CLINICAL UTILITY OF THE AUTISM OBSERVATION SCALE FOR INFANTS (AOSI) FOR PREDICTING AUTISM SPECTRUM DISORDERS (ASD) IN 12-MONTH-OLD INFANTS AT INCREASED LIKELIHOOD OF ASD

by

Kyle Burke Reid

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Medical Sciences – Pediatrics
University of Alberta

Abstract

Background: The Autism Observation Scale for Infants (AOSI) has been used to investigate early features of children with an increased likelihood (IL) of a later diagnosis of autism spectrum disorder (ASD). Though independent research groups have evaluated its use in IL infant siblings (younger siblings of the children with ASD), recent studies have examined the AOSI's use in other IL populations. Since first published in 2005, an assessment of the AOSI's discriminatory and predictive utility in infant siblings and other IL populations is warranted. In addition, growing access to sophisticated computational technology has facilitated increased use of powerful computing techniques such as machine learning in research and healthcare spaces. While common in clinical and research fields, ASD research has yet to fully leverage this technology. Currently, Bussu et al. (2018) is the only study to have generated supervised machine learning classifiers using early AOSI data.

Objectives: (1) examination of research assessing the individual classification properties and group differences of the AOSI across different IL groups from 6-18 months, (2) generation of supervised machine learning classifiers using 12-month AOSI and Mullen Scales of Early Learning (MSEL) data in a cohort of infant siblings (n=373), and (3) assessment of classifier performance at predicting 36-month ASD diagnosis in infant siblings from two Canadian longitudinal studies (n=92; n=90).

Methods: A systematic search for relevant articles was conducted across six databases: CINAHL, EMBASE-OVID, ERIC, JSTOR, PubMed, and Web of Science, with articles independently reviewed for inclusion and exclusion criteria by two reviewers. Supervised

machine learning classifiers using logistic regression (with and without regularization) and support vector machines using linear, polynomial, and radial basis function kernels were generated in R/RStudio using combinations of participant biological sex, 12-month MSEL standard scores (Visual Reception, Receptive Language, Expressive Language, Fine Motor, and Early Learning Composite), and 12-month AOSI item-level and Total Score data. Factor analysis (informed by principal axis parallel analysis) was used as a means of reducing item-level AOSI data dimensionality during modelling to mitigate model overfitting. Classifiers were assessed by their ability to predict 36-month ASD diagnosis in subsets of infant siblings (n=92; n=90) from two Canadian longitudinal cohorts.

Results: The systematic search identified 354 articles with 17 meeting inclusion criteria. Four IL infant populations were assessed: younger siblings of children diagnosed with ASD, and infants with Fragile X Syndrome (FXS), Tuberous Sclerosis Complex (TSC), and Down Syndrome (DS). The systematic review had three main findings. First, five studies reported individual classification properties, although they did not use a consistent approach. Second, stable group differences emerged between IL non-ASD, IL-ASD, and infants at low likelihood of ASD (i.e., no family history) beginning at 12 months. Third, meta-analyses resulted in a large effect size for comparisons between low likelihood and IL-ASD samples and a moderate effect size for comparisons of IL non-ASD and IL-ASD samples. For supervised machine learning classifiers built with 12-month data, best-performing classifiers across all algorithm types were between 76-77% accurate and had areas under the curve (AUC) between 0.73 and 0.76. Though their specificity was excellent (0.94-1.0), they were characterized by extremely poor sensitivity (0-0.19). Relative to the performance of a 12-month AOSI Total Score cut point of 7 at predicting

36-month ASD diagnosis (informed by Youden index assessment; AUC = 0.66, sensitivity = 0.52, specificity = 0.74), machine learning classifiers had enhanced AUC and specificity, but significantly decreased sensitivity. The best-performing classifiers in this study yielded higher accuracy, AUC, and specificity (but not sensitivity) relative to the best performing classifier generated by Bussu et al. (2018) using 14-month data (accuracy = 64%, AUC = 0.71, sensitivity = 0.61, specificity = 0.67) using similar machine-learning methodology.

Conclusion: Utility of the AOSI to identify early signs of ASD in IL populations, including infant siblings of children diagnosed with ASD, FXS, TSC, and DS was demonstrated. Though the best-performing supervised learning classifiers performed below levels recommended for early screening, accuracy, AUC, and specificity were moderately improved relative to those generated by Bussu et al. (2018) using 14-month AOSI data. Further exploration into feature selection, extraction, or inclusion of 12-month AOSI and MSEL data may allow continued refinement of machine learning models built using 12-month clinical data and capable of predicting ASD at 36-months.

Preface

This thesis is an original work by Kyle Reid. The systematic review and meta-analysis reported in Chapter 2 did not require ethics approval as it was conducted on publicly accessible de-identified information. No informed consent was required for its conduct as it construes a review of publicly available research on de-identified participants. The study using machine learning detailed in Chapter 3 was conducted using human research data drawn from two large Canadian longitudinal studies: (1) *Early Development in Infants at Risk of Autism Spectrum Disorder* (CISS-1; Pro00047347), and (2) *Early Symptom Development in ASD: Role of Attention Control and Emotional Regulation* (CISS-2; Pro00061228). Both CISS-1 and CISS-2 were approved by the research ethics boards at each participating institution, and all participating families gave written informed consent upon enrollment.

For the systematic review reported in Chapter 2, Dr. Lori-Ann Sacrey, Dr. Lonnie Zwaigenbaum, Dr. Isabel M. Smith, Dr. Jessica Brian, and I all developed the search question and terms. Lori-Ann Sacrey and I completed the systematic search and functioned as both primary and secondary reviewers for review screening purposes. Though I took the lead role in the writeup and interpretation of the review, Lori-Ann Sacrey conducted and interpreted the meta-analyses. While Dr. Lonnie Zwaigenbaum and Dr. Jessica Brian co-developed the Autism Observation Scale for Infants (AOSI) together with Dr. Susan Bryson, and Dr. Lori-Ann Sacrey and Dr. Isabel M. Smith have previously published papers either on or using the AOSI, I took lead in the interpretation and write up for the systematic review – the discussion reflects my view and interpretations of the results. No part of this thesis has been previously published.

For the machine learning study reported in Chapter 3, while Lori-Ann Sacrey, Lonnie Zwaigenbaum, and I designed the study, I was wholly responsible for its conduct. I generated all code used in R/RStudio to generate the supervised learning classifiers. Similarly, I was wholly responsible for data analysis, interpretation, and write up of the results. In an effort for total transparency, the R/Rstudio code and machine learning models developed over the course of the study described in Chapter 3 are publicly available in a public GitHub repository.

My work has been supported by funding from Alberta Innovates – Health Solutions, the Canadian Institutes for Health Research (CIHR), Kids Brain Health Network (formerly NeuroDevNet), Brain Canada, the Azrieli Foundation, the Stollery Children's Hospital Foundation Chair in Autism, and the Transforming Autism Care Consortium (TACC) / Quebec Autism Research Training (QART) Program. Funders had no role in study design, data collection, analysis, manuscript preparation, or the decision to submit any results for formal publication.

Dedication

This work is dedicated to all individuals and families whose lives are impacted by autism spectrum disorders. The lives you lead are not easy; the hours you work often long and hard. The combined efforts that you and people like you make in trying to bring forward a better future is a resounding endorsement of what it means to be human.

Acknowledgements

As the common proverb 'it takes a village to raise a child' goes, I would like to argue that 'it takes a community to complete a thesis – especially *my* thesis.' The work you are reading today would, quite simply, not have been possible without the cumulative care, support, and guidance I have received throughout my life from many different amazing people.

Chief among them is my family. They have all been the champions who have supported me through the many ups and downs of my life. Thank you all so, *so* much. I know it wasn't always easy – especially when I was at my most stressed (some might have even gone so far as to described me at times as *grumpy*) – but I truly could not have done this without all of you in my life. Without you Colin, I would likely not even been here doing what I am today. A large part of who I am and what I aim to be has been shaped and moulded by the experiences I've had being your brother… and I wouldn't change it for anything. Dad, you are truly one of the strongest people I know. You have dealt with so much in your life and continually picked yourself right back up off the floor again and again *and again* such that I hope to have but a mere fraction of the strength you have shown. With the strength and resiliency I have learned from you over my tenure on this planet, I know I will be able to go far. Last but certainly not least, thank you Mumsey for absolutely *everything*. Words really do not do this acknowledgement justice for I am no wordsmith such as thee. Trying to boil down all the myriad thoughts and feelings I have about everything you have helped and supported me with over the years would require a *significantly* larger acknowledgements section – probably on the order of a few hundred more pages (it would definitely be a touching tear-jerker though!). You have been the rock that I've been able to anchor my life on. Suffice to say, all three of you have shaped different facets of who I am today… and I wouldn't have it any other way. Thank you all so, *so* much. I love you all from the bottom of my heart.

I'd also like to thank my supervisors Dr. Lonnie Zwaigenbaum and Dr. Lori Sacrey. Lonnie has been an inspirational leader in the field of autism. In addition to his personal support for me, I am grateful for how his efforts continue to support families like mine. Lori is a brilliant and generous mentor who, above all else, is an amazing human being. She has gone over and above the call of duty in her support for me during my undergraduate and graduate studies and I am

forever grateful for that. I would also like to thank Dr. Francois Bolduc and Dr. Sandy Thompson-Hodgetts for their support and guidance as members of my committee, and Dr. Frederick Shic for graciously serving as the external member of my committee.

Thanks to everyone (past and present) from the Edmonton Autism Research Centre. I have learned and grown so much over these past few years. It would be remiss to not recognize that the work we do would not be possible without the families who so generously give of their time to be participate in research like ours.

My interest in research was in part ignited by the inspired teachings and guidance of Dr. Katie Biittner. Her scholarship continues to influence how I engage with the world not only as a researcher and scientist, but as a human being trying to make the world a better, kinder place.

Finally, thanks to Dr. Miguel-Cruz who so generously took the time to help a lost graduate student find his way.

To all of you who have helped me get to where I am today, from the bottom of my heart, *thank you*.

I truly do stand on the shoulders of giants.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADI-R	Autism Diagnostic Interview-Revised
ADOS	Autism Diagnostic Observation Schedule
AIC	Akaike information criterion
AOSI	Autism Observation Scale for Infants
ASD	Autism Spectrum Disorders
AUC	Area under the ROC curve
CDC	Centres for Disease Control and Prevention
CI	Confidence interval
CISS-1	Canadian Infant Sibling Study
CISS-2	The new Canadian Infant Sibling Study
DS	Down Syndrome
DSM	Diagnostic and Statistical Manual of Mental Disorders
DSM-5	DSM, Fifth Edition
DSM-IV-TR	DSM, Fourth Edition with Text Revised
EL	Expressive Language
ELC	Early Learning Composite
EM	Expectation Maximization
FM	Fine Motor
FXS	Fragile X Syndrome
IL	Increased Likelihood
IL-ASD	Infant siblings diagnosed with ASD at 36-months
IL-N	Infant siblings *not* diagnosed with ASD at 36-months
IL-Siblings	Infant siblings at increased likelihood for ASD diagnosis
LL	Infants at low likelihood for ASD
MCAR	Missing completely at random
MSEL	Mullen Scales of Early Learning
PDD	Pervasive developmental disorder
PDD-NOS	Pervasive Developmental Disorder–Not Otherwise Specified
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta Analyses
RL	Receptive Language
ROC	Receiver operator characteristic
SES	Socioeconomic status
SVM	Support vector machines
TSC	Tuberous Sclerosis Complex
TSC-ASD	Children with TSC who are diagnosed with ASD
VABS	Vineland Adaptive Behavior Scales
VR	Visual Reception

Chapter 1: Introduction

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition characterized by differences in social-communication, the presence of restricted interests, repetitive behaviours, and/or atypical responses to sensory input (APA, 2013). The Centre for Disease Control and Prevention (CDC) estimates that the community prevalence rate of ASD in the United States is 1 in every 44 children by age 8 (Maenner et al., 2021). Though sex differences in ASD diagnostics have been observed (males being four times as likely than females to be diagnosed; Baio et al., 2018), sex difference in older youth and adults may reflect camouflaging of ASD symptoms or phenotypic differences in females that may lead to delayed or missed diagnoses (Tubío-Fungueiriño et al., 2021). Some populations are at increased likelihood (IL) of being diagnosed with ASD due to environmental and/or genetic factors such as increasing paternal age, premature birth, Fragile X Syndrome, Down syndrome, and Tuberous Sclerosis Complex (Hultman et al., 2011, Capal et al., 2017, Agrawal et al., 2018, Abbeduto et al., 2014). Relative to a general community population, ASD prevalence in these different IL contexts is considerably elevated (Numis et al., 2011, Ozonoff et al., 2011, Abbeduto et al., 2014, Szatmari et al., 2016, Hahn et al., 2020). While studying these IL can be beneficial due to the increased prevalence rates, it is important to note that IL infants do not constitute a homogenous group; IL infants who go on to receive an ASD diagnosis have been shown to be distinct from IL infants who do not (Brian et al., 2014, Estes et al., 2015).

ASD diagnoses carry with them a lifetime of direct and indirect costs related to medical and healthcare expenses, therapeutics, (special) education, productivity loss for family or caregivers, accommodations, respite care, and out-of-pocket expenses (Rogge & Janssen, 2019). It is estimated that by 2029, the societal lifetime social cost of ASD is expected to grow from $7 trillion to US $11.5 trillion (Cakir et al., 2020). Given that the annual costs that families spend to support individuals with ASD are enormous (often well beyond what the annual income for families), provision of supportive care necessitates access to external resources and supports (Dudley et al., 2014). However, in Canada, access to supports or funding for ASD individuals is often restricted based on eligibility criteria including symptom severity, intelligence quotient cut-offs, or intellectual disability (Dudley et al., 2014). If supports can be obtained, a complicating issue is that they may be substandard or inadequate (Dudley et al., 2014). When time, energy,

and funding is limited and/or existing community infrastructure falter, supportive care for individuals falls to families or caregivers (Dudley et al., 2014). Though there are many barriers to ASD-related care (e.g., shortage of healthcare services, physician knowledge, cost of services, family and individual knowledge, language, and stigma; Malik-Soni et al., 2021), there is a growing awareness of the benefit and long term impact early detection has on individuals and their families with ASD.

Early identification of children with developmental conditions can facilitate early intervention services which can improve functional outcomes of the child and decrease the impact of the delay or disability on the child and family (AAP, 2001). Though the logistic and financial challenges associated with early screening and detection of ASD are still being explored, these studies point to the benefits of early detection and diagnosis. The importance of early ASD diagnosis (Gardner et al., 2013, Fuller & Kaiser, 2019, Towle et al., 2020) is underscored by the increasing recognition of the benefits early intervention has for children with ASD at improving functional outcomes (Fuller & Kaiser, 2019, Dawson et al., 2010, Pickles et al., 2016, Noyes-Grosser et al., 2018, Kodak & Bergmann, 2020). Though enhanced early ASD screening and detection is not *the* answer to all the issues faced by individuals and families with ASD, it potentially has compounding benefits – especially given that early ASD diagnosis is very stable from early to mid-childhood (Brian et al., 2016). Increased early access to developmental monitoring and, in the case of ASD diagnosis, earlier access to intervention resources has the potential to serve a dual role of supporting individuals with ASD as well as their families. This is an important consideration as families or caregivers of children with ASD often experience challenges when navigating multiple systems in search of or access to support services (Crossman et al., 2020). Enhanced early screening and detection technology therefore has the potential to not just facilitate access to support services, but also kickstart families and caregivers with respect to the knowledge and education required to navigate healthcare and support service services systems to best support their child (Towle et al., 2020). Though there is evidence to support the utility of ASD-specific screens at 18 and 24 months, screening prior to 24 months old may be associated with higher false-positive rates than screening after 24 months (Zwaigenbaum et al., 2015). While research has already established the long-term stability of ASD diagnosis in children ≥ 24 months of age, further research is required to substantiate stability of ASD

diagnoses prior to 24 months of age – especially in the context of early screening Zwaigenbaum et al., 2015). Given that early screening initiatives are only as effective as the access to resources and appropriate follow-up allows them to be (Zwaigenbaum et al., 2015), development of new tools or further exploration and refinement of existing assessment tools such as the Autism Observation Scale for Infants (AOSI) is necessary if future early screening efforts are to be expanded.

The AOSI is a brief, 19-item observational measure designed to characterize early behavioural signs of ASD between 6 and 18 months in a familial cohort of infants at increased likelihood of the disorder (i.e., are infant siblings of children diagnosed with ASD; Bryson et al., 2008). The AOSI assesses multiple overlapping constructs that characterize prodromal ASD (e.g., social communication, emotional regulation, atypical sensory-motor behaviours, repetitive behaviours, etc.) within an interactive, play-based context in which behaviour can be systematically elicited by trained examiners (Bryson et al., 2008). Parents are encouraged to assist in making their child feel comfortable during assessment (and are present in the room during assessment), however they are otherwise encouraged to assume an observer role (Bryson et al., 2008). The AOSI takes approximately 20 minutes to complete with sessions videotaped to both assist in behavioural ratings and provide a database for future analysis (Bryson et al., 2008). Though the AOSI has been validated for use in IL infant siblings, research groups are just starting to assess the tool's utility in identifying early signs of ASD in other IL populations even though early signs of ASD may be expressed differently across them. Since first published in 2005, there has been no critical examination of the AOSI's performance in infant siblings or any other IL population. Accordingly, an in-depth examination of published research assessing the AOSI's classification properties and group differences when applied to IL populations 6-18 months old is warranted.

The American Academy of Pediatrics (AAP) recommends universal ASD screening for consistent practice and optimal detection of young children with early ASD symptomology across varied contexts. However, pediatricians are limited by an economy of scale and time; they are already required to complete an increasing number of tasks including screening for non-ASD conditions during child visits (Zwaigenbaum et al., 2015). Major barriers to ASD screening include (1) lack of time and inadequate reimbursement, (2) logistical issues including work flow,

lack of familiarity, scoring difficulties, (3) lack of office base systems needed for referrals / outcome monitoring (Zwaigenbaum et al, 2015). In addition, identifying developmental delays in children can be problematic if they only become apparent at ages when specific developmental milestones expected to be reached are missed (AAP, 2001). Screening for a condition based on missed major developmental milestones can lead to late recognition and intervention (AAP, 2001). Given that provision of accurate and accessible diagnoses is a fundamental challenge of not just autism-specific healthcare but global healthcare in general, modern techniques like artificial intelligence and machine learning are showing increasing promise in revolutionizing healthcare by facilitating provision of precise and personalized diagnoses (Richens et al., 2020).

Machine learning is a branch of artificial intelligence and computer science focused on utilizing data and computer algorithms to imitate the way humans think and learn to gradually improve their accuracy and performance (IBM, 2020). There are numerous advantages to using machine learning in health research. For instance, machine learning can be flexible and scalable, allowing for deployment in many different areas including risk stratification, diagnosis, classification, and survival predictions (Ngiam & Kor, 2019). When applied to clinical or research contexts, machine learning can enable analysis of increasingly diverse types of data that can summarily be incorporated into models used to help predict disease risk, diagnosis, and even treatment modalities (Ngiam & Kor, 2019). One major benefit of applying machine learning to data analysis modalities is that it can allow for rapid evaluation of different combinations of factors or variables as a means of determining which combination provides the best accuracy and predictive power when assessing for a given outcome. In ASD machine learning research, one of the primary goals is developing strategies to minimize diagnostic time with improved accuracy (Eman et al., 2020). Leveraging the use of machine learning techniques on large existing clinical or research datasets can be one means by which we better our understanding on the early emergence or characteristics of ASD. This information, in turn, could prove invaluable for future efforts into the refinement of existing (or the creation of new) tools used for early ASD screening or detection capacities.

This thesis is organized into two main sections. The first section (Chapter 2; Study 1) is a systematic review and meta-analyses investigating the use of the AOSI with particular focus on

exploring the AOSI's classification properties across different IL groups of 6-18 months old infants. The second section comprises a machine learning study (Chapter 3; Study 2) exploring the utility and performance of AOSI data in the context of generating predictive supervised machine learning classifiers. The content reported in Chapter 2 and 3 are reported in a paper-based format. Accordingly, both Chapter 2 and 3 contain details pertaining to background information, methods, results, and discussion. While Study 1 and 2 are stand-alone, the results of Study 1 informed critical methodological decisions for Study 2. In particular, the choice of building predictive models using 12-month data was predicated on the meta-analytical results from Study 1.

Chapter 2: Study One

Background

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by differences or impairments in social-communication and the presence of restricted interests, repetitive behaviours, and/or atypical responses to sensory input (American Psychiatric Association 2013). The current community prevalence rate of ASD as reported in the United States by the Centers for Disease Control and Prevention (CDC) is 1 in every 44 children by age 8 (Maenner et al., 2021). There are some populations who are at an increased likelihood (IL) for developing ASD due to environmental or genetic factors such as increasing paternal age, children with premature birth, fragile X syndrome (FXS), and tuberous sclerosis complex (TSC; Hultman et al., 2011, Capal et al., 2017, Agrawal et al., 2018, Abbeduto et al., 2014). Because ASD is characterized by highly complex and variable phenotypic presentation, it is important to assess the utility of any measure attempting to investigate early features of ASD.

The importance of early diagnosis (Gardner et al., 2013, Fuller & Kaiser, 2019, Towle et al., 2020) is highlighted by the increasing recognition of the benefits of early intervention for children with ASD (Fuller & Kaiser, 2019, Towle et al., 2020, Dawson et al., 2010, Bonis 2015, Pickles et al., 2016, Noyes-Grosser et al., 2018). It is important for primary care practitioners to provide referrals to specialists and early intervention services (Zwaigenbaum et al., 2015). Given that diagnosis of ASD is very stable from early to mid-childhood (94% of infant siblings of children with ASD followed from ages 3 to 9 years retained a diagnosis in Brian et al., 2016's study), tools that aid in early identification of ASD have potential utility to facilitate access to early intervention services.

The Autism Observation Scale for Infants (AOSI) is a brief, 19-item observational measure that was initially designed to characterize early behavioural signs of ASD between 6 and 18 months in a familial cohort of infants at increased likelihood of the disorder (i.e., are infant siblings of children diagnosed with ASD; Bryson et al., 2008). The AOSI assesses multiple overlapping constructs that characterize prodromal ASD (e.g., social communication, emotional regulation, atypical sensory-motor behaviours, repetitive behaviours, etc.) within an interactive, play-based context in which behaviour can be systematically elicited by trained examiners (Bryson et al.,

2008). The AOSI has been validated in IL infant siblings and research groups are now assessing the tool for use in identifying early signs of ASD in other populations of infants at IL for ASD including infants who were born premature, or who have underlying genetic or neurological conditions such as Down Syndrome (DS; Sanderson 2016, Hahn et al., 2020). Yet, early signs of ASD may be expressed differently across these populations. The purpose of this systematic review and meta-analysis is to provide an in-depth examination of research assessing the individual classification properties and group differences of the AOSI across different IL groups from 6-18 months to examine if early manifestations of ASD are similar across different IL populations.

Methods

Search strategy

A systematic review was completed following the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA; Moher et al., 2009) checklist. Searches were performed on July 4th, 2022, in six databases: CINAHL, EMBASE-OVID, ERIC, JSTOR, PubMed, and Web of Science. Search terms and strategies were refined following discussion between two reviewers (KR and LS) using the terms “Autism Observation Scale for Infants,” “AOSI,” and "autism", “autism spectrum disorder,” and “autistic disorder.” No published search filters were used. Because the AOSI was first published in 2005, date limits restricted the identification of articles from 2005 and onwards. Although no language limits were used to allow for capture of any non-English publications (as the AOSI has been translated into other languages, such as Hebrew; Ben-Sasson & Carter, 2012), no non-English studies were identified. Primary database searches identified 453 articles. Grey literature databases (opengrey.eu, worldcat.org, greylit.org) were surveyed using identical search terms used in primary database searches to identify relevant unpublished data and identified 27 articles. The same search terms were employed in the primary and grey literature searches. One article (Zwaigenbaum et al., 2005) was manually imported for primary screening as study authors knew it was the first paper published on the AOSI but was not captured in either the primary or grey literature search. IL groups were not pre-specified for the search to be as inclusive as possible and not potentially exclude articles from IL populations unknown to study authors. In total, 481 articles were imported into Covidence (covidence.org) for review. Following de-duplication, 354 articles were identified for further screening. The

complete search strategy as run and PRISMA checklist can be found in Appendix 1. Though no PROSPERO protocol for this review was registered, the PROSPERO database was searched to ensure no other similar review had been registered or conducted prior to this study. As of July 4th, 2022, while one protocol was identified that used the AOSI (CRD42020158688), the AOSI was used as an outcome measure in a systematic review of ASD-related interventions in the first 2 years of life and thusly does not conflict with this review focusing on AOSI classification properties.

Screening for Inclusion and Exclusion Criteria

To be included in this review, a paper (1) used the AOSI in a population of IL infants characterized by a specific factor known to be associated with increased likelihood of ASD diagnosis (e.g., infant siblings of a child with ASD, infants with FSX, TSC, or DS) and a sample of control infants (low likelihood [LL] or IL infants not diagnosed with ASD), (2) either reported at least one AOSI cut point and its corresponding sensitivity and specificity or compared AOSI Total Scores between two or more groups, and (3) included original data. A paper was excluded from analysis if it (1) did not use the AOSI, (2) did not include AOSI Total Scores, number of AOSI Risk Markers, or sensitivity and specificity data, (3) lacked a comparison group (IL-not diagnosed/IL-N; LL controls), or (4) was a review article, commentary, conference abstract, or conference presentation. Titles and abstracts of 354 articles were screened using the reported inclusion and exclusion criteria in Covidence by two independent reviewers (KR and LS) to identify the studies meriting full-text review. Both reviewers assessed the 33 articles meriting full-text review and had 97% agreement for studies meeting inclusion criteria. The one disagreement was resolved by consensus following discussion between reviewers. In total, 17 articles were selected for full-text extraction, with nine included in meta-analyses. Reasons for exclusion at the full-text review level are described in Figure 2.01.

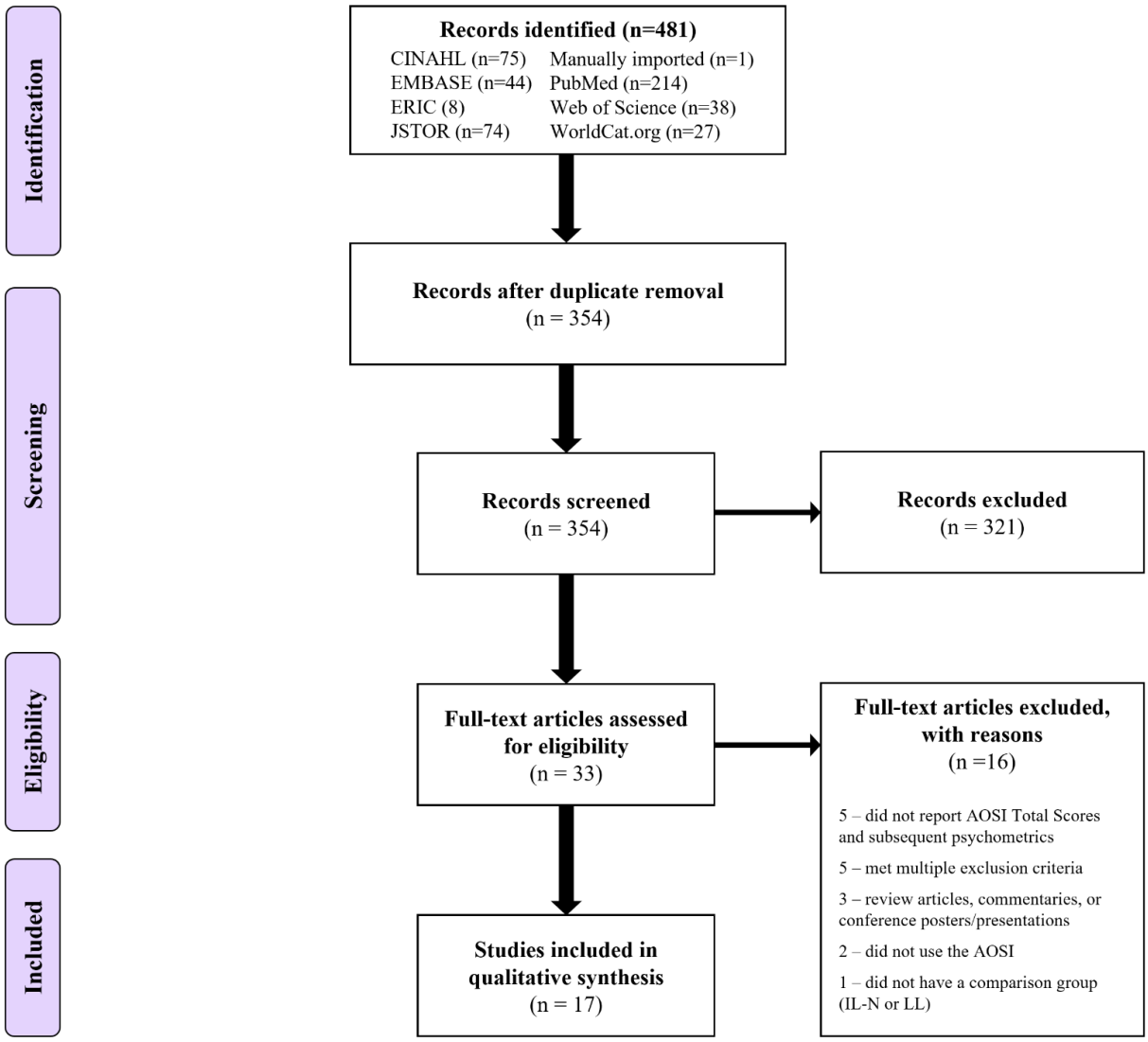

Figure 2.01 | Systematic Review Search Strategy

Assessment of Study Quality and Risk of Bias

Authors of the *Cochrane Handbook for Systematic Reviews of Interventions* recommend focusing on assessing risk of bias over methodological quality (Stang 2018). For this reason, study quality, methodology, and potential sources of bias were assessed using a composite form generated using items from the National Institute of Health's Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies (Nhlbi 2022), the Joanna Briggs Institute's Checklist for Systematic Reviews and Research Syntheses (Moola et al., 2017), the Critical Appraisal Skills Programme Checklist for Cohort Studies (CASP 2018), and the Scottish Intercollegiate Guidelines Network's Methodology Checklist 3 for Cohort Studies (British Thoracic Society 2016). KR and LS generated the table assessing for potential sources of bias in

the included studies. Both KR and LS independently scored each of the 17 studies by each category in the risk of bias table before comparing scores. Any disagreements were resolved via consensus. One intent behind this review and meta-analysis was to highlight potential sources of bias that may warrant further investigation or consideration as it relates to study quality and validity, as well as to facilitate a discussion of the generalizability of results.

Data Extraction

Two primary reviewers (KR, LS) developed a standardized data extraction form. Extracted demographic information included sample size, the IL population being examined, sex ratio, ethnicity, parental age, and socioeconomic status (SES). AOSI-relevant information and potential biasing factors extracted from the studies included inclusion/exclusion criteria; chronological age at assessment; statistical method; covariates; ASD classification/diagnostic assessment; AOSI cut points, sensitivity, and specificity; group comparisons; data required to calculate effect sizes (IL/LL sample sizes, AOSI Total Score and standard deviation data); and study limitations. The data extraction form was iteratively developed to allow for flexibility and comprehensiveness (Colquhoun et al., 2014). Reviewers KR and LS each extracted data from a portion of the 17 included studies and cross-checked the other's work for validation purposes.

Statistical Analysis

Meta-analyses on AOSI Total Score were completed in Stata using the *metan* command (Sterne 2009). Eight separate meta-analyses were conducted: (1) studies with LL controls versus IL infants diagnosed with ASD (IL-ASD) between 6-10 months, (2) studies with LL controls versus IL IL-ASD between 12-14 months, (3) studies with IL infants not diagnosed with ASD versus IL-ASD between 6-10 months, with differences between the comparison group explored using the subgrouping command, (4) studies with IL infants not diagnosed with ASD versus IL-ASD between 12-14 months, with differences between the comparison group explored using the subgrouping command, (5) studies with IL infants not diagnosed with ASD but who were classified as having a developmental delay versus IL-ASD between 6-10 months, (6) studies with IL infants not diagnosed with ASD but who were classified as having a developmental delay versus IL-ASD between 12-14 months, (7) studies with IL infants with typical development

(e.g., IL infants without any developmental concerns) versus IL-ASD between 6-10 months, and (8) studies with IL infants with typical development versus IL-ASD between 12-14 months.

Cohen's d effect sizes (calculated using the following formula: $d = M_1 - M_2/\sigma_{spooled}$ where $\sigma_{spooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2)/2]}$) and standard errors were computed for each study (for which data were available) and used in the meta-analyses, with d = 0.2 – 0.49 = small effect, d = 0.5 – 0.79 = medium effect, and d ≥ 0.8 = large effect (Cohen 1988). Heterogeneity was examined using confidence intervals (CI), the I^2 statistic, and forest plots. The I^2 statistic, which ranges from 0 to 100%, is a measure of the variability in effect estimates resulting from heterogeneity between studies rather than chance (e.g., sampling error) (Higgins et al., 2019). Statistical heterogeneity can be considered unimportant between 0-40%, moderate between 30-60%, substantial between 50-90%, and considerable between 75-100% (Higgins et al., 2019). Preliminary analyses suggested our meta-analyses had I^2 statistics > 50%, thus random effects modelling was used. Funnel plot, trim and fill analyses, and Egger's tests for small study effects were completed using the *metafunnel*, *metatrim*, and *metabias* commands in Stata (Sterne 2009) to investigate publication bias and heterogeneity through visual and statistical examination of the data (Egger et al., 1997).

Overall, 9 of the 17 articles were included in the meta-analyses (Capal et al., 2017, Hahn et al., 2020, Gammer et al., 2015, Estes et al., 2015, McDonald et al., 2017, Bussu et al., 2018, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, Hahn et al., 2017). The remaining 8 articles were not included as they were earlier studies from the same research groups or were conducted using the same study population (i.e., overlapping participants between studies). For studies conducted on the same infant cohort or published from the same research group, studies with the highest sample sizes were chosen for inclusion in meta-analyses. In addition, no study was included in the same meta-analysis more than once to prevent unduly weighting or biasing analyses.

Ethics Statement

Ethics approval was not required for this study as it is a systematic review conducted on publicly accessible de-identified information. No informed consent was required as this article is a review and no individual participants have identifying information.

Results

This systematic review examining the utility of the AOSI to identify early signs of ASD across different IL populations included 17 peer-reviewed articles. The results are organized as follows: a descriptive overview of the included articles with location, sample size, age, and participant demographics; an overview of the IL group status; an overview of study design and methodology; description of how and at what age(s) the AOSI was used; statistical analyses employed; AOSI cut points and their associated psychometric data, and risk of bias assessment.

Study and Participant Demographics

Overview of Included Articles. Although no language limits were used in the search, all articles meeting inclusion criteria were published in English. The earliest article meeting criteria was published in 2005 and the most recent in 2021. The articles originated from three countries: Canada (n=4), the United Kingdom (n=6), and the United States (n=7). Fifteen were longitudinal cohort studies (participants assessed at multiple time points) and two were cross-sectional. Total sample sizes ranged from N=36 (Hahn et al., 2017) to N=681 (Zwaigenbaum et al., 2020). IL subsamples ranged from n=15 (FXS; Roberts et al., 2016) to n=501 (infant Siblings; Zwaigenbaum et al., 2020). Several studies were either conducted by the same research group (the British Autism Study of Infant Siblings [BASIS; Gammer et al., 2015, Bussu et al., 2018, Gilga et al., 2015, Bedford et al., 2016, Bedford et al., 2017, Bedford et al., 2019]; the Canadian Infant Sibling Study [CISS-1; Zwaigenbaum et al., 2005, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, Sacrey et al., 2018]) or using overlapping participants (see [Hahn et al., 2020, Hahn et al., 2017, Roberts et al., 2016] or [McDonald et al., 2017, Jeste et al., 2014]).

Participant Demographics. Of the 17 included studies, four assessed infants at multiple times between ages 3 and 24 months (Zwaigenbaum et al., 2005, Estes et al., 2015, Roberts et al., 2016, Gilga et al., 2015), eleven assessed infants at multiple times between 6 and 36 months

(Capal et al., 2017, Gammer et al., 2015, McDonald et al., 2017, Bussu et al., 2018, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, Bedford et al., 2016, Bedford et al., 2017, Bedford et al., 2019, Sacrey et al., 2018, Jeste et al., 2014) and two assessed infants at one time point, between 7 and 18 months (Hahn et al., 2020, Hahn et al., 2017). Detailed participant demographic data (including both ethnicity and SES) were only reported by three studies (McDonald et al., 2017, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021) which consisted of study populations of middle-to-higher SES families of largely Caucasian ancestry. Three studies (Hahn et al., 2020, Estes et al., 2015, Hahn et al., 2017) only report ethnicity data, and likewise feature largely Caucasian study populations (with two-thirds of participants or more being Caucasian). Two studies, Hahn et al., 2017 and Sacrey et al., 2018, use SES or family demographic data in their analyses but do not directly report the results or descriptive statistics in their paper. The remaining nine studies reported no participant demographic data outside of the biological sex of the participant. Descriptive characteristics of included studies can be seen in Table 2.01.

Table 2.01 | Included Study Characteristics

Article (by publication year)	Country	IL group	Sample size (IL)	Sample size (LL)	Sex Ratio	Participant or Family Ethnicity	Participant / Family SES
2005 Zwaigenbaum	Canada	Infant siblings	n=65	n=23	ns	ns	ns
2014 Jeste	United States	Infants with TSC	n=40	ns	ns	ns	ns
2015 Estes	United States	Infant siblings	n=210	n=98	**Sibs**: 129 male, 81 female **LL:** 55 male, 43 female	86.4% Caucasian overall. $^{\delta *}$	ns
2015 Gammer	United Kingdom	Infant siblings	n=53	n=50	**Sibs:** 21 male, 32 female **LL**: 21 males, 29 females	ns	ns
2015 Gilga	United Kingdom	Infant siblings	n= 82	n=27	**Sibs:** 45 male, 37 female **LL**: 14 male, 13 female	ns	ns
2016 Bedford	United Kingdom	Infant siblings	n=54	n=48	**Sibs**:21 male, 33 female **LL**: 21 male, 29 female	ns	ns
2016 Roberts	United States	Infants with FXS, Infant siblings	**FXS**: n=15 **Sibs:** n=23	n=17	**FXS**: 15 males **Sibs**: ns **LL**: ns	ns	ns
2017 Capal	United States	Infants with TSC	n=79	n/a	**TSC:** 43 males, 36 female	ns	ns
2017 Hahn	United States	Infants with FXS, Infant siblings	**FXS:** n=18 **Sibs:** n=21	n=22	**FXS**: 14 male, 4 female **Sibs:** 17 male, 4 female **LL**: 18 male, 4 female	73.9% Caucasian overall. $^{\delta *}$	Overall, 51.7% of mothers had a college degree or higher overall. $^{\delta *}$ No group difference observed between family income, race, or maternal education between IL/LL groups. Mean family income = $57,851.50 USD. *
2017 Bedford	United Kingdom	Infant siblings	n=42	n=37	**Sibs:** 15 males, 27 female **LL**: 15 males, 22 female	ns	Family income was used in analyses, but data was not directly reported.
2017 McDonald	United States	Infants with TSC	n=23	n=21	**TSC:** 16 male, 7 female **LL:** 9 male, 12 female	73.8% Caucasian overall. $^{\delta *}$	Overall, 78.1% of mothers had a 4-year college/some graduate school or an advanced or professional degree. *
2018 Bussu	United Kingdom	Infant siblings	n=161	n=71	**Sibs:** 85 male, 76 female **LL**: 31 males, 40 females	ns	ns
2018 Sacrey	Canada	Infant siblings	n=188	n/a	**Sibs:** 111 male, 77 female	ns	Family demographics (participant's birth order, number of children in the family, father's and mother's age at

							participant's birth, and family SES) were used in analyses but not reported.
2019 Bedford	United Kingdom	Infant siblings	n=54[α]	n=50[α]	**Sibs:** 21 male, 33 female **LL**: 21 male, 29 female	ns	ns
2020 Hahn	United States	Infants with DS	n=18	n=18	**DS:** 14 male, 4 female **LL**: 14 male, 4 female	66.6% Caucasian overall. *	ns
2020 Zwaigenbaum	Canada	Infant siblings	n=501[α]	n=180[α]	**Sibs:** 281 male, 220 female * **LL:** 97 male, 83 female *	Overall, 84.8% of fathers and 82.3% of mothers were Caucasian. *	51.36% of participants families had a Hollingshead Four-Factor Index between 51 and 66.
2021 Zwaigenbaum	Canada	Infant Siblings	n=500	n=177	**Sibs:** 280 male, 220 female **LL**: 94 male, 83 female	Overall, 84.8% of fathers and 85.0% of mothers were Caucasian. *	52.07% of participants families had a Hollingshead Four-Factor Index between 51 and 66.

