

PROJECT SUMMARY

Autism spectrum disorder (ASD) is a developmental disability associated with significant social and behavioral challenges, and there is a distinct need for tools that help identify children with ASD as early as possible. To that effect, we introduce and propose to validate the ASD Co-morbid Risk (ACoR) score in a limited clinical study. The ACoR is computed via sophisticated pattern discovery on longitudinal history of diagnostic codes for individual patients, and potentially signals a future ASD diagnosis within 16-26 months of age. Computation of ACoR requires no new blood-work, or questionnaires, and uses data already available on patient file, with no demand for any particular test or demographic information. Thus, ACoR is positioned to be a universal screening tool, that can estimate the risk of autism for all children in a pediatric facility near-instantaneously, potentially outperforming existing tools. Despite being highly heritable, our current incomplete understanding of ASD pathogenesis and the lack of reliable biomarkers hampers early detection, intervention and patient outcomes. The currently available questionnaire based screening tools suffer from vast number of false positives which create long wait-times for diagnostic evaluations. Additionally, standardized checklists are vulnerable to socio-economic and interpretational biases that disproportionately impact diagnosis in diverse communities. Borderline cases with children with average to above average cognitive abilities might be left undiagnosed till start of school, which negatively impact effectiveness of interventions. The ACoR score is designed to address the aforementioned complicated challenges of ASD screening by distilling incipient patterns predicting elevated risk from past medical history of individual patients. Thus to compute ACoR, 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. In the setting of a pediatric primary care clinic at the Department of Pediatrics, University of Chicago, we plan to carry out a comparative study of ACoR with M-CHAT/F, which is the most common screening tool in current use. Via a direct comparison, we specifically aim to 1) estimate prospectively to what extent we can reduce false positives, 2) the possibility of combining the scores for significant improvements in either Positive Predictive Value or the sensitivity while not losing specificity, 3) the superior performance in ethnically and demographically diverse cohorts, and 4) shed light on the ASD pathobiology by classifying patterns of co-morbidities that map to distinct presentations. We have extensively validated our results in retrospective studies on two independent databases of patient records with over four million children. These results indicate superior performance to existing tools, achieving out-of-sample AUC exceeding 80% for either sex from just over 2 years of age. Unlike standard machine learning applications, ACoR represents a novel screening modality for ASD, functionaly independent of questionnaires, and potentially can address documented language, cultural and social barriers of the existing tools.

PROJECT NARRATIVE

In contrast to the current questionnaire based screening tools for autism spectrum disorders (ASD) that suffer from vast amounts of false positives, and a host of demographic, socio-economic and interpretative biases, we aim to validate the ASD Co-morbid Risk (ACoR) score, that estimates ASD risk via sophisticated pattern discovery on the longitudinal medical history of individual patients. Computation of ACoR requires no new blood-work, laboratory tests, questionnaires or psychiatric/cognitive consults, and may be carried out purely from the history of past medical encounters at no additional administrative burden or resource utilization. ACoR outperforms the current tool M-CHAT/F in preliminary studies, and on account of functional independence, the two scores may be combined to further boost performance to either boost positive predictive value up to 100% or sensitivity up to 50% with no loss in current specificity.

FACILITIES AND OTHER RESOURCES

ACADEMIC PEDIATRICS AND DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS

University of Chicago Comer Children's Hospital is a 172-bed acute care facility founded in 2005, uniting advanced technology with a family-centered, child-friendly philosophy to provide state-of-the-art care in a six-floor, 242,000-square-feet building. As a major tertiary referral center, Comer Children's sees children with common as well as the most complex medical problems. It admits about 5,000 patients annually from the Chicago area, the Midwest, and around the world. Each year its outpatient clinics accommodate nearly 37,000 general pediatric and specialty visits. Comer Children's is staffed by close to 170 highly trained pediatricians, as well as specially trained nurses and caring support staff, who work together to provide general and specialty medical care for newborns to young adults.

SECTION OF ACADEMIC PEDIATRICS: The University of Chicago is situated on Chicago's South Side, a medically underserved area. Our Primary Service area comprises a population of three-quarter of a million people. There is an acute need for primary care practitioners to improve health in the state of Illinois (FY2025 projected shortage -4.0% and specifically among this population (Chicago South Side Primary Care HPSA score =19). Physicians in this area deliver care that addresses complex medical, psychosocial, and behavioral needs. They lead efficient, patient-centered primary care practices in underserved communities.

The Section of Academic Pediatrics is dedicated to managing the unique primary care needs of newborns, children and adolescents. The scope of its clinical service covers both the outpatient and inpatient areas located in Comer Children's Hospital and community sites. Its clinical programs include Comer and community hospital medicine, general care nursery, pediatric primary care, Sports Medicine, Child Advocacy and Protective Services and the Comer Mobile Unit. To support this grant, we will leverage our academic institution and its expanding clinical network across Chicago's South Side, as well as neighboring federally qualified health centers, to provide diverse and complementary experiences in team-based care. Dr. James Mitchell is the Medical Director for patient services.

