3. Research Strategy

Abbreviations Used

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

3.1. Significance: Autism spectrum disorder is a developmental disability associated with significant social 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, to 1-in-44 in 2021, with males diagnosed at nearly four times the rate of females.[13], [14] There is a current lack of consensus on whether increased awareness and recent changes in diagnostic practices[1] can fully explain this trend.[15] Nevertheless, with possibly over 1% of individuals affected worldwide,[16] ASD is a human condition with potentially serious negative impacts on individuals, families, and communities. Early detection improves outcomes,[1] is of paramount importance when designing interventions, and is aligned with priorities identified in National Advisory Mental Health Council (NAMHC) publications.

Even though ASD may be reliably diagnosed as early as the age of two,[14] children frequently remain undiagnosed until after the fourth birthday.[17] At this time, there are no standardized laboratory tests for ASD, so a careful review of behavioral history and social interactions is necessary for a clinical diagnosis.[1], [18] 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,[19] 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.[19] 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.[19], [20]

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 aim to validate a novel screening tool ACoR, which operationalizes a documented aspect of ASD symptomology in that it has a wide range of co-morbidities occurring at higher rates than in the general population.[1] ACoR can potentially address the aforementioned challenges of ASD screening, by leveraging predictive signatures of elevated risk gleaned from past medical history of individual patients alone which are available at the point-of-care, and using no questionnaires, or additional bloodwork or laboratory tests.

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.[19] 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.

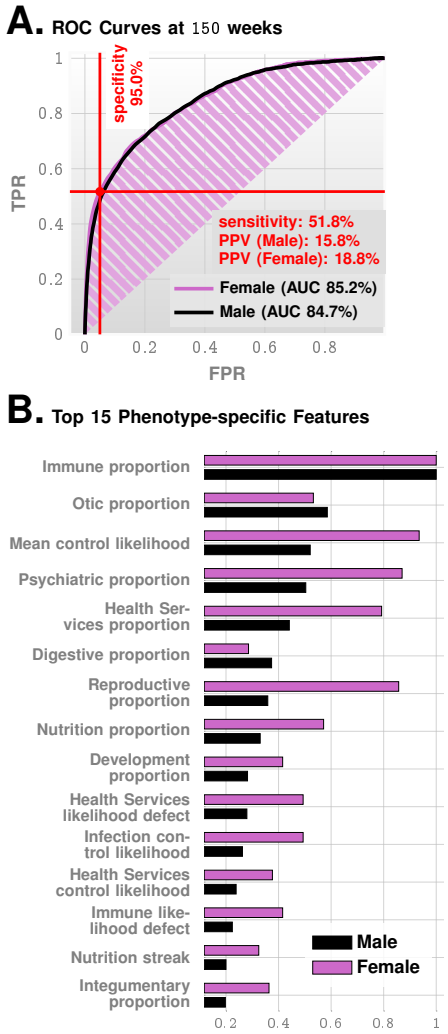

Fig. 1. Panel A: ROC curves. Panel B: feature importance inferred by our prediction pipeline. The most import feature is related to immunologic disorders, and we note that in addition to features related to individual disease categories, we also have the mean control likelihood (rank 3), which may be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

Significance of Specific Aim 2: In our retrospective preliminary studies,[8] ACoR supersedes M-CHAT/F performance[21] 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,[22], [23] efforts to identify causal biomarkers have had limited success.[24], [25] While $100-1000$ genes might modulate ASD risk,[22], [26]–[28] genetics have accounted for a limited number of cases.[29] 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.[24], [25], [30]–[35] 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. Our ML algorithm is fundamentally novel, specifically analyzing sparse, noisy categorical diagnostic sequences.

TABLE 2
ASD ICD10 codes used in training

F84	**Pervasive developmental disorders**
F84.0	**Autistic disorder**
F84.2	**Rett's syndrome**
F84.3	**Other childhood disintegrative disord.**
F84.5	**Asperger's syndrome**
F84.8	**Other pervasive develop. disorders**
F84.9	**Pervasive develop. disorder, unspec**

Sophisticated analytics to identify children at risk is of substantial interest, with progress being made by several groups.[36]–[42] However, the focus is often on analyzing questionnaires, and more recently video clips of toddler behavior. Emerging biomraker-based tools[43]–[45] are not mature enough. Additionally, the inclusion of older children and small cohort sizes in these studies is problematic. More importantly, the use of standard ML on well-established modalities focuses on mimicing the physician. In contrast, ACoR exploits under-utilized diagnostic modalities, aiming to model the disease itself, and not the physician.