DS = Down Syndrome, FXS = Fragile X Syndrome, IL = infants at increased likelihood of being diagnosed with ASD, LL = infants at low likelihood of being diagnosed with ASD, SES = Socioeconomic status, Sibs = Infant siblings of children with ASD, TSC = Tuberous Sclerosis Complex, USD = US Dollar
ns = not specified
[α] = sample sizes varied depending on analysis
* = calculated from data provided in the paper
δ = No delineation between parent and child ethnicity – ethnicity was presented without context as to if it was the parents or child/study participant

Increased Likelihood Group Status

Four IL groups were assessed: (1) younger siblings of children formally diagnosed with ASD (hereafter infant siblings), (2) infants with FXS, (3) infants with TSC, and (4) infants with DS. All four populations have elevated rates of ASD diagnoses relative to the general population, with the prevalence rate of ASD in infant siblings, FXS infants, TSC infants, and DS infants reported to be as high as 20%, 50%, 40%, and 42%, respectively (Abbeduto et al., 2014, Hahn et al., 2020, Szatmari et al., 2016, Numis et al., 2011, Ozonoff et al., 2011). Infant siblings comprised part or all of the IL sample in 13 of the 17 studies (Zwaigenbaum et al., 2005, Gammer et al., 2015, Estes et al., 2015, Bussu et al., 2018, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, Hahn et al., 2017, Roberts et al., 2016, Gilga et al., 2015, Gilga et al., 2015, Bedford et al., 2017, Bedford et al., 2019, Sacrey et al., 2018). Descriptions of how ASD diagnoses were confirmed in the probands (older siblings diagnosed with ASD), study inclusion/exclusion criteria, and reliability assessment can be found in Appendix 1. Three studies included infants with TSC (Capal et al., 2017, McDonald et al., 2017, Jeste et al., 2014), two included infants with FXS (Hahn et al., 2017, Roberts et al., 2016), and one included infants with DS (Hahn et al., 2020).

Study Design and Methodology

An overview of study design, including study objectives, inclusion criteria, and exclusion criteria is provided in Table 2.02.

Table 2.02 | Included Study Methodologies

Article	Study objective	Inclusion criteria	Exclusion criteria	ASD diagnostic / outcome assessment
2005 Zwaigenbaum α	Characterization of behavioural manifestations of ASD in the first year of life of IL infant siblings of children with ASD.	**IL-Sibs:** Have an older sibling formally diagnosed with ASD confirmed by the ADOS and a clinical interview using DSM-IV criteria. **LL:** Term gestation, birth weight >2500g.	**IL-Sibs:** ns **LL**: No 1st or 2nd degree relative with ASD.	24-month ADOS classification for ASD.
2014 Jeste β	Defining early clinical, behavioral, and biological markers of ASD in IL infants with TSC.	**IL-TSC:** Recruited through TSC specialty clinics, newborn nurseries, pediatrician's offices, or the TSC alliance in the United States. **LL**: Recruited through IRB-approved infant databases in the greater Los Angeles and Boston area.	**IL-TSC:** ns **LL**: Prematurity (<37 weeks gestational age), birth trauma, developmental concerns, or immediate family history of ASD or intellectual disability.	Diagnoses were based on the convergence of ADOS scores (at 18-, 24- and 36-months) and clinical judgement of a board-certified pediatric neurologist.
2015 Estes	Compare IL-sibs diagnosed with ASD to those who are not with respect to (1) longitudinal trajectories of cognitive development and adaptive functioning from 6-24 months and cross-sectional differences at 6, 12, and 24 months, and (2) behavioural features at 6 and 12 months.	**IL-Sibs**: Have an older sibling that met criteria for ASD on the SCQ and ADI-R, confirmed by medical records. **LL:** Typically developing older sibling who did not meet for ASD on the SCQ or Family Interview for Genetic Studies (FIGS) and had no first-degree relative with ASD or intellectual disability.	**All participants**: (1) Genetic conditions/syndromes, (2) medical/neurological conditions affecting growth, development, or cognition (e.g. seizure disorders) or significant sensory impairments (vision/hearing loss), (3) birth weight <2,000g, gestational age <36 weeks, significant perinatal adversity and/or exposure in utero to neurotoxins, (4) contraindications for MRI, (5) predominant home language other than English, (6) adopted children or half-siblings, (7) 1st degree relative with psychosis, schizophrenia, or bipolar disorder, and (8) twins.	24-month clinical best estimate diagnosis using DSM-IV-TR criteria assessing for ASD or PDDNOS. Two clinicians assigned diagnoses: one who conducted the diagnostic assessment, and the other (a clinical psychologist or psychiatrist) was naïve to previous examinations and IL status but reviewed testing results to provide an independent DSM diagnosis confirmation.
2015 Gammer δ	To investigate if (1) AOSI scores differ between IL-Sibs and LL controls at 7/14-months, (2) if AOSI scores differ between IL-Sibs who diagnosed with ASD and those who are not, and (3) investigate any associations between ~7/14-month AOSI scores and later 24/36-month ADOS-G scores.	**IL-Sibs**: have an older full/half sibling with a community clinical ASD diagnosis confirmed using the DAWBA and SCQ by expert clinicians. **LL**: No 1st degree relative with ASD, have at least one older full/half sibling who does not meet criteria SCQ criteria for ASD.	**IL-sibs:** Significant conditions (e.g., FXS, TSC). **LL:** Full term birth (gestational age 37-42 weeks).	36-month diagnostic assessment conducted by four clinical researchers using ICD-10 criteria (childhood autism, PDD) based on 24- and 36-month ADOS and ADI-R.
2015 Gilga δ	Investigate whether perceptual and social interaction atypicalities in IL-Sibs reflect co-expressed but biologically independent pathologies (measured by	**IL-Sibs**: Have at least one older sibling with a community clinical diagnosis of ASD confirmed by an expert clinician using the DAWBA and SCQ.	**IL-Sibs:** Significant medical conditions in probands or extended family members. **LL:** ns	24-month ADOS scores were used as the primary outcome measure for ASD symptomology.

	eye tracking of spontaneous orienting to letter targets presented among distractors) as suggested by a 'fractionable' phenotype model of autism.	**LL**: Have at least one older sibling, full-term birth, normal birth weight, and lack of ASD diagnoses in any 1st degree family members (confirmed by a parent interview of family medical history).		
2016 Bedford δ	To assess whether sex differences are apparent in early autism markers (attention disengagement speed, gaze-following behaviour, the AOSI) or in the relationships between these early markers and later autistic traits.	ns	**All participants**: Medical or developmental conditions.	36-month consensus diagnostic assessment using ICD-10 criteria (ASD-sibs, childhood autism, atypical autism, PDD) using all available study data (AOSI, ADOS, SCQ, attention disengagement speed, gaze-following behaviour) made by experienced researchers.
2016 Roberts δ	Contrast the profile of ASD symptoms in 12-month IL-Sibs, IL-FXS, and LL controls to identify (1) risk factors for ASD in infants with FXS, (2) the concordance rate of risk factors in IL-FXS vs IL-Sibs, and (3) to document potentially etiologically distinct ASD risk profiles across IL-FXS and IL-Sib groups.	**IL-FXS**: Have a confirmed genetic report of FXS. **IL-Sibs**: Have an older full sibling with a confirmed community clinical ASD diagnosis. **LL**: Absence of known or suspected delays, no history or indicator of ASD.	**All participants**: Neurological conditions or gestation <37 weeks. **LL**: Infants with developmental composite scores >1 SD away from the mean.	24-month ADOS-2 toddler module scores for 39/55 participants (10/15 IL-FXS, 16/23 IL-Sibs, and 13/17 LL controls).
2017 Capal	To determine early predictors of autism risk in infants with TSC to identify children in most need of accessing autism-specific interventions.	**IL-TSC:** Between the 3 and 12 months old at study enrollment, met clinical or genetic criteria for TSC diagnosis.	**IL-TSC:** Gestational age <36 weeks at birth with significant perinatal complications (respiratory support, confirmed infection, intraventricular hemorrhage, cardiac compromise), had taken an investigational drug as part of another research study within 30 days prior to enrollment, were taking an mTOR inhibitor (rapamycin, sirolimus, or everolimus) at study enrollment, had a Subependymal Giant Cell Astrocytoma requiring medical or surgical treatment, had a history of epilepsy surgery, or had any contraindications to completion of study procedures such as MRI.	24-month classification of ASD based on ADOS-2 classification.
2017 Hahn	To identify common and unique aspects of early social communication by investigating descriptive patterns, differences, and the relationship with ASD risk and early social communication complexity across,	**IL-FXS**: Confirmed diagnosis of FXS by genetic report. **IL-Sibs**: Documentation of ASD diagnosis for an older full sibling. **LL:** ns	ns	No ASD outcome assessment was conducted; study was cross-sectional and assessed study participants between 7.5 and 14.5 months. ASD symptomology was measured using the AOSI.

	within, and between IL-FXS, IL-Sibs, and LL controls.			
2017 Bedford δ	To test whether infant neurocognitive markers (indexing eye-gaze processing and attention control) and 7/14-month AOSI scores can distinguish between IL-Sib 7-year ASD diagnostic status.	**IL-Sibs**: Have an older full/half sibling with a community clinical diagnosis of ASD confirmed using the DAWBA and SCQ by expert clinicians. **LL**: Have an older full/half sibling that does not meet criteria for ASD on the SCQ (does not meet instrument cut-off of ≥ 15).	ns	**36-month ASD assessment**: ns **7-year ASD assessment**: Non-blinded diagnostic assessment using DSM-5 criteria for ASD based on all previous study information (ADOS-2, ADI-R, WASI-II, VABS-II) conducted by four experienced researchers.
2017 McDonald β	To investigate whether (1) delays in social communication as measured by the AOSI may be observed within the first year of life for IL-TSC infants, and (2) if such delays are related to later ASD diagnostic status.	**For all participants**: Availability of AOSI and cognitive functioning data at 9 and/or 12 months, clinical outcome data at 18, 24, and/or 36 months. **IL-TSC**: Be diagnosed or present with TSC (based on clinical presentation). **LL**: ns	**IL-TSC**: ns **LL**: prematurity (<37 weeks gestational age), birth trauma, developmental concerns, or close family history of ASD or intellectual disability.	Clinical best estimate diagnosis based on ADOS scores. ASD outcome assessment was made at either 18-, 24-, or 36-months depending on availability of ADOS data (if a child had multiple clinical ASD outcome visits, the most recent ADOS score was used in ASD determination).
2018 Bussu δ	To (1) investigate longitudinal differences from 8-36 months between IL-Sibs with different developmental outcomes (typical, atypical, ASD) and LL controls, and (2) predict ASD or atypical development at 36-months an individual level for IL-Sibs using supervised machine learning classifier analysis based on 8- and 14-month study data.	**IL-Sibs**: Have an older biological sibling with ASD. **LL**: Have an older full sibling with typical development.	**All participants**: Lack of 36-month ADOS and/or 36-month clinical outcome evaluation.	36-month clinical consensus best estimate diagnosis considering 24 (ADOS, MSEL, VABS) and 36-month study data (ADOS, ADI-R, MSEL, VABS) using ICD-10 or DSM-5 criteria (dependent on study phase). Categorization of ASD using ICD-10 (atypical autism, PDD-unspecified, PDD-other) and DSM-5 criteria was considered similar following a review of ASD diagnoses by the clinical research lead.
2018 Sacrey α	To examine the agreement between parent and clinician ratings (on the APSI and AOSI respectively) regarding early symptoms of ASD in a sample of IL-Sibs.	**IL-Sibs**: Have an older sibling formally diagnosed with ASD that was confirmed by clinical assessment or a review of diagnostic records using DSM-IV-TR criteria, have undergone a 36-month diagnostic assessment for ASD, have APSI and AOSI outcome data at 12 and/or 18-months.	**All participants:** Born prior to 36 weeks gestation, birth weight <2500g, identifiable neurological or genetic conditions, or severe sensory or motor impairments.	36-month blind, independent best judgement diagnostic assessment using DSM-IV-TR criteria that considered ADOS, ADI-R, MSEL, and VABS data. Diagnoses were assigned by an expert clinician (developmental pediatrician, child psychiatrist, or clinical psychologist) with 10+ years of diagnostic experience.
2019 Bedford δ	To test the hypothesis that (1) infant regulatory function is negatively associated with traits of ASD, ADHD, but not callous unemotional traits, and	**IL-Sibs**: Have an older full/half sibling with a community clinical diagnosis of ASD, confirmed using the DAWBA and SCQ by expert clinicians.	ns	**7-year ASD assessment**: Diagnostic assessment using DSM-5 criteria was based on ADOS-2, ADI-R, SCQ, VABS-II, and WASI-II study data.

	(2) that regulatory function moderates the association between known infant markers (activity level for ADHD, early autism-like behaviours measured on the AOSI) with later traits of ADHD and ASD.	**LL**: Have an older full/half sibling that does not meet criteria for ASD on the SCQ (does not meet instrument cut-off of ≥ 15).		Diagnoses were assigned by four experienced researchers following a review on ASD symptomatology.
2020 Hahn γ	To describe ASD-associated behaviours in IL-DS infants 7-18 months old relative to LL controls 9-14 months old.	**IL-DS**: Recruited from three pilot studies examining infant phenotype in neurogenetic syndromes who recruited participants by flyers shared with parent groups, DS clinics, and/or other research studies. **LL**: Recruited from another study on the emergence of ASD in FXS. LL controls were matched to IL-DS based on sex at an individual level, and age at a group level.	**IL-DS**: ns **LL**: No ASD or other developmental disability (how this determination was made was not specified).	No ASD outcome assessment was conducted; study was a case-control cross-sectional study that assessed IL-DS 7-18 months old and LL controls 9-14 months old. ASD symptomology was instead measured using the AOSI.
2020 Zwaigenbaum α	To characterize behavioural signs of ASD in IL younger siblings of children with ASD and examine classification features of the AOSI.	**IL-Sibs**: Have an older sibling formally diagnosed with ASD confirmed by clinical assessment or a review of records using DSM-IV-TR criteria. **LL**: No 1st or 2nd degree relative with ASD.	**All participants**: Born <36 weeks gestation, birth weight <2500g, identifiable neurological or genetic conditions, severe sensory or motor impairments.	36-month blind, independent diagnostic assessment using DSM-IV-TR criteria was based on ADOS, ADI-R data and a review of other developmental assessments (MSEL, VABS).
2021 Zwaigenbaum α	To (1) identify distinct trajectories of ASD symptoms indexed by AOSI data from 6-18 months assessments, (2) examine the relationship between AOSI-informed trajectory group membership and 3-year clinical outcomes, and (3) to compare clinical features among IL-Sibs diagnosed with ASD across each trajectory with respect to sex ratio, language, cognitive and adaptive skills, and ASD symptom severity.	**IL-Sibs**: Have an older sibling diagnosed with ASD confirmed by clinical assessment or a review of diagnostic records using DSM-IV-TR criteria. **LL**: No 1st or 2nd degree relative with ASD.	**All participants**: Born <36 weeks gestation, birth weight <2500g, identifiable neurological or genetic conditions, severe sensory or motor impairments.	36-month independent, clinical best estimate diagnostic assessment using DSM-IV-TR criteria was based on all available study data (ADOS, ADI-R, MSEL, VABS) were conducted by an expert clinician blinded to prior study assessments. IL-Sibs and LL controls not meeting diagnostic criteria for ASD were further stratified into a 'delays or differences' category if they scored >1.5 SD below the mean on ≥1 MSEL subscales and/or if they scored >3 on the ADOS calibrated severity score.

ADI-R = Autism Diagnostic Interview-Revised, ADHD = Attention Deficit Hyperactivity Disorder, ADOS = Autism Diagnostic Observation Schedule, APSI = Autism Parent Screen for Infants, ASD = Autism Spectrum Disorders, CCS = communication complexity scale, DAWBA = Developmental and Wellbeing Assessment, DS = Down Syndrome, DSM = Diagnostic and Statistical Manual of Mental Disorders, FXS = Fragile X Syndrome, ICD-10 = International Classification of Diseases, 10th revision, IL = Infants at increased likelihood for ASD, IL-FXS = Infants diagnosed with Fragile X Syndrome, IL-TSC = Infants diagnosed with Tuberous Sclerosis Complex, IL-DS = Infants diagnosed with Down Syndrome, IRB = Institutional Review Board, ns = not specified, MRI = magnetic resonance imaging, PDDNOS = Pervasive Developmental Disorder Not Otherwise Specified, Proband = an IL-Sibs older sibling who is either diagnosed with or meets criteria for ASD, SCQ = Social Communication Questionnaire, SD = Standard deviation, TSC = tuberous sclerosis complex, VABS = Vineland Adaptive Behaviour Scale, WASI-II = Wechsler Abbreviated Scale of Intelligence-Second Edition
α = Conducted on Canadian Infant Sibling Study [CISS-1] participants

β = Conducted using some of the same study participants
γ = Conducted using some of the same study participants
δ = Conducted on British Autism Study in Infant Siblings [BASIS] participants

ASD outcome assessment. The assessment of ASD varied across the 17 included studies. Of the five studies (Capal et al., 2017, Zwaigenbaum et al., 2005, Estes et al., 2015, Roberts et al., 2016, Gilga et al., 2015) using 24-month ADOS scores as an outcome measure of ASD symptoms, only (Estes et al., 2015) conducted 24-month clinical best estimate diagnostic assessments using 24-month ADOS, ADI-R scores, and DSM-IV-TR criteria (ASD or pervasive developmental disorder [PDD] not otherwise specified). Eight studies (Gammer et al., 2015, McDonald et al., 2017, Bussu et al., 2018, Bussu et al., 2018, Zwaigenbaum et al., 2021, Bedford et al., 2016, Sacrey et al., 2018, Jeste et al., 2014) conducted 36-month ASD diagnostic assessments, though their assessment modalities varied. Bussu et al., 2018, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, and Sacrey et al., 2018 conducted independent or clinical consensus best estimate ASD diagnostic assessments based on ADOS, ADI-R, and cognitive, language, or developmental scales (MSEL, VABS) using ICD-10 (atypical autism, PDD-unspecified, PDD-other; Bussu et al., 2018) or DSM diagnostic criteria (Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, Sacrey et al., 2018). Gammer et al., 2015 conducted assessments based on ADOS and ADI-R data using ICD-10 diagnostic criteria (childhood autism, PDD), Bedford et al., 2016 based on ADOS and Social Communication Questionnaire (SCQ) data using ICD-10 criteria for autism (childhood autism, PDD), and McDonald et al., 2017 made clinical best estimate diagnoses based on ADOS data with no mention of using DSM or ICD-10 criterion. Jeste et al., 2014 assigned ASD diagnoses based on convergence of ADOS scores (taken at 18-, 24-, and 36-month assessments) and clinical judgement with no mention of ICD-10 or DSM criterion. Two studies (Bedford et al., 2017, Bedford et al., 2019) focused on ASD outcomes in early-to-mid childhood and conducted seven-year ASD diagnostic assessments using ADOS, ADI-R, and cognitive, language, or developmental scales (VABS-II, WASI-II). Finally, the remaining two studies (Hahn et al., 2017, Hahn et al., 2017) were cross-sectional in nature and did not assess for ASD outcomes (ASD diagnoses were not applicable based on their study objectives).

Age at AOSI Administration. Three studies administered the AOSI at 12- or 14-month time points (Capal et al., 2017, Roberts et al., 2016, Bedford et al., 2016). Two studies, Hahn et al., 2020 and Hahn et al., 2017, administered the AOSI over a wide range of ages (7-18 months) instead of at a specified time point. The remaining 12 studies administered the AOSI over multiple time points between 6 and 18 months.

Calculating AOSI Total Scores or AOSI Risk Markers. The AOSI can be scored using two different metrics: the AOSI Total Score constituting a summed score of items 1 to 18 on the scale, and AOSI Risk Markers constituting a tally of AOSI items 1 to 18 that score at least a 1 or higher (Bryson et al., 2008, Zwaigenbaum et al., 2005). It is important to note that these metrics are *not* the same thing. While 15 of 17 studies in this review calculate AOSI Total Scores for IL or LL study participants (barring Hahn et al., 2020 and Zwaigenbaum et al., 2005), only 2 of 17 studies report calculated AOSI Risk Marker scores (Hahn et al., 2020, Roberts et al., 2016).

AOSI Metrics Used in Sensitivity and Specificity Estimates. Overall, only six studies report whether or not they employed or calculated AOSI Total Score (Capal et al., 2017, Zwaigenbaum et al., 2020, Hahn et al., 2017) or AOSI Risk Marker cut points (Hahn et al., 2020, Zwaigenbaum et al., 2005, Roberts et al., 2016). Of these six studies, only four (Capal et al., 2017, Zwaigenbaum et al., 2005, Zwaigenbaum et al., 2020, Roberts et al., 2016) directly report their corresponding psychometric estimates (sensitivity/specificity) or the data needed to calculate them. Two studies (Zwaigenbaum et al., 2005, Roberts et al., 2016) used AOSI Risk Markers for their psychometric estimates, and two (Capal et al., 2017, Zwaigenbaum et al., 2020) used AOSI Total Scores.

How AOSI Total Scores or AOSI Risk Markers have been used in these four studies varied as no consistent cut point for either metric was employed. Two studies (Zwaigenbaum et al., 2005, Roberts et al., 2016) used a cut-point of ≥7 or >7 AOSI Risk Markers respectively to predict 24-month ASD classification, whereas Capal et al., 2017 and Zwaigenbaum et al., 2020 computed multiple AOSI Total Score cut points to predict 24-month or 36-month ASD classification or diagnosis respectively. That is, Capal et al., 2017 provided a range of possible Total Score cut points based on 12-month assessment data while Zwaigenbaum et al., 2020 computed a range of possible Total Score cut points for each time point they administered the AOSI (6, 9, 12, 15, and 18 months).

Though not reporting AOSI cut points and their corresponding psychometric estimates, Zwaigenbaum et al., 2021 imported AOSI Total Score data from participants assessed at 6, 9, 12,

15, and 18 months into STATA to generate semi-parametric group-based trajectory models that reflect sub-populations of participants. After selecting for a 3-group quadratic model, Zwaigenbaum et al., 2021 compared participant membership in these groups (Group 1 = 'Low and stable,' Group 2 = 'Intermediate and stable,' and Group 3 = 'Inclining') in their trajectory model against later 36-month ASD diagnostic outcomes (IL siblings diagnosed with ASD, IL siblings not diagnosed with ASD, LL controls). While not reporting AOSI cut points and their corresponding psychometric estimates, the sensitivity and specificity of these trajectory models relative to 36-month ASD outcomes was documented. Table 2.03 provides more details.

Table 2.03 | AOSI Analyses and Psychometric Estimates

Article	How was the AOSI used/applied?	Total Score or Risk Marker?	IL Group	Timepoint	Cut Point	Sensitivity	Specificity
2005 Zwaigenbaum *	AOSI data taken from 6 and 12-month assessments was compared against IL/LL study participants based on 24-month ADOS classification using One-way ANOVA analysis with follow-up multiple comparisons. 12-month AOSI scores were used to predict 24-month ADOS classification.	Risk Markers	IL-Sibs	12 months	≥7	0.84	0.98
2016 Roberts	12-month AOSI scores were explored across IL/LL groups using one-way Kruskal Wallis analysis with Dunn post hoc pairwise comparisons. Fischer's exact test was used to investigate (1) group differences in the proportion of IL/LL infants who flagged positive on the AOSI relative to those who did not, and (2) item-level group differences. 12-month AOSI scores were used to predict 24-month ADOS classification.	Risk Markers	IL-FXS	12 months	>7	0.57	1.00
			IL-Sibs	12 months	>7	1.00	0.57
2017 Capal	12-month AOSI scores were used as a predictor variable in logistical regression models against 24-month ADOS-2 and ADI-R outcome data (IL-TSC participants classified with/without ASD). 12-month AOSI Total Score cut points were examined with respect to later 24-month ADOS classification.	Total Score	IL-TSC	12 months	8	0.67	0.70
					9	0.67	0.73
					10	0.58	0.77
					11	0.51	0.82
					12	0.48	0.82
					13	0.39	0.89
					14	0.36	0.91
2020 Zwaigenbaum *	AOSI data taken from 6, 9, 12, 15, and 18-month assessments were compared using linear mixed modelling. ROC curve analysis assessed longitudinal associations between AOSI Total Score data at each timepoint with later 36-month clinical outcomes. Optimal Total Score cut points were calculated using Youden indices. AOSI scoring data was compared across IL-Sibs with/without ASD, with Fischer's exact test calculated to compare the percentage of IL-Sibs correctly identified at 24- and 36-month assessments.	Total Score	IL-Sibs	6 months	7	0.57	0.51
				9 months	8	0.60	0.53
				12 months	7	0.52	0.74
				15 months	10	0.41	0.90
				18 months	6	0.73	0.65
2021 Zwaigenbaum *	Trajectory modeling based AOSI Total Scores data was derived using Stata group-based modelling approach on data taken at 6, 9, 12, 15, and 18-month assessments. The relationship between AOSI trajectory group membership in the finalized trajectory model and 36-month clinical outcomes was examined to assess accuracy of group membership relative to ASD diagnosis. Clinical features of participants diagnosed with ASD at 36 months were compared by trajectory group using one-way ANOVAs.	Longitudinal Total Score data from 6-18 months	IL-Sibs	n/a	Group 1	0.28	0.94
				n/a	Group 2	0.68	0.59

ADI-R = Autism Diagnostic Interview-Revised, ADOS = Autism Diagnostic Observation Schedule, ANOVA = Analysis of Variance, AOSI = Autism Observation Scale for Infants, IL = infants at increased likelihood for ASD diagnosis, IL-FXS = IL infants with Fragile X Syndrome, IL-TSC = IL infants with Tuberous Sclerosis Complex, IL-Sibs = IL infant siblings, LL = infants at low likelihood for ASD diagnosis, ROC = receiver operating characteristic.
* = Conducted on Canadian Infant Sibling Study [CISS-1] participants

For additional methodological considerations including article study design, AOSI reliability data (inter-rater, item-level agreement between coders, etc.), how infant sibling studies defined the older sibling (proband) as having ASD, and what inclusion/exclusion criteria were employed, see Appendix 1.

Main Findings

AOSI Sensitivity and Specificity Estimates for Infant Siblings. The cut points and AOSI metrics used varied across studies which makes it difficult to compare sensitivity, as described in Table 2.03. Of the four studies which assessed infant siblings, two studies used AOSI Risk Marker cut points of ≥7 or >7 (Zwaigenbaum et al., 2005, Roberts et al., 2016) had sensitivity estimates of 0.84 and 1.00 respectively. Zwaigenbaum et al., 2020, who assessed different AOSI Total Score cut points across a range of time points, had sensitivity values ranging between 0.41 and 0.73. For Zwaigenbaum et al., 2021 who used trajectory-based grouping based on AOSI Total Scores, sensitivity estimates for the inclining trajectory and inclining + intermediate trajectory groups were 0.28 and 0.68 respectively.

Though specificity estimates were largely higher than sensitivity estimates for infant siblings, variation was still noted. Two studies that used AOSI Risk Marker cut points of ≥7 or >7 (Zwaigenbaum et al., 2005, Roberts et al., 2016) reported specificity estimates of 0.98 and 0.57 respectively. Zwaigenbaum et al., 2020 who assessed different AOSI Total Score cut points across a range of time points reported specificity estimates ranging between 0.51 and 0.90. For Zwaigenbaum et al., 2021 who used trajectory-based grouping based on AOSI Total Scores, specificity estimates for the inclining trajectory and inclining + intermediate trajectory groups were 0.94 and 0.59 respectively.

AOSI Sensitivity and Specificity Estimates for FXS and TSC Infants. In addition to there being fewer psychometric estimates available for FXS and TSC infants, cut points and metric used varied relative to infant siblings as described in Table 2.03. Using the AOSI Risk Marker cut point of >7, Roberts et al., 2016's data led to a single calculated sensitivity estimate of 0.57 for FXS infants. For Capal et al., 2017 who report a range of 12-month AOSI Total Score cut points in TSC infants, sensitivity estimates ranged between 0.36 and 0.67.

Specificity estimates for infants with FXS and TSC resembled those for infant siblings. Using the AOSI Risk Marker cut point of >7, Roberts et al., 2016's data led to a single calculated specificity estimate of 1.00 for FXS infants. For Capal et al., 2017, specificity estimates for a variety of 12-month AOSI Total Score cut points ranged between 0.70 and 0.91.

AOSI Total Score Comparison. As shown in Figure 2.2 (scatterplot), a consistent pattern of AOSI Total Scores emerges at 12 months of age, with IL-ASD groups (TSC, FXS, DS, and Infant Siblings with ASD) consistently showing higher scores compared to LL and IL non-ASD comparison groups.

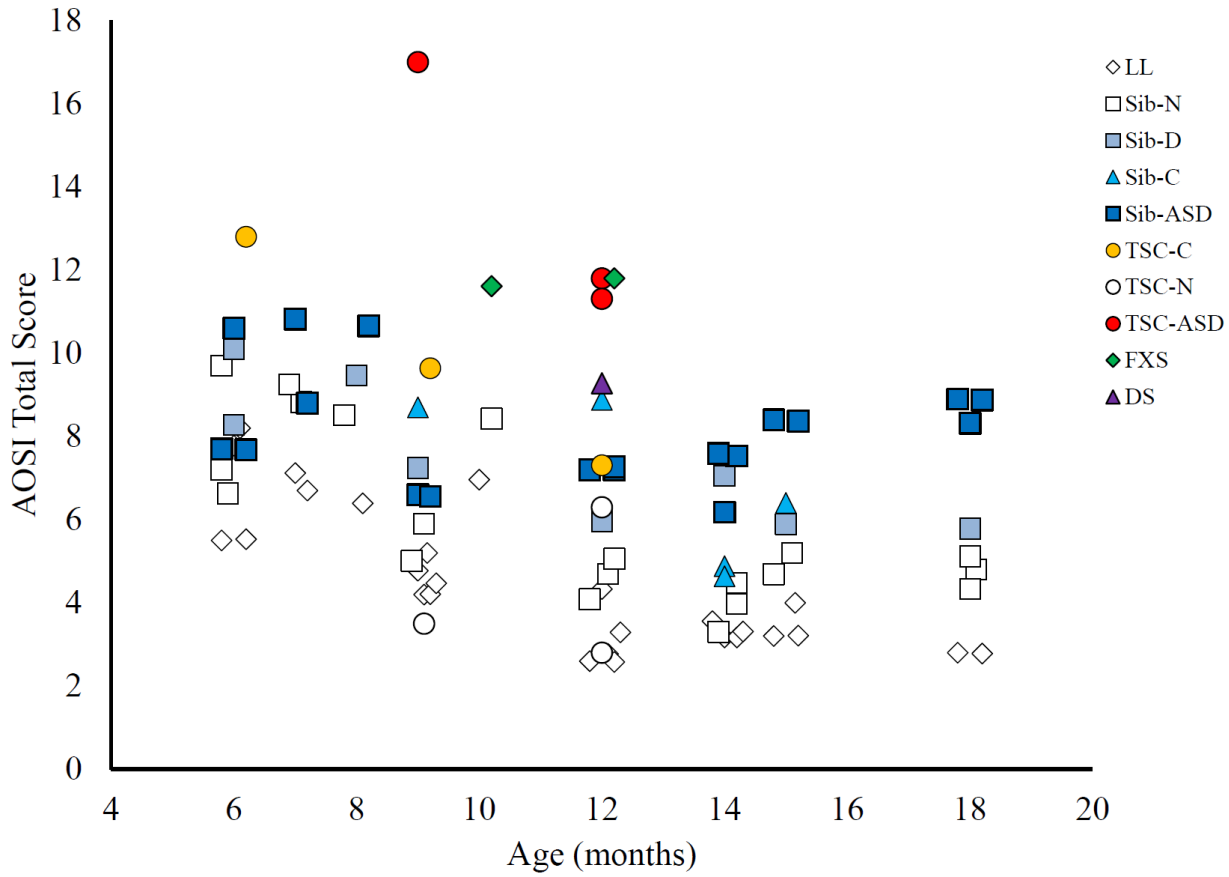

Figure 2.02 | Scatterplot of Age (in Months) by AOSI Total Score. Note that while different IL-ASD groups are denoted by the filled symbols, LL and IL non-ASD groups are denoted by the open symbols. DS = infants with Down Syndrome, FXS = infants with Fragile X Syndrome, IL = infants at increased likelihood for ASD, LL = Low likelihood controls, Sibs-N = IL-siblings not diagnosed with ASD, Sibs-D = developmentally delayed IL-siblings, Sibs-C = combined grouping of Sibs-D and Sibs-N for studies which report it, Sibs-ASD = IL-siblings diagnosed with ASD, TSC = infants with tuberous sclerosis complex, TSC-N = LL controls in TSC studies not diagnosed with ASD, TSC-C = combined grouping of all TSC participants (ASD not separated out), TSC-ASD = IL-TSC infants diagnosed with ASD

Meta-Analyses

LL Controls and IL-ASD

Between 6 and 10 Months. A total of five comparisons of AOSI Total Scores were included in this meta-analysis. There was a significant effect of AOSI Total Score, suggesting that the IL-ASD group had higher AOSI Total Scores compared to the LL control group (Cohen's $d = 1.01$, 95% CI = 0.49 - 1.52, $z = 3.82$, $p < 0.001$, Figure 2.03a). High heterogeneity was seen among the included studies (I^2 heterogeneity statistic = 81.2%); thus, a random effects model was adopted to pool the relevant data and explore subgrouping analyses to determine any differential effects of the IL-ASD subgroup on AOSI Total Score. As shown in Figure 2.03a, all three IL-ASD groups (Sib-ASD, FXS, and TSC-ASD) produced significant effects (all p's < 0.01), resulting in higher AOSI Total Scores compared to LL controls. Funnel plot analyses on Cohen's d for AOSI Total Score demonstrated symmetry, but we still assessed for the presence of bias (Figure 2.03a). Trimming the set of data systematically removes each 'outlier' one at a time and recalculates the resulting Cohen's d. The resultant value was changed following the trim and fill analyses, suggesting 2 missing studies. Evaluation of the Egger test provided little evidence of small study effects impacting Cohen's d (bias coefficient = 5.43, standard error = 2.36; $t = 2.30$, $p = 0.15$).

Between 12 and 14 Months. A total of four comparisons of AOSI Total Scores were included in this meta-analysis. There was a significant effect of AOSI Total Score, suggesting that the IL-ASD+ group had higher AOSI Total Scores compared to the LL control group (Cohen's $d = 1.15$, 95% CI = 0.90 - 1.40, $z = 8.96$, $p < 0.001$, Figure 2.03b). Though low heterogeneity was seen among the included studies (I^2 heterogeneity statistic = 14.8%); we still adopted a random effects model to pool relevant data and explore subgrouping analyses to determine any differential effects of the IL-ASD subgroup on AOSI Total Score. As shown in Figure 2.03b, all three IL-ASD+ groups (Sib-ASD, DS, and TSC-ASD) produced significant effects (all p's < 0.03), resulting in higher AOSI Total Scores compared to LL controls. Though funnel plot analyses on Cohen's d for AOSI Total Score demonstrated symmetry, we still assessed for the presence of bias (Figure 2.03b). The Cohen's d value was unchanged following the trim and fill analyses, suggesting no bias. Evaluation of the Egger test provided little evidence of small study effects impacting Cohen's d (bias coefficient = -0.01, standard error = 1.37; $t = 0.00$, $p = 0.99$).

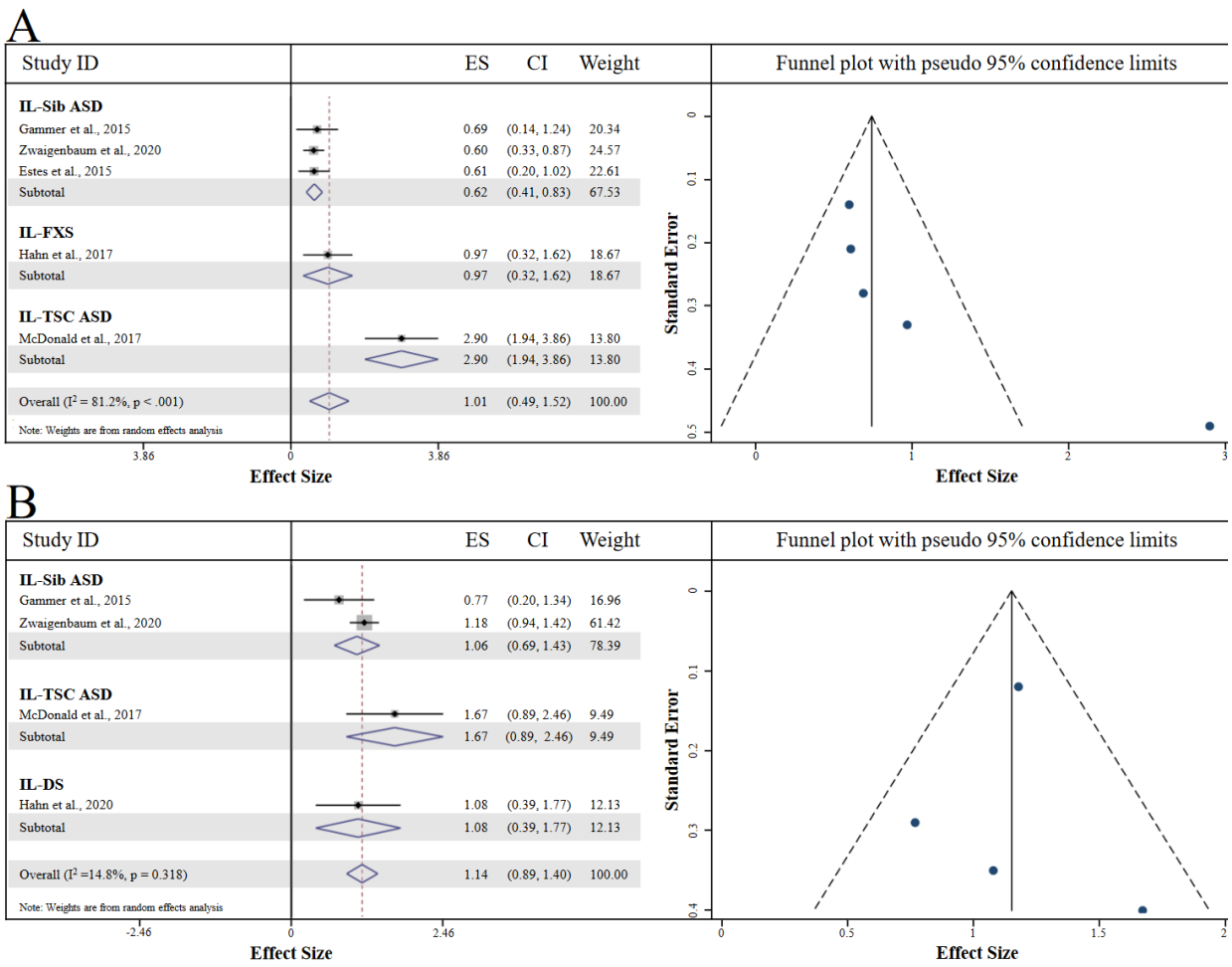

Figure 2.03a, b | Meta Analyses Comparing LL Controls to IL-ASD Samples (left) with the Trim and Fill Plot (right). A = for ages 6-10 months, B = for ages 12-14 months.