Our provider network includes community health centers and Federally Qualified Health Centers (FQHCs). Sites will include: Comer Children's Hospital General Pediatrics Clinic, Friend Family Health Center, Erie Family Health Center, ACCESS Community Health Network, and Asian Human Services. All sites have Health Professional Shortage Area designation with scores of 19-21. Some of patient volume statistics are provided as follows:

Comer Children's Hospital General Pediatrics Clinic, located on the University of Chicago Medical campus, serves approximately 12,000 pediatric patients over 22,000 visits annually. Approximately 56

The Friend Family Health Center is a large, multi-site Federally Qualified Health Center. It has served as the primary ambulatory training site for University of Chicago Pediatrics Residency Program since the 1990s. Friend Health serves over 27,000 patients from Chicago's South Side as part of its mission to provide access to high quality, comprehensive health care

The Erie Family Health Center, a Federally Qualified Health Center with seven primary care centers and five school-based health centers, provides care to more than 74,000 patients over 300,000 visits annually. These health centers cares for patients across 62 languages. The patient population is 71

ACCESS Community Health Network, a large Federally Qualified Health Center with 35 health centers across Chicago, cares for more than 183,000 patients per year. ACCESS services the largest proportion of Medicaid beneficiaries in Illinois.

Asian Human Services (AHS) is a multi-site Federally Qualified Health Center which services a largely immigrant population. AHS provides comprehensive care to more than 30,000 people from more than 55 countries.

THE SECTION OF DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS (DBP): The Section of DBP is a Center of Excellence in developmental diagnosis, biomedical management and family supports for children with motor, communicative, sensory, developmental, genetic, neurological, learning and behavior disorders. The section's goals are to:

- Promote the highest quality interdisciplinary assessment and biopsychosocial management practices to optimize child functioning, support families and maximize prevention strategies across health, education

and community care systems.

- Provide the necessary leadership to ensure that children with complex challenges have a high quality medical home and that best practices are used to improve their ability to communicate, move, regulate behavior, interact socially, learn functional and adaptive skills, perform in school and participate in the community.
- Serve as a resource for primary care practitioners and pediatric subspecialists for children with the highest biomedical and psychosocial risks associated with suboptimal educational outcomes. •
- Serve as a resource for community professionals and agencies for children with developmental delays, complex disorders after prematurity, genetic, neurological, or cardiopulmonary disorders or experiencing complex behavioral challenges after life-threatening illnesses. .
- Enhance training and research so families will benefit from the best clinical and scientific advances with the highest standards of ethics, professionalism and advocacy.
- For children with complex medical and/or behavioral challenges and for families with psycho-social or economic stress, DBP provides interventions and assistance. Many of these children and families require teamwork and leadership to promote their health, development, and social competencies. DBP physicians take into account the entire family dynamic and provide guidance in securing ancillary therapies, support services and special school considerations as needed.
- We have the clinical expertise coupled with research, educational; advocacy and public policy activities needed for insuring that children receiving translational technologies have information and supports for long term success. Our scope and leadership roles are recognized at the local, regional, national and international levels. We have established clinical, advocacy and research cooperative activities with specialists in community pediatrics, child and adolescent psychiatry, neonatology, critical care medicine, genetics, neurology, child protective services, developmental psychology, public policy and social sciences. These collaborations are not limited to the University of Chicago campus, but extend to other institutions nationally and internationally. Our activities are complementary to those of our collaborators—bringing cross-fertilization of ideas to bear on topics of life course outcomes of children with prematurity. Our Section's activities are interdisciplinary and include faculty from Public Policy, Social Sciences and Economics to work towards our common interests in improving the quality of life for children with disabilities.

The Woodlawn Center is the clinical home of DBP. It is a site for clinical evaluation and research that can accommodate patients 6 days a week with two exam rooms, equipped to provide developmental assessments from 0-18 years of age. The focus at the Woodlawn Center is on early developmental and cognitive assessments that provide early definition of school-age diagnosis and both clinical and educational needs.

Interdisciplinary professionals at this center include speech and occupational therapists, social work, developmental psychology, and community outreach professionals. We provide Autism Diagnostic Observation Schedule (ADOS) testing for accurately assessing and diagnosing autism spectrum disorders. ADOS is a standardized diagnostic test which provides information on communication, social interaction, and play (or imaginative use of materials) for individuals suspected of having autism or other pervasive developmental disorders, and, as appropriate, provide access to services which may be recommended for each child. .

DBP and the **Pediatric Neuropsychology Service** both maintain up to date assessment protocols and manuals for the interdisciplinary assessment of children's motor, manipulative, conceptual, cognitive, behavioral, and educational skills. These include all the assessment instruments specified in the neurodevelopmental assessment protocol for children in this study. These assessment tools have been used by health professionals representing developmental pediatrics, developmental psychology, physical therapy, occupational therapy, speech language pathology, and special education. A specialized library of current assessment references as well as audiovisual and computer assisted training materials are available. The Pediatric Neuropsychology Service is a clinical, training and research program within the Section of Child and Adolescent Psychiatry, in the Department of Psychiatry. The Service provides assessment and Diagnostic services for infants and children with neurodevelopmental learning, emotional regulation and medical disorders and conditions. The service is also involved in collaborative clinical research, examining the neurocognitive and behavioral sequelae of a wide range of disorders.

