efficacy of ACoR when M-CHAT/F is borderline (and does not trigger downstream diagnostic evaluation by itself).

3.4. Study Interventions: No intervention is planned. Outcomes are efficacy and applicability of ACoR.

Risk To Patients: The design of the study guarantees that patients suffer no negative impact from the added ACoR screen. Indeed, pursuant to available resources, patients might be expedited for ADOS evaluation which reduces 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. There is a small possibility that ACoR, due to its own false positives different from that of M-CHAT/F, might schedule some children for ADOS, who do not have autism, and might cause some stress in parents and families if they are . 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 access to their past medical history for carrying out the ACoR screen. If there is a flag either in M-CHAT/F or if M-CHAT/F is borderline with a flag in ACoR, the pediatrician will inform parents of a potential elevated risk of ASD, and offer to schedule for an ADOS-2 evaluation. The ADOS-2 evaluation triggered by 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 and Prof. Chattopadhyay. Demographic and clinical history data will be collected and entered into an HIPAA compliant secure databases. Monthly reports will be generated to monitor progress.