IL Non-ASD Combined Controls and IL-ASD

Between 6 and 10 Months. A total of four comparisons of AOSI Total Scores were included in this meta-analysis. There was a significant effect of AOSI Total Score, suggesting that the IL-ASD group had higher AOSI Total Scores compared to the IL control group (Cohen's $d = 0.89$, 95% CI = 0.03 - 1.75, $z = 2.02$, $p = 0.004$, Figure 2.04a). High heterogeneity was seen among the included studies (I^2 heterogeneity statistic = 90.1%); thus, a random effects model was adopted to pool the relevant data and explore subgrouping analyses to determine any differential effects of the IL-ASD subgroup on AOSI Total Score. As shown in Figure 2.04a, two of the three IL-ASD groups produced significant effects resulting in higher AOSI Total Scores compared to IL controls for FXS ($p = 0.05$) and TSC-ASD ($p < 0.001$). Funnel plot analyses on Cohen's d for

AOSI Total Score demonstrated symmetry, but we assessed for the presence of bias regardless (Figure 2.04a). The Cohen's *d* value was changed following trim and fill analyses, suggesting two missing studies. Evaluation of the Egger test provided little evidence of small study effects impacting Cohen's *d* (bias coefficient = 5.43, standard error = 2.36; $t = 2.30$, $p = 0.15$).

Between 12 and 14 Months. A total of four comparisons of AOSI Total Scores were included in the meta-analysis. There was a significant effect of AOSI Total Score, suggesting that the IL-ASD group had higher AOSI Total Scores compared to the IL control group (Cohen's $d = 0.79$, 95% CI = 0.42 - 1.17, $z = 4.15$, $p < 0.001$, Figure 2.04b). Moderate heterogeneity was seen among the included studies (I^2 heterogeneity statistic = 59.9%); thus, a random effects model was adopted to pool relevant data and explore subgrouping analyses to determine any differential effects of the IL-ASD subgroup on AOSI Total Score. As shown in Figure 2.04b, both IL-ASD groups (Sib-ASD and TSC-ASD) produced significant effects (all *p*'s < 0.01), resulting in higher AOSI Total Scores compared to IL controls. Funnel plot analyses on Cohen's *d* for AOSI Total Score demonstrated symmetry, but we assessed for the presence of bias regardless (Figure 2.04b). The Cohen's *d* value was unchanged following the trim analyses, but the fill analysis suggested there was 1 missing study. Evaluation of the Egger test provided little evidence of small study effects impacting Cohen's *d* (bias coefficient = 1.91, standard error = 1.26; $t = 1.52$, $p = 0.27$).

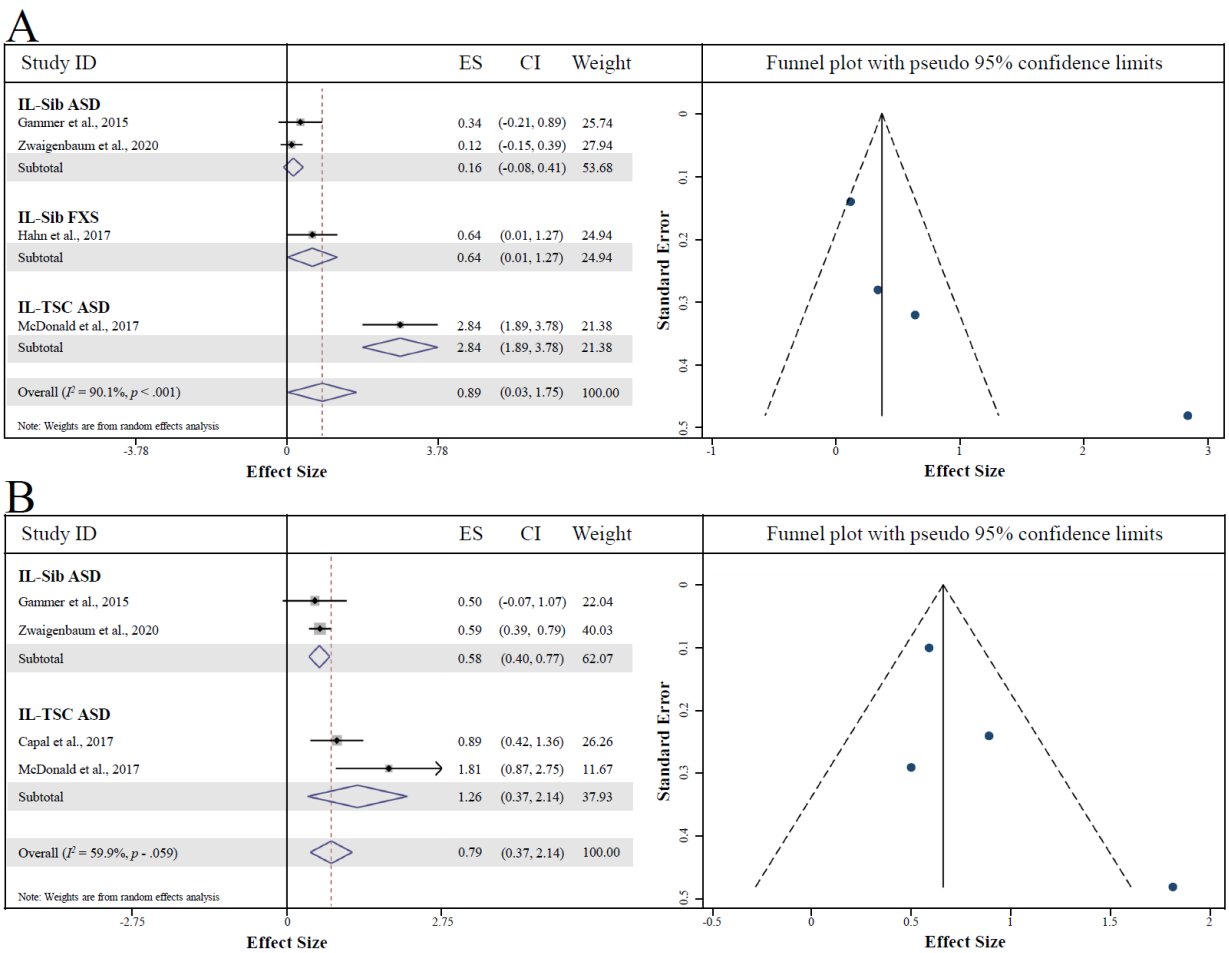

Figure 2.04a, b | Meta Analyses Comparing IL Non-ASD Combined Controls to IL-ASD Samples (left) with the Trim and Fill Plot (right). A = for ages 6-10 months, B = for ages 12-14 months.

IL-DD/IL-Typical and IL-ASD

Meta-analyses were also performed on studies that broke the IL-N ASD group into those who met criteria for developmental delay (IL-DD) and those who showed typical development (IL-Typical). These data are presented in Appendix 1.

Checklist of Bias and Quality of Study Methodology

Table 2.04 provides a visual overview of the methodological strengths and weaknesses of the 17 studies included in this review. Overall, there was no consistent approach with respect to classification or diagnosis of ASD (both for age and measures used), inclusion or exclusion criteria for participants, choice of comparison groups (or lack thereof), whether AOSI item-level, Risk Marker, or Total Score data are reported, and participant demographics (age, SES, ethnicity,

parental age, etc.). A consideration of each of these factors is important when making methodological decisions.

Table 2.04 | Bias and Quality Checklist for Included Studies

	Zwaigenbaum 2005	Jeste 2014	Estes 2015	Gammer 2015	Gilga 2015	Bedford 2016	Roberts 2016	Capal 2017	Hahn 2017	Bedford 2017	McDonald 2017	Bussu 2018	Sacrey 2018	Bedford 2019	Hahn 2020	Zwaigenbaum 2020	Zwaigenbaum 2021
Objective / purpose																	
Question	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓
Hypothesis	-	-	-	-	✓	-	-	-	-	-	✓	✓	-	✓	-	-	-
Study Design																	
Cross-sectional	-	-	-	-		-	-	-	✓	-	-	-	-	-	✓	-	-
Longitudinal	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	-	✓	✓
Exclusion Criteria																	
Birth weight	-	-	✓	-	✓	-	-	-	-	-	-	-	✓	-	-	✓	✓
Term birth	-	✓	✓	-	✓Θ	-	✓	✓	-	✓Θ	✓	-	✓	-	-	✓	✓
Genetic causes	-	-	-	✓	-	-	✓	-	-	-	-	-	✓	-	-	✓	✓
Other conditions	-	✓	✓	✓	-	-	✓	✓	-	-	✓	-	✓	-	-	✓	✓
Recruitment																	
Same cohort	✓	✓	-	✓	✓*	✓*	✓	✓	✓	-	-	-	✓	✓	-	✓	✓
Sample calculation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Control group																	
LL controls	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	✓
IL controls	✓	✓	✓	✓	-	-	✓	✓	✓	✓	✓	✓	✓	-	-	✓	✓
Experimental group																	
Infant siblings	✓	-	✓	✓	✓	✓	✓	-	-	✓	-	✓	✓	✓	-	✓	✓
Infants with FXS	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-	-	-	-
Infants with TSC	-	✓	-	-	-	-	-	✓	-	-	✓	-	-	-	-	-	-
Infants with DS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Demographics																	
Sex	-	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SES	-	-	-	✓	-	-	-	-	-	✓#	-	-	✓#	-	-	✓	✓
Ethnicity	-	-	✓	-	-	-	-	-	✓	-	✓	-	-	-	✓	✓	✓
Parental age	-	-	-	-	-	-	-	-	-	-	-	-	✓#	-	-	-	-
Outcome Assessment																	
2 years	✓	-	✓	-	✓	-	✓	✓	-	✓	✓	✓		✓Δ	-	-	-
3 years	-	✓	-	✓	-	✓	-	-	-	✓	✓	✓	✓	✓Δ	-	✓	✓
7 years	-	-	-	-	-	-	-	-	-	✓	-	-	-	✓	-	-	-
Gold-standard?	-	-	-	✓	-	-	-	✓	-	✓	✓	✓	✓	✓	-	✓	✓
Blinded?	-	-	✓α	-	-	-	-	✓α	-	✓	-	✓	✓	-	-	✓	✓
Diagnostic Criteria																	
DSM	-	-	✓	-	-	-	-	-	-	✓	-	✓	✓	✓	-	✓	✓

ICD	-	-	-	✓	-	✓	-	-	-	-	-	✓	-	-	-	-	-
Age of AOSI Administration																	
< 12 months	-	✓	✓	✓	✓		-	-	-	✓	✓	✓	-	✓	-	✓	✓
≥ 12 months	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	-	✓	✓
Range <12 >	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	✓	-	-
AOSI Administrations																	
One	-	-	-	-	-	✓	✓	✓	✓	-	-	-	-	✓	✓	-	-
Two +	✓	✓	✓	✓	✓	-	-	-	-	✓	✓	✓	✓	-	-	✓	✓
Statistical analyses																	
Covariates	-	-	✓	✓	-	✓	✓	✓	✓	-	✓	✓	✓	✓	-	-	-
Post-hoc	✓	-	✓	✓	-	-	✓	-	✓	-	✓	✓	-	-	-	✓	✓
Included in meta-analysis?	-		✓	✓	-	-		✓	✓	-	✓	✓	-	-	✓	✓	✓
AOSI content used																	
Item-level data	-	-	-	✓	-	-	✓	✓	-	-	✓	-	✓	-	-	✓	-
Total Score data	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Psychometric properties																	
Sensitivity	✓	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-	✓	✓
Specificity	✓	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-	✓	✓
Cut-off score	✓	-	-	-	-	-	✓	✓	-	-	-	-	✓	-	-	✓	✓
IRR assessment?	✓	-	-	-	-	✓	✓	-	✓	✓	-	-	-	✓	-	-	-

AOSI = Autism Observation Scale for Infants, DS = Down Syndrome, DSM = Diagnostic and Statistical Manual of Mental Disorders, FXS = Fragile X Syndrome, ICD = International Classification of Diseases, IRR = inter-rater reliability, SES = socioeconomic status, TSC = Tuberous Sclerosis Complex. Note: checkmarks for each item are not weighted and instead simply denote the presence or absence of the item.

$^{\alpha}$ = Not all clinicians responsible for assigning a diagnosis were blinded group status

* = Control sample came from volunteer databases, but it is not clear whether they were recruited from the same cohort as IL participants

Θ = Term birth was described only for control participants

= Family income was calculated and used in statistical calculations but not described or reported (i.e., no direct reporting of SES for IL vs LL groups)

Δ = While participants were assessed these ages, the assessment procedure was not detailed or described outside of being mentioned.

Discussion

This systematic review and meta-analysis focused on previous studies assessing classification properties and group differences on the AOSI across different IL infant populations. Four IL populations were identified in this review: infants with FXS, TSC, DS, and infant siblings of children with ASD. The review had three main findings. First, although five studies reported individual classification properties, sensitivity and specificity estimates were not comparable due to the different metrics, methodologies, and cut point scores used. Second, stable group differences emerged between LL and IL non-ASD control groups and IL-ASD groups by 12 months of age. Third, meta-analyses identified a large effect size for comparisons between LL control and IL-ASD samples, and a moderate effect size for comparisons of IL non-ASD and IL samples with signs or diagnoses of ASD. While the AOSI was used as a measure of ASD symptomology in these different populations, it is currently unclear whether ASD diagnosis in FXS, TSC, DS, and infant siblings all reflect the same underlying and neurobiological impairments seen in non-syndromic or idiopathic ASD (Abbeduto et al., 2014). Gaining a better understanding of how the AOSI performs across different populations of infants who are at increased likelihood for ASD is vital to our understanding and characterization of the emergence of ASD during early childhood.

Classifying and Diagnosing ASD in IL Samples

ASD outcomes were assigned based on either 24-month ADOS classification or 36-month blinded diagnostic assessments. When assessing for ASD, the age of the child and the comprehensiveness of the assessment are important. Infant behaviour can be affected by situational factors, such as their state of alertness (Jones et al., 2014), time of day, and biological state (e.g., hunger or sleepiness; McNally et al., 2015). Gold-standard ASD diagnostic assessments (defined as use of validated observational and interview measures such as the ADOS and ADI-R in conjunction with expert clinical judgement; Kaufman 2022) utilize a broad scope of clinical information before assigning a diagnosis. Use of a single observational measure to determine ASD outcome is therefore a poor proxy and likely suffers from decreased sensitivity, specificity, and diagnostic stability (Jones et al., 2014). Furthermore, in IL infant siblings, although diagnostic stability of early ASD diagnosis at 18- and 24-months is high at 93% and 82% respectively, early classification suffers from low sensitivity (Ozonoff et al.,

2015). At 18- and 24-month assessments, 63% and 41% of children who are later diagnosed with ASD at 36-months are missed (Ozonoff et al., 2015). Since 24-month clinical best estimate ASD diagnosis can miss such a substantial percentage of children later diagnosed at 36-months, 24-month classification of ASD based on ADOS scores alone are likely even less informative.

As reviewed here, ASD classification using 24-month ADOS scores are also plagued with inconsistency. Studies using a single observational measure like the ADOS to classify ASD need to provide details about what scoring algorithms were used. Of the four studies in this review using 24-month ADOS scores as their main ASD outcome determinant (Capal et al., 2017, Zwaigenbaum et al., 2005, Roberts et al., 2016, Gilga et al., 2015), only Roberts et al., 2016 states what ADOS severity score thresholds was employed. The ADOS calibrated severity score is a metric with values ranging between 1 and 10, with higher scores representing increasing severity of ASD-related symptoms (Duda et al., 2014). The ADOS-2's scoring algorithm indicates that a severity score of 7 equates to the broader autism spectrum, and 8 or higher associated with autism (Lord et al., 2012). Reporting what ADOS severity score threshold is used is crucial as it may influence the characteristics of ASD being classified in study participants. For instance, while Roberts et al., 2016 employed a relatively inclusive severity score threshold of >4 flagging participants with at least mild-to-moderate ASD symptomology (Lord et al., 2012), other studies have either used more stringent severity score cut thresholds of >8 (Sedgewick et al., 2019). In this review, while Capal et al., 2017 and Zwaigenbaum et al., 2005 classified infants with ASD based on 24-month ADOS scores, they did not detail what scoring thresholds were used. Given that we do not know if Capal et al., 2017, Zwaigenbaum et al., 2005, Roberts et al., 2016, or Gilga et al., 2015 all used the same ADOS scoring threshold, it is plausible their study results are influenced by labelling participants with ASD of differing severity levels. Studies relying on the ADOS alone to classify ASD should specify what scoring thresholds are used to allow comparison of similarly characterized outcome groups.

Validation of the AOSI in Different IL Samples

When extending the use of an established scale to a new context, caution must be practiced; it cannot be assumed that a scale validated in one population can be equally applied in a different population without initial validation (Streiner et al., 2014). Each time a scale is used in a new

context, it is necessary to establish psychometric properties and validity of the inferences drawn from them (Streiner et al., 2014). In addition, in pursuit of optimal reliability and validity, scales often need to be revised – changes may be subtle or substantial (Streiner et al., 2014). For example, FXS infants with ASD have significantly higher motor impairments relative to infant siblings with ASD (Roberts et al., 2016). Whether such variance in item-level scoring is present across the different IL populations is not clear. Possible alterations to the AOSI may be warranted to capture population differences that may be indicative of later ASD diagnoses. We suggest that item-level data should be reported to assist this effort.

Sensitivity, the ability of a test to correctly identify an individual as having a particular condition, and specificity, the capability of a test to correctly identify individuals as not having that condition, are inversely proportional (Parikh et al. 2008). The AOSI cut point should optimize both sensitivity and specificity (Akobeng 2007). Although the best tests are both highly sensitive and specific, this is not always feasible in practice (Akobeng 2007) as trade-offs may exist between valuing high sensitivity over specificity (or vice versa, Trevethan 2017). In situations where it is vital that a diagnosis is not missed (e.g., diseases with high mortality), high sensitivity is sought. In contrast, if the consequences of false positives are serious (e.g., psychological implications of a false HIV diagnosis), high specificity is sought (Akobeng 2007).

AOSI sensitivity and specificity estimates for infant siblings varied across the papers reviewed here. Although Zwaigenbaum et al., 2005 and Roberts et al., 2016 used a similar cut point (≥7 and >7 AOSI Risk Markers respectively), their estimates of specificity differed. This likely stemmed from study differences in inclusion/exclusion criteria, participant demographics, and use of 24-month classification assessment (which may be less sensitive to children with milder ASD presentation). The issue of psychometric properties is further muddied by the AOSI metric used. Rather than AOSI Risk Markers, sensitivity and specificity estimates from Zwaigenbaum et al., 2020 were calculated using the AOSI Total Score, which may account for differences in sensitivity and specificity. The original Zwaigenbaum et al., 2005 article introducing the AOSI published preliminary psychometric estimates based on a cut point of *≥7 AOSI Risk Markers*, not the AOSI Total Score. The two metrics are not comparable. AOSI Risk Markers denote the total number of AOSI items that scored '1' or higher and range from 0-16 (Zwaigenbaum et al.,

2005). This differs from the AOSI Total Score, the summed score of all AOSI items and ranges from 0-38 (Bryson et al., 2008). While there are many studies exploring group differences using the AOSI in IL infant sibling populations, few studies directly report the scale's psychometric properties, or the data required to calculate them. This leads to challenges with evaluating what the optimal cut points are for the scale based on currently available evidence. Given that clinical measures should have cut points yielding sensitivity and specificity values exceeding 0.70 (Zwaigenbaum et al., 2015) and ideally between 0.80 and 0.90 if ascribing to Bayes Theorem (Medow & Lucey, 2011), determination of what cut point sensitivity and specificity thresholds are acceptable or even achievable given the cost of false positives and negatives should be considered when the AOSI is used in different IL infant contexts.

Considerations for Future Data Collection and Analyses

First, when assessing the utility of a scale in a novel context, it is paramount to control for demographic factors that can confound results. For example, low SES is linked to poor outcomes in many areas of early development (Bradley & Corwyn, 2002, Chen et al., 2019, Lawson et al., 2018, Freitas et al., 2013) and can be affected by other related cofactors, such as ethnicity (Bradley & Corwyn, 2002). Papers included in this review may be biased due to a failure to control for the potential impacts of factors such as family SES and ethnicity. Finally, while ASD has been known to be related to advancing paternal age (Puelo et al., 2012), none of the studies in this review included it as a possible covariate. Future studies should include family demographics in their analysis to promote generalizability of findings.

Second, reliability and validity need to be reassessed in novel contexts**.** The presentation of ASD in FXS, TSC, DS, and infant siblings may manifest differently (Abbeduto et al., 2014). Thus, assessment of reliability and validity of ASD symptom assessment tools is warranted in novel IL populations. Reporting item-level data may aid in the identification of emergent patterns across IL populations (e.g., FXS infants with ASD have increased motor impairments relative to ASD infant siblings; Roberts et al., 2016).

Third, more stringent and explicitly stated inclusion and exclusion criteria are needed. Differences in exclusion criteria, for example, gestational age, birthweight, and the other

neurological conditions, impact comparability and generalizability of results. Inclusion and exclusion criteria should be selected based on the study question. For example, preterm infants are at 3-4 times increased likelihood for ASD diagnosis relative to the general population (7% vs 0.76% respectively; (Agrawal et al., 2018, Chen et al., 2006) and thus, should be considered a separate IL group. Premature infants also experience cognitive impairment that have a developmental interaction with SES (Tong et al., 2006, Torche & Echevarría, 2011).

Fourth, AOSI cut points (for the Total Score or number of Risk Markers) need to be reported. A paucity of literature addresses the AOSI's prediction of ASD in FXS, TSC, and DS populations. When using the AOSI, it is imperative to describe explicitly how the measure was used, including cut points (both the actual cut point used and the metric [AOSI Total Score or Risk Markers]). Failure to do so can draw into question the validity of study results and undermine the generalizability of findings to other contexts.

Fifth, non-ASD or IL control groups are needed. Lack of appropriate control group(s) negates the possibility of investigating whether patterns of results are group or syndrome-specific (i.e., associated with IL status or ASD diagnosis) or reflect typical child development. Are the reported results which attempt to characterize ASD features specific to a particular IL population (e.g., infant siblings, FXS, TSC, DS) or is it possible that the reported findings are not specific to ASD or IL populations and instead are a feature of typical development? Future studies should include non-clinical comparison groups when using the AOSI with IL infant populations.

Sixth, it is important to consider age at outcome assessment. It is imperative when investigating early features of a condition like ASD that results are accurately attributed to the condition of interest. Diagnostic assessments at 24-months are less sensitive (Ozonoff et al., 2015). This is likely due to different groups of children being identified at 24- and 36-months (i.e., children diagnosed with ASD at 24-months generally have more severe symptom presentation than children diagnosed at 36-months; Zwaigenbaum et al., 2020). Since the goal of these studies is early detection, using 24-month outcome assessments (although likely to only capture a specific group of ASD children) is still pertinent.

Seventh, the age at which the AOSI is administered should be determined by the research question. AOSI Total Scores were not able to distinguish between IL and LL infants when administered at 6 and 9 months across the included studies. Given that meta-analyses report clear evidence of group differences emerging by 12-months of age and older among IL-ASD and LL or IL non-ASD infant populations, reliance of AOSI scores before 12-months for classification purposes is not recommended. If studies aimed to investigate the emergence of ASD symptoms across the developmental timespan from infancy to age at diagnosis, earlier AOSI administrations (at 6 and/or 9 months) could be warranted.

Limitations

This is the first systematic review and meta-analysis evaluating the use and classification properties of the AOSI across IL infant populations. This review has several limitations. Though we conducted a thorough search for studies using the AOSI in IL infants in six databases, it is possible that we still may have missed some AOSI papers. In addition, although most studies identified using the AOSI were on IL infant siblings, few studies have applied the measure to FXS, TSC, DS, and other IL populations. It is important to note, however, that several of the studies included in this review were the first to use the AOSI in their non-infant sibling IL cohort.

Conclusion

This review summarized the results of research that assessed group differences and psychometric performance of the AOSI in populations of infants at IL for a diagnosis of ASD. Overall, group differences on the AOSI were consistently found by 12 months of age between IL-ASD and LL or IL non-ASD groups. However, individual classification properties were less promising, likely due to methodological differences. As such, it is critical to investigate further the psychometric properties (i.e., sensitivity and specificity) of the AOSI across different IL populations in which phenotypic differences may exist. Ensuring study design and methodology are robust and transparent to not only protect against biasing factors, but also allow for comparison with similar or follow-up studies is important. Understanding the differences in methodology can inform future studies as researchers continue to investigate the early presentation of signs of ASD across

diverse IL populations. Overall, the AOSI shows promise as an early detection tool for different infant groups at IL for ASD.

Chapter 3: Study Two

Background

Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition characterized by differences in social-communication and the presence of restricted interests, repetitive behaviours, and/or atypical responses to sensory input (APA, 2013). The Centre for Disease Control and Prevention's (CDC) estimate for the community prevalence rate of ASD in the United States is 1 in every 44 children by age 8 (Maenner et al., 2021). Though sex differences in ASD diagnostics have been observed (males being four times as likely than females to be diagnosed; Baio et al., 2018), sex difference in older youth and adults may reflect camouflaging of ASD symptoms or phenotypic differences in females that may lead to delayed or missed diagnoses (Tubío-Fungueiriño et al., 2021). Some populations are at increased likelihood (IL) of being diagnosed with ASD due to environmental and/or genetic exposures such as increasing paternal age, premature birth, Fragile X Syndrome, Down syndrome, and Tuberous Sclerosis Complex (Hultman et al., 2011, Capal et al., 2017, Agrawal et al., 2018, Abbeduto et al., 2014). Relative to a general community population, ASD prevalence in these different IL contexts is considerably elevated (Numis et al., 2011, Ozonoff et al., 2011, Abbeduto et al., 2014, Szatmari et al., 2016, Hahn et al., 2020).

One defining feature of ASD is its complexity; the condition is characterized by extreme phenotypic and etiological heterogeneity (Mottron & Bzdok 2020). Individuals on the spectrum can vary tremendously in how their symptoms manifest and present (Wozniak et al., 2016), and in the supports they require to enhance personal independence, productivity, participation in society, and increased community integration and/or improved quality of life (Thompson et al., 2002). Often however, access to supports require a formal diagnosis. The vital importance of early detection and diagnosis (Gardner et al., 2013, Fuller and Kaiser, 2019, Towle et al., 2020) is further highlighted by the benefits of early intervention (Fuller and Kaiser, 2019, Towle et al., 2020, Dawson et al., 2010, Bonis, 2015, Pickles et al., 2016, Noyes-Grosser et al., 2018). Current meta-analytical estimates for the mean age of ASD diagnosis across 40 countries (considering data from 56 studies encompassing 120,540 ASD individuals) is 5 years old (99% CI = 4.18-5.90; van't Hof et al., 2020). Given that (1) early ASD diagnosis and intervention is associated with improved social, communication, brain function, as well as decreased costs of specialized

therapies and education services (Bonnis et al., 2016) and (2) ASD diagnosis is very stable from early to mid-childhood (Brian et al., 2016), tools that aid in early identification of ASD have potential utility to facilitate access to early intervention services. This is particularly relevant to IL children given documentation of the emergence of ASD symptoms at 6 to 18 months old (Tanner & Dounavi et al., 2021). Some evidence supports the utility of ASD-specific screens at 18 and 24 months, however screening prior to 24 months old may be associated with higher false-positive rates compared to screening after 24 months (Zwaigenbaum et al., 2015). Unfortunately, limited data supports screening tools (e.g., the Screening Tool for Autism in Two-Year-Olds [STAT], Systematic Observation of Red Flags [SORF], or Infant-Toddler Checklist [ITC]) that can identify with a high degree of accuracy children <18-month-old who are later diagnosed with ASD (Stone et al, 2008, Wetherby et al., 2008, Dow et al., 2020). With growing access to increasingly powerful computational technology, investigation into the development of new tools or the refinement of existing instruments is becoming increasingly viable.

Since the late 1960s, performance and functionality of digital devices has doubled roughly every 2 years (Shalf, 2022). This observation, dubbed Moore's Law, is built on how exponential growth in integrated-circuits and transistor component density has allowed for new innovations, applications, and technological possibilities at decreasing prices (Khan et al., 2018). Since 1965, Moore's Law has become one of the most durable technological forecasts ever made and has since become an emblem of the information age (Denning & Lewis, 2017). Though the technological underpinnings of Moore's law may end in the coming decades as electronic manufacturing approach the limit of atomic scaling (Shalf 2020), Moore's law has coincided with an unprecedented reduction in cost of electronic storage (Keyes 2006) and growing availability of increasingly powerful computational devices (Bini, 2018). Access to progressively more powerful computers has facilitated the use of increasingly sophisticated data analysis modalities that, historically, have been intractable or infeasible given technological limitations or lack of computational power (Cohen, 2021). For modern researchers and scientists, the ready access to powerful technology has led to widespread use and adoption of applying machine learning algorithms to complex problems and datasets.

Given that provision of accurate and accessible diagnoses is a fundamental challenge of global healthcare, artificial intelligence and machine learning tools show promise in revolutionizing healthcare by facilitating provision of precise and personalized diagnoses (Richens et al., 2020). Machine learning is a branch of artificial intelligence and computer science that focuses on using data and computer algorithms to imitate the way humans think and learn to gradually improve their accuracy and performance (IBM, 2020). Machine learning algorithms are commonly divided into four main categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning algorithms (Sarker et al., 2021). In most healthcare settings, focus is placed on supervised learning algorithms which are built and trained using labelled (input-output) patient or research data (Burkov 2019, Doupe et al., 2019, Mahesh, 2020). Amongst the different types of supervised learning algorithms, the most common types employed are classification and regression algorithms (Garg & Mago, 2020). Though any classification-type machine learning algorithm implicitly or explicitly generates a decision boundary in data, *how* the decision boundary is calculated delineates the different learning algorithm types (Burkov 2019). There are numerous advantages to using machine learning in health research. For instance, machine learning can be flexible and scalable, allowing for deployment in many different areas including risk stratification, diagnosis, classification, and survival predictions (Ngiam & Kor, 2019). Machine learning enables analysis of increasingly diverse types of data that can be incorporated into models used to help predict disease risk, diagnosis, and even treatment modalities (Ngiam & Kor, 2019). Some notable disadvantages of machine learning include the requirement for data pre-processing, model training, and the need for model refinement with respect to the clinical problem being assessed (Ngiam & Kor, 2019).

Current diagnostic learning algorithms continue to struggle to achieve the high accuracy required for differential diagnosis (Richens et al., 2020). Though machine learning is becoming increasingly common in clinical and research sciences (often via machine learning toolkits such as Hadoop, TensorFlow, Spark, R; Hyde et al., 2019) the field of ASD research has yet to fully leverage the technology to the same degree as embraced in other fields (Hyde et al., 2019). One major drawback to traditional ASD diagnostics is that administration and interpretation of standardized ASD diagnostic tools can be costly and time-intensive (Eman et al., 2020). Given that one of the primary goals for machine learning in ASD research is to minimize diagnostic

time with improved accuracy (Eman et al., 2020), progress in this area has huge potential to expedite early screening and detection efforts which, in turn, can facilitate increased access to early intervention services for ASD individuals and their families.

A major benefit of machine learning in data analysis is that it can be used to rapidly evaluate different combinations of factors to determine which provides the best accuracy and predictive power when assessing for a given outcome. Today, the volume, velocity, and variety of data that can be generated has substantially increased due to the availability and affordability of requisite infrastructure and technology needed to create it (Hyde et al., 2019). This increase is coupled with a rise in the amount of data with high dimensionality (i.e., data where the number of features or variables approach or exceed the number of observations in the dataset; Buhlmann et al., 2014).

Taken together, the benefits of highly accurate machine learning classifier models in clinical and research settings are tantalizing. Though existing classifier models struggle to achieve high accuracy (Richens et al., 2020), in the context of ASD diagnostics, the potential is huge. A highly accurate learning algorithm trained to distinguish early signs of ASD from neurotypical individuals has a huge potential to facilitate increased access to early intervention services which, in turn, can have a profound impact on the quality of life of individuals and families with ASD.

This study aims to build on previous research into early screening and detection of ASD. The primary objectives of this study are (1) to conduct a cross-domain supervised learning classifier analysis using 12-month Autism Observation Schedule for Infants (AOSI), Mullen Scales of Early Learning (MSEL), and demographic data (biological sex), and (2) to assess classifier performance at predicting 36-month ASD diagnostic outcomes in a cohort of Canadian infant siblings at increased likelihood for ASD.

Methods

Participants

The Canadian Infant Sibling Study (CISS-1) is a longitudinal study with 937 participants, 699 of which are infant siblings at increased likelihood for ASD (IL-siblings). Starting in the early-to-

mid 2000s, participants were recruited from one of five ASD diagnostic and treatment centers in Canada: The Glenrose Rehabilitation Hospital (Edmonton), McMaster Children's Hospital (Hamilton), the Hospital for Sick Children (Toronto), Holland Bloorview Kids Rehabilitation Hospital (Toronto), and the IWK Health Centre (Halifax).

Confirmation of IL-Sibling Status

To verify participant status as IL-siblings, diagnosis of ASD in the older sibling was confirmed through clinical assessment or a review of diagnostic records using the Diagnostic and Statistical Manual of Mental Disorders (DSM) fourth edition, text-revised (DSM-IV-TR) criteria. All IL-sibling participants were characterized by the lack of identifiable neurological or genetic conditions, nor any severe sensory or motor impairments. All participants were born at 36-42 weeks gestation and had a birth weight greater than 2500g.

Ethics

This study was approved by the research ethics boards at each institution, and all participating families gave written informed consent upon enrollment. No personally identifying information was used, considered, or incorporated into any of the statistical models generated in this paper. As such, all statistics and data reported in this study stem from de-identified, anonymized IL-sibling data.

Canadian Infant Sibling Study (CISS-1) Participant Data

As with all longitudinal studies, participant attrition and missing data present methodological problems that become increasingly prevalent as time goes on (Gustavson et al., 2012). Of the 699 IL-siblings enrolled in CISS-1, 482 have 36-month diagnostic outcome data (i.e., underwent a comprehensive 36-month diagnostic assessment for ASD). Of these 482 participants, 465 had 12-month AOSI data and are the principal participants of this paper. Participants included in this study were restricted to IL-siblings only as a means of investigating the ability to correctly classify ASD in a sample of infant siblings at increased likelihood for a diagnosis of ASD – not in a general community context. Data on low likelihood (LL) control participants was therefore excluded from analysis.

Participants were assessed at 6-, 9-, 12-, 15-, 18-, 24-, and 36-months of age. Adaptive behaviour, intellectual and language abilities, temperament, and ASD symptomatology were assessed at different timepoints using the Autism Diagnostic Observation Schedule (ADOS), Autism Diagnostic Interview-Revised (ADI-R), Mullen Scales of Early Learning (MSEL), and the Vineland Adaptive Behavior Scales (VABS; Mullen 1995, Sparrow et al., 1989, Lord et al., 1989, Lord et al., 1994). Early behavioural markers of ASD were investigated when participants were 6-, 9-, 12-, 15-, and 18-months old using the AOSI. Clinicians or trained research staff administered and collected ADOS, ADI-R, AOSI, MSEL, and VABS data at their respective timepoint per the CISS-1 study protocol.

IL-Siblings Diagnostic Procedure

At 36-months old, each participant underwent an independent diagnostic evaluation conducted by an expert clinician blind to previous study assessments. Clinicians assigned ASD diagnosis using DSM-IV-TR criteria, based on the best clinical judgment of the clinician (developmental pediatrician, child psychiatrist, or clinical psychologist with 10 or more years of diagnostic experience) using information from ADI-R, ADOS, and concurrent developmental information from the MSEL and VABS.

Measures

The Autism Observation Scale for Infants (AOSI) is a brief, 19-item observational measure designed to characterize early behavioural signs of ASD between 6 and 18 months in a familial cohort of infant siblings at IL for ASD (i.e., infant siblings of children who already have a diagnosis of ASD; Bryson et al., 2008). The AOSI assesses multiple overlapping constructs that characterize prodromal ASD (e.g., social communication, emotional regulation, atypical sensory-motor behaviours, repetitive behaviours, etc.) within an interactive, play-based context where behaviour can be systematically elicited by trained examiners (Bryson et al., 2008). AOSI Items 1-19 are scored on an integer scale from 0 to 2-3, with 0 = typical behaviour, 1 = inconsistent, partial, or questionable behaviour, 2 = atypical behaviour, and 3 = a total lack or absence of behaviour (Bryson et al., 2008). Using this data, two different AOSI scoring metrics can be calculated: (1) the AOSI Total Score by summing Items 1 to 18 on the scale (ranging in value from 0-38), and (2) the number of AOSI Risk Markers constituting Items 1 to 19 that score one or higher (ranging in value from 0-16; Bryson et al., 2008, Zwaigenbaum et al., 2005).

The Mullen Scales of Early Learning (MSEL) is a developmental measure consisting of five scales, four of which (Visual Reception [VR], Receptive Language [RL], Expressive Language [EL], and Fine Motor [EL]) assess nonverbal cognitive and language abilities, whereas the fifth measures Gross Motor development (from 0 to 29 months only; Mullen 1995). An Early Learning Composite [ELC] is calculated from scores on all but the Gross Motor domain for children aged 0-69 months (Mullen 1995). The Mullen has exhibited excellent inter-rater and test-retest reliability (Mullen 1995).

The Hollingshead Four-Factor Index is a measure of socioeconomic status (SES) in which individual raw scores are collapsed into four distinct classes according to standardized cut-off criteria (Hollingshead, 1975). A composite family socioeconomic status score can be calculated ranging between <20 and 66, with higher values indicating higher family SES (Hollingshead, 1975).

Why Build Predictive Models Using 12-month Data?

As reported in Chapter 2, group differences on the AOSI emerge at 12-months of age between IL-siblings diagnosed with ASD at 36-months (IL-ASD), IL-siblings *not* diagnosed at 36-months (IL-N), and LL controls. This study aims to build on that work by developing and assessing the performance of predictive classifiers built using 12-month IL-sibling clinical data.

Variable Selection: Assessment of 12-month MSEL and VABS data completeness was sought as potential moderating variables in later statistical or machine learning model generation as previous studies have used similar predictor variables during learning classifier analysis (Bussu et al., 2018). Of the 465 IL-sibling participants from CISS-1 with 36-month diagnostic outcomes and 12-month AOSI administrations, 409 and 262 participants had recorded 12-month MSEL and VABS data reflecting a missing data rate of 12.04% and 43.66% respectively. Though there is no established cut off criterion pertaining to acceptable percentages of missing data required for statistical inferences (Dong & Peng, 2013) since less than 15% of the 12-month MSEL data was missing, it was deemed eligible for missing data imputation – assuming data missingness was completely at random. As such a large proportion of participants were missing 12-month

VABS data, the VABS was eliminated from consideration in predictive modelling as including it would require severely reducing the total sample of participants. Therefore, all predictive classifier models in this study are built using combinations of participant demographic (i.e., biological sex), 12-month AOSI (item level, Total Score), and MSEL standard score (ELC, VR, FM, RL, EL) data.

Data Preprocessing

Data was first cleaned and preprocessed to allow for assessment of (1) the proportion and randomness of missing data of the 465 IL-siblings relative to 12-month clinical data, (2) determination of what variables/features to use in statistical modelling, (3) feasibility of using missing data imputation techniques such as expectation maximization (EM; which take into account conditions under which missing data occurred; Dong & Peng, 2013) to address missing data, and (4) the impact of imputation techniques on 12-month dataset statistics for IL-ASD and IL-N participants.