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.

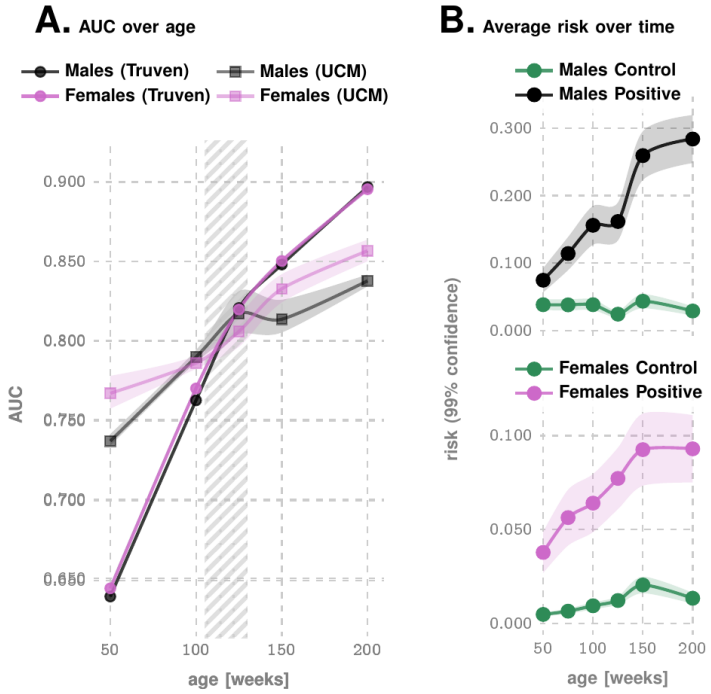

Fig. 2. ACoR performance in retrospective studies. Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets: we achieve $> 80\%$ AUC for either gender from shortly after 2 years. Panel B illustrates risk variation with time for the control and the positive cohorts.

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[46] (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).

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, n=38,012) consisting of de-identified diagnostic records for children treated at the University of Chicago Medical Center between the years of 2006 to 2018, aided 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, See Table 2). 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 or known comorbidity with ASD.

Modeling & Prediction: The significant diversity of diagnostic codes, their typical sparsity, leads to very few consistent repeats for straightforward probability calculations, making 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.[47] 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.[48], [49] 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 the retrosspective studies.[8] Variation in these inferred models across positive and control groups quantify the divergence of comorbidity patterns with increasing risk.[8], [50] 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[51] aiming to map individual patients to a raw risk score. 50% 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. 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 dataset, our performance is comparable: 83.1% and 81.3% for males and females respectively at 125 weeks of age. Predictive performance was observed to increase with patient age (ROC curve obtained at 150 weeks shown in Fig. 1A, and AUC variation with age with 99% confidence bounds is shown in Fig. 2A), reaching close to 90% at around 4 years in the national database. 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 3. We enumerate the top 15 predictive features in Fig. 1B. We also computed the county-specific performance of ACoR, and we got nearly uniform performance across the country for both sexes.[8] We find that while the AUC gradients are slightly different in the two datasets are comparable.

TABLE 3
Standalone ACoR performance (M-CHAT/F: sensitivity=38.8%,specificity=95%, PPV=14.6%)

spec.	sens.	PPV	sex	dataset
0.93	0.39	0.16	F	UCM
0.95	0.39	0.20	M	UCM
0.96	0.39	0.22	F	Truven
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.*[21] has demonstrated that as a nearly universal screening tool (n=20,375) M-CHAT/F (between 18-26 months) has a sensitivity of 38.8%, specificity of 94.9% and PPV of 14.6%, which suggests (See Table 3) 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 ($\approx$ 112 weeks).

TABLE 4
ACoR Performance at 26 months Conditioned on M-CHAT/F

M-CHAT/F Outcome				perf. (Truven)			perf. (UCM)		
0-2 NEG	3-7 NEG	3-7 POS	>8 POS	specificity	sensitivity	PPV	specificity	sensitivity	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

In Aim 2, we aim to combine ACoR amd M-CHAT/F via a conditional choice of sensitivity/specificity trade-offs. In our preliminary studies, we use Guthrie *et al.*[21]'s estimate of the population distribution of M-CHAT/F scores to estimate this conditional trade-off, and find 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 4).