COMPUTATIONAL FACILITIES

The principal investigators have access to extensive computational facilities available at the University of Chicago to carry out the tasks described.

Access to Clinical Data for AI-enabled Analytics: The ZeD lab (overseen by Professor Chattopadhyay) is housed within the Department of Medicine at the University of Chicago, and has access to the full range of high end computing resources offered by the University of Chicago. In addition, Prof. Chattpadyay's laboratory has access to the HIPAA compliant clinical data warehouse maintained by the Biological Sciences Division as detailed below:

The Clinical Research Data Warehouse: (CRDW) within the Biomedical Sciences Division of the University of Chicago is one of the deepest, richest, and most research-ready data repositories of its kind. Containing more than a decade of University of Chicago medical data, it seamlessly brings together multiple internal and external data sources to provide researchers with access to more than 12 million encounters for 2.3 million patients. The associated diagnoses, labs, medications, and procedures number in the tens of millions each. The CRDW is run on IBM Netezza Pure Data System for Analytics servers, a patented Asymmetric Massively Parallel Processing architecture designed to deliver exceptional query performance and modular scalability on highly complex mixed workloads.

In order to meet the acute need for data related to COVID-19, the CRDW team has constructed three data marts (de-identified, limited, and identified) to provide the most commonly requested data elements for this patient population. The initial instance of the COVID-19 data mart includes de-identified structured data on patient demographics, encounters, diagnoses, labs, medications, flow sheets, and procedures. Additional data will be added based on resource availability and urgency.

Cohort Discovery Tool: The purpose of this tool (SEE Cohorts) is to provide a secure web-based tool for the initial exploration of de-identified data. It allows researchers to search available data, build a cohort of patients, and view actual de-identified data within the interface. The data in SEE Cohorts is refreshed weekly.

Research Computing Center: The University of Chicago Research Computing Center (RCC) provides high-end research computing resources to researchers at the University of Chicago, which include high-performance computing and visualization resources; high-capacity storage and backup; software; high-speed networking; and hosted data sets. Resources are centrally managed by RCC staff who ensure the accessibility, reliability, and security of the compute and storage systems. A high-throughput network connects the Midway Compute Cluster to the UChicago campus network and the public internet through a number of high-bandwidth uplinks. To support data-driven research RCC hosts a number of large datasets to be accessed within the RCC compute environment.

RCC maintains three pools of servers for distributed high-performance computing. Ideal for tightly coupled parallel calculations, tightly-coupled nodes are linked by a fully non-blocking FDR-10 Infiniband interconnect. Loosely-coupled nodes are similar to the tightly-coupled nodes, but are connected with GigE rather than Infiniband and are best suited for high-throughput jobs. Finally, shared memory nodes contain much larger main memories (up to 1 TB) and are ideal for memory-bound computations. The types of CPU architectures RCC maintains are tabulated in Table 1.

RCC also maintains a number of specialty nodes:

- *Large shared memory nodes* - up to 1 TB of memory per node with either 16 or 32 Intel CPU cores. Midway is always expanding, but at time of writing RCC contains a total of 13,500 cores across 792 nodes, and 1.5 PB of storage.
- *Hadoop*: Originally developed at Google, Hadoop is a framework for large-scale data processing.
- *GPU Computing*: Scientific computing on graphics cards can unlock even greater amounts of parallelism from code. RCC GPU nodes each include two Nvidia Tesla-class accelerator cards and are integrated in the Infiniband network. RCC currently provides access to Fermi-generation M2090 GPU devices and Kepler-generation K20 and K40 devices.
- *Xeon Phi*: The Many Integrated-Core architecture (MIC) is Intel's newest approach to manycore computing. Researchers can experiment with these accelerators by using MIC nodes, each of which have two Xeon Phi cards, and are integrated into the Infiniband network.

Persistent and High-Capacity Storage. Storage is accessible from all compute nodes on Midway1 and Midway2 as well as outside of the RCC compute environment through various mechanisms, such as mounting directories as network drives on your personal computer or accessing data as a Globus Online endpoint (at the time of this writing, Globus Online is supported on Midway1). RCC takes snapshots of all home directories (users' private storage space) at regular intervals so that if any data is lost or corrupted, it can