Dataset Partitioning into Training and Testing Sets

Cleaned/EM-imputed IL-sibling data was imported into R/RStudio using the *read_sav* function from the *haven* R package (Wickam et al., 2022). Cleaned data was then randomly partitioned into training and testing datasets that contained 80 and 20% of the total data using the *createDataPartition* function from the R package *caret* (Khun et al., 2022) before being exported as a comma delimited *.csv* file using the *write.table* R function (R v4.2.1, 2022). Partitioned data was imported back into a R/RStudio environment using the *read_csv* function from the *readr* package (Wickam et al., 2022) any time statistical modelling or testing was performed. The R/RStudio code used to partition study data is described in the public GitHub repository.

Generation of an Independent Validation Dataset

Enrollment and participant follow-up in CISS-1 is concluded and a new CISS-1 conducted by the same principal investigators is currently underway (CISS-2) in the same Canadian ASD diagnostic and treatment centers. Since CISS-2 is still actively recruiting participants, fewer IL-sibling data is available. As of October 2022, data on 133 IL-siblings with 12-month AOSI and 36-month diagnostic outcomes was available for use in this study. Identical data preprocessing and cleaning steps relative to the CISS-1 IL-sibling data was employed as described in this

methods section. The rationale surrounding use of this dataset in this study, as well as data preprocessing results are described in detail in the Appendix 2.

Assessing Distribution/Normality of 12-Month AOSI and MSEL Data

Given how the AOSI is scored (scores across the range of 0 to 2-3 represent increasing degrees of impairment; Bryson et al., 2008), it was deemed unlikely that item-level AOSI or Total Score data would be normally distributed. Within an IL sample, a relatively small proportion are expected to have ASD features based on the known ASD recurrence rates. Accordingly, scores are expected to be non-normal. Given that parametric statistical techniques require normally distributed data, thus Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were conducted in SPSS Version 28.0.1.1 (14) (IBM 2022) using the *Explore* command, with data factored by 36-month diagnostic outcome (IL-ASD / IL-N) prior to any follow-up statistical analyses.

Assessing for Group Differences in 12-Month AOSI and MSEL Data

Group differences between CISS-1 IL-ASD / IL-N siblings (and between training and testing datasets partitions to assess for robustness of the randomization code) with respect to 12-month AOSI and MSEL data were explored using nonparametric Mann-Whitney U tests with post hoc Benjamini & Hochberg 1995 corrections. In this method, p-values are ordered smallest to largest (Benjamini & Hochberg 1995). The α level for each test is then set at $(k*\alpha)/m$, with k corresponding to the p-value's rank (lowest $p = 1$), and m corresponding to the number of comparisons (Benjamini & Hochberg 1995). This method decreases the chance of false positives; comparisons stop once one of the tests are rejected (this method uses 'q*' rather than 'p' to denote the critical alpha level; Benjamini & Hochberg 1995).

Assessing AOSI Item-Level Data

AOSI Item-Level Correlations: Exploratory bivariate one-tailed Pearson correlations of 12-month item-level AOSI data (items 1-19) were conducted on cleaned CISS-1 IL-sibling participant data to determine if any AOSI Item was significantly correlated with one another – potentially indicative of AOSI items measuring a similar feature or construct characteristic of ASD. Correlations were conducted in SPSS Version 28.0.1.1 (14) using the *Analyze* → *Correlate* → *Bivariate* → *One-tailed* command.

Reducing AOSI Data Dimensionality

Use of highly dimensional data during machine learning can increase not just computational complexity, but also the risk of overfitting (Ayesha et al., 2020). Overfitting refers to when models or algorithm have poor performance when applied to new, unseen data (Ying 2019). To mitigate against this, dimensionality reduction methods are often employed via methods like feature extraction (transformation of high dimensional data into lower dimensionality through techniques like principal component analysis), and/or feature selection (selection of features that are most relevant for a given problem; Ayesha et al., 2020). A benefit when using dimensionality reduction techniques is the reduction in number of input variables (i.e., data dimensions) which can reduce computational time and enable more efficient use of available computing resources (Ayesha et al., 2020). In this study, factor analysis (informed by principal axis parallel analysis) was used as a means of reducing item-level AOSI data dimensionality.

Principal Axis Parallel Analysis and Follow-Up Factor Analysis of Item-Level AOSI Data

Factor analysis of 12-month item-level AOSI data was conducted to determine (1) if different AOSI items factored together, and (2) to identify possible predictor variable combinations of AOSI items for use during predictive classifier generation as a dimensionality reduction technique. Principal axis parallel analysis employing a Monte Carlo simulation was conducted to identify the number of statistically significant eigenvalues in item-level AOSI data for extraction during follow-up factor analysis. Principal axis analysis was conducted using 5000 parallel datasets based on permutations of cleaned/expectation maximization-imputed item-level AOSI data (items 1-19) in SPSS Version 28.0.1.1 (14) utilizing O'Connor, 2000's parallel analysis *code rawpar*. In this method, eigenvalues calculated for each AOSI item are compared against their simulated eigenvalues based on random permutations of the original dataset containing 465 participants in a Monte Carlo simulation (O'Connor, 2000). Factors or components are retained if the observed eigenvalues calculated from the raw data are greater than 95th percentile simulated eigenvalues derived from the random permutations of the original dataset (O'Connor, 2000). This approach was chosen as the K1 rule (which retains all factors with Eigenvalues greater than one) can potentially over- or under-estimate the true number of factors or components that should be extracted from a dataset (O'Connor, 2000).

Factor Analysis Items for Group Differences in IL-ASD/IL-N Participants

Factor analysis of 12-month item-level AOSI data was run in IBM SPSS Version 28.0.1.1 (14) using the *Analyze* → *Dimension Reduction* → *Factor Analysis* commands with maximum likelihood, a fixed number of factors being extracted (informed by principal axis parallel analysis), and direct oblimin data rotation.

Assessing for Group Differences Between IL-ASD/IL-N Participants on Factor Analysis AOSI Items

Independent Mann-Whitney U tests were conducted to explore if factor analysis-identified AOSI items differed significantly between CISS-1 IL-ASD/IL-N participants. To control for multiple comparisons, Benjamini & Hochberg (1995) corrections were employed.

Benchmark ROC Curve Performance of 12-month AOSI Total Score

Predictive performance of the 12-month AOSI Total Score with respect to 36-month ASD diagnostic status was assessed using receiver operator characteristic (ROC) curves in the CISS-1 IL-sibling data. This analysis effectively serves as a benchmark against which all classifiers generated in this study can be compared. After generating the ROC curve, area under the curve (AUC) was calculated which represents an accuracy index where higher AUC values indicate better predictive ability (0.50 = chance, 0.70-0.90 = moderate, ≥ 0.90 = high accuracy; Akobeng 2007). Youden indexes (the maximum vertical distance between the ROC curve and the chance/diagonal line [Youden's Index (*J*) = (sensitivity + specificity) -1]) were calculated to determine optimal cutoff for 12-month AOSI Total Scores (Akobeng 2007). In this method, the highest *J* value represents the optimal cut score (Akobeng 2007).

Statistical Modelling

R/RStudio

All predictive classifier models in this study were built using R version 4.2.1 (2022-06-23 ucrt; R v4.2.1, 2022) run in conjunction with RStudio Desktop 2022.07.1 Build 554 (Rstudio, 2022) on a computer running a 64-bit Windows 10 operating system. All models were generated using 10-fold cross-validation – a data resampling method that is used to not only assess model generalizability, but to prevent overfitting (Berrar, 2019).

For all statistical models generated in this study, training, testing, and independent validation datasets were imported into R/Rstudio's working environment as comma-delimited files using the *read_csv* function from the *readr* package (Wickam et al., 2022). All statistical models were generated on the training dataset utilizing various combinations of participant (biological sex), AOSI (item-level, Total Score), and MSEL standard score (ELC, VR, FM, RL, EL) data. Specific AOSI variable combinations were generated following factor analysis and post hoc assessment in CISS-1 IL-sibling data. Factor analysis identified AOSI Items 6, 8, 14, 16, 18. The factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrections were Items 8, 14, and 18. After predictive classifier generation, accuracy and performance was assessed across the training, testing, and independent validation datasets by comparing the predicted vs real 36-month diagnostic outcomes.

For classifiers built using logistic and regularized logistic regression, model performance at correctly predicting 36-month IL-ASD diagnosis in the training, testing, and independent validation sets is reported using the (1) default logistic regression decision threshold of 0.500 and (2) a decision threshold optimized for maximal combined sensitivity/specificity (a modification to Youden's indexes where instead of calculating J = sensitivity + specificity -1 [Akobeng 2007], maximal combined sensitivity + specificity is sought). Both thresholds were reported because by changing the logistic regression decision threshold between 0 and 1, model performance can be heavily affected. The benefit of model optimization is trading off a small amount of specificity for a potentially large amount of sensitivity (or vice versa depending on context). Moreover, depending on the model being generated, the optimum decision threshold may be on either side of the default 0.500 boundary. By reporting model performance using both the default and optimized threshold, a better assessment of the model's performance metric could be ascertained.

Multivariate Logistic Regression Modelling

Logistic regression is a common probabilistic statistical model used in machine learning on classification problems (Sarker 2021). This technique employs use of a logistic function to estimate classification probabilities that a given event will or will not occur (output probability ranges between 0 and 1) based on input variables (Ray 2020, Sarker 2021). Logistic regression is

simple to implement, computationally efficient, and is easy to employ regularization on (Ray 2020).

Various 10-fold cross-validated multivariate logistic regression models with (Models L1-L15) and without biological sex (Models L16-L30) were generated in R/Rstudio using the *caret* package's *train* function (Khun et al., 2022) with method = "glm", metric = "ROC", family = "binomial," and a customized trControl function to generate more detailed model performance metrics. All logistic regression models were built on the training dataset. The R code used to generate, assess, and extract logistic regression model performance is described in the public GitHub repository.

For logistic regression models built with/without biological sex, three models from each grouping characterized by the highest AUC when applied to the testing dataset were assessed for predictor variable importance. Selection of models with the highest AUC was based on testing set performance only due to differences between CISS-1 and CISS-2 methodology; CISS-1 uses DSM-IV-TR ASD criteria to diagnoses ASD while CISS-2 uses the fifth edition of the DSM (DSM-5). Of the AOSI and MSEL data used to generate the models, variables which were significant (as determined by R's *summary* function; R v4.2.1, 2022) were retained. Non-significant predictor variables were systematically removed to assess if model performance increased (defined as decreased Akaike information criterion (AIC) and increasing AUC relative to the last best-performing model). If predictor variable removal had no impact on AIC and AUC relative to the last best-performing model, the variable was removed in accordance with the principle of Occam's Razor (the simplest hypothesis is usually the best one) and maximum model parsimony. Following non-significant variable pruning, truncated logistic regression model performance on the training, testing, and independent validation set was assessed.

Regularized Logistic Regression Modelling

In machine learning, regularization refers to the process of facilitating increased model generalization to new data (Zhu et al., 2018, Tian & Zhang, 2022). Many different types of regularization algorithms have been developed and vary in use depending on the machine learning algorithm being employed (Zhu et al., 2018). In the context of logistic regression, while

there is a tendency to overfit higher-dimensionality data (Sarker 2021), techniques like L_1 and L_2 regularization are relatively easy to apply (Ray 2020, Sarker 2021). L_1 regularization applies a penalty term to the model during generation encouraging parameters to be small while L_2 regularization encourages the sum of parameter squares to be small (Ng, 2004).

Various regularized, 10-fold cross-validated multivariate logistic regression models with (Models R1-R15) and without biological sex (Models R16-R30) were generated using various combinations of predictor variables (biological sex, AOSI and MSEL data) in R/Rstudio using the *caret* package's *train* function (Khun et al. 2022) with method = "regLogistic", metric = "ROC", tuneLength = 10, and a customized trControl function to generate more detailed model performance metrics. To function properly, the regLogistic method employed in *caret* required the R package *LiblineaR* to be installed (Helleputte et al., 2021). The benefit of this approach to model generation is when the code to generate models is executed, model performance across the different model tuning parameters (cost, loss function [L_1/L_2 regularization], and epsilon [tolerance]) are evaluated across a range of different values (defined by tuneLength) before the final model is selected based on the combination of tuning parameters yielding the highest ROC performance (defined by metric = "ROC"). All regularized logistic regression models were built on the training dataset. The R code used to generate, assess, and extract regularized logistic regression model performance is described in the public GitHub repository.

Support Vector Machines (SVM) Modelling

SVMs are commonly used in machine learning for classification and regression problems (Sarker 2021). SVMs map input data into a higher dimensional space where an optimal separating surface (i.e. the hyperplane) is generated that maximally separates data (Bhavsar et al., 2012). Two parallel hyperplanes constructed on either side and form the support vectors (Bhavsar et al., 2012). SVM analysis is predicated on finding the hyperplane orientated such that the margin (i.e., distance) between the support vectors is maximized (Bhavsar et al., 2012). In the case where there is no linear divide between data, SVMs can employ kernel functions that map data into a higher-dimensional space where non-linear decision boundaries can be generated (Bhavsar et al., 2012). The choice of SVM kernel varies depending on the problem: linear kernels allow for linear separation of data, polynomial allow for separation of more complex, non-linearly

separable data up to the order of polynomial hyperplanes, and radial basis functions allow separation of circular data (Bhavsar et al., 2012).

Using the *caret* package, multiple 10-fold cross-validated SVM models with and without biological sex were generated using various combinations of predictor variables (biological sex, AOSI and MSEL data). Three types of SVMs were generated in R: SVMs with linear kernels (method = "svmLinear" that are characterized by the tuning parameters *cost* and *class weights*), SVMs with a polynomial kernel (method = "svmPoly" that are characterized by the tuning parameters *polynomial degree*, *scale*, and *cost*), and SVMs using a radial basis function kernel (method = "svmRadial" that are characterized by the tuning parameters *sigma* and *cost*). All models were generated using the *train* function (Khun et al., 2022) with method = "svmLinear, svmPoly, or svmRadial," metric = "ROC", tuneLength = 10 (except for svmPoly; tuneLength was set at 6 due to the exorbitant computational processing requirement required to fit the model with tuneLength = 10), and a customized trControl function to generate more detailed model performance metrics. To function properly, SVM modelling in *caret* required the R package *kernlab* to be installed (Karatzoglou et al., 2022). The R code used to generate, assess, and extract SVM model performance is described in the public GitHub repository.

Assessing Model Performance

Performance of all predictive classifiers generated in this study was determined by their ability to predict IL-ASD at 36-months when applied to training, testing and validation datasets using R's *predict* function (R v4.2.1, 2022). More specifically, model ROC performance metrics (ROC curves; AUC) was assessed using the *roc* and *auc* functions from the *pROC* R package (Robin et al., 2021). Model performance metrics (including, but not limited to, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value) were extracted from a confusion matrix generated for each model using the *confusionMatrix* function from the R package *caret* (Khun et al., 2022). The R code used to find the optimal logistic regression decision threshold to maximize combined sensitivity and specificity is described in the public GitHub repository. The entire data preprocessing, partitioning, and model testing process is depicted visually in Figure 3.01.

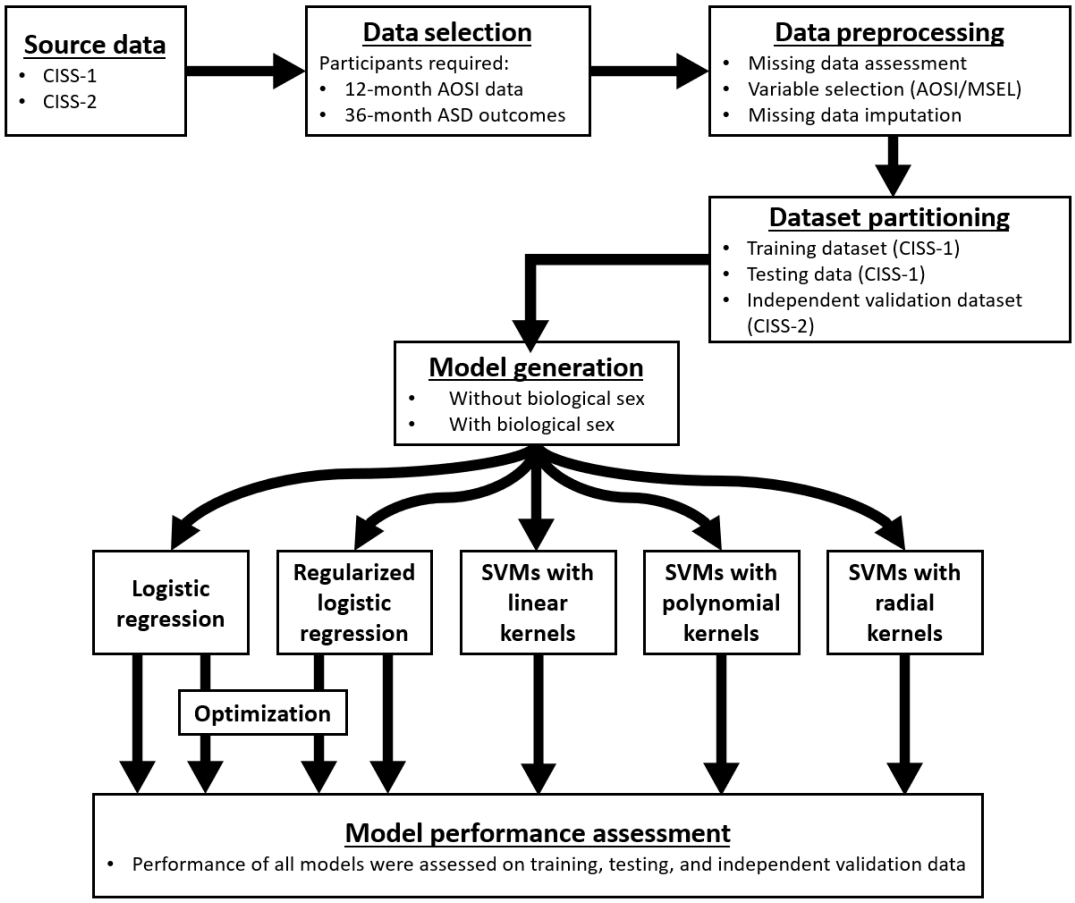

Figure 3.01: Procedural Diagram of Study Methods | A procedural flow diagram depicting an overview of the study's methods is reported above. AOSI = Autism Observation Scale for Infants, CISS-1 = Canadian Infant Sibling Study, CISS-2 = the new Canadian Infant Sibling Study, MSEL = Mullen Scales of Early Learning, SVM = Support Vector Machines

Results

Participant Demographics of CISS-1 IL-Siblings/Families

Family demographic data for 465 IL-siblings/families were compared by 36-month diagnostic status (IL-ASD / IL-N). Demographic data assessed included IL-sibling biological sex, birth order, number of children in the family, study site assessments venue, parent age at IL-sibling birth, highest level of parental education, parental occupation, parental relationship, and family SES in SPSS using chi-squared analyses. Overall, 67.52% of Fathers and 65.59% of Mothers were Caucasian, with 67.31% of participant's families characterized by a Hollingshead Four-Factor Index between 36 and 66. A detailed breakdown of demographic data for IL-ASD and IL-N infants from the old CISS-1 study are detailed in Appendix 2 Table A2.1.

Data Preprocessing

IL-ASD vs IL-N Infants: Of the 465 CISS-1 IL-siblings with 12-month AOSI data and 36-month diagnostic outcomes, 125 (26.9%) were diagnosed with ASD at 36-months (IL-ASD), and 340 (73.1%) were not (IL-N).

Assessing Randomness of Missing AOSI and MSEL Data: Missing data imputation techniques such as EM requires that missing data which is being replaced is randomly distributed (Scheffer, 2002). To test the hypothesis that 12-month AOSI and MSEL data was missing completely at random, Little's Missing Completely at Random (MCAR) test was conducted. In effect, if data is missing completely at random, the mechanism behind data missingness is not dependent on variables under consideration; data is collected randomly and does not depend on any other variable in the dataset (Scheffer, 2002). Little's MCAR test was conducted on data split according to 36-month diagnostic outcome as IL-ASD / IL-N groups are not homogenous; IL-ASD infants score higher on the AOSI and have greater range of impairments on the MSEL relative to IL-N infants.

For the 125 IL-ASD infants Little's MCAR test was non-significant for item-level AOSI (χ^2 = 33.470, DF = 31, p = 0.348) and MSEL standard score data (χ^2 = 2.778, DF = 4, p = 0.596) indicating data was missing completely at random. Though minimal AOSI data were missing (between 0 and 1.6% across all AOSI items), moderate amounts of MSEL data were missing (between 11.2 and 12.0% of ELC, VR, FM, RP, or EL standard scores).

For the 340 IL-N infants Little's MCAR tests was non-significant for item-level AOSI (χ^2 = 22.440, DF = 58, p = 1.000) and MSEL standard score data (χ^2 = 4.643, DF = 9, p = 0.864) indicating data was missing completely at random. Though minimal AOSI data were missing (between 0 and 2.6% across all AOSI items), moderate amounts of MSEL standard score data were missing (between 11.5 and 12.4% of ELC, VR, FM, RL, or EL standard scores).

Missing Data Imputation via Expectation Maximization

Since missing 12-month AOSI and MSEL data were randomly distributed for both IL-ASD and IL-N populations, missing data were eligible for data imputation. Prior to imputation, it is important to note that all AOSI items (1 – 19) have the option of being coded an '8' signifying

the item in question was 'not applicable' (e.g. item 1 assesses visual tracking; a score of 8 would be appropriate if the child has vision problems) or 'or unable to code' (i.e. the child did not engage or react in a manner during assessment that facilitated coding on that item; Bryson et al., 2008). When calculating the AOSI Total Score, any items scoring an '8' are treated as '0.' Accordingly, all IL-ASD/N participants that scored an '8' for any AOSI item were replaced with '0's prior to EM missing data imputation to not unduly bias the imputation process. Missing data was imputed using EM in SPSS GradPack Version 28 for IL-ASD and IL-N populations using *Analyze → Missing Value Analysis → Estimation → EM*.

Assessing for Differences in Raw vs Cleaned/Imputed Dataset Statistics

Relative to the original AOSI and MSEL dataset of 465 participants, EM imputation had minimal impact on dataset AOSI and MSEL statistics for IL-ASD / IL-N participants.

For 12-month IL-ASD participants, EM on item-level AOSI data had either no impact (signifying no missing items being replaced) or minimal impact on mean, standard deviation, and standard error (max absolute difference in means, standard deviation, and standard error between the raw and EM-imputed data was 0.008, 0.010, and 0.428 respectively). For 12-month IL-ASD MSEL standard score data, EM had minimal impact on mean, standard deviation, and standard error (max absolute difference in mean, standard deviation, and standard error between the raw and EM-imputed data was 0.159, 0.969, and 1.490 respectively).

For 12-month IL-N participants, EM on item-level AOSI data had either no impact (signifying no missing items being replaced) or minimal difference in mean, standard deviation, and standard error (max absolute difference in means, standard deviation, and standard error between the raw and EM-imputed data was 0.005, 0.004, and <0.001 respectively). For IL-N MSEL standard score data, EM had minimal impact on mean, standard deviation, and standard error (max absolute difference in mean, standard deviation, and standard error between the raw and EM-imputed data was 0.038, 1.048, and 0.057 respectively).

Dataset Partitioning and Characteristics of the Training and Testing Set

Of the 465 IL-siblings from CISS-1 with 12-month AOSI data and 36-month diagnostic outcomes, 80% were randomly partitioned into the training dataset. In total, the training dataset

contains data on 373 IL-siblings, is 58.7% male (219 male, 154 female), and has 27.8% (n=104) of participants diagnosed with ASD. The remaining 20% of CISS-1 IL-ASD / IL-N participant data was used to generate the testing set that classifier performance was assessed against. The testing dataset contains data on 92 IL-siblings (43 male, 49 female) and has 22.8% (n=21) of participants diagnosed with ASD.

Independent Validation Set Characteristics

In total, 90 of 133 IL-siblings from CISS-2 were retained following data preprocessing and represent an independent validation set that classifier performance could be assessed against. Of the 90 IL-siblings that were retained, 25.5% (n=23) were diagnosed with ASD at 36-months, and 74.5% (n=67) were not. Details pertaining to randomness and missingness of data, eligibility and impact of EM imputation, and distribution of cleaned data is described in Appendix 2.

Distribution/Normality of 12-Month AOSI and MSEL Data

For 12-month IL-ASD/N CISS-1 participants: For the 465 IL-siblings from CISS-1, AOSI item-level and Total Score data was right-skewed for both IL-ASD and IL-N groups (ps < 0.001 for all AOSI data on both Kolmogorov-Smirnov and Shapiro-Wilk tests of normality). MSEL data was similarly non-normal for all subscales (ps < 0.01) on both the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality barring the ELC scores for IL-ASD participants (p = 0.227). Normality test results for IL-ASD / IL-N data are described in Appendix 2 Table A2.2.

For 12-month AOSI and MSEL scores in the training and testing datasets: In the training and testing set, AOSI item-level and Total Score data was right-skewed for both IL-ASD and IL-N groups (ps < 0.001 for all AOSI data on both Kolmogorov-Smirnov and Shapiro-Wilk tests of normality). All MSEL data (barring the ELC score in the testing set) were non-normal (p < 0.01) on the Kolmogorov-Smirnov test of normality. Similarly, all MSEL data (barring VR and FM scores in the testing set) were non-normal (p < 0.04) on the Shapiro-Wilk test of normality. Normality test results for IL-ASD / IL-N data are described in Appendix 2 Table A2.3.

Group Differences in 12-Month AOSI and MSEL Data

For 12-month IL-ASD/N CISS-1 participants: Nonparametric Mann-Whitney U-tests yielded significant differences between IL-ASD / IL-N participants for eleven AOSI items (ps < 0.05;

items 3, 4, 7-14, 17, 18, Total Score), with only nine (items 3, 4, 7-10, 14, 18, and the Total Score) surviving the follow up post hoc Benjamini & Hochberg 1995 corrected significance level of $q^* = 0.029$. All MSEL subscales (ELC, VR, FM, RL, EL) were significantly different between IL-ASD/N participants and survived Benjamini & Hochberg 1995 corrections. Results are described in Table 3.01.

For 12-month AOSI and MSEL scores in the training and testing datasets: Nonparametric Mann-Whitney U-tests yielded no significant differences between the training and testing set with respect to 12-month AOSI or MSEL data (all ps > 0.1). Results are described in Table 3.02.

Table 3.01 | IL-ASD vs IL-N Characteristics

Measures	IL-ASD			IL-N			*U*	*Z*-score	*p*-value
	Mean	SD	Mean Rank	Mean	SD	Mean Rank			
Participant Characteristics									
n	125			340					
Gender	87M:38F			175M:165F			$\chi^2 = 12.213$		<0.001
Autism Observation Scale for Infants									
Age at assessment	12.398	0.587	237.728	12.400	0.582	231.262	20659.00	-0.460	0.645
Item 1	0.160	0.410	233.348	0.179	0.480	232.872	21206.50	-0.056	0.955
Item 2	0.184	0.559	234.660	0.168	0.536	232.390	21032.50	-0.315	0.753
Item 3 [αβ]	0.720	0.809	277.040	0.332	0.589	216.809	15475.00	-5.158	<0.001
Item 4 [αβ]	0.448	0.665	250.932	0.315	0.588	226.407	19008.50	-2.213	0.027
Item 5	0.320	0.736	245.040	0.168	0.465	228.574	19745.00	-1.875	0.061
Item 6	0.360	0.689	245.840	0.238	0.569	228.279	19645.00	-1.840	0.066
Item 7 [αβ]	1.184	0.928	264.344	0.882	0.868	221.476	17332.00	-3.235	0.001
Item 8 [αβ]	0.752	0.973	263.420	0.394	0.797	221.816	17447.50	-3.972	<0.001
Item 9 [αβ]	0.744	0.822	258.952	0.521	0.726	223.459	18006.00	-2.841	0.004
Item 10 [αβ]	0.120	0.326	247.900	0.032	0.177	227.522	19387.50	-3.643	<0.001
Item 11 [α]	0.384	0.669	248.256	0.244	0.506	227.391	19343.00	-2.007	0.045
Item 14 [αβ]	0.496	0.591	258.060	0.324	0.499	223.787	18117.50	-2.942	0.003
Item 15	0.288	0.579	243.132	0.182	0.409	229.275	19983.50	-1.456	0.145
Item 16	0.120	0.326	232.900	0.121	0.326	233.037	21237.50	-0.017	0.986
Item 17 [α]	0.720	0.964	250.200	0.518	0.877	226.676	19100.00	-2.139	0.033
Item 18 [αβ]	0.368	0.778	256.280	0.094	0.424	224.441	18340.00	-4.718	<0.001
Total Score [αβ]	7.368	4.780	289.124	4.712	3.575	212.366	14234.50	-5.482	<0.001
Mullen Scale of Early Learning									
Age at assessment	12.357	0.491	227.452	12.396	0.526	235.040	20556.50	-0.540	0.589
ELC [αβ]	98.022	14.370	189.076	104.256	14.197	249.149	15759.50	-4.276	<0.001
Visual Reception [αβ]	102.024	13.491	186.720	107.591	13.179	250.015	15465.00	-4.531	<0.001
Fine Motor [αβ]	108.340	14.161	205.860	112.788	13.894	242.978	17857.60	-2.658	<0.001
Receptive Language [αβ]	90.405	13.774	191.036	94.402	12.543	248.428	16004.50	-4.120	<0.001
Expressive Language [αβ]	91.238	15.699	194.928	96.963	16.635	246.997	16491.00	-3.723	<0.001

ASD = Autism spectrum disorders, ELC = Early Learning Composite, IL-ASD = Infant siblings diagnosed with autism at 36-months, IL-N = Infant siblings not diagnosed with autism at 36-months, SD = Standard deviation, SS = standard scores

[α] = significantly different based on 2-tailed Mann-Whitney U score

[β] = survived Benjamini & Hochberg 1995 corrected significance levels for multiple comparisons ($q^* = 0.029$)

Table 3.02 |Training and Test set 12-month clinical characteristics using Mann-Whitney non-parametric analysis

Measures	Training Set (80%)			Test Set (20%)			*U*	*Z*-score	*p*-value
	Mean	SD	Mean Rank	Mean	SD	Mean Rank			
Participant Characteristics									
n	373			92					
Gender	219:154			43:49			$\chi^2 = 4.302$		0.038
ASD diagnosis	104			21			$\chi^2 = 0.960$		0.327
Autism Observation Scale for Infants									
Age at assessment	12.397	0.577	232.247	12.408	0.610	236.054	16877.0	-0.244	0.808
Item 1	0.161	0.447	230.436	0.228	0.516	243.397	16201.5	-1.377	0.169
Item 2	0.182	0.557	234.213	0.130	0.474	228.082	16705.5	-0.765	0.444
Item 3	0.410	0.656	228.843	0.543	0.747	249.853	15607.5	-1.617	0.106
Item 4	0.362	0.618	235.076	0.304	0.588	224.582	16383.5	-0.851	0.395
Item 5	0.217	0.571	233.787	0.174	0.483	229.810	16864.5	-0.407	0.684
Item 6	0.292	0.625	236.294	0.185	0.512	219.647	15929.5	-1.567	0.117
Item 7	0.981	0.899	235.394	0.891	0.870	223.293	16265.0	-0.821	0.412
Item 8	0.483	0.857	232.099	0.522	0.883	236.652	16822.0	-0.391	0.696
Item 9	0.590	0.759	234.780	0.543	0.762	225.783	16494.0	-0.647	0.518
Item 10	0.059	0.236	233.713	0.043	0.205	230.109	16892.0	-0.579	0.563
Item 11	0.265	0.525	231.255	0.348	0.670	240.076	16507.0	-0.762	0.446
Item 14	0.367	0.530	232.340	0.380	0.531	235.674	16912.0	-0.257	0.797
Item 15	0.209	0.457	233.034	0.217	0.488	232.864	17145.5	-0.016	0.987
Item 16	0.123	0.329	233.673	0.109	0.313	230.272	16907.0	-0.386	0.700
Item 17	0.584	0.911	234.442	0.522	0.883	227.152	16620.0	-0.595	0.552
Item 18	0.172	0.561	233.446	0.152	0.533	231.190	16991.5	-0.300	0.764
Total Score	5.458	4.151	233.720	5.293	3.924	230.082	16889.5	-0.233	0.815
Mullen Scale of Early Learning									
Age at assessment	12.387	0.513	233.824	12.380	0.531	229.658	16850.5	-0.267	0.790
ELC	102.301	14.538	232.638	103.712	14.341	234.467	17023.0	-0.117	0.907
Visual Reception	106.033	13.483	232.682	106.342	13.525	234.288	17039.5	-0.103	0.918
Fine Motor	111.449	14.288	232.930	112.175	13.316	233.283	17132.0	-0.023	0.982
Receptive Language	92.802	12.877	228.878	95.460	13.311	249.712	15620.5	-1.344	0.179
Expressive Language	95.062	16.748	230.480	96.890	15.822	243.217	16218.0	-0.818	0.413

ASD = Autism spectrum disorders, ELC = Early Learning Composite, IL-ASD = Infant siblings diagnosed with autism at 36-months, IL-N = Infant siblings not diagnosed with autism at 36-months, SD = Standard deviation, SS = standard scores

Assessing AOSI Item-Level Data and Dimensionality Reduction

AOSI Item-Level Correlations: Though most correlations between AOSI items were negligible (r = 0.00-0.10) or weak (r = 0.10 to 0.39), several moderate (r = 0.40 - 0.69; Patrick 2018) correlations were identified between AOSI items. Results are described in Appendix 2 Table A2.4.

Principal Axis and Factor Analysis of Item-Level AOSI Data: Since several weak or moderate correlations were found amongst item-level AOSI data, it was plausible that several AOSI items may be measuring the same feature or construct of ASD. Follow-up principal axis analysis of CISS-1 item-level AOSI data in SPSS identified five statistically significant eigenvalues which exceeded their simulated 95th-percentile eigenvalues based on a Monte Carlo simulation, indicating that factor analysis should extract for five factors. Parallel analysis results are described in Appendix 2 Table A2.5.

Follow-up factor analysis in SPSS (extracting for a fixed number of five factors/components) produced the pattern matrix described in Appendix 2 Table A2.6. Though many negligible or minor factor loading scores were identified (factor loading scores between 0 and |0.25|), between two and four AOSI items with moderate factor loading scores (>|0.25|) were loaded onto the five factors/components being extracted from the data. For the purposes of dimensionality reduction during classifier generation, a stringent factor loading cut-off was employed that excluded any AOSI items with factor loading score less than |0.6| to ensure only items that were strongly loaded to each factor were retained. This criterion resulted in only one AOSI item loading onto each of the five extracted factors/components: AOSI item 6: Imitation of Action, Item 8: Eye Contact, Item 14: Social Interest and Shared Affect, Item 16: Motor Control and Behaviour, and Item 18: Atypical Sensory Behaviours.

Assessing Factor Analysis Items for Group Differences in IL-ASD/N Participants: To explore if the AOSI items (6, 8, 14, 16, and 18) identified via factor analysis differed significantly between IL-ASD / IL-N participants in the CISS-1 dataset, independent Mann-Whitney U tests were conducted with Benjamini & Hochberg 1995 corrections. Three AOSI

items survived the corrected significance level of q* = 0.03: item 8: Eye Contact, Item 14: Social Interest and Affect, and Item 18: Atypical Sensory Behaviours.

ROC Curve Performance of 12-Month AOSI Total Score

ROC curve analyses of the entire cleaned CISS-1 IL-sibling dataset (n=465) were conducted to (1) assess the predictive utility of 12-month AOSI Total Score by itself at predicting 36-month ASD diagnostic status, and (2) provide a benchmark performance metric that learning classifier models can be compared against. The AUC for 12-month AOSI Total Score was 0.66 (99% CI 0.61, 0.72). The optimal Total Score cut point (as informed by Youden index calculations) was 7. These results are nearly identical to Zwaigenbaum et al., 2020 who conducted similar analyses in a slightly larger sample of the same CISS-1 IL-siblings than those used in this study (n=501 vs n=465). This sample discrepancy stemmed from Zwaigenbaum et al., 2020 not limiting IL-sibling data used during ROC curve analysis to only IL-siblings with 12-month data, as was done here. Results are detailed in Table 3.03.

Table 3.03 | ROC Characteristics for 12-month AOSI Total Score predicting IL-ASD at 36-months

Cutoff	Sens	Spec	PPV	NPV	Youden (J)
0	1.000	0.000	1.000	0.000	0.000
1	0.984	0.076	0.826	0.929	0.060
2	0.936	0.191	0.643	0.890	0.127
3	0.832	0.326	0.484	0.841	0.158
4	0.712	0.429	0.379	0.802	0.141
5	0.664	0.553	0.306	0.817	0.217
6	0.568	0.650	0.243	0.804	0.218
7[α]	0.528	0.741	0.208	0.810	0.269
8	0.464	0.797	0.176	0.802	0.261
9	0.368	0.850	0.137	0.785	0.218
10	0.312	0.900	0.113	0.781	0.212
11	0.256	0.929	0.092	0.773	0.185
12	0.208	0.947	0.075	0.765	0.155
13	0.176	0.959	0.063	0.760	0.135
14	0.128	0.968	0.046	0.751	0.096
15	0.080	0.985	0.029	0.744	0.065
16	0.056	0.991	0.020	0.741	0.047
17	0.016	0.997	0.006	0.734	0.013
18	0.016	1.000	0.006	0.734	0.016
19	0.008	1.000	0.003	0.733	0.008
20	0.000	1.000	0.000	0.731	0.000

J = Youden Index, NPV = negative predictive value, PPV = Positive predictive value, Sens = Sensitivity, Spec = Specificity.
Area under the ROC Curve = 0.665 (99% Confidence Interval = 0.608, 0.722)
[α] = Optimal AOSI Total Score cutoff (via Youden's *J*)

Statistical Analyses

Multivariate Logistic Regression Using the Default Decision Threshold

For logistic regression models employing the default decision threshold of 0.500, model performance varied across the three datasets. For models built without biological sex (L1-L15), average AUC in the training, testing, and independent validation datasets was 0.685, 0.682, and 0.638 respectively. Inclusion of biological sex (Models L16-L30) in regression modelling increased average AUC to 0.702, 0.704, and 0.686 when models were applied to training, testing, and independent validation datasets respectively. With respect to logistic regression model performance on testing and independent validation data, AUC values were largely consistent with each other. The average absolute difference in AUC values for models L1-L30 between the testing and independent validation set was 0.051 (minimum AUC difference = 0.001, maximum = 0.159).

Though all logistic regression models were characterized by poor sensitivity, all were defined by high specificity. For models built without biological sex (Models L1-L15), average sensitivity was 0.242, 0.159, and 0.223 while average specificity was 0.951, 0.936, and 0.889 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models L16-L30), average sensitivity was 0.244, 0.168, and 0.194 while average specificity was 0.953, 0.938, and 0.873 when applied to training, testing, and independent validation datasets. All logistic regression performance results using the default decision threshold of 0.500 are described in Table 3.04.

Multivariate Logistic Regression Using an Optimized Decision Thresholds

When logistic regression decision thresholds were optimized for maximum combined sensitivity and specificity, model sensitivity increased substantially while specificity was commensurately reduced. For models built with biological sex (Models L1-L15) using optimized decision thresholds, average sensitivity was 0.562, 0.838, and 0.612 while average specificity was 0.755, 0.522, and 0.662 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models L16-L30) using optimized decision thresholds, average sensitivity was 0.557, 0.806, and 0.826 while average specificity was 0.770, 0.588, and 0.549 when applied to training, testing, and independent validation datasets. Since only the classifier's

decision threshold (ranging in value between 0 and 1) was changed, AUC for these models remained unchanged. All results are described in detail in Table 3.05.

Truncated Multivariate Logistic Regression Model Performance

Three logistic regression models built with (Models L10, L11, and L14) and without biological sex (Models L25, L26, and L30) that were characterized by the highest AUC when applied to the testing dataset were systematically assessed for predictor variable importance on testing data. Across all six models, removal of non-significant variables led to an average increase in AUC of 0.015 and an average decrease of 4.920 in AIC. Variable importance assessment results for these truncated models are described in Table 3.06.