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. 3 for males.

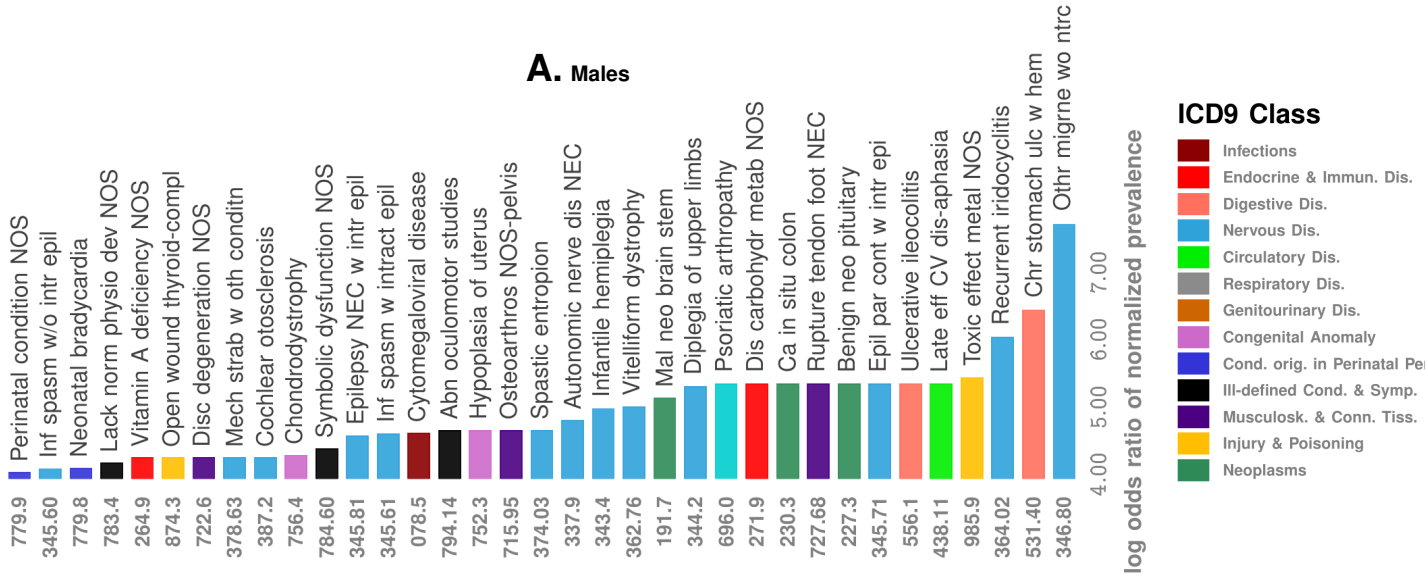

Fig. 3. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions in males. The color coding shows the disease categories of the co-morbidities.

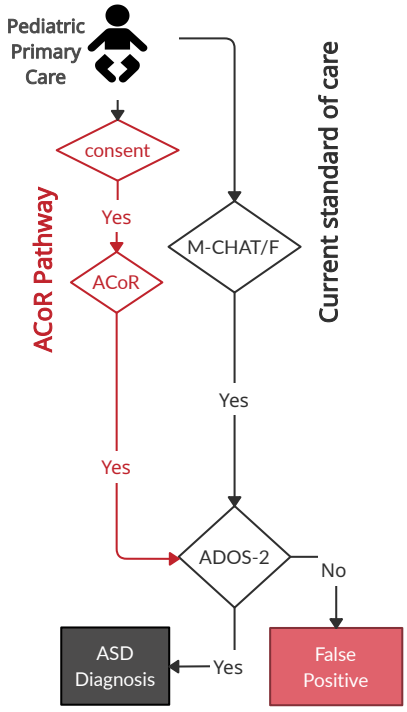

Fig. 4. Patient processing logic: Note we have parallel pathways through M-CHAT/F and ACoR, and a flag in either triggers the ADOS-2 evaluation.

Disambiguation From Unrelated Psychiatric Phenotypes: In our retrospective analyses, we investigated if we can discriminate between ASD and other unrelated psychiatric phenotypes, 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,[8] which establishes that our pipeline is indeed largely specific to ASD.