TABLE 1
University of Chicago Research Computing Center Capabilities Summary

Cluster	Partition	Compute cores (CPUs)	Memory	Other configuration details
midway1	westmere	12 x Intel X5675 3.07 GHz	24 GB	
	sandyb	16 x Intel E5-2670 2.6GHz	32 GB	
	bigmem	16 x Intel E5-2670 2.6GHz	256 GB	
		32 x Intel E7-8837 2.67GHz	1 TB	
	gpu	16 x Intel E5-2670 2.6GHz	32 GB	2 x Nvidia M2090 or K20 GPU
		20 x Intel E5-2680v2 2.8GHz	64 GB	2 x Nvidia K40 GPU
	mic	16 x Intel E5-2670 2.6GHz	32 GB	2 x Intel Xeon Phi 5100 coprocessor
	amd	64 x AMD Opteron 6386 SE	256 GB	
	ivyb	20 x Intel E5-2680v2 2.8GHz	64 GB	
	broadwl	28 x Intel E5-2680v4 2.4GHz	64 GB	
midway2	bigmem2	28 x Intel E5-2680v4 @ 2.4 GHz	512 GB	
		28 x Intel E5-2680v4 @ 2.4 GHz	64 GB	4 x Nvidia K80 GPU

easily be recovered. RCC maintains GPFS Filesystem Snapshots for quick and easy data recovery. In the event of catastrophic storage failure, archival tape backups can be used to recover data from persistent storage locations on Midway. Automated snapshots of the home and project directories are available in case of accidental file deletion or other problems. Currently snapshots are available for these time periods: 1) 7 daily snapshots, 2) 4 weekly snapshots.

Tape Backups. Backups are performed on a nightly basis to a tape machine located in a different data center than the main storage system. These backups are meant to safeguard against events such as hardware failure or disasters that could result in the complete loss of RCC's primary data center.

Data Sharing. All data in RCC's storage environment is accessible through a wide range of tools and protocols. Because RCC provides centralized infrastructure, all resources are accessible by multiple users simultaneously, which makes RCC's storage system ideal for sharing data among your research group members. Additionally, data access and restriction levels can be put in place on an extremely granular level.

Data Security & Management. The HIPAA compliant security of the Research Computing Center's storage infrastructure, protected by two-factor authentication, gives users peace of mind that their data is stored, managed, and protected by HPC professionals. Midway's file management system allows researchers to control access to their data. RCC has the ability to develop data access portals for different labs and groups.

1. COVER PAGE

2. SPECIFIC AIMS

Early diagnosis of Autism Spectrum Disorder (ASD) and timely intervention is widely recognized as critical for achieving improved cognitive, behavioral and social outcomes.^[1] Despite a growing list of suspected risk factors,^{[2]-[5]} the etiology of Autism is still unclear. Even with increasingly widespread adoption of screening with standardized checklists at 18 and 24 months, the median age of diagnosis for ASD remains at over 4 years. Starting with a positive initial screen, a clinical diagnosis of ASD is a frustrating multi-step process spanning 3 months to 1 year, often delaying time-critical interventions. One obvious driver for these delays is the vast number of false positives encountered in the current initial screening. For example, the M-CHAT/F, the most widely used screener,^{[1], [6]} produces over 85 false positives out of every 100 flagged for diagnostic evaluation, significantly inflating wait times^[5] especially in rural and underserved communities. Further, current screening tools are sensitive to language barriers and cultural issues, and are particularly ineffective for children with milder symptoms with average or above-average cognitive abilities until about school age,^{[1], [7]} often due to a “wait and see” approach adopted at the primary care. The need for better screening tools is thus paramount.

In this study, we plan to develop and validate the efficacy of machine inferred **digital biomarkers** for autism, mined automatically from past medical encounters. Using individual diagnostic codes already present in individual patient files, we plan to engineer a reliable risk estimator (ASD Co-morbid Risk: ACoR) enabled by novel machine learning algorithms, and compare its effectiveness against existing tools such as the M-CHAT/F in a limited clinical study, with children between 16-26 months of age. Our rationale is informed by the extensively documented comorbidities of ASD ranging from dysregulation of immune pathways such as eczema, allergies, asthma, as well as ear and respiratory infections, gastrointestinal problems, developmental issues, severe headaches, migraines, and seizures.^{[8], [9]} While ASD presentation is highly variable, sophisticated pattern recognition on the longitudinal history of diagnostic codes is expected to reveal uncharted associations that allow precise screening for at-risk patients. Orthogonal to questionnaire based detection of behavioral signals, the proposed tool potentially reduces socio-economic, ethnic and demographic biases to elicit more objective and stable results — with zero administrative burden on clinicians and parents. With a team comprising machine learning experts, pediatric clinicians and developmental experts with extensive history of research and service in the field, the ACoR score can significantly improve outcomes by either boosting sensitivity of current screening or slashing false positives by half. Thus, the principal aims this study are the following:

- **Aim 1: Reduce false positives in current screening protocols.** The current high false positive rate exacerbates post screen wait-time, increasing diagnostic delay. To evaluate the *hypothesis: ACoR reduces up to 50% of false positives*, we will track the cases in which MCHAT-F triggers a flag, but ACoR does not. Our aim to evaluate the positive predictive value (PPV) of ACoR, under high specificity conditions (> 95%). Additionally, evaluate if ACoR replicates high sensitivity observed in preliminary studies without losing specificity.
- **Aim 2. Evaluate the statistical relationship between the ACoR score and M/CHAT-F, and formalize a joint or conditional operational protocol.** We will characterize statistical association, if any, between the test scores. *Hypothesis: The uncertainties or errors in the two tests are statistically independent.* Additionally, we will evaluate our ability to boost performance by conditioning the sensitivity-specificity trade-offs on the M-CHAT/F score of individual patients.
- **Aim 3. Evaluate the effectiveness of ACoR in a demographically diverse population with a range of socio-economic confounders.** *Hypothesis: A questionnaire-free approach has the potential to mitigate biases that arise from limitation of language, cultural barriers, and demographic diversity, e.g. disproportionately failing to diagnose children with average to above-average intelligence in diverse populations,^[10] and under-reporting of symptoms by parents or primary care-givers due to cultural differences.^[11]*
- **Aim 4. Characterize heterogeneity of ASD presentation by relating it to patterns in medical history, and predictive co-morbidities.** Heterogeneous presentation is a key barrier in the mechanistic understanding of ASD pathobiology. *Hypothesis: We can characterize the distinct classes and/or hierarchies of co-morbidities, by leveraging our ability to disambiguate them from individual medical histories.* This will shed light into the potentially intrinsic classes of the underlying disease processes, and refine/inform intervention design.

Thus, we are proposing to exploit observed co-morbidities in children who ultimately meet the criteria for ASD to develop a risk estimation pipeline, and predict future clinical diagnosis under 2 years of age. Orthogonal to checklists, we aim to reduce the median diagnostic age for ASD, by reducing the long post-screen wait times,^[5] by significant boosts in positive predictive value, reduction in false positives, and increased sensitivities at little or no loss of specificity, and at no additional administrative burden or resource utilization.

3. RESEARCH STRATEGY

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, 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 in 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.

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

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).

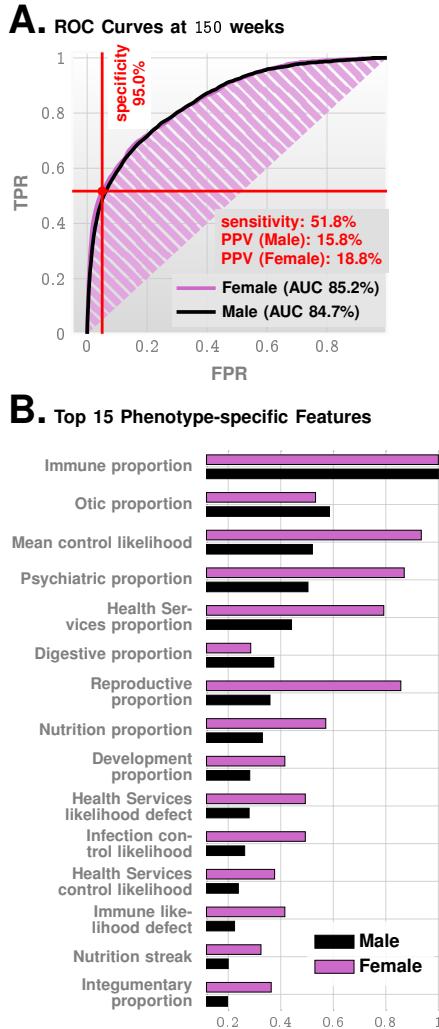


Fig. 1. Panel A: ROC curves. Panel B: feature importance inferred by our prediction pipeline. The most import feature is related to immunologic disorders, and we note that in addition to features related to individual disease categories, we also have the mean control likelihood (rank 3), which may be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

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

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 endocrinological 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).