Table 3.04 | Multivariate Logistical Regression Model Performance

Model #	Variable Combination	Dataset	Threshold	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
Predictor Variable Combinations										
1	IL	Training	0.500	0.775	(0.729, 0.816)	0.725	0.356	0.937	0.685	0.790
		Testing	0.500	0.717	(0.614, 0.806)	0.657	0.143	0.887	0.273	0.778
		Independent	0.500	0.722	(0.618, 0.811)	0.685	0.304	0.866	0.438	0.784
2	IL + TS	Training	0.500	0.775	(0.729, 0.816)	0.725	0.356	0.937	0.685	0.790
		Testing	0.500	0.717	(0.614, 0.806)	0.657	0.143	0.887	0.273	0.778
		Independent	0.500	0.722	(0.618, 0.811)	0.685	0.304	0.866	0.438	0.784
3	IL + MSEL	Training	0.500	0.788	(0.743, 0.829)	0.736	0.356	0.955	0.755	0.793
		Testing	0.500	0.750	(0.649, 0.834)	0.681	0.286	0.887	0.429	0.808
		Independent	0.500	0.744	(0.642, 0.831)	0.701	0.435	0.851	0.500	0.814
4	IL + TS + MSEL	Training	0.500	0.788	(0.743, 0.829)	0.736	0.356	0.955	0.755	0.793
		Testing	0.500	0.750	(0.649, 0.834)	0.681	0.286	0.887	0.429	0.808
		Independent	0.500	0.744	(0.642, 0.831)	0.701	0.435	0.851	0.500	0.814
5	TS + MSEL	Training	0.500	0.743	(0.695, 0.786)	0.677	0.192	0.955	0.625	0.754
		Testing	0.500	0.783	(0.684, 0.862)	0.681	0.190	0.958	0.571	0.800
		Independent	0.500	0.733	(0.630, 0.821)	0.681	0.174	0.925	0.444	0.765
6	MSEL	Training	0.500	0.713	(0.664, 0.759)	0.635	0.029	0.978	0.333	0.723
		Testing	0.500	0.772	(0.672, 0.853)	0.690	0.000	1.000	0.000	0.772
		Independent	0.500	0.767	(0.666, 0.849)	0.624	0.130	0.985	0.750	0.767
7	TS	Training	0.500	0.745	(0.698, 0.789)	0.666	0.221	0.948	0.622	0.759
		Testing	0.500	0.761	(0.661, 0.844)	0.661	0.143	0.944	0.429	0.788
		Independent	0.500	0.722	(0.618, 0.811)	0.659	0.174	0.910	0.400	0.763
8	FA	Training	0.500	0.740	(0.692, 0.784)	0.659	0.192	0.952	0.606	0.753
		Testing	0.500	0.761	(0.661, 0.844)	0.650	0.143	0.944	0.429	0.788
		Independent	0.500	0.700	(0.594, 0.792)	0.496	0.087	0.910	0.250	0.744
9	FA + TS	Training	0.500	0.753	(0.706, 0.796)	0.677	0.240	0.952	0.658	0.764
		Testing	0.500	0.761	(0.661, 0.844)	0.690	0.190	0.930	0.444	0.795
		Independent	0.500	0.711	(0.606, 0.802)	0.674	0.174	0.896	0.364	0.759
10	FA + MSEL	Training	0.500	0.740	(0.692, 0.784)	0.680	0.212	0.944	0.595	0.756
		Testing	0.500	0.772	(0.672, 0.853)	0.728	0.143	0.958	0.500	0.791
		Independent	0.500	0.700	(0.594, 0.792)	0.569	0.174	0.881	0.333	0.756
11	FA + TS + MSEL	Training	0.500	0.751	(0.704, 0.794)	0.690	0.240	0.948	0.641	0.763
		Testing	0.500	0.761	(0.661, 0.844)	0.723	0.143	0.944	0.429	0.788
		Independent	0.500	0.689	(0.583, 0.782)	0.687	0.261	0.836	0.353	0.767
12	TRA	Training	0.500	0.743	(0.695, 0.786)	0.639	0.192	0.955	0.625	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.641	0.143	0.944	0.429	0.788
		Independent	0.500	0.700	(0.594, 0.792)	0.507	0.130	0.896	0.300	0.750
13	TRA + TS	Training	0.500	0.748	(0.701, 0.791)	0.670	0.212	0.955	0.647	0.758
		Testing	0.500	0.772	(0.672, 0.853)	0.669	0.143	0.958	0.500	0.791
		Independent	0.500	0.711	(0.606, 0.802)	0.653	0.174	0.896	0.364	0.759

14	TRA + MSEL	Training	0.500	0.737	(0.689, 0.781)	0.673	0.212	0.941	0.579	0.755
		Testing	0.500	0.772	(0.672, 0.853)	0.724	0.143	0.958	0.500	0.791
		Independent	0.500	0.711	(0.606, 0.802)	0.580	0.217	0.881	0.385	0.766
15	TRA + TS + MSEL	Training	0.500	0.759	(0.712, 0.801)	0.682	0.260	0.952	0.675	0.769
		Testing	0.500	0.772	(0.672, 0.853)	0.705	0.143	0.958	0.500	0.791
		Independent	0.500	0.700	(0.594, 0.792)	0.673	0.174	0.881	0.333	0.756
Predictor Variable Combinations + Gender										
16	Gender + IL	Training	0.500	0.783	(0.737, 0.824)	0.740	0.365	0.944	0.717	0.794
		Testing	0.500	0.717	(0.614, 0.806)	0.683	0.143	0.887	0.273	0.778
		Independent	0.500	0.711	(0.606, 0.802)	0.728	0.304	0.851	0.412	0.781
17	Gender + IL + TS	Training	0.500	0.783	(0.737, 0.824)	0.740	0.365	0.944	0.717	0.794
		Testing	0.500	0.717	(0.614, 0.806)	0.683	0.143	0.887	0.273	0.778
		Independent	0.500	0.711	(0.606, 0.802)	0.728	0.304	0.851	0.412	0.781
18	Gender + IL + MSEL	Training	0.500	0.780	(0.735, 0.821)	0.752	0.346	0.948	0.720	0.789
		Testing	0.500	0.750	(0.649, 0.834)	0.702	0.286	0.887	0.429	0.808
		Independent	0.500	0.700	(0.594, 0.792)	0.730	0.304	0.836	0.389	0.778
19	Gender + IL + TS + MSEL	Training	0.500	0.780	(0.735, 0.821)	0.752	0.346	0.948	0.720	0.789
		Testing	0.500	0.750	(0.649, 0.834)	0.702	0.286	0.887	0.429	0.808
		Independent	0.500	0.700	(0.594, 0.792)	0.730	0.304	0.836	0.389	0.778
20	Gender + TS + MSEL	Training	0.500	0.743	(0.685, 0.786)	0.689	0.212	0.948	0.611	0.757
		Testing	0.500	0.793	(0.696, 0.871)	0.707	0.238	0.958	0.625	0.810
		Independent	0.500	0.722	(0.618, 0.811)	0.727	0.174	0.910	0.400	0.763
21	Gender + MSEL*	Training	0.500	0.718	(0.670, 0.764)	0.649	0.058	0.974	0.462	0.728
		Testing	0.500	0.772	(0.672, 0.853)	0.714	0.000	1.000	0.000	0.772
		Independent	0.500	0.756	(0.654, 0.840)	0.644	0.174	0.955	0.571	0.771
22	Gender + TS	Training	0.500	0.753	(0.706, 0.796)	0.681	0.231	0.955	0.667	0.763
		Testing	0.500	0.804	(0.709, 0.880)	0.696	0.238	0.972	0.714	0.812
		Independent	0.500	0.733	(0.630, 0.821)	0.713	0.174	0.925	0.444	0.765
23	Gender + FA	Training	0.500	0.748	(0.701, 0.761)	0.680	0.192	0.963	0.667	0.755
		Testing	0.500	0.761	(0.661, 0.844)	0.672	0.095	0.958	0.400	0.782
		Independent	0.500	0.689	(0.583, 0.782)	0.610	0.087	0.896	0.222	0.741
24	Gender + FA + TS*	Training	0.500	0.761	(0.715, 0.804)	0.691	0.260	0.955	0.692	0.769
		Testing	0.500	0.783	(0.684, 0.862)	0.701	0.238	0.944	0.556	0.807
		Independent	0.500	0.644	(0.537, 0.743)	0.706	0.087	0.836	0.154	0.727
25	Gender + FA + MSEL	Training	0.500	0.737	(0.689, 0.781)	0.701	0.202	0.944	0.583	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.730	0.095	0.958	0.400	0.782
		Independent	0.500	0.689	(0.583, 0.782)	0.611	0.174	0.866	0.308	0.753
26	Gender + FA + TS + MSEL	Training	0.500	0.764	(0.718, 0.806)	0.701	0.260	0.959	0.711	0.770
		Testing	0.500	0.772	(0.672, 0.853)	0.739	0.190	0.944	0.500	0.798
		Independent	0.500	0.678	(0.571, 0.772)	0.713	0.217	0.836	0.313	0.757
27	Gender + TRA	Training	0.500	0.743	(0.695, 0.786)	0.671	0.163	0.967	0.654	0.749
		Testing	0.500	0.761	(0.661, 0.844)	0.675	0.095	0.958	0.400	0.782
		Independent	0.500	0.700	(0.594, 0.792)	0.613	0.087	0.910	0.250	0.744

28	Gender + TRA + TS	Training	0.500	0.748	(0.701, 0.791)	0.687	0.231	0.948	0.632	0.761
		Testing	0.500	0.783	(0.684, 0.862)	0.696	0.238	0.944	0.556	0.807
		Independent	0.500	0.689	(0.583, 0.782)	0.710	0.130	0.881	0.273	0.747
29	Gender + TRA + MSEL	Training	0.500	0.740	(0.692, 0.784)	0.694	0.202	0.948	0.600	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.723	0.095	0.958	0.400	0.782
		Independent	0.500	0.689	(0.583, 0.782)	0.617	0.174	0.866	0.308	0.753
30	Gender + TRA + TS + MSEL	Training	0.500	0.745	(0.698, 0.789)	0.700	0.231	0.944	0.615	0.760
		Testing	0.500	0.750	(0.649, 0.834)	0.734	0.143	0.930	0.375	0.786
		Independent	0.500	0.678	(0.571, 0.772)	0.709	0.217	0.836	0.313	0.757

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), Threshold = logistic regression decision threshold, TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Table 3.05 | Multivariate Logistical Regression Model Performance Optimized for Maximal Combined Sensitivity and Specificity

Model #	Variable Combination	Dataset	Threshold	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
Predictor Variable Combinations without sex										
1	IL	Training	0.352	0.777	(0.732, 0.819)	0.725	0.548	0.866	0.613	0.832
		Testing	0.162	0.489	(0.383, 0.596)	0.657	0.952	0.352	0.303	0.962
		Independent	0.225	0.622	(0.514, 0.722)	0.685	0.783	0.567	0.383	0.884
2	IL + TS	Training	0.352	0.777	(0.732, 0.819)	0.725	0.548	0.866	0.613	0.832
		Testing	0.162	0.489	(0.383, 0.596)	0.657	0.952	0.352	0.303	0.962
		Independent	0.225	0.622	(0.514, 0.722)	0.685	0.783	0.567	0.383	0.884
3	IL + MSEL	Training	0.291	0.735	(0.687, 0.779)	0.736	0.606	0.784	0.521	0.837
		Testing	0.177	0.609	(0.501, 0.709)	0.681	0.905	0.521	0.358	0.949
		Independent	0.340	0.722	(0.618, 0.811)	0.701	0.652	0.746	0.469	0.862
4	IL + TS + MSEL	Training	0.291	0.735	(0.687, 0.779)	0.736	0.606	0.784	0.521	0.837
		Testing	0.177	0.609	(0.501, 0.709)	0.681	0.905	0.521	0.358	0.949
		Independent	0.340	0.722	(0.618, 0.811)	0.701	0.652	0.746	0.469	0.862
5	TS + MSEL	Training	0.299	0.705	(0.656, 0.751)	0.677	0.538	0.770	0.475	0.812
		Testing	0.180	0.478	(0.373, 0.585)	0.681	1.000	0.324	0.304	1.000
		Independent	0.444	0.767	(0.666, 0.849)	0.681	0.391	0.896	0.563	0.811
6	MSEL	Training	0.260	0.622	(0.571, 0.671)	0.635	0.702	0.591	0.399	0.837
		Testing	0.292	0.717	(0.614, 0.806)	0.690	0.714	0.718	0.429	0.895
		Independent	0.339	0.678	(0.751, 0.772)	0.624	0.522	0.731	0.400	0.817
7	TS	Training	0.331	0.718	(0.670, 0.764)	0.666	0.481	0.810	0.495	0.801
		Testing	0.236	0.620	(0.512, 0.719)	0.661	0.667	0.606	0.333	0.860
		Independent	0.297	0.644	(0.537, 0.743)	0.659	0.565	0.672	0.371	0.818
8	FA	Training	0.239	0.665	(0.614, 0.713)	0.659	0.538	0.714	0.421	0.800
		Testing	0.234	0.609	(0.501, 0.709)	0.650	0.714	0.577	0.333	0.872
		Independent	0.752	0.767	(0.666, 0.849)	0.496	0.087	1.000	1.000	0.761
9	FA + TS	Training	0.320	0.718	(0.670, 0.764)	0.677	0.500	0.803	0.495	0.806
		Testing	0.198	0.543	(0.436, 0.648)	0.690	0.905	0.437	0.322	0.939
		Independent	0.204	0.567	(0.458, 0.671)	0.674	0.913	0.448	0.362	0.938
10	FA + MSEL	Training	0.220	0.598	(0.546, 0.648)	0.680	0.712	0.554	0.381	0.832
		Testing	0.257	0.696	(0.591, 0.787)	0.728	0.762	0.676	0.410	0.906
		Independent	0.308	0.622	(0.514, 0.722)	0.569	0.478	0.672	0.333	0.789
11	FA + TS + MSEL	Training	0.376	0.745	(0.698, 0.789)	0.690	0.433	0.866	0.556	0.798
		Testing	0.191	0.565	(0.458, 0.668)	0.723	1.000	0.437	0.344	1.000
		Independent	0.236	0.611	(0.503, 0.712)	0.687	0.826	0.537	0.380	0.900
12	TRA	Training	0.278	0.689	(0.639, 0.736)	0.639	0.462	0.777	0.444	0.789
		Testing	0.225	0.598	(0.490, 0.699)	0.641	0.667	0.577	0.318	0.854
		Independent	0.552	0.722	(0.618, 0.811)	0.507	0.130	0.925	0.375	0.756
13	TRA + TS	Training	0.309	0.713	(0.664, 0.759)	0.670	0.500	0.796	0.486	0.805
		Testing	0.236	0.641	(0.535, 0.739)	0.669	0.714	0.620	0.357	0.880
		Independent	0.188	0.500	(0.393, 0.607)	0.653	0.913	0.358	0.328	0.923

14	TRA + MSEL	Training	0.221	0.611	(0.560, 0.661)	0.673	0.731	0.565	0.394	0.844
		Testing	0.271	0.739	(0.637, 0.825)	0.724	0.714	0.746	0.455	0.898
		Independent	0.273	0.578	(0.469, 0.681)	0.580	0.609	0.567	0.326	0.809
15	TRA + TS + MSEL	Training	0.304	0.713	(0.664, 0.759)	0.682	0.529	0.784	0.487	0.812
		Testing	0.184	0.511	(0.404, 0.617)	0.705	1.000	0.366	0.318	1.000
		Independent	0.221	0.589	(0.480, 0.692)	0.673	0.870	0.493	0.370	0.917
Predictor Variable Combinations With Sex										
16	Gender + IL	Training	0.393	0.788	(0.743, 0.829)	0.740	0.500	0.900	0.658	0.823
		Testing	0.166	0.543	(0.436, 0.648)	0.683	0.952	0.423	0.328	0.968
		Independent	0.181	0.633	(0.525, 0.732)	0.728	0.913	0.537	0.404	0.947
17	Gender + IL + TS	Training	0.393	0.788	(0.743, 0.829)	0.740	0.500	0.900	0.658	0.823
		Testing	0.166	0.543	(0.436, 0.648)	0.683	0.952	0.423	0.328	0.968
		Independent	0.181	0.633	(0.525, 0.732)	0.728	0.913	0.537	0.404	0.947
18	Gender + IL + MSEL	Training	0.338	0.761	(0.715, 0.804)	0.752	0.577	0.833	0.571	0.836
		Testing	0.194	0.630	(0.523, 0.729)	0.702	0.857	0.563	0.367	0.930
		Independent	0.255	0.689	(0.583, 0.782)	0.730	0.783	0.657	0.439	0.898
19	Gender + IL + TS + MSEL	Training	0.338	0.761	(0.715, 0.804)	0.752	0.577	0.833	0.571	0.836
		Testing	0.177	0.598	(0.490, 0.699)	0.702	0.857	0.521	0.346	0.925
		Independent	0.255	0.689	(0.583, 0.782)	0.730	0.783	0.657	0.439	0.898
20	Gender + TS + MSEL	Training	0.298	0.689	(0.639, 0.736)	0.689	0.567	0.736	0.454	0.815
		Testing	0.226	0.598	(0.490, 0.699)	0.707	0.810	0.535	0.340	0.905
		Independent	0.212	0.622	(0.514, 0.722)	0.727	0.957	0.507	0.400	0.971
21	Gender + MSEL*	Training	0.219	0.525	(0.473, 0.577)	0.649	0.856	0.398	0.355	0.877
		Testing	0.325	0.783	(0.684, 0.862)	0.714	0.571	0.845	0.522	0.870
		Independent	0.371	0.722	(0.618, 0.811)	0.644	0.478	0.806	0.458	0.818
22	Gender + TS	Training	0.330	0.710	(0.662, 0.756)	0.681	0.462	0.807	0.480	0.795
		Testing	0.175	0.533	(0.426, 0.637)	0.696	0.952	0.408	0.323	0.967
		Independent	0.224	0.622	(0.514, 0.722)	0.713	0.826	0.552	0.388	0.902
23	Gender + FA	Training	0.248	0.657	(0.606, 0.705)	0.680	0.615	0.673	0.421	0.819
		Testing	0.300	0.761	(0.661, 0.844)	0.672	0.476	0.845	0.476	0.845
		Independent	0.210	0.544	(0.436, 0.650)	0.610	0.870	0.433	0.345	0.906
24	Gender + FA + TS*	Training	0.357	0.743	(0.695, 0.786)	0.691	0.433	0.862	0.549	0.797
		Testing	0.183	0.554	(0.447, 0.658)	0.701	0.905	0.451	0.328	0.941
		Independent	0.242	0.678	(0.571, 0.772)	0.706	0.783	0.642	0.429	0.896
25	Gender + FA + MSEL	Training	0.246	0.657	(0.606, 0.705)	0.701	0.731	0.628	0.432	0.858
		Testing	0.227	0.707	(0.602, 0.799)	0.730	0.857	0.662	0.429	0.940
		Independent	0.182	0.478	(0.371, 0.586)	0.611	0.913	0.328	0.318	0.917
26	Gender + FA + TS + MSEL	Training	0.387	0.748	(0.701, 0.791)	0.701	0.433	0.870	0.563	0.799
		Testing	0.226	0.641	(0.535, 0.739)	0.739	0.857	0.577	0.375	0.932
		Independent	0.231	0.633	(0.525, 0.732)	0.713	0.826	0.567	0.396	0.905
27	Gender + TRA	Training	0.243	0.654	(0.603, 0.702)	0.671	0.548	0.695	0.410	0.799
		Testing	0.330	0.761	(0.661, 0.844)	0.675	0.429	0.859	0.474	0.836
		Independent	0.216	0.544	(0.436, 0.650)	0.613	0.826	0.448	0.339	0.882

28	Gender + TRA + TS	Training	0.373	0.759	(0.712, 0.801)	0.687	0.423	0.888	0.595	0.799
		Testing	0.184	0.554	(0.447, 0.658)	0.696	0.952	0.437	0.333	0.969
		Independent	0.232	0.656	(0.548, 0.753)	0.710	0.826	0.597	0.413	0.909
29	Gender + TRA + MSEL	Training	0.258	0.665	(0.614, 0.713)	0.694	0.673	0.662	0.435	0.840
		Testing	0.235	0.696	(0.591, 0.787)	0.723	0.857	0.648	0.419	0.939
		Independent	0.270	0.589	(0.480, 0.692)	0.617	0.696	0.552	0.348	0.841
30	Gender + TRA + TS + MSEL	Training	0.393	0.788	(0.743, 0.829)	0.740	0.500	0.900	0.658	0.823
		Testing	0.166	0.543	(0.436, 0.648)	0.683	0.952	0.423	0.328	0.968
		Independent	0.180	0.567	(0.458, 0.671)	0.709	1.000	0.418	0.371	1.000

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), Threshold = logistic regression decision threshold, TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Table 3.06 | Assessment of 6 Logistic Regression Model (3 with Sex, 3 without) that had the highest AUC values on the testing dataset

Logistic Regression Variable Combination	Acc	95% CI	AUC	Sens	Spec	PPV	NPV	Res. Dev	AIC
Models without biological sex									
Model 10									
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.728	0.143	0.958	0.500	0.791	407.29	429.29
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.736	0.143	0.944	0.429	0.788	407.55	427.55
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.732	0.143	0.944	0.429	0.788	407.75	425.75
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.724	0.143	0.958	0.500	0.791	407.86	425.86
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.736	0.143	0.944	0.429	0.788	407.89	425.89
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.711	0.143	0.944	0.429	0.788	409.38	425.38
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.733	0.143	0.944	0.429	0.788	408.15	424.15
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.747	0.143	0.958	0.500	0.791	408.14	424.15
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.739	0.143	0.944	0.429	0.788	408.232	422.23
Model 11									
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.723	0.143	0.944	0.429	0.788	397.00	421.00
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.783	0.684, 0.862	0.717	0.143	0.972	0.600	0.793	397.96	419.96
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.724	0.143	0.944	0.429	0.788	397.00	419.00
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.783	0.684, 0.862	0.707	0.190	0.958	0.571	0.800	398.51	418.51
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.703	0.143	0.958	0.500	0.791	398.45	418.45
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.722	0.143	0.944	0.429	0.788	397.05	417.05
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.750	0.649, 0.844	0.715	0.143	0.930	0.375	0.786	397.92	417.92
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.722	0.143	0.958	0.500	0.791	397.15	417.15
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.729	0.143	0.944	0.429	0.788	397.19	417.19
AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.741	0.143	0.958	0.500	0.791	397.30	415.30
Model 14									
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.724	0.143	0.958	0.500	0.791	407.86	425.86
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.722	0.143	0.944	0.429	0.944	408.09	424.09
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.823	0.725	0.143	0.958	0.500	0.791	408.21	424.21
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.705	0.143	0.944	0.429	0.788	409.72	423.72
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.823	0.720	0.143	0.958	0.500	0.791	408.40	422.40
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.823	0.736	0.143	0.958	0.500	0.791	408.43	422.43
AQ8 + AQ14 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.823	0.730	0.143	0.958	0.500	0.791	408.49	420.49
Models with biological sex									
Model 25									
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.730	0.095	0.958	0.400	0.782	402.76	426.76
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.823	0.730	0.143	0.958	0.500	0.791	403.15	425.15
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.728	0.095	0.944	0.400	0.782	403.33	423.33
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.723	0.095	0.958	0.400	0.782	403.44	423.44
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.730	0.143	0.958	0.500	0.791	403.68	423.68
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.750	0.649, 0.834	0.709	0.095	0.944	0.333	0.779	405.12	423.12
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.738	0.143	0.958	0.500	0.791	403.90	421.90
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.756	0.143	0.958	0.500	0.791	404.28	420.28

Sex + ~~AQ6 +~~ AQ8 + AQ14 + AQ16 + AQ18 + ~~ELC +~~ VR + ~~FM + RL + EL~~	0.772	0.672, 0.853	0.753	0.143	0.958	0.500	0.791	404.30	418.30
Model 26									
Sex + AQ6 + AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.772	0.672,0.853	0.739	0.190	0.944	0.500	0.798	392.64	418.64
Sex + ~~AQ6 +~~ AQ8 + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.783	0.684, 0.862	0.745	0.190	0.958	0.571	0.800	393.36	417.36
Sex + ~~AQ6 + AQ8 +~~ AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.783	0.684, 0.862	0.745	0.190	0.958	0.571	0.800	393.41	415.41
Sex + ~~AQ6 + AQ8 + AQ14 +~~ AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.793	0.696, 0.871	0.723	0.238	0.958	0.625	0.809	394.83	414.83
Sex + ~~AQ6 + AQ8 +~~ AQ14 + ~~AQ16 +~~ AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.730	0.190	0.930	0.444	0.795	394.89	414.89
Sex + ~~AQ6 + AQ8 +~~ AQ14 + AQ16 + AQ18 + AQTS + ~~ELC +~~ VR + FM + RL + EL	0.783	0.684, 0.862	0.744	0.190	0.958	0.571	0.800	393.48	413.48
Sex + ~~AQ6 + AQ8 +~~ AQ14 + AQ16 + AQ18 + AQTS + ELC + ~~VR +~~ FM + RL + EL	0.772	0.672, 0.853	0.730	0.190	0.944	0.500	0.798	394.33	414.33
Sex + ~~AQ6 + AQ8 +~~ AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + ~~FM +~~ RL + EL	0.783	0.684, 0.862	0.741	0.190	0.958	0.571	0.800	393.54	413.54
Sex + ~~AQ6 + AQ8 +~~ AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + ~~RL +~~ EL	0.783	0.684, 0.862	0.744	0.190	0.958	0.571	0.800	393.67	413.67
Sex + ~~AQ6 + AQ8~~ + AQ14 + AQ16 + AQ18 + AQTS + ELC + VR + FM + RL + ~~EL~~	0.783	0.684, 0.862	0.745	0.190	0.958	0.571	0.800	393.45	413.45
Model 30									
Sex + AQ8 + AQ14 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.750	0.649, 0.834	0.734	0.143	0.930	0.375	0.786	394.82	416.82
Sex + ~~AQ8~~ + AQ14 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.660, 0.844	0.730	0.190	0.930	0.444	0.795	394.89	414.89
Sex + AQ8 + ~~AQ14~~ + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.772	0.672, 0.853	0.722	0.190	0.944	0.500	0.798	395.91	415.91
Sex + AQ8 + AQ14 + AQ18 + AQTS + ~~ELC~~ + VR + FM + RL + EL	0.761	0.660, 0.844	0.732	0.190	0.930	0.444	0.795	394.91	414.91
Sex + AQ8 + AQ14 + AQ18 + AQTS + ELC + ~~VR~~ + FM + RL + EL	0.750	0.649, 0.834	0.721	0.190	0.915	0.400	0.793	395.80	415.80
Sex + AQ8 + AQ14 + AQ18 + AQTS + ELC + VR + ~~FM~~ + RL + EL	0.772	0.672, 0.853	0.732	0.190	0.944	0.500	0.798	394.93	414.93
Sex + AQ8 + AQ14 + AQ18 + AQTS + ELC + VR + FM + ~~RL~~ + EL	0.772	0.672, 0.853	0.738	0.190	0.944	0.500	0.798	395.07	415.07
Sex + AQ8 + AQ14 + AQ18 + AQTS + ELC + VR + FM + RL + EL	0.761	0.661, 0.844	0.745	0.143	0.943	0.429	0.788	395.24	413.24

Acc = accuracy, AUC = Area Under the Curve, NPV = Negative Predictive Value, PPV = Positive Predictive Value, Res. Dev = Residual deviance, AQ = AOSI Question # (e.g., AQ3 = the third question on the AOSI). Variables were pruned if removal resulted in ↓AIC values and ↑AUC values. If removal did not result in this change for both AIC and AUC, it was not removed from the model. Variables with a strikethrough and in grey text indicate removal from the model.
= the predictor variable was significant ($p < 0.05$) in the original logistic regression model prior to pruning attempts

For the truncated regression models using the default decision threshold of 0.500, variable removal on average resulted in increased model performance with respect to AUC, sensitivity, and specificity. The average increase in AUC, sensitivity, and specificity of the truncated models was 0.000, -0.038, and -0.089 (for models applied to training data), 0.030, 0.030, and 0.008 (for models applied to testing data), and 0.019, 0.019, and 0.022 (for models applied to independent validation data).

For the truncated regression models calculated using optimized decision threshold, variable removal on average resulted in increased model performance with respect to AUC, sensitivity, and specificity. The average increase in AUC, sensitivity, and specificity of the truncated models was 0.000, 0.112, and -0.194 (for models applied to training data), 0.013, -0.024, and 0.009 (for models applied to testing data), and 0.015, -0.007, and -0.027 (for models applied to independent validation data). Truncated model performance metrics for the six models with the highest AUC built with/without biological sex (models L10, L11, L14, L25, L26, and L30) pre and post non-significant variable removal are described in Table 3.07.

Table 3.07 | Performance of Pruned Logistic Regression Models Pre- and Post-Optimization

Model #	Variable Combination without biological sex	Data	Threshold	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
10	FA + MSEL	Training	0.500	0.740	(0.692, 0.784)	0.680	0.212	0.944	0.595	0.756
		Training	0.220	0.598	(0.546, 0.648)	0.680	0.712	0.554	0.381	0.832
		Testing	0.500	0.772	(0.672, 0.853)	0.728	0.143	0.958	0.500	0.791
		Testing	0.257	0.696	(0.591, 0.787)	0.728	0.762	0.676	0.410	0.906
		Independent	0.500	0.700	(0.594, 0.792)	0.569	0.174	0.881	0.333	0.756
		Independent	0.308	0.622	(0.514, 0.722)	0.569	0.478	0.672	0.333	0.789
10	FA + MSEL - Pruned/Truncated	Training	0.500	0.700	(0.594, 0.792)	0.678	0.174	0.881	0.333	0.756
		Training	0.226	0.444	(0.340, 0.553)	0.678	0.739	0.343	0.279	0.793
		Testing	0.500	0.772	(0.672, 0.853)	0.747	0.143	0.958	0.500	0.791
		Testing	0.229	0.674	(0.568, 0.768)	0.747	0.857	0.620	0.400	0.936
		Independent	0.500	0.700	(0.594, 0.792)	0.596	0.174	0.881	0.333	0.756
		Independent	0.265	0.600	(0.491, 0.702)	0.596	0.652	0.582	0.349	0.830
11	FA + TS + MSEL	Training	0.500	0.751	(0.704, 0.794)	0.690	0.240	0.948	0.641	0.763
		Training	0.376	0.745	(0.698, 0.789)	0.690	0.433	0.866	0.556	0.798
		Testing	0.500	0.761	(0.661, 0.844)	0.723	0.143	0.944	0.429	0.788
		Testing	0.191	0.565	(0.458, 0.668)	0.723	1.000	0.437	0.344	1.000
		Independent	0.500	0.689	(0.583, 0.782)	0.687	0.261	0.836	0.353	0.767
		Independent	0.236	0.611	(0.503, 0.712)	0.687	0.826	0.537	0.380	0.900
11	FA + TS + MSEL - Pruned/Truncated	Training	0.500	0.678	(0.571, 0.772)	0.691	0.217	0.836	0.313	0.757
		Training	0.350	0.678	(0.571, 0.772)	0.691	0.435	0.761	0.385	0.797
		Testing	0.500	0.772	(0.672, 0.853)	0.741	0.143	0.958	0.500	0.791
		Testing	0.190	0.565	(0.458, 0.668)	0.741	1.000	0.437	0.344	1.000
		Independent	0.500	0.678	(0.571, 0.772)	0.701	0.217	0.836	0.313	0.757
		Independent	0.289	0.578	(0.469, 0.681)	0.701	0.478	0.612	0.297	0.774
14	TRA + MSEL	Training	0.500	0.737	(0.689, 0.781)	0.673	0.212	0.941	0.579	0.755
		Training	0.221	0.611	(0.560, 0.661)	0.673	0.731	0.565	0.394	0.844
		Testing	0.500	0.772	(0.672, 0.853)	0.724	0.143	0.958	0.500	0.791
		Testing	0.271	0.739	(0.637, 0.825)	0.724	0.714	0.746	0.455	0.898
		Independent	0.500	0.711	(0.606, 0.802)	0.580	0.217	0.881	0.385	0.766
		Independent	0.273	0.578	(0.469, 0.681)	0.580	0.609	0.567	0.326	0.809
14	TRA + MSEL - Pruned/Truncated	Training	0.500	0.700	(0.594, 0.792)	0.673	0.174	0.881	0.333	0.756
		Training	0.221	0.467	(0.575, 0.673)	0.673	0.826	0.343	0.302	0.852
		Testing	0.500	0.772	(0.672, 0.853)	0.736	0.143	0.958	0.500	0.791
		Testing	0.269	0.728	(0.626, 0.816)	0.736	0.714	0.732	0.441	0.897
		Independent	0.500	0.700	(0.594, 0.792)	0.604	0.174	0.881	0.333	0.756
		Independent	0.259	0.600	(0.491, 0.702)	0.604	0.652	0.582	0.349	0.830

Model #	Variable Combinations with biological sex	Data	Threshold	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
25	Gender + FA + MSEL	Training	0.500	0.737	(0.689, 0.781)	0.701	0.202	0.944	0.583	0.754
		Training	0.246	0.657	(0.606, 0.705)	0.701	0.731	0.628	0.432	0.858
		Testing	0.500	0.761	(0.661, 0.844)	0.730	0.095	0.958	0.400	0.782

		Testing	0.227	0.707	(0.602, 0.799)	0.730	0.857	0.662	0.429	0.940
		Independent	0.500	0.689	(0.583, 0.782)	0.611	0.174	0.866	0.308	0.753
		Independent	0.182	0.478	(0.371, 0.586)	0.611	0.913	0.328	0.318	0.917
25	Gender + FA + MSEL - Pruned/Truncated	Training	0.500	0.689	(0.583, 0.782)	0.698	0.130	0.881	0.273	0.747
		Training	0.255	0.600	(0.491, 0.702)	0.698	0.652	0.582	0.349	0.830
		Testing	0.500	0.772	(0.672, 0.853)	0.757	0.143	0.958	0.500	0.791
		Testing	0.242	0.739	(0.637, 0.825)	0.757	0.857	0.704	0.462	0.943
		Independent	0.500	0.689	(0.583, 0.782)	0.648	0.130	0.881	0.273	0.747
		Independent	0.199	0.522	(0.414, 0.629)	0.648	0.957	0.373	0.344	0.962
26	Gender + FA + TS + MSEL	Training	0.500	0.764	(0.718, 0.806)	0.701	0.260	0.959	0.711	0.770
		Training	0.387	0.748	(0.701, 0.791)	0.701	0.433	0.870	0.563	0.799
		Testing	0.500	0.772	(0.672, 0.853)	0.739	0.190	0.944	0.500	0.798
		Testing	0.226	0.641	(0.535, 0.739)	0.739	0.857	0.577	0.375	0.932
		Independent	0.500	0.678	(0.571, 0.772)	0.713	0.217	0.836	0.313	0.757
		Independent	0.231	0.633	(0.525, 0.732)	0.713	0.826	0.567	0.396	0.905
26	Gender + FA + TS + MSEL - Pruned/Truncated	Training	0.500	0.678	(0.571, 0.772)	0.702	0.217	0.836	0.313	0.757
		Training	0.255	0.611	(0.503, 0.712)	0.702	0.739	0.567	0.370	0.864
		Testing	0.500	0.783	(0.684, 0.862)	0.745	0.190	0.958	0.571	0.800
		Testing	0.242	0.663	(0.557, 0.758)	0.745	0.714	0.648	0.375	0.885
		Independent	0.500	0.678	(0.571, 0.772)	0.709	0.217	0.836	0.313	0.757
		Independent	0.199	0.522	(0.414, 0.629)	0.709	0.957	0.373	0.344	0.962
30	Gender + TRA + TS + MSEL	Training	0.500	0.783	(0.737, 0.824)	0.740	0.365	0.944	0.717	0.794
		Training	0.393	0.788	(0.743, 0.829)	0.740	0.500	0.900	0.658	0.823
		Testing	0.500	0.717	(0.614, 0.806)	0.683	0.143	0.887	0.273	0.778
		Testing	0.166	0.543	(0.436, 0.648)	0.683	0.952	0.423	0.328	0.968
		Independent	0.500	0.678	(0.571, 0.772)	0.709	0.217	0.836	0.313	0.757
		Independent	0.180	0.567	(0.458, 0.671)	0.709	1.000	0.418	0.371	1.000
30	Gender + TRA + TS + MSEL - Pruned/Truncated	Training	0.500	0.678	(0.571, 0.772)	0.700	0.217	0.836	0.313	0.757
		Training	0.255	0.633	(0.525, 0.732)	0.700	0.783	0.582	0.391	0.886
		Testing	0.500	0.761	(0.661, 0.844)	0.745	0.143	0.944	0.429	0.788
		Testing	0.242	0.641	(0.535, 0.739)	0.745	0.714	0.620	0.357	0.880
		Independent	0.500	0.678	(0.571, 0.772)	0.724	0.217	0.836	0.313	0.757
		Independent	0.199	0.556	(0.447, 0.660)	0.724	0.957	0.418	0.361	0.966

AUC = area under the curve, CI = confidence interval, Data = the data the model in question was being applied to – the randomly partitioned training set or test set, EL = Mullen expressive language standard score, ELC = Mullen Early Learning Composite standard score, FM = Mullen Fine Motor standard score, NPV = negative predictive value, PPV = positive predictive value, RL = Mullen Receptive Language standard score, Sens = sensitivity, Spec = specificity, Threshold = logistic regression decision threshold, VR = Mullen Visual Reception standard score

Regularized Multivariate Logistic Regression Using the Default Decision Threshold

For regularized logistic regression models employing the default decision threshold of 0.500 built without biological sex, average AUC in the training, testing, and independent validation datasets was 0.656, 0.653, and 0.625 respectively. Inclusion of biological sex in regularized regression modelling resulted in a minor increase of average AUC to 0.683, 0.687, and 0.667 when applied to training, testing, and independent validation datasets respectively. With respect to regularized regression model performance on testing and independent validation data, AUC values were largely consistent with each other. The average absolute difference in AUC values for all regularized classifiers (models R1-R30) between the testing and independent validation set was 0.054 (min difference in AUC = 0.002, max difference = 0.162).

Though regularized logistic regression classifiers were characterized by poor sensitivity, all were defined by high specificity. For models built without biological sex (Models R1-R15), average sensitivity was 0.151, 0.076, and 0.122 while average specificity was 0.971, 0.962, and 0.933 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models R16-R30), average sensitivity was 0.206, 0.130, and 0.133 while average specificity was 0.961, 0.950, and 0.905 when applied to training, testing, and independent validation datasets. All regularized logistic regression performance results using the default decision threshold of 0.500 are described in Table 3.08.

Regularized Multivariate Logistic Regression Using an Optimized Decision Threshold

When logistic regression decision thresholds were optimized for maximum combined sensitivity and specificity, model sensitivity increased substantially while specificity was reduced. For models built with biological sex (Models R1-R15) using optimized decision thresholds, average sensitivity was 0.576, 0.819, and 0.620 while average specificity was 0.721, 0.513, and 0.620 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models R16-R30) using optimized decision threshold, average sensitivity was 0.547, 0.806, and 0.853 while average specificity was 0.753, 0.550, and 0.565 when applied to training, testing, and independent validation datasets. Since only the classifier's decision threshold was changed, AUC for these models remained unchanged. All results are described in detail in Table 3.09.