Addressing 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 preliminary 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: To achieve the specific aims, we will gather data from both the child and the primary caregiver (standard care) in the participating primary care clinic. Eligible patients at the Department of Pediatrics, University of Chicago (patients who present for a well-child visit between 16-26 months 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. 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. The sequential steps in the study (Fig. 4) are as follows: 1) Pediatric clinic team (led by Dr. Mitchell) will administer M-CHAT/F to incoming children with 16-26 months and procure consent, 2) The PI's team will compute individual ACoR, 3) If flagged by either M-CHAT/F or ACoR as high risk for ASD, the patients will be scheduled for ADOS-2 evaluation overseen by Dr. Smith, Dr. Msall and supporting team, and finally,

4) The evaluation scores will be analyzed by PI and his team. The feasibility of proposed design has been validated with the same study team under a modest intramural pilot grant, with low sample size (Of 5 ACoR flags produced in this pilot, all were diagnosed with ASD by ADOS-2).

TABLE 5
Expected population demographics

Sex	
Male	53.6%
Female	46.6%
Race	
Black/African-Am	73.2%
White	18.2%
Asian	3%
Ethnicity	
Hispanic	7.9%
Non-Hispanic	91.8%

Cohort Selection: Inclusion and Exclusion Criteria: We are aiming to validate a universal screening protocol for ASD risk. As such, we do not plan to select patients based on gender or demographic criteria, and plan to carry out ongoing recruitment at the pediatric clinic involved in this study throughout the study timeline, as long as the patients are within the target age bracket of 16-26 months. Nevertheless, the population from which our patients will be drawn is diverse (See Fig. 5), with greater than 70% of the patients being non-Caucasian. The split between male and females is expected to be approximately even (54% males to 46% females), as estimated from past patient population characteristics (in the age group of interest) treated at the University of Chicago Medical center. Estimating from the number of well-child visits at the pediatric clinic involved in this study, we estimate that our sample size will be approximately 3000. Hence our inclusion criteria is: age within 16-26 months, and our exclusion criteria are: 1) children who already have a ASD diagnsosis will be excluded. 2) children with less than 5 diganostic codes in their medical history.

Sample Size for ADOS-2 Evaluations: ACoR screening conditioned on M-CHAT/F is expected to have a sensitivity $> 70\%$ and a PPV $> 20\%$, implying that we expect to have prevalence x 0.7 x 3000 x (1/0.2) flags at 95% specificity. Assuming a 10% prevalence (consenting parents might have observed some concerning developmental issues, and thus bias the sample from population prevalence of ASD), we determine that we would need to do about 300 ADOS-2 evaluations per year. We would also investigate higher sensitivity operation for ACoR at 90% (since ACoR allows for a choice of a range of operating points on the ROC curve), where up to 500 ADOS-2 evaluations per year would suffice. From estimated confidence bounds on ACoR performance, this sample size estimates are sufficient to achieve greater than 80% power at 5% signifiance.

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

Risk To Patients: Since no intervention is planned, and outcomes are efficacy and applicability of ACoR, patients are expected to suffer limited negative impact from the additional screen. For some borderline cases might be flagged due to false positives, which may be viewed as a risk. However, the potential benefits, including significant reduction in false positives, outweighs the temporary anxiety on part of the care-givers on a false positive flag. It is conceivable that a false-positive flag gets a positive diagnostic evaluation in the subsequent ADOS-2 evaluation, and actually gets a clinical diagnosis of ASD. This is unlikely, but is theoretically a risk to patients.

The upper bound on the likelihood of a false flag from ACoR, where the M-CHAT/F is negative is estimated to be $\sim$7% of the total study sample. Given that ADOS evaluation has been reported in some studies to have about $\sim$15% false positive rate, the risk of a positive diagnosis from ADOS-2 arising triggered by the false flag from ACoR expected to be $\sim$1% of the total sample size. However, the ADOS-2 evaluation results will be analyzed by Drs. Smith and Msall (clinical evaluation by experienced clinicians decreases false positive rate) to minimize the odds of such false positives being recorded as a clinical diagnosis of ASD for a patient. Thus the actual odds of such an event is expected to be very small.

Procedures: The ADOS-2 evaluations triggered by ACoR will be at no cost to the patient. Patients who are flagged by M-CHAT/F alone will follow standard care, and will not be charged to the project. All study procedures and consent forms will be approved by the University of Chicago Institutional Review Board (IRB). 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 and Resource 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. De-identified data will be deposited to NDA following data harmonization guidelines, and will also be made available in Zenodo.