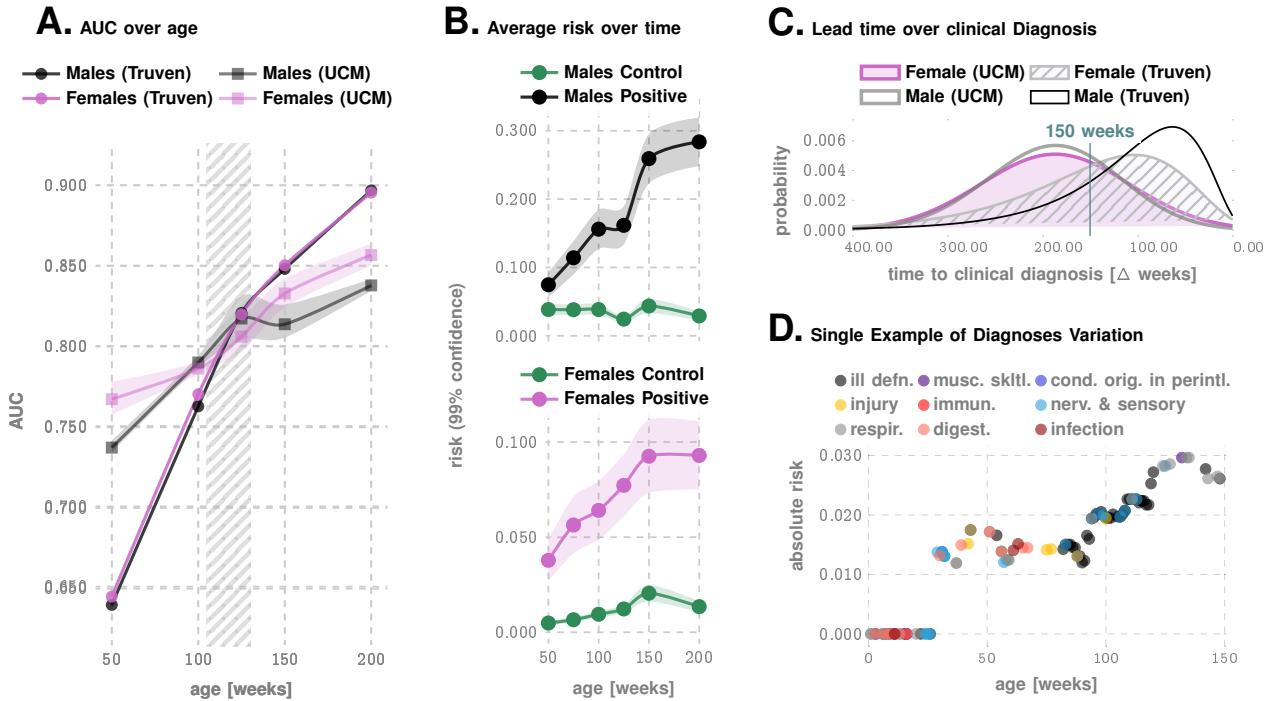


Fig. 2. Variation of Inferred Risk. Panel A illustrates AUC achieved as a function of patient age, for the Truven and UCM datasets: we achieve $> 80\%$ AUC for either gender from shortly after 2 years. Panel B illustrates risk variation with time for the control and the positive cohorts. Panel C shows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

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 2) 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)).

Ultimately, depending on Aim 2, we would attempt to combine ACoR and 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 3).

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

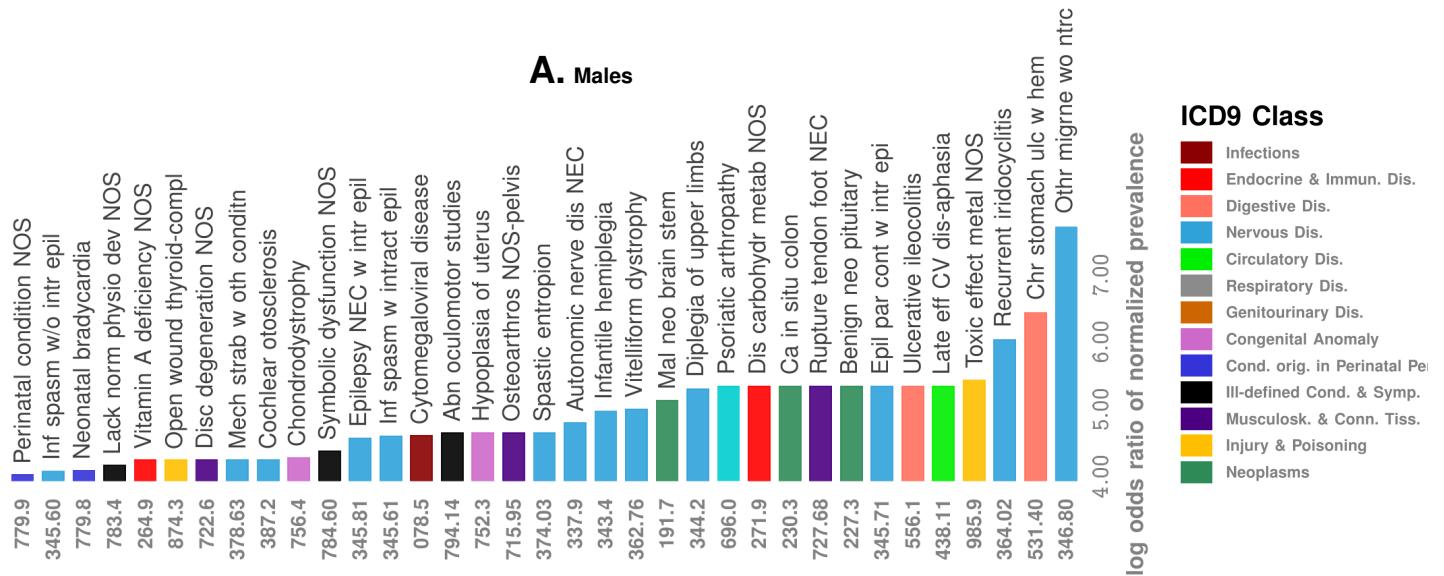


Fig. 3. Difference in occurrence frequencies of diagnostic codes between true positive (TP) and true negative (TN) predictions in males. The color coding shows the disease categories of the co-morbidities.

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.