Table 3.08 | Regularized Logistic Regression Model Performance

Model #	Variable Combination	Dataset	Threshold	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
Predictor Variable Combinations Without Sex										
1	IL	Training	0.500	0.775	(0.729, 0.816)	0.716	0.327	0.948	0.708	0.785
		Testing	0.500	0.739	(0.637, 0.825)	0.639	0.190	0.901	0.364	0.790
		Independent	0.500	0.733	(0.630, 0.821)	0.713	0.304	0.881	0.467	0.787
2	IL + TS	Training	0.500	0.748	(0.701, 0.791)	0.701	0.163	0.974	0.708	0.751
		Testing	0.500	0.761	(0.661, 0.844)	0.657	0.095	0.958	0.400	0.782
		Independent	0.500	0.767	(0.666, 0.849)	0.728	0.174	0.970	0.667	0.774
3	IL + MSEL	Training	0.500	0.775	(0.729, 0.816)	0.705	0.337	0.944	0.700	0.786
		Testing	0.500	0.739	(0.637, 0.825)	0.641	0.048	0.944	0.200	0.770
		Independent	0.500	0.700	(0.594, 0.792)	0.674	0.304	0.836	0.389	0.778
4	IL + TS + MSEL	Training	0.500	0.751	(0.704, 0.794)	0.686	0.231	0.952	0.649	0.762
		Testing	0.500	0.750	(0.649, 0.834)	0.608	0.048	0.958	0.250	0.773
		Independent	0.500	0.689	(0.583, 0.782)	0.596	0.174	0.866	0.308	0.753
5	TS + MSEL	Training	0.500	0.729	(0.681, 0.774)	0.670	0.038	0.996	0.800	0.728
		Testing	0.500	0.761	(0.661, 0.844)	0.722	0.000	0.986	0.000	0.769
		Independent	0.500	0.744	(0.642, 0.831)	0.708	0.043	0.985	0.500	0.750
6	MSEL	Training	0.500	0.721	(0.673, 0.766)	0.583	0.000	1.000	0.000	0.721
		Testing	0.500	0.772	(0.672, 0.853)	0.662	0.000	1.000	0.000	0.772
		Independent	0.500	0.744	(0.642, 0.831)	0.672	0.000	1.000	0.000	0.744
7	TS	Training	0.500	0.735	(0.687, 0.779)	0.666	0.067	0.993	0.778	0.734
		Testing	0.500	0.761	(0.661, 0.844)	0.661	0.000	0.986	0.000	0.769
		Independent	0.500	0.756	(0.654, 0.840)	0.659	0.087	0.985	0.667	0.759
8	FA	Training	0.500	0.745	(0.698, 0.789)	0.654	0.192	0.959	0.645	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.655	0.143	0.944	0.429	0.788
		Independent	0.500	0.700	(0.594, 0.792)	0.493	0.087	0.910	0.250	0.744
9	FA + TS	Training	0.500	0.721	(0.673, 0.766)	0.543	0.000	1.000	0.000	0.721
		Testing	0.500	0.772	(0.672, 0.853)	0.532	0.000	1.000	0.000	0.772
		Independent	0.500	0.744	(0.642, 0.831)	0.579	0.000	1.000	0.000	0.744
10	FA + MSEL	Training	0.500	0.751	(0.704, 0.794)	0.671	0.154	0.981	0.762	0.750
		Testing	0.500	0.772	(0.672, 0.853)	0.716	0.095	0.972	0.500	0.784
		Independent	0.500	0.733	(0.630, 0.821)	0.625	0.130	0.940	0.429	0.759
11	FA + TS + MSEL	Training	0.500	0.732	(0.684, 0.776)	0.614	0.173	0.948	0.563	0.748
		Testing	0.500	0.750	(0.649, 0.834)	0.628	0.095	0.944	0.333	0.779
		Independent	0.500	0.711	(0.606, 0.802)	0.546	0.130	0.910	0.333	0.753
12	TRA	Training	0.500	0.743	(0.695, 0.786)	0.628	0.192	0.955	0.625	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.615	0.143	0.944	0.429	0.788
		Independent	0.500	0.700	(0.594, 0.792)	0.502	0.130	0.896	0.300	0.750
13	TRA + TS	Training	0.500	0.724	(0.675, 0.769)	0.657	0.010	1.000	1.000	0.723
		Testing	0.500	0.772	(0.672, 0.853)	0.623	0.000	1.000	NaN	0.772
		Independent	0.500	0.744	(0.642, 0.831)	0.646	0.000	1.000	NaN	0.744

14	TRA + MSEL	Training	0.500	0.743	(0.695, 0.786)	0.673	0.183	0.959	0.633	0.752
		Testing	0.500	0.761	(0.661, 0.844)	0.714	0.143	0.944	0.429	0.788
		Independent	0.500	0.711	(0.606, 0.802)	0.583	0.130	0.910	0.333	0.753
15	TRA + TS + MSEL	Training	0.500	0.745	(0.698, 0.789)	0.671	0.202	0.955	0.636	0.756
		Testing	0.500	0.772	(0.672, 0.853)	0.716	0.143	0.958	0.500	0.791
		Independent	0.500	0.711	(0.606, 0.802)	0.653	0.130	0.910	0.333	0.753
Predictor Variable Combinations With Sex										
16	Gender + IL	Training	0.500	0.783	(0.737, 0.824)	0.735	0.337	0.955	0.745	0.788
		Testing	0.500	0.728	(0.637, 0.825)	0.664	0.143	0.901	0.300	0.780
		Independent	0.500	0.689	(0.583, 0.782)	0.741	0.217	0.851	0.333	0.760
17	Gender + IL + TS	Training	0.500	0.783	(0.718, 0.806)	0.735	0.365	0.944	0.717	0.794
		Testing	0.500	0.750	(0.672, 0.853)	0.666	0.143	0.930	0.375	0.786
		Independent	0.500	0.689	(0.583, 0.782)	0.742	0.217	0.851	0.333	0.760
18	Gender + IL + MSEL	Training	0.500	0.769	(0.726, 0.814)	0.717	0.288	0.955	0.714	0.776
		Testing	0.500	0.739	(0.626, 0.816)	0.661	0.190	0.901	0.364	0.790
		Independent	0.500	0.733	(0.630, 0.821)	0.680	0.348	0.866	0.471	0.795
19	Gender + IL + TS + MSEL	Training	0.500	0.756	(0.732, 0.819)	0.709	0.288	0.937	0.638	0.773
		Testing	0.500	0.750	(0.637, 0.825)	0.632	0.095	0.944	0.333	0.779
		Independent	0.500	0.733	(0.618, 0.811)	0.623	0.217	0.910	0.455	0.772
20	Gender + TS + MSEL	Training	0.500	0.724	(0.689, 0.781)	0.683	0.010	1.000	1.000	0.723
		Testing	0.500	0.772	(0.649, 0.834)	0.745	0.000	1.000	0.000	0.772
		Independent	0.500	0.744	(0.666, 0.849)	0.737	0.000	1.000	0.000	0.744
21	Gender + MSEL*	Training	0.500	0.721	(0.673, 0.766)	0.598	0.000	1.000	0.000	0.721
		Testing	0.500	0.772	(0.672, 0.853)	0.676	0.000	1.000	0.000	0.772
		Independent	0.500	0.744	(0.642, 0.831)	0.685	0.000	1.000	0.000	0.744
22	Gender + TS	Training	0.500	0.759	(0.706, 0.796)	0.678	0.260	0.952	0.675	0.769
		Testing	0.500	0.793	(0.696, 0.871)	0.703	0.238	0.958	0.625	0.810
		Independent	0.500	0.711	(0.618, 0.811)	0.713	0.174	0.896	0.364	0.759
23	Gender + FA	Training	0.500	0.745	(0.695, 0.786)	0.680	0.192	0.959	0.645	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.692	0.095	0.958	0.400	0.782
		Independent	0.500	0.678	(0.594, 0.792)	0.624	0.087	0.881	0.200	0.738
24	Gender + FA + TS*	Training	0.500	0.756	(0.698, 0.789)	0.686	0.192	0.974	0.741	0.757
		Testing	0.500	0.793	(0.696, 0.871)	0.699	0.190	0.972	0.667	0.802
		Independent	0.500	0.733	(0.571, 0.772)	0.701	0.087	0.955	0.400	0.753
25	Gender + FA + MSEL	Training	0.500	0.745	(0.701, 0.791)	0.692	0.183	0.963	0.655	0.753
		Testing	0.500	0.761	(0.672, 0.853)	0.706	0.095	0.958	0.400	0.782
		Independent	0.500	0.678	(0.618, 0.811)	0.644	0.087	0.881	0.200	0.738
26	Gender + FA + TS + MSEL	Training	0.500	0.737	(0.704, 0.794)	0.627	0.125	0.974	0.650	0.742
		Testing	0.500	0.750	(0.684, 0.862)	0.664	0.048	0.958	0.250	0.773
		Independent	0.500	0.700	(0.594, 0.792)	0.507	0.087	0.910	0.250	0.744
27	Gender + TRA	Training	0.500	0.735	(0.695, 0.786)	0.655	0.212	0.937	0.564	0.754
		Testing	0.500	0.761	(0.661, 0.844)	0.658	0.238	0.915	0.455	0.802
		Independent	0.500	0.689	(0.594, 0.792)	0.602	0.087	0.896	0.222	0.741

28	Gender + TRA + TS	Training	0.500	0.748	(0.698, 0.789)	0.690	0.221	0.952	0.639	0.760
		Testing	0.500	0.783	(0.684, 0.862)	0.690	0.190	0.958	0.571	0.800
		Independent	0.500	0.689	(0.583, 0.782)	0.701	0.130	0.881	0.273	0.747
29	Gender + TRA + MSEL	Training	0.500	0.748	(0.695, 0.786)	0.677	0.192	0.963	0.667	0.755
		Testing	0.500	0.772	(0.661, 0.844)	0.734	0.143	0.958	0.500	0.791
		Independent	0.500	0.700	(0.594, 0.792)	0.642	0.130	0.896	0.300	0.750
30	Gender + TRA + TS + MSEL	Training	0.500	0.751	(0.709, 0.799)	0.676	0.221	0.955	0.657	0.760
		Testing	0.500	0.761	(0.672, 0.853)	0.708	0.143	0.944	0.429	0.788
		Independent	0.500	0.711	(0.594, 0.792)	0.668	0.130	0.910	0.333	0.753

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), Threshold = logistic regression decision threshold, TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Table 3.09 | Regularized logistic regression model performance optimized for maximal combined sensitivity + specificity

Model #	Variable Combination	Dataset	Threshold	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
Predictor Variable Combinations										
1	IL	Training	0.365	0.775	(0.729, 0.816)	0.716	0.519	0.874	0.614	0.825
		Testing	0.184	0.522	(0.415, 0.627)	0.637	0.810	0.437	0.298	0.886
		Independent	0.275	0.656	(0.548, 0.753)	0.713	0.783	0.612	0.409	0.891
2	IL + TS	Training	0.362	0.775	(0.729, 0.816)	0.713	0.519	0.874	0.614	0.825
		Testing	0.189	0.522	(0.415, 0.627)	0.645	0.857	0.423	0.305	0.909
		Independent	0.269	0.656	(0.548, 0.753)	0.710	0.783	0.612	0.409	0.891
3	IL + MSEL	Training	0.377	0.775	(0.729, 0.816)	0.717	0.481	0.888	0.625	0.816
		Testing	0.183	0.500	(0.394, 0.606)	0.662	1.000	0.352	0.313	1.000
		Independent	0.261	0.678	(0.571, 0.772)	0.695	0.696	0.672	0.421	0.865
4	IL + TS + MSEL	Training	0.262	0.710	(0.662, 0.756)	0.714	0.635	0.740	0.485	0.840
		Testing	0.184	0.478	(0.373, 0.585)	0.647	1.000	0.324	0.304	1.000
		Independent	0.249	0.689	(0.583, 0.782)	0.689	0.739	0.672	0.436	0.882
5	TS + MSEL	Training	0.328	0.724	(0.675, 0.769)	0.670	0.452	0.829	0.505	0.796
		Testing	0.232	0.609	(0.501, 0.709)	0.703	0.857	0.535	0.353	0.927
		Independent	0.341	0.733	(0.630, 0.821)	0.683	0.478	0.821	0.478	0.821
6	MSEL	Training	0.273	0.643	(0.592, 0.692)	0.624	0.654	0.639	0.412	0.827
		Testing	0.283	0.728	(0.626, 0.816)	0.682	0.619	0.761	0.433	0.871
		Independent	0.288	0.756	(0.654, 0.840)	0.732	0.609	0.806	0.519	0.857
7	TS	Training	0.334	0.718	(0.670, 0.764)	0.666	0.481	0.810	0.495	0.801
		Testing	0.245	0.620	(0.512, 0.719)	0.661	0.667	0.606	0.333	0.860
		Independent	0.303	0.644	(0.537, 0.743)	0.659	0.565	0.672	0.371	0.818
8	FA	Training	0.240	0.665	(0.614, 0.713)	0.658	0.538	0.714	0.421	0.800
		Testing	0.230	0.598	(0.490, 0.699)	0.648	0.714	0.563	0.326	0.870
		Independent	0.720	0.756	(0.654, 0.840)	0.496	0.087	0.985	0.667	0.759
9	FA + TS	Training	0.309	0.705	(0.656, 0.751)	0.670	0.529	0.773	0.474	0.809
		Testing	0.237	0.630	(0.523, 0.729)	0.671	0.762	0.592	0.356	0.894
		Independent	0.248	0.567	(0.458, 0.671)	0.634	0.739	0.507	0.340	0.850
10	FA + MSEL	Training	0.218	0.550	(0.498, 0.601)	0.653	0.760	0.468	0.356	0.834
		Testing	0.221	0.565	(0.458, 0.668)	0.689	0.905	0.465	0.333	0.943
		Independent	0.225	0.533	(0.425, 0.639)	0.620	0.783	0.448	0.327	0.857
11	FA + TS + MSEL	Training	0.247	0.654	(0.603, 0.702)	0.670	0.596	0.677	0.416	0.813
		Testing	0.210	0.565	(0.458, 0.668)	0.712	0.905	0.465	0.333	0.943
		Independent	0.213	0.533	(0.425, 0.639)	0.638	0.826	0.433	0.333	0.879
12	TRA	Training	0.263	0.689	(0.639, 0.736)	0.639	0.462	0.777	0.444	0.789
		Testing	0.232	0.598	(0.490, 0.699)	0.641	0.667	0.577	0.318	0.854
		Independent	0.521	0.722	(0.618, 0.811)	0.507	0.130	0.925	0.375	0.756
13	TRA + TS	Training	0.316	0.713	(0.664, 0.759)	0.669	0.500	0.796	0.486	0.805
		Testing	0.236	0.630	(0.523, 0.729)	0.675	0.762	0.592	0.356	0.894
		Independent	0.206	0.500	(0.393, 0.607)	0.648	0.913	0.358	0.328	0.923

14	TRA + MSEL	Training	0.236	0.574	(0.522, 0.624)	0.650	0.731	0.513	0.367	0.831
		Testing	0.238	0.598	(0.490, 0.699)	0.699	0.905	0.507	0.352	0.947
		Independent	0.243	0.611	(0.503, 0.712)	0.622	0.696	0.582	0.364	0.848
15	TRA + TS + MSEL	Training	0.202	0.542	(0.489, 0.593)	0.666	0.788	0.446	0.355	0.845
		Testing	0.209	0.576	(0.469, 0.679)	0.688	0.857	0.493	0.333	0.921
		Independent	0.319	0.711	(0.606, 0.802)	0.626	0.478	0.791	0.440	0.815
Predictor Variable Combinations + Gender										
16	Gender + IL	Training	0.357	0.756	(0.709, 0.799)	0.735	0.519	0.848	0.568	0.820
		Testing	0.194	0.511	(0.404, 0.617)	0.662	0.905	0.394	0.306	0.933
		Independent	0.195	0.600	(0.491, 0.702)	0.741	0.913	0.493	0.382	0.943
17	Gender + IL + TS	Training	0.282	0.702	(0.653, 0.748)	0.734	0.644	0.725	0.475	0.841
		Testing	0.208	0.565	(0.458, 0.668)	0.691	0.857	0.479	0.327	0.919
		Independent	0.217	0.633	(0.525, 0.732)	0.739	1.000	0.507	0.411	1.000
18	Gender + IL + MSEL	Training	0.345	0.759	(0.712, 0.801)	0.722	0.519	0.851	0.574	0.821
		Testing	0.172	0.500	(0.394, 0.606)	0.649	1.000	0.352	0.313	1.000
		Independent	0.276	0.711	(0.606, 0.802)	0.711	0.696	0.716	0.457	0.873
19	Gender + IL + TS + MSEL	Training	0.277	0.724	(0.675, 0.769)	0.732	0.615	0.766	0.504	0.837
		Testing	0.182	0.522	(0.415, 0.627)	0.666	0.952	0.394	0.317	0.966
		Independent	0.226	0.689	(0.583, 0.782)	0.755	0.870	0.627	0.444	0.933
20	Gender + TS + MSEL	Training	0.327	0.718	(0.670, 0.764)	0.683	0.462	0.818	0.495	0.797
		Testing	0.259	0.652	(0.546, 0.749)	0.719	0.714	0.634	0.366	0.882
		Independent	0.240	0.611	(0.503, 0.712)	0.711	0.913	0.507	0.389	0.944
21	Gender + MSEL*	Training	0.209	0.499	(0.447, 0.551)	0.628	0.913	0.338	0.348	0.910
		Testing	0.315	0.761	(0.661, 0.844)	0.695	0.524	0.831	0.478	0.855
		Independent	0.312	0.733	(0.630, 0.821)	0.722	0.522	0.806	0.480	0.831
22	Gender + TS	Training	0.366	0.727	(0.678, 0.771)	0.680	0.433	0.840	0.511	0.793
		Testing	0.214	0.554	(0.447, 0.658)	0.700	0.857	0.465	0.321	0.917
		Independent	0.271	0.667	(0.559, 0.763)	0.718	0.826	0.612	0.422	0.911
23	Gender + FA	Training	0.291	0.697	(0.648, 0.743)	0.671	0.490	0.777	0.459	0.798
		Testing	0.305	0.750	(0.649, 0.834)	0.679	0.476	0.831	0.455	0.843
		Independent	0.277	0.589	(0.480, 0.692)	0.639	0.826	0.507	0.365	0.895
24	Gender + FA + TS*	Training	0.369	0.756	(0.709, 0.799)	0.689	0.442	0.877	0.582	0.803
		Testing	0.171	0.511	(0.404, 0.617)	0.699	0.952	0.380	0.313	0.964
		Independent	0.261	0.678	(0.571, 0.772)	0.700	0.826	0.627	0.432	0.913
25	Gender + FA + MSEL	Training	0.222	0.571	(0.519, 0.622)	0.659	0.750	0.502	0.368	0.839
		Testing	0.229	0.641	(0.535, 0.739)	0.710	0.905	0.563	0.380	0.952
		Independent	0.229	0.622	(0.514, 0.722)	0.629	0.739	0.582	0.378	0.867
26	Gender + FA + TS + MSEL	Training	0.339	0.740	(0.692, 0.784)	0.686	0.413	0.866	0.544	0.793
		Testing	0.218	0.630	(0.523, 0.729)	0.732	0.857	0.563	0.367	0.930
		Independent	0.207	0.611	(0.503, 0.712)	0.703	0.913	0.507	0.389	0.944
27	Gender + TRA	Training	0.298	0.678	(0.628, 0.725)	0.664	0.471	0.758	0.430	0.788
		Testing	0.298	0.750	(0.649, 0.834)	0.692	0.524	0.817	0.458	0.853
		Independent	0.279	0.578	(0.469, 0.681)	0.629	0.783	0.507	0.353	0.872

28	Gender + TRA + TS	Training	0.367	0.745	(0.698, 0.789)	0.690	0.452	0.859	0.553	0.802
		Testing	0.171	0.489	(0.383, 0.596)	0.696	0.952	0.352	0.303	0.962
		Independent	0.265	0.678	(0.571, 0.772)	0.694	0.826	0.627	0.432	0.913
29	Gender + TRA + MSEL	Training	0.248	0.630	(0.579, 0.679)	0.687	0.673	0.613	0.402	0.829
		Testing	0.250	0.674	(0.568, 0.768)	0.704	0.762	0.648	0.390	0.902
		Independent	0.182	0.556	(0.447, 0.660)	0.643	0.913	0.433	0.356	0.935
30	Gender + TRA + TS + MSEL	Training	0.358	0.735	(0.687, 0.779)	0.681	0.413	0.859	0.531	0.791
		Testing	0.215	0.620	(0.512, 0.719)	0.720	0.857	0.549	0.360	0.929
		Independent	0.189	0.556	(0.447, 0.660)	0.700	0.957	0.418	0.361	0.966

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), Threshold = logistic regression decision threshold, TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Support Vector Machines Using Linear Kernels

For SVMs generated using linear kernels built without biological sex (Lin1-Lin15), average AUC in the training, testing, and independent validation datasets was 0.657, 0.640, and 0.589 respectively. Inclusion of biological sex in linear SVM modelling (Lin16-Lin30) resulted in a minor increase of average AUC to 0.668, 0.650, and 0.617 when applied to training, testing, and independent validation datasets respectively. Linear SVM AUC performance on testing and independent validation data were largely consistent. The average absolute difference in AUC values for all regularized classifiers (Lin1-Lin30) between the testing and independent validation set was 0.057 (min difference in AUC = 0.001, max difference = 0.198).

Linear SVM classifiers were all characterized by poor sensitivity and high specificity. For models built without biological sex (Models Lin1-Lin15), average sensitivity was 0.177, 0.127, and 0.104 while average specificity was 0.958, 0.947, and 0.916 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models Lin16-Lin30), average sensitivity was 0.177, 0.127, and 0.104 while average specificity was 0.958, 0.947, and 0.916 when applied to training, testing, and independent validation datasets. All linear SVM classifiers performance results are described in Table 3.10.

Support Vector Machines Using Polynomial Kernels

For SVMs generated using polynomial kernels without biological sex as a predictor variable (Poly1-Poly15), average AUC in the training, testing, and independent validation datasets was 0.651, 0.613, and 0.599 respectively. Inclusion of biological sex in polynomial SVM modelling resulted in a minor increase of average AUC to 0.672, 0.677, and 0.632 when applied to training, testing, and independent validation datasets respectively. Polynomial SVM performance on testing and independent validation data was largely consistent. Average absolute difference in AUC values for all polynomial SVMs (Poly1-Poly30) between the testing and independent validation set was 0.047 (min difference in AUC = 0.006, max difference = 0.168).

Polynomial SVM classifiers were all characterized by poor sensitivity and high specificity. For models built without biological sex (Poly1-Poly15), average sensitivity was 0.174, 0.105, and 0.107 while average specificity was 0.967, 0.954, and 0.926 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models Poly16-

Poly30), average sensitivity was 0.178, 0.114, and 0.113 while average specificity was 0.964, 0.952, and 0.917 when applied to training, testing, and independent validation datasets. All polynomial SVM classifiers performance results are described in Table 3.11.

Support Vector Machines Using Radial Basis Function Kernels

For SVMs generated using radial basis function kernels without biological sex (Rad1-Rad15), average AUC in the training, testing, and independent validation datasets was 0.787, 0.621, and 0.567 respectively. Inclusion of biological sex in RBF-SVM modelling resulted in a minor increase to AUC of 0.789, 0.641, and 0.571 when applied to training, testing, and independent validation datasets respectively. With respect to RBF-SVM performance on testing and independent validation data, the average absolute difference in AUC values for all RBF-SVM classifiers (Rad1-Rad30) between the testing and independent validation set was 0.085 (min difference in AUC = 0.001, max difference = 0.235).

RBF-SVM classifiers were all characterized by poor sensitivity and high specificity. For models built without biological sex (Rad1-Rad15), average sensitivity was 0.219, 0.102, and 0.081 while average specificity was 0.989 0.971, and 0.941 when applied to training, testing, and independent validation datasets. For models built with biological sex (Models Rad16-Rad30), average sensitivity was 0.226, 0.111, and 0.096 while average specificity was 0.987, 0.968, and 0.937 when applied to training, testing, and independent validation datasets. All RBF-SVM classifier performance results are described in Table 3.12.

Table 3.10 | Support Vector Machine with linear kernels model performance

Model #	Variable Combination	Dataset	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
	Predictor Variable Combinations								
1	IL	Training	0.751	(0.704, 0.794)	0.718	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.662	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.591	0.130	0.896	0.300	0.750
2	IL + TS	Training	0.751	(0.704, 0.794)	0.693	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.730	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.532	0.130	0.896	0.300	0.750
3	IL + MSEL	Training	0.751	(0.704, 0.794)	0.718	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.669	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.683	0.130	0.896	0.300	0.750
4	IL + TS + MSEL	Training	0.751	(0.704, 0.794)	0.722	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.670	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.668	0.130	0.896	0.300	0.750
5	TS + MSEL	Training	0.721	(0.673, 0.766)	0.660	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.628	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.596	0.000	1.000	NaN	0.744
6	MSEL	Training	0.721	(0.673, 0.766)	0.579	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.678	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.534	0.000	1.000	NaN	0.744
7	TS	Training	0.721	(0.673, 0.766)	0.666	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.661	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.659	0.000	1.000	NaN	0.744
8	FA	Training	0.743	(0.695, 0.786)	0.600	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.568	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.483	0.130	0.896	0.300	0.750
9	FA + TS	Training	0.743	(0.695, 0.786)	0.647	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.623	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.709	0.130	0.896	0.300	0.750
10	FA + MSEL	Training	0.743	(0.695, 0.786)	0.590	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.526	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.496	0.130	0.896	0.300	0.750
11	FA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.627	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.568	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.555	0.130	0.896	0.300	0.750
12	TRA	Training	0.743	(0.695, 0.786)	0.635	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.624	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.504	0.130	0.896	0.300	0.750
13	TRA + TS	Training	0.743	(0.695, 0.786)	0.664	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.618	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.633	0.130	0.896	0.300	0.750

14	TRA + MSEL	Training	0.743	(0.695, 0.786)	0.659	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.685	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.546	0.130	0.896	0.300	0.750
15	TRA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.670	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.693	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.645	0.130	0.896	0.300	0.750
Predictor Variable Combinations + Gender									
16	Gender + IL	Training	0.751	(0.704, 0.794)	0.731	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.671	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.649	0.130	0.896	0.300	0.750
17	Gender + IL + TS	Training	0.751	(0.704, 0.794)	0.721	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.671	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.662	0.130	0.896	0.300	0.750
18	Gender + IL + MSEL	Training	0.751	(0.704, 0.794)	0.733	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.685	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.676	0.130	0.896	0.300	0.750
19	Gender + IL + TS + MSEL	Training	0.751	(0.704, 0.794)	0.731	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.676	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.676	0.130	0.896	0.300	0.750
20	Gender + TS + MSEL	Training	0.721	(0.673, 0.766)	0.685	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.695	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.643	0.000	1.000	NaN	0.744
21	Gender + MSEL	Training	0.721	(0.673, 0.766)	0.606	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.611	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.494	0.000	1.000	NaN	0.744
22	Gender + TS	Training	0.721	(0.673, 0.766)	0.662	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.652	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.646	0.000	1.000	NaN	0.744
23	Gender + FA	Training	0.743	(0.695, 0.786)	0.664	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.702	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.602	0.130	0.896	0.300	0.750
24	Gender + FA + TS	Training	0.743	(0.695, 0.786)	0.640	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.682	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.627	0.130	0.896	0.300	0.750
25	Gender + FA + MSEL	Training	0.743	(0.695, 0.786)	0.584	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.529	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.482	0.130	0.896	0.300	0.750
26	Gender + FA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.698	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.730	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.650	0.130	0.896	0.300	0.750
27	Gender + TRA	Training	0.743	(0.695, 0.786)	0.616	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.626	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.632	0.130	0.896	0.300	0.750

28	Gender + TRA + TS	Training	0.743	(0.695, 0.786)	0.677	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.674	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.725	0.130	0.896	0.300	0.750
29	Gender + TRA + MSEL	Training	0.743	(0.695, 0.786)	0.589	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.433	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.476	0.130	0.896	0.300	0.750
30	Gender + TRA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.689	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.717	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.618	0.130	0.896	0.300	0.750

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Table 3.11 | Performance of Support Vector Machines with Polynomial Kernel

Model #	Variable Combination	Dataset	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
	Predictor Variable Combinations								
1	IL	Training	0.769	(0.723, 0.811)	0.754	0.221	0.981	0.821	0.765
		Testing	0.761	(0.661, 0.844)	0.613	0.048	0.972	0.333	0.775
		Independent	0.733	(0.630, 0.821)	0.625	0.130	0.940	0.429	0.759
2	IL + TS	Training	0.751	(0.704, 0.794)	0.722	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.643	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.616	0.130	0.896	0.300	0.750
3	IL + MSEL	Training	0.772	(0.726, 0.814)	0.713	0.221	0.985	0.852	0.766
		Testing	0.750	(0.649, 0.834)	0.673	0.048	0.958	0.250	0.773
		Independent	0.733	(0.630, 0.821)	0.682	0.087	0.955	0.400	0.753
4	IL + TS + MSEL	Training	0.769	(0.723, 0.811)	0.713	0.212	0.985	0.846	0.764
		Testing	0.750	(0.649, 0.834)	0.671	0.048	0.958	0.250	0.773
		Independent	0.733	(0.630, 0.821)	0.679	0.087	0.955	0.400	0.753
5	TS + MSEL	Training	0.721	(0.673, 0.766)	0.646	0.000	1.000	0.000	0.721
		Testing	0.772	(0.672, 0.853)	0.612	0.000	1.000	0.000	0.772
		Independent	0.744	(0.642, 0.831)	0.579	0.000	1.000	0.000	0.744
6	MSEL	Training	0.721	(0.673, 0.766)	0.514	0.000	1.000	0.000	0.721
		Testing	0.772	(0.672, 0.853)	0.502	0.000	1.000	0.000	0.772
		Independent	0.744	(0.642, 0.831)	0.511	0.000	1.000	0.000	0.744
7	TS	Training	0.737	(0.689, 0.781)	0.698	0.135	0.970	0.636	0.744
		Testing	0.761	(0.661, 0.844)	0.632	0.095	0.958	0.400	0.782
		Independent	0.767	(0.666, 0.849)	0.638	0.130	0.985	0.750	0.767
8	FA	Training	0.743	(0.695, 0.786)	0.548	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.531	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.520	0.130	0.896	0.300	0.750
9	FA + TS	Training	0.743	(0.695, 0.786)	0.647	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.654	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.637	0.130	0.896	0.300	0.750
10	FA + MSEL	Training	0.743	(0.695, 0.786)	0.589	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.530	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.506	0.130	0.896	0.300	0.750
11	FA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.668	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.665	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.596	0.130	0.896	0.300	0.750
12	TRA	Training	0.743	(0.695, 0.786)	0.594	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.498	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.484	0.130	0.896	0.300	0.750
13	TRA + TS	Training	0.743	(0.695, 0.786)	0.656	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.636	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.692	0.130	0.896	0.300	0.750

14	TRA + MSEL	Training	0.743	(0.695, 0.786)	0.661	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.683	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.621	0.130	0.896	0.300	0.750
15	TRA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.640	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.647	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.600	0.130	0.896	0.300	0.750
Predictor Variable Combinations + Gender									
16	Gender + IL	Training	0.751	(0.704, 0.794)	0.715	0.279	0.933	0.617	0.770
		Testing	0.750	(0.649, 0.834)	0.666	0.190	0.915	0.400	0.793
		Independent	0.700	(0.594, 0.792)	0.672	0.130	0.896	0.300	0.750
17	Gender + IL + TS	Training	0.767	(0.720, 0.809)	0.710	0.279	0.955	0.707	0.774
		Testing	0.750	(0.649, 0.834)	0.674	0.143	0.930	0.375	0.786
		Independent	0.700	(0.594, 0.792)	0.699	0.130	0.896	0.300	0.750
18	Gender + IL + MSEL	Training	0.764	(0.718, 0.806)	0.723	0.231	0.970	0.750	0.765
		Testing	0.761	(0.661, 0.844)	0.688	0.095	0.958	0.400	0.782
		Independent	0.722	(0.618, 0.811)	0.703	0.130	0.925	0.375	0.756
19	Gender + IL + TS + MSEL	Training	0.764	(0.718, 0.806)	0.724	0.231	0.970	0.750	0.765
		Testing	0.761	(0.661, 0.844)	0.698	0.095	0.958	0.400	0.782
		Independent	0.722	(0.618, 0.811)	0.708	0.130	0.925	0.375	0.756
20	Gender + TS + MSEL	Training	0.721	(0.673, 0.766)	0.686	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.756	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.653	0.000	1.000	NaN	0.744
21	Gender + MSEL	Training	0.721	(0.673, 0.766)	0.549	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.525	0.000	1.000	NaN	0.772
		Independent	0.733	(0.630, 0.821)	0.555	0.043	0.970	0.333	0.747
22	Gender + TS	Training	0.740	(0.692, 0.784)	0.665	0.115	0.981	0.706	0.742
		Testing	0.761	(0.661, 0.844)	0.693	0.048	0.972	0.333	0.775
		Independent	0.756	(0.654, 0.840)	0.630	0.087	0.985	0.667	0.759
23	Gender + FA	Training	0.743	(0.695, 0.786)	0.645	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.630	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.547	0.130	0.896	0.300	0.750
24	Gender + FA + TS	Training	0.743	(0.695, 0.786)	0.660	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.641	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.690	0.130	0.896	0.300	0.750
25	Gender + FA + MSEL	Training	0.743	(0.695, 0.786)	0.653	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.682	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.584	0.130	0.896	0.300	0.750
26	Gender + FA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.695	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.765	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.635	0.130	0.896	0.300	0.750
27	Gender + TRA	Training	0.743	(0.695, 0.786)	0.637	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.642	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.498	0.130	0.896	0.300	0.750

28	Gender + TRA + TS	Training	0.743	(0.695, 0.786)	0.673	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.686	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.715	0.130	0.896	0.300	0.750
29	Gender + TRA + MSEL	Training	0.743	(0.695, 0.786)	0.664	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.728	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.559	0.130	0.896	0.300	0.750
30	Gender + TRA + TS + MSEL	Training	0.743	(0.695, 0.786)	0.682	0.192	0.955	0.625	0.754
		Testing	0.761	(0.661, 0.844)	0.685	0.143	0.944	0.429	0.788
		Independent	0.700	(0.594, 0.792)	0.635	0.130	0.896	0.300	0.750

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Table 3.12 | Performance of Support Vector Machines with Radial Basis Function Kernel

Model #	Variable Combination	Dataset	Accuracy	95% CI	AUC	Sens	Spec	PPV	NPV
	Predictor Variable Combinations								
1	IL	Training	0.831	(0.789, 0.868)	0.887	0.423	0.989	0.936	0.816
		Testing	0.783	(0.684, 0.862)	0.633	0.238	0.944	0.556	0.807
		Independent	0.733	(0.630, 0.821)	0.632	0.130	0.940	0.429	0.759
2	IL + TS	Training	0.831	(0.789, 0.868)	0.887	0.413	0.993	0.956	0.814
		Testing	0.772	(0.672, 0.853)	0.632	0.190	0.944	0.500	0.798
		Independent	0.733	(0.630, 0.821)	0.616	0.130	0.940	0.429	0.759
3	IL + MSEL	Training	0.818	(0.775, 0.856)	0.887	0.375	0.989	0.929	0.804
		Testing	0.750	(0.649, 0.834)	0.630	0.000	0.972	0.000	0.767
		Independent	0.744	(0.642, 0.831)	0.660	0.174	0.940	0.500	0.768
4	IL + TS + MSEL	Training	0.818	(0.775, 0.856)	0.880	0.375	0.989	0.929	0.804
		Testing	0.750	(0.649, 0.834)	0.632	0.048	0.958	0.250	0.773
		Independent	0.744	(0.642, 0.831)	0.658	0.174	0.940	0.500	0.768
5	TS + MSEL	Training	0.745	(0.698, 0.789)	0.811	0.115	0.989	0.800	0.743
		Testing	0.783	(0.684, 0.862)	0.608	0.095	0.986	0.667	0.787
		Independent	0.744	(0.642, 0.831)	0.596	0.000	1.000	NaN	0.744
6	MSEL	Training	0.721	(0.673, 0.766)	0.825	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.693	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.564	0.000	1.000	NaN	0.744
7	TS	Training	0.737	(0.689, 0.781)	0.599	0.163	0.959	0.607	0.748
		Testing	0.783	(0.684, 0.862)	0.542	0.143	0.972	0.600	0.793
		Independent	0.722	(0.618, 0.811)	0.601	0.087	0.940	0.333	0.750
8	FA	Training	0.756	(0.709, 0.790)	0.655	0.154	0.989	0.842	0.751
		Testing	0.761	(0.661, 0.844)	0.618	0.048	0.972	0.333	0.775
		Independent	0.711	(0.606, 0.802)	0.537	0.087	0.925	0.286	0.747
9	FA + TS	Training	0.791	(0.746, 0.831)	0.730	0.250	1.000	1.000	0.775
		Testing	0.761	(0.661, 0.844)	0.557	0.048	0.972	0.333	0.775
		Independent	0.656	(0.548, 0.753)	0.559	0.000	0.881	0.000	0.720
10	FA + MSEL	Training	0.761	(0.715, 0.804)	0.862	0.144	1.000	1.000	0.751
		Testing	0.772	(0.672, 0.853)	0.618	0.048	0.986	0.500	0.778
		Independent	0.722	(0.618, 0.811)	0.551	0.000	0.970	0.000	0.739
11	FA + TS + MSEL	Training	0.756	(0.709, 0.799)	0.848	0.135	0.996	0.933	0.749
		Testing	0.793	(0.696, 0.871)	0.596	0.143	0.986	0.750	0.795
		Independent	0.711	(0.606, 0.802)	0.457	0.000	0.955	0.000	0.736
12	TRA	Training	0.745	(0.698, 0.789)	0.630	0.144	0.978	0.714	0.747
		Testing	0.772	(0.672, 0.853)	0.660	0.143	0.958	0.500	0.791
		Independent	0.722	(0.618, 0.811)	0.520	0.130	0.925	0.375	0.756
13	TRA + TS	Training	0.777	(0.732, 0.819)	0.675	0.231	0.989	0.889	0.769
		Testing	0.783	(0.684, 0.862)	0.543	0.143	0.972	0.600	0.793
		Independent	0.700	(0.594, 0.792)	0.556	0.130	0.896	0.300	0.750

14	TRA + MSEL	Training	0.756	(0.709, 0.799)	0.815	0.163	0.985	0.810	0.753
		Testing	0.772	(0.672, 0.853)	0.709	0.095	0.972	0.500	0.784
		Independent	0.722	(0.618, 0.811)	0.507	0.087	0.940	0.333	0.750
15	TRA + TS + MSEL	Training	0.764	(0.718, 0.806)	0.806	0.192	0.985	0.833	0.759
		Testing	0.783	(0.684, 0.862)	0.638	0.143	0.972	0.600	0.793
		Independent	0.711	(0.606, 0.802)	0.491	0.087	0.925	0.286	0.747
Predictor Variable Combinations + Gender									
16	Gender + IL	Training	0.831	(0.789, 0.868)	0.886	0.423	0.989	0.936	0.816
		Testing	0.772	(0.672, 0.853)	0.655	0.190	0.944	0.500	0.798
		Independent	0.722	(0.618, 0.811)	0.678	0.174	0.910	0.400	0.763
17	Gender + IL + TS	Training	0.831	(0.789, 0.868)	0.886	0.423	0.989	0.936	0.816
		Testing	0.772	(0.672, 0.853)	0.651	0.190	0.944	0.500	0.798
		Independent	0.733	(0.63, 0.821)	0.673	0.174	0.925	0.444	0.765
18	Gender + IL + MSEL	Training	0.818	(0.775, 0.856)	0.890	0.375	0.989	0.929	0.804
		Testing	0.761	(0.661, 0.844)	0.655	0.095	0.958	0.400	0.782
		Independent	0.733	(0.63, 0.821)	0.685	0.174	0.925	0.444	0.765
19	Gender + IL + TS + MSEL	Training	0.818	(0.775, 0.856)	0.881	0.375	0.989	0.929	0.804
		Testing	0.761	(0.661, 0.844)	0.647	0.095	0.958	0.400	0.782
		Independent	0.733	(0.63, 0.821)	0.687	0.174	0.925	0.444	0.765
20	Gender + TS + MSEL	Training	0.751	(0.704, 0.794)	0.801	0.154	0.981	0.762	0.750
		Testing	0.772	(0.672, 0.853)	0.631	0.095	0.972	0.500	0.784
		Independent	0.733	(0.63, 0.821)	0.594	0.043	0.970	0.333	0.747
21	Gender + MSEL	Training	0.721	(0.673, 0.766)	0.838	0.000	1.000	NaN	0.721
		Testing	0.772	(0.672, 0.853)	0.737	0.000	1.000	NaN	0.772
		Independent	0.744	(0.642, 0.831)	0.555	0.000	1.000	NaN	0.744
22	Gender + TS	Training	0.753	(0.706, 0.796)	0.641	0.221	0.959	0.676	0.761
		Testing	0.793	(0.696, 0.871)	0.682	0.190	0.972	0.667	0.802
		Independent	0.733	(0.63, 0.821)	0.447	0.130	0.940	0.429	0.759
23	Gender + FA	Training	0.753	(0.706, 0.796)	0.647	0.144	0.989	0.833	0.749
		Testing	0.761	(0.661, 0.844)	0.633	0.048	0.972	0.333	0.775
		Independent	0.711	(0.606, 0.802)	0.541	0.087	0.925	0.286	0.747
24	Gender + FA + TS	Training	0.820	(0.778, 0.858)	0.766	0.385	0.989	0.930	0.806
		Testing	0.761	(0.661, 0.844)	0.650	0.095	0.958	0.400	0.782
		Independent	0.700	(0.594, 0.792)	0.503	0.087	0.910	0.250	0.744
25	Gender + FA + MSEL	Training	0.756	(0.709, 0.799)	0.840	0.144	0.993	0.882	0.750
		Testing	0.761	(0.661, 0.844)	0.651	0.048	0.972	0.333	0.775
		Independent	0.700	(0.594, 0.792)	0.561	0.000	0.940	0.000	0.733
26	Gender + FA + TS + MSEL	Training	0.764	(0.718, 0.806)	0.830	0.163	0.996	0.944	0.755
		Testing	0.783	(0.684, 0.862)	0.601	0.143	0.972	0.600	0.793
		Independent	0.711	(0.606, 0.802)	0.534	0.043	0.940	0.200	0.741
27	Gender + TRA	Training	0.748	(0.701, 0.791)	0.597	0.144	0.981	0.750	0.748
		Testing	0.772	(0.672, 0.853)	0.519	0.143	0.958	0.500	0.791
		Independent	0.722	(0.618, 0.811)	0.619	0.130	0.925	0.375	0.756

28	Gender + TRA + TS	Training	0.753	(0.706, 0.796)	0.688	0.135	0.993	0.875	0.748
		Testing	0.783	(0.684, 0.862)	0.620	0.048	1.000	1.000	0.780
		Independent	0.722	(0.618, 0.811)	0.508	0.043	0.955	0.250	0.744
29	Gender + TRA + MSEL	Training	0.751	(0.704, 0.794)	0.824	0.135	0.989	0.824	0.747
		Testing	0.783	(0.684, 0.862)	0.668	0.143	0.972	0.600	0.793
		Independent	0.722	(0.618, 0.811)	0.481	0.087	0.940	0.333	0.750
30	Gender + TRA + TS + MSEL	Training	0.759	(0.712, 0.801)	0.826	0.173	0.985	0.818	0.755
		Testing	0.783	(0.684, 0.862)	0.617	0.143	0.972	0.600	0.793
		Independent	0.711	(0.606, 0.802)	0.497	0.087	0.925	0.286	0.747

AUC = Area Under the Curve, FA = AOSI items identified by factor analysis (6, 8, 14, 16, 18), IL = AOSI items 1-18, Independent = new ASIB independent validation dataset (contains n=90 participants), MSEL = Mullen Scales of Early Learning subscales (Early Learning Composite, Visual Reception, Fine Motor, Receptive Language, Expressive Language), NPV = Negative Predictive Value, PPV = Positive Predictive Value, Sens = Sensitivity, Spec = Specificity, Testing = old ASIB testing dataset (contains n=92 participants; 20% of old ASIB data), TRA = Factor analysis AOSI items that survived Benjamini & Hochberg 1995 corrected multiple comparisons when compared using IL-ASD and IL-N groups, Training = old ASIB training dataset (contains n=373 participants; 80% of old ASIB data), TS = AOSI Total Score

Discussion

This study focused on the generation and assessment of supervised logistic regression and support vector machine classifiers built using 12-month IL-sibling data and their ability at correctly predicting 36-month ASD diagnosis. Supervised learning classifiers were successfully generated using different combinations of 12-month AOSI and MSEL data, with their performance assessed across training, testing, and independent validation IL-sibling datasets. Overall, this study has three main findings: (1) AUC for all models ranged from low (0.500-0.700) to moderate (0.700-0.766), with several models (using combinations of AOSI and MSEL predictor variables) exceeding the benchmark performance of the 12-month AOSI Total Score with a cut off of 7 (AUC = 0.665) alone, (2) all classifiers (barring regression models optimized for maximal combined sensitivity and specificity) were characterized by low sensitivity and extremely high specificity, and (3) classifier performance across testing and independent validation sets was largely comparable for all models.

Though modest, the best-performing classifiers had an increase in AUC of between 0.070 and 0.100 when built using combinations of AOSI item-level, Total Score, and MSEL standard score data relative to the benchmark 12-month AOSI Total Score cut point of 7. While no single classifier had standout performance, the consistently moderate predictive performance of models built using 12-month data at predicting ASD diagnosis at 36-months – especially in relation to the predictive ability of the 12-month AOSI Total Score alone – suggests that additional refinement or selection of variables used during predictive modelling may, in future, yield promising results. Further investigation into what features on the AOSI, MSEL, and other developmental or ASD-specific measures during early childhood development are the most important at predicting later ASD diagnostic status needs to be elucidated. Such work has the potential to not just contribute to our understanding and characterization of the emergence of ASD during early childhood, but also to build or refine better early ASD screening or detection tools for children at or around their first year of life.

Though most classifiers were characterized by low accuracy (AUC = 0.500-0.700; Akobeng 2007), the best performing classifiers were characterized by moderate accuracy with AUC between 0.700 and 0.800 on the testing set. The classifier with the highest performance in the

testing dataset using logistic regression, regularized logistic regression, and SVMs with linear, polynomial, and radial kernels were models L26, R20, Lin2/26, Poly26, and Rad21. These classifiers had testing set AUCs between 0.730 (Lin2/26) and 0.765 (Poly26). Although the debate continues as to whether the 4:1 male to female ASD sex ratio accurately represents the real distribution of ASD in males and females or is instead potentially reflective of aspects like social camouflaging in females (Dean et al., 2016, Tubío-Fungueiriño et al., 2021) under-detection due to clinician bias (Zwaigenbaum et al., 2012), or represent genuine differences in clinical presentation (Sacrey et al., 2017). Whichever the reason, in this study inclusion of biological sex as a predictor variable across statistical classifier types resulted a modest increase in classifier AUC when applied across training, testing, and independent validation datasets for all learning algorithms that were tested.

All models in this study (barring optimized regression models) were characterized by poor sensitivity and extremely high specificity. Although the best clinical or diagnostic tests are both highly sensitive and specific, in practice this is not always feasible (Akobeng 2007). Trade-offs exist between valuing high sensitivity over specificity (or vice versa; Trevethan 2017). Tests with high sensitivity are useful in 'ruling out' participants if they test negative, while high specificity is useful for 'ruling in' participants if they test positive (Akobeng 2007). Although none of the algorithms reported are both highly sensitive and specific, they still have potential utility. In the case of ASD diagnostics, a test that is only highly specific (i.e., correctly identifies individuals who *do not* have ASD) still can provide valuable information to families who already have a child diagnosed with ASD. From a caregiving perspective, neurodevelopmental disorders like ASD can exert tremendous social, economic, and health burdens on families – though many caregivers simply look at it as trying to give the best possible quality of life to their loved one (Dudley & Emery, 2014). An ASD diagnosis carries with it a lifetime of direct and indirect costs related to medical and healthcare expenses, therapeutics, (special) education, productivity loss for family or caregivers, accommodations, respite care, and out-of-pocket expenses (Rogge & Janssen, 2019). Based on cost estimate studies of expenditures or productivity loss, it is estimated that the lifetime cost of an ASD diagnosis was approximately US $3.6 million dollars in 2019 (Cakir et al., 2020). For families having their first child (or for infant sibling families who already have a child diagnosed with ASD), reassurance in the form of a negative result on a

highly specific screening tool that their child likely does not have ASD based on the available information may be profoundly reassuring. As such, tests characterized by poor sensitivity and high specificity still may have the potential to provide meaningful information to families – though perhaps not in a clinical or diagnostic capacity.

Across all classifiers generated in this study, model performance in the testing set was, by-and-large, corroborated in the independent validation set. However, classifier performance on the independent validation data was somewhat reduced for several models relative to the testing dataset across all algorithms tested. This result may stem from differences in ASD diagnostic determination between IL-sibling participants in the testing and independent validation datasets. In this study, the training and testing sets were generated from CISS-1 IL-sibling participants that were assessed for ASD at 36-months using DSM-IV-TR criteria. In contrast, while using a nearly identical study protocol, CISS-2 assesses participants for ASD at 36-months using *DSM-5* criteria (Sacrey et al., 2021). Since its release, several studies have since been published indicating that the DSM-5 is less sensitive to individuals who previously meet criteria on the DSM-IV or DSM-IV-TR – especially individuals with pervasive developmental disorder not otherwise specified (PDD-NOS) or Asperger's syndrome (Bennett & Goodall 2016, Yaylaci & Miral, 2016, Mazurek et al., 2017). It is entirely plausible the change in DSM diagnostic criteria for ASD between CISS-1 and CISS-2 participants (and the corresponding change in sensitivity of the DSM-5 to individuals previously classified as PDD-NOS and Asperger's syndrome – though Asperger's is rarely diagnosed at 3 years old) may account in part for the reduction in classifier performance from the testing to independent validation sets.

The best performing supervised learning classifiers generated in this study all had higher performance than the predictive AUC benchmark and corresponding 99% CI set by the 12-month AOSI Total Score using a cut point of 7 alone. Interestingly, the highest performing classifiers in this study (Model L26, R20, Lin2/Lin26 [tied AUC], Poly26, Rad21) all incorporated additional predictor variables in addition to (or in lieu of; see Rad21) the AOSI Total Score during model generation. In addition, outside of linear SVM model Lin2 that was generated using item-level and AOSI Total Score predictor variables, the best performing models were generated using biological sex, MSEL standard score data, and/or factor-analysis-identified

AOSI items. These results suggest that when aiming to develop early predictive models or screening tools for ASD in individuals at or around 12-months old, the AOSI Total Score alone is likely not sufficient in fully capturing all features and dimensions that are characteristic of early ASD emergence. This is further corroborated by performance of logistic regression models that underwent variable importance analysis – for all six of the logistic regression models assessed (L10, L11, L14, L25, L26, L30) removal of AOSI and MSEL variables from the model resulted in increased model performance (measured by increasing AUC and decreasing AIC). Whether this tendency for removing AOSI and/or MSEL variables from consideration holds true for SVM modelling is unclear. As such, future research in early detection would likely benefit from not just investigating variable or feature importance on the AOSI, MSEL at 12-months, but other ASD or developmental tool while simultaneously assessing the impact dimensionality reduction techniques like principal component analysis have on supervised classifiers performance.

Since most classification algorithms exhibit degraded performance when built using irrelevant or redundant features (Morán-Fernández et al., 2022), assessing variable or feature importance as a means of reducing data dimensionality (via techniques like feature extraction or selection) can be critical to development of high performance early ASD predictive classifiers. In this study, this was especially prominent during variable importance assessment of logistic regression predictor variables. It is, however, important to note that these results are all drawn from IL-sibling participants; no LL control participants were included in this study. Future studies should consider whether the importance of features under consideration are truly specific to IL group status (and/or to later ASD diagnostic status) or are instead aspects of early neurotypical childhood development that is being erroneously associated with ASD status when predictive modelling is conducted only in IL samples. Inclusion of additional variables in predictive modelling that measure different constructs of early ASD emergence (or refinement of existing AOSI and MSEL predictor variables) are necessary for future early ASD screening or detection efforts. For instance, 12-month-old infants later diagnosed with ASD are more likely to exhibit more symptoms (e.g. abnormal social communication, unusual eye contact, failure to orient to names, lack of gestures, language delays, visual abnormalities, etc.) relative to infants who do not meet diagnostic criteria (Gieserman & Carter, 2017). Further exploration of these differences

(between IL-ASD and IL-N, as well as between IL and LL populations) and which are the most predictive is necessary. It should be noted that context is likewise important; the results presented here are all anchored in a Canadian IL-sibling context. Given that previous research have shown that differences in IL ASD populations such as infants with Fragile X Syndrome have significantly higher atypical motor impairments relative to infant siblings on the AOSI (Roberts et al., 2016), future efforts in develop ASD screening tools may need to consider IL context and the differences in early symptom profiles that may exist between them. However, care should also be taken to ensure that the data being used to generate and refine predictive machine learning models does not inadvertently run the risk of exacerbating existing disparities in the care or diagnosis of ASD between rural and urban regions (Haritos 2017). By considering data in predictive modeling which requires in-person assessment by trained staff or personnel to obtain (such as the AOSI, MSEL, VABS, ADOS etc.) risks potentially disenfranchising people to whom accessible predictive ASD screening models may be best poised to support. The inability to obtain the data required by predictive models to function (by virtue of a lack of trained personnel or the inability to go to urban centers where such assessments are more likely to take place) is something that must be considered when choosing what variables or features predictive models are going to be built on. Consideration of widely available parent-report or questionnaire data should therefore be explored for use as alternate or replacement features in future machine learning studies.

Though more work needs to be done to generate high performance predictive classifiers for infants <18 months old, the potential impact of such technology is tremendous. There is a broad consensus that early interventions for ASD are associated with improved functional outcomes (Gardner et al., 2013, (Bonnis et al., 2016, Landa 2017, McDonald et al., 2018, Noyes-Gosser, 2018, Fuller and Kaiser, 2019). If high performance predictive classifiers can be developed that reliably screen for ASD status or risk of later diagnosis, more resources could be directed to developmental surveillance of screen-positive individuals. This has potentially compounding benefits; increased access to early developmental monitoring and, in the case of early ASD diagnosis, early intervention resources would benefit not just individuals with ASD, but their families as well. It is important to note that while a diagnosis of ASD can open doors to early intervention and support services, access to broader supports to ASD individuals and their

families in Canada vary province-to-province with access to supports or funding often restricted based on intelligence quotient eligibility or intellectual disability (Dudley et al., 2014). The importance of supports is also reflected in current DSM diagnostic criteria. The DSM-5 assigns severity levels for ASD predicated on the anticipated level of supports needed based upon the symptom domains of social communication and restricted and repetitive behaviour (APA, 2013, Mehling et al., 2016, Eman et al., 2020). Severity levels range from Level 1 (individuals require some support) to Level 3 (individuals require substantial support; APA, 2013, Eman et al., 2020). This focus on supports is especially salient given the considerable social and economic burden of ASD (Dudley et al., 2014, Lavelle et al., 2014, Cakir et al., 2020). In Canada, there is currently unequal or incomplete access to supports for individuals with ASD and/or their families/caregivers; all suffer from varying levels of inadequacy (Dudley et al., 2014). In addition, there is an unsustainable overreliance on family or caregivers to provide supports as community services are fragmented or unavailable (Dudley et al., 2014). With such a tremendous social and economic burden associated with caregiving for individuals on the spectrum (Dudley et al., 2014), change is imperative.

Given that the annual costs that families spend to support individuals with ASD are enormous (often well beyond what the annual income for families), provision of supportive care necessitates access to external resources and supports (Dudley et al., 2014). Enhanced early ASD screening and detection, while not *the* answer to all the issues faced by individuals and families with ASD, can certainly help. Though multiple studies have shown early intervention efforts *are* efficacious, research is just starting to explore the relationship between the age of intervention and developmental outcomes for children with ASD (Towle et al., 2020). Even if 'the earlier, the better' is not as clear cut as currently viewed with respect to early ASD intervention modalities, from a family or caregiver perspective the earlier an individual can be identified with ASD, the earlier their family or caregivers can learn where to go and what to do when navigating the healthcare and human services systems to access services that can best support their child (Towle et al., 2020). This is an important consideration as families or caregivers of children with ASD often experience challenges when navigating multiple systems in search of or access to support services (Crossman et al., 2020). Just as early identification and interventions are important for individuals with ASD, supporting parent or caregiver education is also vital.

Though many studies have started investigating the utility of developing predictive classifiers using various types of machine learning algorithms, the ethical considerations of this technology are rarely discussed – in recent systematic and non-systematic reviews of machine learning in ASD research, ethics or ethical implications of the technology were never discussed (Pagnozzi et al., 2018, Hyde et al., 2019, Moon et al., 2019, Kollias et al., 2021, Siddiqui et al., 2021). There is an increasing trend of leveraging machine learning for use in high-stakes predictive applications in healthcare and the criminal justice system, as well as the tendency for many of these algorithms to be "black boxes" (i.e., models that are too complicated for humans to fully comprehend, or are proprietary and not transparent; Rudin 2019). Part of the problem with black box models is that if they are flawed, there is no way of investigating or detecting it (due to the lack of transparency); it becomes difficult or impossible to troubleshoot or correct (Rudin 2019). In healthcare, there is a tendency towards blind acceptance of black box models (Rudin 2019). While this can open the door for companies to develop and sell models for use in clinical applications, adopting automated systems require trust on the database they were built, processed, and refined on (Rudin 2019). In computer sciences, a common saying is 'garbage in, garbage out.' It is well known that in the field of machine learning, biased training data can lead to biased models which can have unintended or deleterious consequences and outcomes that threaten model validity (Sanders & Axe, 2017, DeBrusk, 2018, Aleyani, 2021, Geiger et al., 2021). Biases in training data (especially with respect to human-labelled data) can reflect historical cognitive biases in thinking and behaviour that manifest in collected data (Aleyani, 2021). For example, consider a machine learning algorithm designed to assess how hireable a new applicant is for a job in the financial sector that has been trained on a company's historical hiring history. If the algorithm disproportionately suggests hiring men over women (educational experience, past credentials, and everything else all being equal) it could be a manifestation of the historical gender bias against women in financial positions (Esser & Swalve, 2022); these factors are reflected as biased training data from the company upon which the algorithm was trained. This issue can further be exacerbated if the learning algorithm in question was a black box and lacked transparency. There is a fundamental conflict of responsibility when using black box models in high stakes decision making; companies can profit from developing and selling these models (especially to hospitals in a healthcare setting) while simultaneously not being

responsible for the quality and consequences of the individual predictions being made (Rudin 2019). This issue becomes very important in healthcare settings where the consequences of ill-performing or biased predictive algorithms can be especially deleterious on patients and their families/caregivers.

Though early screening technology for ASD is not yet performing at a level acceptable for clinical or healthcare implementation, use of artificial intelligence or machine learning-based tools should be relegated to only that: screening or early detection. Such technology should *never* be used in place of in-person diagnostic assessments. If machine learning-based technology expands in scope in clinical or healthcare settings, careful consideration is merited regarding the end goal of implementing the technology. In the advent of high-performance early screening tools, steps should be taken to ensure that access to supports, resources, or follow-up assessments is not solely contingent on screening results. If a decision is made by an algorithm that is not transparent or uninterpretable (i.e., a black box) and an individual does not screen positive, they or their families may have no idea by what metric the decision was based. Given the propensity for cognitive biases in thinking and behaviour to manifest in data we use to train machine learning algorithms (Aleyani, 2021), such a scenario is rife with potential to inadvertently exacerbate and/or perpetuate inequities in society. This is unacceptable at the best of times; it is doubly so given the population, individuals with ASD and their families or caregivers who are a traditionally marginalized population that often experience discrimination in various areas of life (Como et al., 2019, Saldago, 2020, Niles & Monaco, 2019, Cascio et al. 2020, Botha et al., 2022). Given the significant burden on caregivers (Cadman et al., 2012, Dudley et al., 2014, Lavelle et al., 2014, Marsack-Topolewski & Church, 2019, Ortiz-Rubio et al., 2021), when access to follow up or resources are contingent upon an algorithmic decision, people have the right to know the metrics upon which they are being evaluated. If access to supports or resources is contingent upon an algorithmic decision, transparency is critical as failure to access them can have a negative impact on the quality of life for individuals with ASD and/or their families and caregivers.

The vehicle to promote transparency may be an ethical guideline or framework for artificial intelligence and machine learning. Multiple organizations and institutions have developed

guidelines or frameworks including (but not limited to) *Asilomar AI Principles* (Future of Life Institute, 2017), *The Montreal Declaration* (Dilhac 2018), *IEEE's Ethically Aligned Design* (IEEE.org, 2017), *Recommendations of the Council on Artificial Intelligence* (Yeung, 2019). However, no single unified standard has been codified and widely adopted. In a field where a governing ethical framework is still being established and formalized, moving forwards without due consideration to the implications of artificial intelligence or machine learning technology is rife with the potential to potentially perpetuate or exacerbate pre-existing inequalities in society.

There are several perceived strengths of this study. First, all classifiers were built, trained, and assessed on the same training, testing, and independent validation datasets. Accordingly, variance in model performance results stem from the different machine learning techniques employed and not the data source. Second, findings are similar to Bussu et al., 2018 who investigated the utility of using AOSI data to generate predictive ASD classifiers. Bussu et al., 2018 generated classifiers for participants and their best-performing models at 8- and 14-months old had AUC of 0.69 and 0.71 respectively. The best performing results in this study, while slightly higher in performance and generated at 12-months, are similar in magnitude; AUC for the best performing models range between 0.73 and 0.76. Third, this study's comprehensive report of model performance across *all* datasets (training, testing, independent validation datasets) is a strong mechanism where model overfitting can be identified and addressed. In this study, the majority of SVMs generated using radial kernels had much stronger performance (as measured by AUC) on the training set data relative to the testing or independent validation datasets pointing towards issues with overfitting. Though the root causes of model overfitting (i.e., poor model generalization to new, unseen data) are often complicated (Ying 2019), if model performance in training data was not reported, overfitting in radial SVM classifiers in this study might have been missed. Finally, this study greatly benefitted from model performance being assessed in two datasets: a testing dataset sourced from CISS-1, and a new, independent validation dataset sourced from CISS-2 – both of which had nearly identical number of participants (n=92 and n=90 respectively). This access to new IL-sibling participant data collected using a nearly identical study protocol (barring different 36-month ASD outcome criteria) allows for strong assessment of model generalizability in new IL-sibling contexts.

This study also has some notable limitations. First, both CISS-1 and CISS-2 participants predominantly are Caucasian with middle-to-upper class SES. These data are not representative of the IL-sibling community at large. Historically there has been an underrepresentation of minorities and overrepresentation of Caucasian participants in health research (Redwood & Gill, 2013) and especially in ASD research (West et al., 2016, Robertson et al., 2017). In the future, researchers should strive towards having socially, economically, and ethnically diverse participant pools to increase generalizability and applicability of research findings. Generalization of this study's findings to external contexts should be done with caution. Second, while study results are similar to Bussu et al., 2018, results are not directly comparable. The classifiers used by Bussu et al., 2018 were generated via *k*-fold cross-validation ten times using least squares SVM from the toolbox *LS-LSVMlab* in MATLAB 9.1. Though not mentioned directly what kernel functions (linear, polynomial, radial) were employed, contact with the paper authors confirmed use of radial basis function kernel (G. Bussu, personal communication, November 3rd, 2022). In this study, least squares SVM were not generated. Not only was a different coding platform used to generate the models (MATLAB vs R), but limitations in the R package *caret* used to generate the models in this study precluded them from being assessed. Though least squares SVMs *can* be generated in R using the *caret* and *kernlab* packages, class probabilities (i.e., the probability that predictions made for each participant belong to one class [IL-ASD] or the other [IL-N]) are not implemented for them and thusly rendering ROC and AUC assessment moot. Third, overfitting (indicated by disproportionate performance in the training set vs testing and independent validation sets) appeared to be an issue for several of the predictive classifiers that were generated – especially for radial SVMs with and without biological sex as a predictor. Further work in the future should be conducted in refining model parameters to correct for this issue. Fourth, ASD diagnostic determination was not the same for all participants; CISS-1 (used to generate training and testing datasets) employed DSM-IV-TR criteria during 36-month ASD diagnostic assessments while CISS-2 uses DSM-5 criteria. This difference in ASD classification may be the main driver behind some of the disparate classifier performance results in testing vs independent validation data. Finally, the data presented in this paper is from IL-siblings. While IL-siblings are a great population to study as a means of investigating early emergence and manifestation of ASD due to the increased prevalence rate in these populations (upwards of 20%; Ozonoff et al., 2011, Szatmari et al., 2016), results drawn

from these populations are not wholly representative of ASD in general. Other IL populations such as infants with FXS are at IL of being diagnosed with ASD can present differently on early ASD screening tools such as the AOSI. For instance, infants with FXS who are diagnosed with ASD are characterized by significantly higher motor impairments relative to infant siblings diagnosed with ASD on the AOSI (Roberts et al., 2016).

Conclusion

In this study, predictive classifiers using various supervised learning algorithms were successfully generated in R using 12-month AOSI and MSEL data. Though the best performing models were characterized by moderate AUC values, poor sensitivity, and extremely high specificity, they all had enhanced predictive performance relative to use of the 12-month AOSI Total Score alone. Overall, the best performing models in this study had performance below the recommended levels for screening (Zwaigenbaum et al., 2015). Further exploration into feature extraction or refinement of 12-month clinical data characteristic of ASD emergence in infant siblings 12-months old is necessary. Though there is an increasing trend towards use of machine learning in predictive healthcare applications (Rudin 2019), early machine learning-based ASD detection algorithms in IL-sibling populations are not yet performing at a level acceptable level for clinical or healthcare implementation. Moreover, there is unfortunately a severe dearth in ASD machine learning literature pertaining to the impact of machine learning in clinical or healthcare settings. In future, as early screening technology progresses, considerations about the implications and use of this technology need to be addressed to prevent exacerbation or perpetuation of inequalities in a population that has traditionally been marginalized or the object of discrimination (Como et al., 2019, Saldago, 2020, Niles & Monaco, 2019, Cascio et al. 2020, Botha et al., 2022).

References

Abbeduto, L., McDuffie, A., & Thurman, A. J. (2014). The fragile X syndrome–autism comorbidity: what do we really know?. *Frontiers in genetics*, *5*, 355.

Agrawal, S., Rao, S. C., Bulsara, M. K., & Patole, S. K. (2018). Prevalence of autism spectrum disorder in preterm infants: a meta-analysis. *Pediatrics*, *142*(3).

AI Principles (2017). futureoflife.org. https://futureoflife.org/open-letter/ai-principles/

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica*, *96*(3), 338-341.

Alelyani, S. (2021). Detection and Evaluation of Machine Learning Bias. *Applied Sciences*, *11*(14), 6271.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th edition). https://doi.org/10.1176/appi.books.9780890425596

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44-58.

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., ... & Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries*, *67*(6), 1.

Bedford, R., Gliga, T., Hendry, A., Jones, E. J., Pasco, G., Charman, T., ... & Pickles, A. (2019). Infant regulatory function acts as a protective factor for later traits of autism spectrum disorder and attention deficit/hyperactivity disorder but not callous unemotional traits. *Journal of Neurodevelopmental Disorders*, *11*(1), 1-11.

Bedford, R., Gliga, T., Shephard, E., Elsabbagh, M., Pickles, A., Charman, T., & Johnson, M. H. (2017). Neurocognitive and observational markers: prediction of autism spectrum disorder from infancy to mid-childhood. *Molecular autism*, *8*(1), 1-10.

Bedford, R., Jones, E. J., Johnson, M. H., Pickles, A., Charman, T., & Gliga, T. (2016). Sex differences in the association between infant markers and later autistic traits. *Molecular autism*, *7*(1), 1-11.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Bennett, M., & Goodall, E. (2016). A meta-analysis of DSM-5 autism diagnoses in relation to DSM-IV and DSM-IV-TR. *Review Journal of Autism and Developmental Disorders*, *3*(2), 119-124.

Ben-Sasson, A., & Carter, A. S. (2012). The application of the first year inventory for ASD screening in Israel. *Journal of autism and developmental disorders*, *42*(9), 1906-1916.

Berrar, D. (2018) *Cross-Validation*. Data Science Laboratory, Tokyo Institute of Technology.

Bhavsar, H., & Panchal, M. H. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *1*(10), 185-189.

Bini, S. A. (2018). Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?. *The Journal of arthroplasty*, *33*(8), 2358-2361.

Bonis, S. (2016). Stress and parents of children with autism: A review of literature. *Issues in mental health nursing*, *37*(3), 153-163.

Botha, M., Dibb, B., & Frost, D. M. (2022). " Autism is me": an investigation of how autistic individuals make sense of autism and stigma. *Disability & Society*, *37*(3), 427-453.
Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual review of psychology*, *53*(1), 371-399.
Brian, J., Bryson, S. E., Smith, I. M., Roberts, W., Roncadin, C., Szatmari, P., & Zwaigenbaum, L. (2016). Stability and change in autism spectrum disorder diagnosis from age 3 to middle childhood in a high-risk sibling cohort. *Autism*, *20*(7), 888-892.
Brian, A. J., Roncadin, C., Duku, E., Bryson, S. E., Smith, I. M., Roberts, W., ... & Zwaigenbaum, L. (2014). Emerging cognitive profiles in high-risk infants with and without autism spectrum disorder. *Research in Autism Spectrum Disorders*, *8*(11), 1557-1566.
British Thoracic Society. (2016). Scottish intercollegiate guidelines network. *British guideline on the management of asthma*, 58.
Bryson, S. E., Zwaigenbaum, L., McDermott, C., Rombough, V., & Brian, J. (2008). The Autism Observation Scale for Infants: scale development and reliability data. *Journal of autism and developmental disorders*, *38*(4), 731-738.
Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, *1*(1), 255-278.
Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.
Bussu, G., Jones, E. J., Charman, T., Johnson, M. H., & Buitelaar, J. K. (2018). Prediction of autism at 3 years from behavioural and developmental measures in high-risk infants: a longitudinal cross-domain classifier analysis. *Journal of Autism and Developmental Disorders*, *48*(7), 2418-2433.
Cadman, T., Eklund, H., Howley, D., Hayward, H., Clarke, H., Findon, J., ... & Glaser, K. (2012). Caregiver burden as people with autism spectrum disorder and attention-deficit/hyperactivity disorder transition into adolescence and adulthood in the United Kingdom. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(9), 879-888.
Cakir, J., Frye, R. E., & Walker, S. J. (2020). The lifetime social cost of autism: 1990–2029. *Research in Autism Spectrum Disorders*, *72*, 101502.
Capal, J. K., Horn, P. S., Murray, D. S., Byars, A. W., Bing, N. M., Kent, B., ... & TACERN Study Group. (2017). Utility of the autism observation scale for infants in early identification of autism in tuberous sclerosis complex. *Pediatric Neurology*, *75*, 80-86.
Cascio, M. A., Weiss, J. A., & Racine, E. (2021). Making autism research inclusive by attending to intersectionality: a review of the research ethics literature. *Review journal of autism and developmental disorders*, *8*(1), 22-36.
Chen, E., Martin, A. D., & Matthews, K. A. (2006). Socioeconomic status and health: do gradients differ within childhood and adolescence?. *Social science & medicine*, *62*(9), 2161-2170.
Chen, L. W., Wang, S. T., Wang, L. W., Kao, Y. C., Chu, C. L., Wu, C. C., ... & Huang, C. C. (2019). Behavioral characteristics of autism spectrum disorder in very preterm birth children. *Molecular autism*, *10*(1), 1-9.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates. *Hillsdale, NJ*, 20-26.
Cohen, S. (2021). The evolution of machine learning: Past, present, and future. In *Artificial Intelligence and Deep Learning in Pathology* (pp. 1-12). Elsevier.

Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., ... & Moher, D. (2014). Scoping reviews: time for clarity in definition, methods, and reporting. *Journal of clinical epidemiology*, *67*(12), 1291-1294.

Committee on Children with Disabilities. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics*, *108*(1), 192-195.

Como, D. H., Florídez, L. I., Tran, C. F., Cermak, S. A., & Stein Duker, L. I. (2020). Examining unconscious bias embedded in provider language regarding children with autism. *Nursing & Health Sciences*, *22*(2), 197-204.

Critical Appraisal Skills Programme. (2018). *CASP Checklists for Cohort Studies (cop. 2018)*.

Crossman, M. K., Lindly, O. J., Chan, J., Eaves, M., Kuhlthau, K. A., Parker, R. A., ... & Murray, D. S. (2020). Families' experiences with family navigation services in the autism treatment network. *Pediatrics*, *145*(Supplement_1), S60-S71.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., ... & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics*, *125*(1), e17-e23.

Dean, M., Harwood, R., & Kasari, C. (2017). The art of camouflage: Gender differences in the social behaviors of girls and boys with autism spectrum disorder. *Autism*, *21*(6), 678-689.

DeBrusk, C. (2018). The risk of machine-learning bias (and how to prevent it). *MIT Sloan Management Review*.

Denning, P. J., & Lewis, T. G. (2016). Exponential laws of computing growth. *Communications of the ACM*, *60*(1), 54-65.

Dilhac, M. A., Abrassart, C., & Voarino, N. (2018). The Montreal declaration for a responsible development of artificial intelligence. *Université de Montréal, Montréal, QC*.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1-17.

Doupe, P., Faghmous, J., & Basu, S. (2019). Machine learning for health services researchers. *Value in Health*, *22*(7), 808-815.

Dow, D., Day, T. N., Kutta, T. J., Nottke, C., & Wetherby, A. M. (2020). Screening for autism spectrum disorder in a naturalistic home setting using the systematic observation of red flags (SORF) at 18–24 months. *Autism Research*, *13*(1), 122-133.

Duda, M., Kosmicki, J. A., & Wall, D. P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational psychiatry*, *4*(8), e424-e424.

Dudley, C., & Emery, J. C. (2014). The value of caregiver time: Costs of support and care for individuals living with autism spectrum disorder. *SPP Research Paper*, (7-1).

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629-634.

Eman, D., & Emanuel, A. W. (2019, November). Machine learning classifiers for autism spectrum disorder: a review. In *2019 4th international conference on information technology, information systems and electrical engineering (icitisee)* (pp. 255-260). IEEE.

Esser, E., & Swalve, N. (2022). Implicit gender bias in perceptions of financial jobs. *SN Social Sciences*, *2*(8), 1-11.

Estes, A., Zwaigenbaum, L., Gu, H., St John, T., Paterson, S., Elison, J. T., ... & Piven, J. (2015). Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. *Journal of neurodevelopmental disorders*, *7*(1), 1-10.

Freitas, T. C., Gabbard, C., Caçola, P., Montebelo, M. I., & Santos, D. C. (2013). Family socioeconomic status and the provision of motor affordances in the home. *Brazilian journal of physical therapy*, *17*, 319-327.

Fuller, E. A., & Kaiser, A. P. (2020). The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis. *Journal of autism and developmental disorders*, *50*(5), 1683-1700.

Gammer, I., Bedford, R., Elsabbagh, M., Garwood, H., Pasco, G., Tucker, L., ... & BASIS Team. (2015). Behavioural markers for autism in infancy: scores on the Autism Observational Scale for Infants in a prospective study of at-risk siblings. *Infant Behavior and Development*, *38*, 107-115.

Gardner, L. M., Murphy, L., Campbell, J. M., Tylavsky, F., Palmer, F. B., & Graff, J. C. (2013). Screening accuracy for risk of autism spectrum disorder using the Brief Infant-Toddler Social and Emotional Assessment (BITSEA). *Research in Autism Spectrum Disorders*, *7*(5), 591-600.

Garg, A., & Mago, V. (2021). Role of machine learning in medical research: A survey. *Computer Science Review*, *40*, 100370.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?. *Quantitative Science Studies*, *2*(3), 795-827.

Giserman Kiss, I., Feldman, M. S., Sheldrick, R. C., & Carter, A. S. (2017). Developing autism screening criteria for the brief infant toddler social emotional assessment (BITSEA). *Journal of autism and developmental disorders*, *47*(5), 1269-1277.

Gliga, T., Bedford, R., Charman, T., Johnson, M. H., Baron-Cohen, S., Bolton, P., ... & Tucker, L. (2015). Enhanced visual search in infancy predicts emerging autism symptoms. *Current Biology*, *25*(13), 1727-1730.

Gustavson, K., von Soest, T., Karevold, E., & Røysamb, E. (2012). Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a Monte Carlo simulation study. *BMC public health*, *12*(1), 1-11.

Hahn, L. J., Brady, N. C., McCary, L., Rague, L., & Roberts, J. E. (2017). Early social communication in infants with fragile X syndrome and infant siblings of children with autism spectrum disorder. *Research in developmental disabilities*, *71*, 169-180.

Hahn, L. J., Hamrick, L. M., Kelleher, B. L., & Roberts, J. E. (2020). Autism spectrum disorder-associated behaviour in infants with down syndrome. *Journal of health science & education*, *4*(2).

Helleputte , T., Paul , J., & Gramme , P. (2021). LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library. cran.r-project.org. https://cran.r-project.org/web/packages/LiblineaR/index.html

Higgins, J. P., Li, T., & Deeks, J. J. (2019). Choosing effect measures and computing estimates of effect. *Cochrane handbook for systematic reviews of interventions*, 143-176.

Hollingshead, A. B. (1975). Four factor index of social status.

Hultman, C. M., Sandin, S., Levine, S. Z., Lichtenstein, P., & Reichenberg, A. (2011). Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies. *Molecular psychiatry*, *16*(12), 1203-1212.

Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. (2019). Applications of supervised machine learning in autism spectrum

disorder research: a review. *Review Journal of Autism and Developmental Disorders*, *6*(2), 128-146.
IBM (2020). Machine Learning. ibm.com. https://www.ibm.com/cloud/learn/machine-learning
IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 28.0.1.1 (14). Armonk, NY: IBM Corp
Jeste, S. S., Wu, J. Y., Senturk, D., Varcin, K., Ko, J., McCarthy, B., ... & Nelson, C. A. (2014). Early developmental trajectories associated with ASD in infants with tuberous sclerosis complex. *Neurology*, *83*(2), 160-168.
Jones, E. J., Gliga, T., Bedford, R., Charman, T., & Johnson, M. H. (2014). Developmental pathways to autism: a review of prospective studies of infants at risk. *Neuroscience & Biobehavioral Reviews*, *39*, 1-33.
Karatzoglou , A., Smola , A., & Hornik , K. (2022). kernlab: Kernel-Based Machine Learning Lab. cran.r-project.org. https://cran.r-project.org/web/packages/kernlab/index.html
Kaufman, N. K. (2022). Rethinking "gold standards" and "best practices" in the assessment of autism. *Applied Neuropsychology: Child*, *11*(3), 529-540.
Keyes, R. W. (2006). The impact of Moore's Law. *IEEE solid-state circuits society newsletter*, *11*(3), 25-27.
Khan, H. N., Hounshell, D. A., & Fuchs, E. R. (2018). Science and research policy at the end of Moore's law. *Nature Electronics*, *1*(1), 14-21.
Khun, M., Wing , J., Weston , S., Williams , A., Keefer, C., Engelhardt, … & Hunt , T. (2022). caret: Classification and Regression Training. cran.r-project.org. https://cran.r-project.org/web/packages/caret/index.html
Kollias, K. F., Syriopoulou-Delli, C. K., Sarigiannidis, P., & Fragulis, G. F. (2021). The Contribution of Machine Learning and Eye-Tracking Technology in Autism Spectrum Disorder Research: A Systematic Review. *Electronics*, *10*(23), 2982.
Kodak, T., & Bergmann, S. (2020). Autism spectrum disorder: Characteristics, associated behaviors, and early intervention. *Pediatric Clinics*, *67*(3), 525-535.
Landa, R. J. (2018). Efficacy of early interventions for infants and young children with, and at risk for, autism spectrum disorders. *International Review of Psychiatry*, *30*(1), 25-39.
Lavelle, T. A., Weinstein, M. C., Newhouse, J. P., Munir, K., Kuhlthau, K. A., & Prosser, L. A. (2014). Economic burden of childhood autism spectrum disorders. *Pediatrics*, *133*(3), e520-e529.
Lawson, G. M., Hook, C. J., & Farah, M. J. (2018). A meta-analysis of the relationship between socioeconomic status and executive function performance among children. *Developmental science*, *21*(2), e12529.
Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), 659-685.
Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). Autism diagnostic observation schedule–2nd edition (ADOS-2). *Los Angeles, CA: Western Psychological Corporation*, *284*.
Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, *19*(2), 185-212.