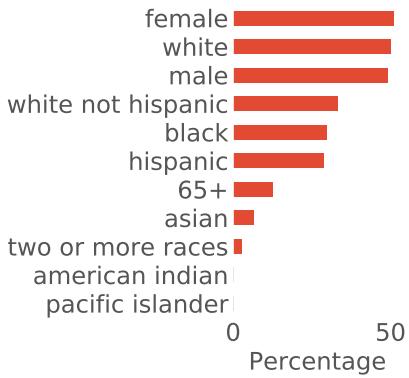


Fig. 4. Expected demographic makeup in proposed study

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 Fig. 5).
- 4. The evaluation scores will be analyzed by PI and his team.

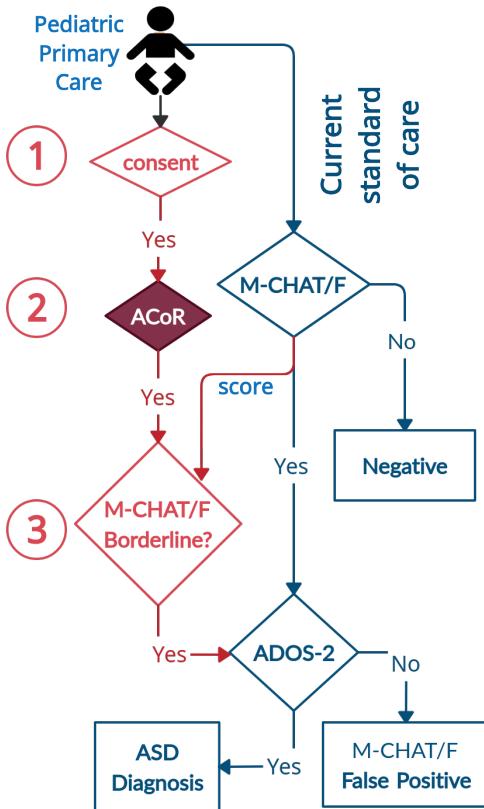


Fig. 5. Patient processing logic

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.

REFERENCES

- [1] Hyman, S. L., Levy, S. E., Myers, S. M. *et al.* Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics* **145** (2020).
- [2] Kalb, L. G. *et al.* Determinants of appointment absenteeism at an outpatient pediatric autism clinic. *Journal of Developmental & Behavioral Pediatrics* **33**, 685–697 (2012).
- [3] Bisgaier, J., Levinson, D., Cutts, D. B. & Rhodes, K. V. Access to autism evaluation appointments with developmental-behavioral and neurodevelopmental subspecialists. *Archives of pediatrics & adolescent medicine* **165**, 673–674 (2011).
- [4] Fenikilé, T. S., Ellerbeck, K., Filippi, M. K. & Daley, C. M. Barriers to autism screening in family medicine practice: a qualitative study. *Primary health care research & development* **16**, 356–366 (2015).
- [5] Gordon-Lipkin, E., Foster, J. & Peacock, G. Whittling Down the Wait Time: Exploring Models to Minimize the Delay from Initial Concern to Diagnosis and Treatment of Autism Spectrum Disorder. *Pediatr. Clin. North Am.* **63**, 851–859 (2016).
- [6] Robins, D. L. *et al.* Validation of the modified checklist for autism in toddlers, revised with follow-up (m-chat-r/f). *Pediatrics* **133**, 37–45 (2014).
- [7] Jashar, D. T., Brennan, L. A., Barton, M. L. & Fein, D. Cognitive and adaptive skills in toddlers who meet criteria for autism in dsm-iv but not dsm-5. *Journal of autism and developmental disorders* **46**, 3667–3677 (2016).
- [8] Tye, C., Runicles, A. K., Whitehouse, A. J. O. & Alvares, G. A. Characterizing the Interplay Between Autism Spectrum Disorder and Comorbid Medical Conditions: An Integrative Review. *Front Psychiatry* **9**, 751 (2018).
- [9] Kohane, I. S. *et al.* The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS ONE* **7**, e33224 (2012).
- [10] Christensen, D. L. *et al.* Prevalence and characteristics of autism spectrum disorder among children aged 4 years—early autism and developmental disabilities monitoring network, seven sites, united states, 2010, 2012, and 2014. *MMWR Surveillance Summaries* **68**, 1 (2019).
- [11] Burkett, K., Morris, E., Manning-Courtney, P., Anthony, J. & Shambley-Ebron, D. African american families on autism diagnosis and treatment: The influence of culture. *Journal of Autism and Developmental Disorders* **45**, 3244–3254 (2015).
- [12] Baio, J. *et al.* Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR Surveill Summ* **67**, 1–23 (2018).
- [13] Data & statistics on autism spectrum disorder — cdc (2019). URL <https://www.cdc.gov/hcbddd/autism/data.html>.
- [14] King, M. & Bearman, P. Diagnostic change and the increased prevalence of autism. *Int J Epidemiol* **38**, 1224–1234 (2009).
- [15] Elsabbagh, M. *et al.* Global prevalence of autism and other pervasive developmental disorders. *Autism Res* **5**, 160–179 (2012).
- [16] Schieve, L. A. *et al.* Population attributable fractions for three perinatal risk factors for autism spectrum disorders, 2002 and 2008 autism and developmental disabilities monitoring network. *Ann Epidemiol* **24**, 260–266 (2014).
- [17] Volkmar, F. *et al.* Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* **53**, 237–257 (2014).
- [18] Gordon-Lipkin, E., Foster, J. & Peacock, G. Whittling down the wait time: exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. *Pediatric Clinics* **63**, 851–859 (2016).
- [19] Althouse, L. A. & Stockman, J. A. Pediatric workforce: A look at pediatric nephrology data from the american board of pediatrics. *The Journal of pediatrics* **148**, 575–576 (2006).
- [20] Guthrie, W. *et al.* Accuracy of Autism Screening in a Large Pediatric Network. *Pediatrics* **144** (2019).
- [21] Satterstrom, F. K. *et al.* Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *bioRxiv* (2019). <https://www.biorxiv.org/content/early/2019/04/24/484113.full.pdf>.
- [22] Sandin, S. *et al.* The Heritability of Autism Spectrum DisorderReassessing the Heritability of Autism Spectrum DisordersLetters. *JAMA* **318**, 1182–1184 (2017). URL <https://doi.org/10.1001/jama.2017.12141>. https://jamanetwork.com/journals/jama/articlepdf/2654804/jama_sandin_2017_Id_170037.pdf.

- [23] Ohja, K. *et al.* Neuroimmunologic and Neurotrophic Interactions in Autism Spectrum Disorders: Relationship to Neuroinflammation. *Neuromolecular Med.* **20**, 161–173 (2018).
- [24] Gadysz, D., Krzywdziska, A. & Hozyasz, K. K. Immune Abnormalities in Autism Spectrum Disorder—Could They Hold Promise for Causative Treatment? *Mol. Neurobiol.* **55**, 6387–6435 (2018).
- [25] Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
- [26] Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
- [27] Werling, D. M. The role of sex-differential biology in risk for autism spectrum disorder. *Biol Sex Differ* **7**, 58 (2016).
- [28] Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat. Rev. Genet.* **9**, 341–355 (2008).
- [29] Yamashita, Y. *et al.* Anti-inflammatory Effect of Ghrelin in Lymphoblastoid Cell Lines From Children With Autism Spectrum Disorder. *Front Psychiatry* **10**, 152 (2019).
- [30] Shen, L. *et al.* Proteomics Study of Peripheral Blood Mononuclear Cells (PBMCs) in Autistic Children. *Front Cell Neurosci* **13**, 105 (2019).
- [31] Theoharides, T. C., Tsilioni, I., Patel, A. B. & Doyle, R. Atopic diseases and inflammation of the brain in the pathogenesis of autism spectrum disorders. *Transl Psychiatry* **6**, e844 (2016).
- [32] Young, A. M. *et al.* From molecules to neural morphology: understanding neuroinflammation in autism spectrum condition. *Mol Autism* **7**, 9 (2016).
- [33] Croen, L. A. *et al.* Family history of immune conditions and autism spectrum and developmental disorders: Findings from the study to explore early development. *Autism Res* **12**, 123–135 (2019).
- [34] Zerbo, O. *et al.* Immune mediated conditions in autism spectrum disorders. *Brain Behav. Immun.* **46**, 232–236 (2015).
- [35] Hyde, K. K. *et al.* Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders* **6**, 128–146 (2019).
- [36] Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E. & Wall, D. P. Multi-modular ai approach to streamline autism diagnosis in young children. *Scientific reports* **10**, 1–8 (2020).
- [37] Duda, M., Daniels, J. & Wall, D. P. Clinical evaluation of a novel and mobile autism risk assessment. *Journal of autism and developmental disorders* **46**, 1953–1961 (2016).
- [38] Duda, M., Kosmicki, J. & Wall, D. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational psychiatry* **4**, e424–e424 (2014).
- [39] Fusaro, V. A. *et al.* The potential of accelerating early detection of autism through content analysis of youtube videos. *PLOS one* **9**, e93533 (2014).
- [40] Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y. & DeLuca, T. F. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PloS one* **7**, e43855 (2012).
- [41] Wall, D. P., Kosmicki, J., Deluca, T., Harstad, E. & Fusaro, V. A. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational psychiatry* **2**, e100–e100 (2012).
- [42] Hansen, L. The truven health marketscan databases for life sciences researchers. *Truven Health Analytics IBM Watson Health* (2017).
- [43] Hedegaard, H., Johnson, R. L., Garnett, M. & Thomas, K. E. The international classification of diseases, 10th revision, clinical modification (icd–10–cm): external cause-of-injury framework for categorizing mechanism and intent of injury (2019).
- [44] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).
- [45] Chattopadhyay, I. & Lipson, H. Data smashing: uncovering lurking order in data. *Journal of The Royal Society Interface* **11** (2014). URL <http://rsif.royalsocietypublishing.org/content/11/101/20140826>.
- [46] Huang, Y. & Chattopadhyay, I. Data smashing 2.0: Sequence likelihood (sl) divergence for fast time series comparison. *arXiv preprint arXiv:1909.12243* (2019).
- [47] Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30**, 3146–3154 (2017).