Maenner, M. J., Shaw, K. A., Bakian, A. V., Bilder, D. A., Durkin, M. S., Esler, A., ... & Cogswell, M. E. (2021). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2018. *MMWR Surveillance Summaries*, *70*(11), 1.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381-386.

Malik-Soni, N., Shaker, A., Luck, H., Mullin, A. E., Wiley, R. E., Lewis, M. E., ... & Frazier, T. W. (2022). Tackling healthcare access barriers for individuals with autism from diagnosis to adulthood. *Pediatric Research*, *91*(5), 1028-1035.

Marsack-Topolewski, C. N., & Church, H. L. (2019). Impact of caregiver burden on quality of life for parents of adult children with autism spectrum disorder. *American journal on intellectual and developmental disabilities*, *124*(2), 145-156.

Mazurek, M. O., Lu, F., Symecko, H., Butter, E., Bing, N. M., Hundley, R. J., ... & Handen, B. L. (2017). A prospective study of the concordance of DSM-IV and DSM-5 diagnostic criteria for autism spectrum disorder. *Journal of autism and developmental disorders*, *47*(9), 2783-2794.

McDonald, N. M., Varcin, K. J., Bhatt, R., Wu, J. Y., Sahin, M., Nelson III, C. A., & Jeste, S. S. (2017). Early autism symptoms in infants with tuberous sclerosis complex. *Autism Research*, *10*(12), 1981-1990.

McDonald, S. W., Kehler, H. L., & Tough, S. C. (2018). Risk factors for delayed social-emotional development and behavior problems at age two: Results from the All Our Babies/Families (AOB/F) cohort. *Health science reports*, *1*(10), e82.

McNally, J., Hugh-Jones, S., Caton, S., Vereijken, C., Weenen, H., & Hetherington, M. (2016). Communicating hunger and satiation in the first 2 years of life: a systematic review. *Maternal & child nutrition*, *12*(2), 205-228.

Medow, M. A., & Lucey, C. R. (2011). A qualitative approach to Bayes' theorem. *Bmj evidence-based medicine*, *16*(6), 163-167.

Mehling, M. H., & Tassé, M. J. (2016). Severity of autism spectrum disorders: Current conceptualization, and transition to DSM-5. *Journal of Autism and Developmental Disorders*, *46*(6), 2000-2016.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269.

Moola, S., Munn, Z., Tufanaru, C., Aromataris, E., Sears, K., Sfetcu, R., ... & Mu, P. F. (2017). Chapter 7: Systematic reviews of etiology and risk. *Joanna briggs institute reviewer's manual. The Joanna Briggs Institute*, *5*.

Moon, S. J., Hwang, J., Kana, R., Torous, J., & Kim, J. W. (2019). Accuracy of machine learning algorithms for the diagnosis of autism spectrum disorder: Systematic review and meta-analysis of brain magnetic resonance imaging studies. *JMIR mental health*, *6*(12), e14108.

Morán-Fernández, L., Bólon-Canedo, V., & Alonso-Betanzos, A. (2022). How important is data quality? Best classifiers vs best features. *Neurocomputing*, *470*, 365-375.

Mottron, L., & Bzdok, D. (2020). Autism spectrum heterogeneity: fact or artifact?. *Molecular Psychiatry*, *25*(12), 3178-3185.

Mullen, E. M. (1995). *Mullen scales of early learning* (pp. 58-64). Circle Pines, MN: AGS.

Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (p. 78).

Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262-e273.
Nhlbi, N. (2022). Study Quality Assessment Tools National Heart. *Lung and Blood Institute.[(accessed on 16 January 2022)]*.
Niles, G., & Harkins Monaco, E. A. (2019). Privilege, social identity and autism: Preparing preservice practitioners for intersectional pedagogy. *Division on Autism and Developmental Disabilities Online Journal*, *6*(1), 112-113.
Noyes-Grosser, D. M., Elbaum, B., Wu, Y., Siegenthaler, K. M., Cavalari, R. S., Gillis, J. M., & Romanczyk, R. G. (2018). Early intervention outcomes for toddlers with autism spectrum disorder and their families. *Infants & Young Children*, *31*(3), 177-199.
Numis, A. L., Major, P., Montenegro, M. A., Muzykewicz, D. A., Pulsifer, M. B., & Thiele, E. A. (2011). Identification of risk factors for autism spectrum disorders in tuberous sclerosis complex. *Neurology*, *76*(11), 981-987.
O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior research methods, instruments, & computers*, *32*(3), 396-402.
Ortiz-Rubio, A., Torres-Sánchez, I., Cabrera-Martos, I., Rodríguez-Torres, J., López-López, L., Prados-Román, E., & Valenza, M. C. (2021). The Caregiver Burden Inventory as a sleep disturbance screening tool for parents of children with autism spectrum disorder. *Journal of pediatric nursing*, *61*, 166-172.
Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., ... & Stone, W. L. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics*, *128*(3), e488-e495.
Ozonoff, S., Young, G. S., Landa, R. J., Brian, J., Bryson, S., Charman, T., ... & Iosif, A. M. (2015). Diagnostic stability in young children at risk for autism spectrum disorder: a baby siblings research consortium study. *Journal of Child Psychology and Psychiatry*, *56*(9), 988-998.
Pagnozzi, A. M., Conti, E., Calderoni, S., Fripp, J., & Rose, S. E. (2018). A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective. *International Journal of Developmental Neuroscience*, *71*, 68-82.
Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, *56*(1), 45.
Paul, M. H. (2017). *Rural-urban disparities in the diagnosis and treatment of children with Autism Spectrum Disorders (ASD)* (Doctoral dissertation, The Ohio State University).
Pickles, A., Le Couteur, A., Leadbitter, K., Salomone, E., Cole-Fletcher, R., Tobin, H., ... & Green, J. (2016). Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial. *The Lancet*, *388*(10059), 2501-2509.
Puleo, C. M., Schmeidler, J., Reichenberg, A., Kolevzon, A., Soorya, L. V., Buxbaum, J. D., & Silverman, J. M. (2012). Advancing paternal age and simplex autism. *Autism*, *16*(4), 367-380.
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.

Redwood, S., & Gill, P. S. (2013). Under-representation of minority ethnic groups in research—call for action. *British Journal of General Practice*, *63*(612), 342-343.
Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, *11*(1), 1-9.
Roberts, J. E., Tonnsen, B. L., McCary, L. M., Caravella, K. E., & Shinkareva, S. V. (2016). Brief report: Autism symptoms in infants with fragile X syndrome. *Journal of Autism and Developmental Disorders*, *46*(12), 3830-3837.
Robertson, R. E., Sobeck, E. E., Wynkoop, K., & Schwartz, R. (2017). Participant diversity in special education research: Parent-implemented behavior interventions for children with autism. *Remedial and Special Education*, *38*(5), 259-271.
Robin , X., Turck , N., Hainard , A., Tiberti , N., Lisacek , F., Sanchez , J. C., … & Billings , Z. (2022). pROC: Display and Analyze ROC Curves. cran.r-project.org. https://cran.r-project.org/web/packages/pROC/index.html
Rogge, N., & Janssen, J. (2019). The economic costs of autism spectrum disorder: A literature review. *Journal of Autism and Developmental Disorders*, *49*(7), 2873-2900.
RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.
Sacrey, L. A. R., Zwaigenbaum, L., Brian, J. A., Smith, I. M., Armstrong, V., Raza, S., ... & Schmidt, L. A. (2021). Affect and gaze responses during an Emotion-Evoking Task in infants at an increased likelihood for autism spectrum disorder. *Molecular autism*, *12*(1), 1-19.
Sacrey, L. A. R., Zwaigenbaum, L., Bryson, S., Brian, J., Smith, I. M., Roberts, W., ... & Garon, N. (2018). Parent and clinician agreement regarding early behavioral signs in 12-and 18-month-old infants at-risk of autism spectrum disorder. *Autism Research*, *11*(3), 539-547.
Sacrey, L. A. R., Zwaigenbaum, L., Szatmari, P., Bryson, S., Georgiades, S., Brian, J., ... & Elsabbagh, M. (2017). Brief Report: Characteristics of preschool children with ASD vary by ascertainment. *Journal of Autism and Developmental Disorders*, *47*(5), 1542-1550.
Salgado, R. G. (2020). *The Roles of Marginalization and Empowerment on Indicators of System Navigation and Mental Health for Parents of Children with Autism Spectrum Disorder* (Doctoral dissertation, University of Oregon).
Sanders, H., & Saxe, J. (2017). Garbage in, garbage out: how purportedly great ML models can be screwed up by bad data. *Proceedings of Blackhat*, *2017*.
Sanderson, C. (2016). *Early Detection of Autism Spectrum Symptomatology in Very Preterm Infants* (Doctoral dissertation, Royal Holloway, University of London).
Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1-21.
Scheffer, J. (2002). Dealing with missing data.
Sedgewick, F., Kerr-Gaffney, J., Leppanen, J., & Tchanturia, K. (2019). Anorexia nervosa, autism, and the ADOS: how appropriate is the new algorithm in identifying cases?. *Frontiers in Psychiatry*, *10*, 507.
Shalf, J. (2020). The future of computing beyond Moore's Law. Philosophical Transactions of the Royal Society A, 378(2166), 20190061.

Siddiqui, S., Gunaseelan, L., Shaikh, R., Khan, A., Mankad, D., & Hamid, M. A. (2021). Food for Thought: Machine Learning in Autism Spectrum Disorder Screening of Infants. *Cureus*, *13*(10).

Sparrow, S. S., & Cicchetti, D. V. (1985). Diagnostic uses of the vineland adaptive behavior scales. *Journal of Pediatric Psychology*, *10*(2), 215-225.

Stang, A., Jonas, S., & Poole, C. (2018). Case study in major quotation errors: a critical commentary on the Newcastle–Ottawa scale. *European Journal of Epidemiology*, *33*(11), 1025-1031.

Sterne, J. A. (2009). *Meta-analysis in Stata: an updated collection from the Stata Journal*. StataCorp LP.

Stone, W. L., McMahon, C. R., & Henderson, L. M. (2008). Use of the screening tool for autism in two-year-olds (STAT) for children under 24 months: An exploratory study. *Autism*, *12*(5), 557-573.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use*. Oxford University Press, USA.

Szatmari, P., Chawarska, K., Dawson, G., Georgiades, S., Landa, R., Lord, C., ... & Halladay, A. (2016). Prospective longitudinal studies of infant siblings of children with autism: lessons learned and future directions. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(3), 179-187.

Tanner, A., & Dounavi, K. (2021). The emergence of autism symptoms prior to 18 months of age: A systematic literature review. *Journal of Autism and Developmental Disorders*, *51*(3), 973-993.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017. http://standards.ieee.org/develop/indconn/ec/autonomous_ systems.html.

Thompson, J. R., Hughes, C., Schalock, R. L., Silverman, W., Tassé, M. J., Bryant, B., ... & Campbell, E. M. (2002). Integrating supports in assessment and planning. *Mental Retardation*, *40*(5), 390-405.

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, *80*, 146-166.

Tong, S., Baghurst, P., & McMichael, A. (2006). Birthweight and cognitive development during childhood. *Journal of paediatrics and child health*, *42*(3), 98-103.

Torche, F., & Echevarría, G. (2011). The effect of birthweight on childhood cognitive development in a middle-income country. *International journal of epidemiology*, *40*(4), 1008-1018.

Towle, P. O., Patrick, P. A., Ridgard, T., Pham, S., & Marrus, J. (2020). Is earlier better? The relationship between age when starting early intervention and outcomes for children with autism spectrum disorder: a selective review. *Autism Research and Treatment*, *2020*.

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health*, *5*, 307.

Tubío-Fungueiriño, M., Cruz, S., Sampaio, A., Carracedo, A., & Fernández-Prieto, M. (2021). Social camouflaging in females with autism spectrum disorder: A systematic review. *Journal of Autism and Developmental Disorders*, *51*(7), 2190-2199.

van't Hof, M., Tisseur, C., van Berckelear-Onnes, I., van Nieuwenhuyzen, A., Daniels, A. M., Deen, M., ... & Ester, W. A. (2021). Age at autism spectrum disorder diagnosis: A systematic review and meta-analysis from 2012 to 2019. *Autism*, *25*(4), 862-873.

West, E. A., Travers, J. C., Kemper, T. D., Liberty, L. M., Cote, D. L., McCollow, M. M., & Stansberry Brusnahan, L. L. (2016). Racial and ethnic diversity of participants in research supporting evidence-based practices for learners with autism spectrum disorder. *The Journal of Special Education*, *50*(3), 151-163.

Wetherby, A. M., Brosnan-Maddox, S., Peace, V., & Newton, L. (2008). Validation of the Infant—Toddler Checklist as a broadband screener for autism spectrum disorders from 9 to 24 months of age. *Autism*, *12*(5), 487-511.

Wickham , H., Hester , J., Francois , R., Bryan , J., Bearrows , S., Jylänki , J., & Jørgensen , M. (2022). readr: Read Rectangular Text Data. cran.r-project.org. https://cran.r-project.org/web/packages/readr/index.html

Wickham , H., Miller , E., & Smith , D. (2022). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. cran.r-project.org. https://cran.r-project.org/web/packages/haven/index.html

Wozniak, R. H., Leezenbaum, N. B., Northrup, J. B., West, K. L., & Iverson, J. M. (2017). The development of autism spectrum disorders: variability and causal complexity. Wiley Interdisciplinary Reviews: Cognitive Science, 8(1-2), e1426.

Yaylaci, F., & Miral, S. (2017). A comparison of DSM-IV-TR and DSM-5 diagnostic classifications in the clinical diagnosis of autistic spectrum disorder. *Journal of autism and developmental disorders*, *47*(1), 101-109.

Yeung, K. (2020). Recommendation of the council on artificial intelligence (OECD). *International Legal Materials*, *59*(1), 27-34.

Ying, X. (2019, February). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168, No. 2, p. 022022). IOP Publishing.

Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. *Big data and cognitive computing*, *2*(1), 5.

Zwaigenbaum, L., Bauman, M. L., Choueiri, R., Kasari, C., Carter, A., Granpeesheh, D., ... & Natowicz, M. R. (2015). Early intervention for children with autism spectrum disorder under 3 years of age: recommendations for practice and research. *Pediatrics*, *136*(Supplement_1), S60-S81.

Zwaigenbaum, L., Bauman, M. L., Fein, D., Pierce, K., Buie, T., Davis, P. A., ... & Wagner, S. (2015). Early screening of autism spectrum disorder: recommendations for practice and research. *Pediatrics*, *136*(Supplement_1), S41-S59.

Zwaigenbaum, L., Brian, J., Smith, I. M., Sacrey, L. A. R., Franchini, M., Bryson, S. E., ... & Roncadin, C. (2021). Symptom trajectories in the first 18 months and autism risk in a prospective high-risk cohort. *Journal of Child Psychology and Psychiatry*, *62*(12), 1435-1443.

Zwaigenbaum, L., Bryson, S. E., Brian, J., Smith, I. M., Sacrey, L., Armstrong, V., ... & Roncadin, C. (2021). Assessment of Autism Symptoms From 6 to 18 Months of Age Using the Autism Observation Scale for Infants in a Prospective High-Risk Cohort. *Child Development*, *92*(3), 1187-1198.

Zwaigenbaum, L., Bryson, S. E., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. *International journal of developmental neuroscience*, *23*(2-3), 143-152.

Zwaigenbaum, L., Bryson, S. E., Szatmari, P., Brian, J., Smith, I. M., Roberts, W., ... & Roncadin, C. (2012). Sex differences in children with autism spectrum disorder identified within a high-risk infant cohort. *Journal of autism and developmental disorders*, *42*(12), 2585-2596.

Appendix 1: Supplemental Content to Study One

Systematic Review Database Searches as Run

CINAHL Plus / ERIC Search via OVID – July 2022

Search as run: ("autism observation scale for infants" OR "AOSI") AND ("autism" OR "autism spectrum disorders" OR "autistic disorder")

Databases being searched: CINAHL Plus with Full Text, ERIC		
	("autism" OR "autism spectrum disorders" OR "autistic")	TX All Text
AND	("autism observation scale for infants" OR "AOSI")	TX All Text
Limit: From January 2005 – July 2022		

Results:
83 results found July 4th, 2022 (75 for CINAHL, 8 for ERIC). Search limited to publications between January 2005 and July 2022.

JSTOR Search – July 2022

Search as run:

1st Keyword		"autism"	All fields
2nd Keyword	OR	"autism spectrum disorders"	All fields
3rd Keyword	OR	"autistic disorder"	All fields
4th Keyword	AND	"autism observation scale for infants"	All fields
5th Keyword	OR	"AOSI"	All fields
Access type: All content **Item type**: n/a **Language:** All languages **Publication date**: From 2005/01/01 to 2022/07/04 **Journal or book title:** n/a **ISBN:** n/a **Journal filter**: n/a			
Date	From: 2005-01-01	To: 2022-07-04	

Results:
74 results found July 4th, 2022. Search limited to publications between January 2005 and July 2022.

PubMed Search – July 2022

Search as run:
((autism) OR (autism spectrum disorders) OR (autistic disorder)) AND ((autism observation scale for infants) OR (AOSI))
Filter: From 2005/1/1 to 2022/7/4
Results:
214 results found July 4th, 2022. Search limited to publications between January 2005 and July 2022.

Web of Science – July 2022

Search as run:

	Topic	("autism" OR "autism spectrum disorder" OR "autistic disorder")	All fields
AND	TOPIC	("Autism Observation Scale for Infants" OR "AOSI")	All fields
Index date	From: 2005-01-01	To: 2022-07-04	

Results

38 results found July 4th, 2022. Search limited to publications between January 2005 and July 2022.

EMBASE/OVID – July 2022

Search as run:

#	**Searches**	**Results**
1	Exp autism/	84582
2	"autism spectrum disorder".mp	29729
3	"autistic disorder"	2618
4	1 or 2 or 3	87479
5	"autism observation scale for infants".mp	36
6	"AOSI".mp	60
7	5 or 6	75
8	4 and 7	44
9	Limit 8 to yr="2005 -Current"	44

Results:

44 results found July 4th, 2022. Search limited to publications between January 2005 and July 2022.

Table A1.01 | Systematic Review PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Pg. 5, 7
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Pg. ii
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Pg. 6-7
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Pg. 7
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Pg. 8
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Pg. 8
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Pg. 7-8
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix 1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Pg. 8, 11
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Pg. 10
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Pg. 10
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Pg. 9-10
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Pg. 10-11
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Pg. 11
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Pg. 11
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Pg. 10-11

Section/topic	#	Checklist item	Reported on page #
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Pg. 7, Figure 2.01
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Pg. 12-13, 16, Table 2.01, 2.02
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Pg. 31, Table 2.04
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Pg. 15-19, Table 3
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Pg. 26-31, Figure 2.03, 2.04, Pg., 126-128, Appendix 1 Figure A1.01, Figure A1.02
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Pg. 31, Table 2.04
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Pg. 31, Appendix 1, Figure A1.01, Figure A1.02
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Pg. 35-40
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Pg. 40
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Pg. 40-41
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data), role of funders for the systematic review.	Pg. vi

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

Confirmation of ASD in Proband of Infant Siblings

IL Infant sibling status varied across the 13 studies that had IL infant sibling study participants: Zwaigenbaum et al., 2005, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, and Sacrey et al., 2018 required confirmation of the older sibling's ASD diagnosis via clinical assessment or review of diagnostic records using DSM criteria, Estes et al., 2015 required older siblings to meet criteria on the Social Communication Questionnaire (SCQ) and ADI-R, Gammer et al., 2015, Gilga et al., 2015, Bedford et al., 2017, and Bedford et al., 2019 confirmed the community clinical diagnosis of ASD using the Development and Well Being Assessment (DAWBA) and SCQ, and Hahn et al., 2017 via documentation of an ASD diagnosis (though what this entails was not specified). Though Bussu et al., 2018, Roberts et al., 2016, and Bedford et al., 2016 included IL infant siblings participants, they did not specify how the older sibling's diagnosis of ASD was confirmed by study authors.

Systematic Review Inclusion criteria

Explicit inclusion criteria for participants were detailed for all seventeen studies but varied according to the goals and objectives of each study. For instance, of the 13 studies that included infant sibling participants, designation as an infant sibling was the most stringent for Zwaigenbaum et al., 2005, Zwaigenbaum et al., 2020, Zwaigenbaum et al., 2021, and Sacrey et al., 2018 due to their required confirmation of the older sibling's ASD diagnosis through either clinical assessment or review of diagnostic records using DSM-IV-TR criteria. All other studies with infant sibling participants either required the older sibling to meet criteria for ASD on some ASD measure (SCQ, ADI-R, DAWBA) or made no mention of how the older sibling received an ASD diagnosis. For the two studies with FXS infants (Hahn et al., 2017, Roberts et al., 2016), infant status with FXS required confirmed by genetic report. For the three studies with TSC infants, infant TSC status was based on clinical presentation or genetic workups (McDonald et al., 2017), meeting clinical or genetic criteria for TSC (Capal et al., 2017), or via being recruited from TSC specialty clinics, newborn nurseries, pediatrician offices, or met genetic criteria for TSC diagnosis (Jeste et al., 2014). For the one study of infants with DS, while no confirmatory testing was done by study authors, infants were recruited from three other pilot studies examining the infant neurogenetic syndromes who themselves recruited participants based on flyers with local parent groups, DS clinics, and ongoing research studies in the United States

(Hahn et al., 2020). Overall, none of the studies that recruited infants with FXS, TSC, or DS described how infant IL status was confirmed relative to any DSM or ICD criterion.

Inclusion criteria for LL controls varied considerably across the 15 studies which employed them. LL inclusion criteria ranged from explicitly detailed and descriptive to increasingly sparse and lacking detail. Gilga et al., 2015 had the most robust inclusion criteria for LL controls; they required (A) control infants to have an older sibling who was born full-term with a normal birth weight, and (B) control infants to lack ASD diagnoses in any first-degree family members confirmed by either parent interview or family medical history. Similarly, Estes et al., 2015 required LL control infants to (A) have an older sibling who did not meet criteria SCQ or Family Interview for Genetic Studies (FIGS) criteria for ASD, and (B) a lack of first-degree relatives with ASD or intellectual disability. Roberts et al., 2016 was less descriptive and simply required an absence of suspected delays and no familial history or indicator of ASD in LL controls. Gammer et al., 2015 required LL controls to (A) not have first degree relatives with ASD, and (B) have an older sibling (full or half) that did not meet criteria for ASD on the SCQ (aka did not meet the cut-off of ≥ 15 on the SCQ). Bedford et al., 2017 and Bedford et al., 2019 both had similar inclusion criteria, but only required LL controls to have an older sibling that did not meet criteria for ASD on the SCQ (did not score ≥ 15). Zwaigenbaum et al., 2005, Zwaigenbaum et al., 2020, and Zwaigenbaum et al., 2021 all required LL controls to lack a first- or second-degree relative with an ASD diagnosis. Bussu et al., 2018 had the least restrictive LL control inclusion criteria and only required control infants to have an older full sibling with typical development. While Jeste et al., 2014 and Hahn et al., 2020 both recruited LL control infants from either IRB-approved infant databases or from other studies respectively. Three studies (McDonald et al., 2017, Hahn et al., 2017, Bedford et al., 2016) did not specify how LL controls were recruited or the criterion used to do so. Only two studies (Capal et al., 2017, Sacrey et al., 2018) did not employ LL control groups .

Systematic Review Exclusion criteria

Explicit exclusion criteria for IL participants were detailed for only 11 studies and varied according to their respective goals and objectives. Estes et al., 2015 excluded participants if they had genetic conditions or syndromes, sensory impairments (e.g., vision or hearing loss), had a

birth weight <2,000g, gestational age of <36 weeks at birth, or suffered significant perinatal adversity and/or were exposed *in utero* to neurotoxins, had any MRI contraindications, had a predominant household language that was not English, were adopted or half-siblings, had a first degree relative with psychosis, schizophrenia, or bipolar disorder, or were twins. Likewise, Capal et al., 2017 excluded IL infants if they were born preterm (<36 weeks), suffered significant perinatal complications, were administered investigational drugs as part of other research studies, were taking an mTOR inhibitor at the time of enrollment, had Subependymal giant Cell Astrocytoma necessitating medical/surgical treatment, had a history of epilepsy, or had any MRI contraindications. The remaining 9 studies employed exclusion criteria of varying detail and robustness. Zwaigenbaum et al., 2020, Sacrey et al., 2018, and Zwaigenbaum et al., 2021 all broadly excluded IL participants if they were not born full-term, had a birth weight <2,500g, and/or had significant neurologic, genetic, or sensory-motor conditions. While similar, Zwaigenbaum et al., 2005 only excluded participants if they did not have term gestation or had a birth weight <2,500g. Roberts et al., 2016 and Gammer et al., 2015 excluded IL participants if they were not born full-term and had significant neurological or developmental conditions. Gilga et al., 2015 and Bedford et al., 2016 both excluded IL participants if they had significant medical or developmental conditions. Bussu et al., 2018 excluded participants who lacked 36-month clinical ASD outcome evaluation. The remaining 6 studies (Hahn et al., 2020, McDonald et al., 2017, Hahn et al., 2017, Bedford et al., 2017, Bedford et al., 2019, Jeste et al., 2014) did not detail exclusion criteria for IL participants.

Exclusion criteria for LL controls were explicitly detailed for only 10 studies and varied considerably. Jeste et al., 2014 and McDonald et al., 2017 both excluded LL controls if they were born preterm (<37 weeks of gestation), suffered birth trauma, had developmental concerns, or had any family history of ASD or intellectual disability. Roberts et al., 2016 excluded LL controls if they were not born full-term, had significant neurological or developmental conditions, and/or if they had developmental composite scores >1 standard deviation away from the mean. The exclusion criteria for LL controls was identical to that of IL participants for Zwaigenbaum et al., 2005, Gammer et al., 2015, Estes et al., 2015, Bussu et al., 2018, Zwaigenbaum et al., 2020, Bedford et al., 2016, and Zwaigenbaum et al., 2021 as described in the paragraph above. Five studies did not report LL control exclusion criteria (Hahn et al., 2020,

Hahn et al., 2017, Hahn et al., 2017, Bedford et al., 2017, Bedford et al., 2019) while two (Capal et al., 2017, Sacrey et al., 2018) did not employ LL control comparison groups in their study design.

Study Design and Interrater Reliability

Though 15 studies were longitudinal and 2 cross-sectional, each administered the AOSI at a single or multiple timepoints (ranging between 6, 9, 12, 15, or 18 months) and either (A) compared AOSI scores against a later ASD classification at 24-months or diagnostic assessment at 36-months, (B) compared AOSI scores across IL/LL study groups, (C) compared AOSI scores against scores on other early measures of autism symptoms, or (D) used AOSI scores in various statistical models (logistical regression, trajectory analysis, mixed modelling, multilevel modelling, autoregression, and machine learning). AOSI Total Scores, Risk Markers, or item-level data were analyzed by group membership (infant siblings, FXS, TSC, DS) against 24-month, 36-month, or 7-year ASD outcomes.

Though AOSI reliability data has been previously reported Bryson et al., 2008, AOSI reliability assessments were conducted by 6 of studies included in this review (Zwaigenbaum et al., 2005, Hahn et al., 2017, Roberts et al., 2016, Bedford et al., 2016, Bedford et al., 2017, Bedford et al., 2019). Zwaigenbaum et al., 2005 reports three main reliability estimates: (1) absolute agreement between raters of >90% for each AOSI item, (2) interrater agreement of the AOSI total score of 0.71, 0.90, and 0.92 for 6-, 12-, and 18-month AOSI administrations, and (3) a test-retest reliability at 12-month AOSI administrations of 0.63. Roberts et al., 2016 and Hahn et al., 2017 both double-coded 20% of AOSI assessments and report an item-level inter-rater reliability of 0.89. While Bedford et al., 2016, Bedford et al., 2017, and Bedford et al., 2019 all report double-coding the majority of AOSI assessments and report an intraclass correlation coefficient of 0.95, these studies all focus on the same sample of IL and LL infants and should be considered as one reliability estimate, not three.

IL-Developmentally Delayed (IL-DD) vs IL-ASD Meta Analyses

Between 6 and 10 Months. A total of three comparisons of AOSI Total Scores were included in this meta-analysis. There was no effect of AOSI Total Score, suggesting that the IL-ASD group

did not differ from the IL-DD control group (Cohen's *d*= 0.16, 95% CI = -0.09 - 0.41, *z* = 1.25, *p* = 0.21, Figure A1.1a).

Between 12 and 14 months. A total of two comparisons of AOSI Total Scores were included in the meta-analysis. There was no effect of AOSI Total Score, suggesting that the IL-ASD group did not differ from IL-DD for AOSI Total Scores (Cohen's *d* = 0.21, 95% CI = -0.05 - 0.47, *z* = 1.61, *p* = 0.11, Figure A1.1b)**.**

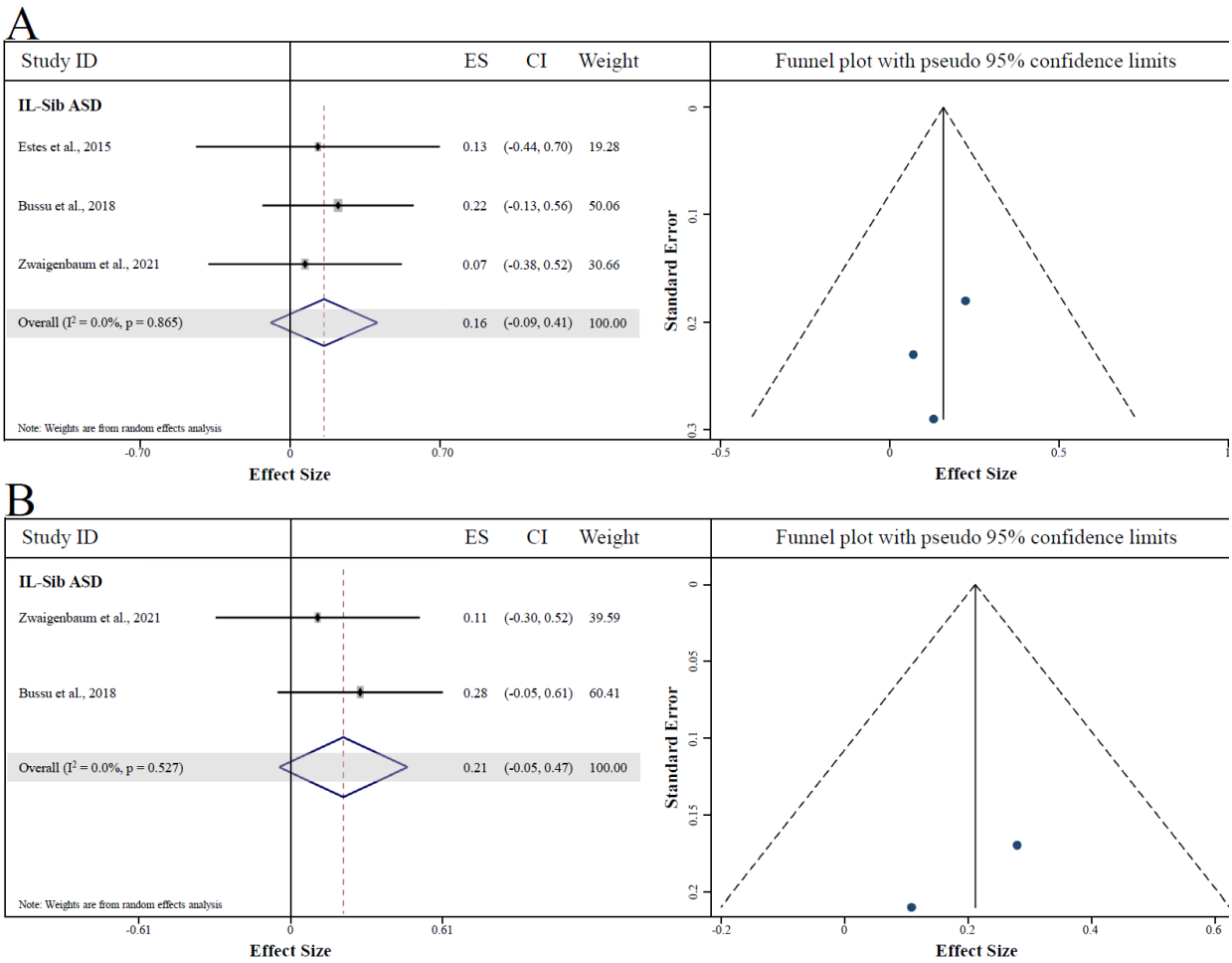

Figure A1.01a, b | Meta-Analysis comparing IL-DD to IL-ASD Samples (left) with the Trim and Fill Plot (right). A = for ages 6-10 months, B = for ages 12-14 months.

IL-Infants with Typical Development vs IL-ASD

Between 6 and 10 Months. A total of three comparisons of AOSI Total Scores were included in this meta-analysis. There was an effect of AOSI Total Score, suggesting that the IL-ASD group differ from the IL-typical control group (Cohen's *d*= 0.29, 95% CI = 0.10 - 0.49, *z* = 2.91, *p* = 0.004, Figure A1.2a).

Between 12 and 14 months. A total of two comparisons of AOSI Total Scores were included in the meta-analysis. There was an effect of AOSI Total Score, suggesting that the IL-ASD group differed from IL-typical for AOSI Total Scores (Cohen's $d = 0.74$, 95% CI = 0.51 - 0.97, $z = 6.35$, $p < 0.001$, Figure A1.2b)**.**

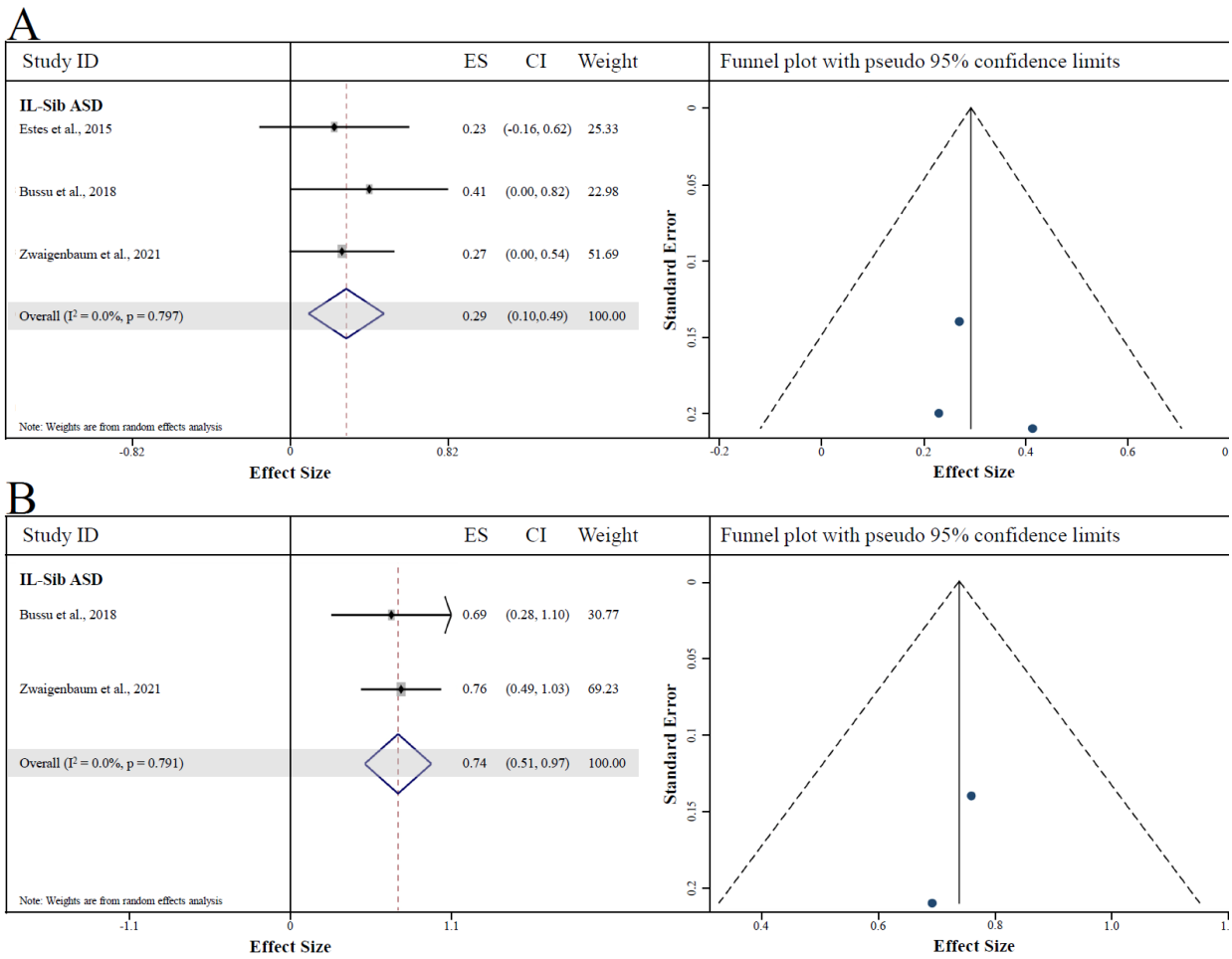

Figure A1.02a, b | Meta-Analysis comparing IL-typical to IL-ASD Samples (left) with the Trim and Fill Plot (right). A = for ages 6-10 months, B = for ages 12-14 months.

Appendix 2: Supplemental Content to Study Two

Use of an Additional, Independent Testing Set of IL-Siblings

While 465 IL-sibling data from CISS-1 was used to generate the training and test datasets used for statistical modelling, participant data on 133 new IL-siblings from CISS-2 were available to be used as an additional independent testing set for model performance. This was desirable, as participants in this new study were assessed using a nearly identical study similar protocol across the same autism ASD diagnostic and treatment centers in Canada (Sacrey et al., 2021). Participants in CISS-2 were assessed using many of the same autism and developmental measures. Germane to this study, participants in CISS-2 were assessed using the AOSI and MSEL at 12-months and underwent a 36-month diagnostic assessment for ASD.

Confirmation of IL-Sibling Status

To confirm their status as IL-siblings, diagnosis of ASD in the older siblings was confirmed through clinical assessment or a review of diagnostic records using DSM-5 criteria. The IL infants from the independent validation set no identifiable neurological or genetic conditions, nor any severe sensory or motor impairments. All IL-sibling participants were born at 36-42 weeks gestation and had a birth weight greater than 2500g.

New IL-Sibling Data 36-Month Diagnostic Procedure

At 36 months old, each participant underwent an independent diagnostic evaluation conducted by an expert clinician blind to results from previous study visits. Unlike the CISS-1 which used the DSM-IV-TR, clinicians in CISS-2 assigned ASD diagnosis using DSM-5 criteria based on best judgment of the clinician (developmental pediatrician, child psychiatrist, or clinical psychologist, all with years of diagnostic experience) while considering information from the ADI-R and ADOS and concurrent developmental information from the MSEL and VABS.

Handling Missing Data

Prior to use as an independent test set, data completeness of the AOSI and MSEL at 12-months was sought. For the 133 IL-siblings in CISS-2, 36-month diagnostic outcomes were characterized by the lowest missing data rate of 1.5%. Missingness of AOSI data (age of administration, items 1-18) ranged between 3.8 and 6.0% for all participants. In contrast, missingness of MSEL data was more considerable, with between 7.5 and 33.1% of MSEL

standard score data missing. MSEL data missingness was not uniform across all MSEL subscales; RL and ELC scores had missing data rates of only 7.5 and 8.3% respectively. VR, FM, and EL standard scores in contrast had missing data rates of 30.8, 30.1, and 33.1% respectively.

Since an unacceptably high number of participants were missing 12-month MSEL VR, FM, and EL data, new participant data was trimmed based on if CISS-2 participants (1) lacked 36-month diagnostic outcomes, (2) were missing data on 12-month AOSI and MSEL assessments, or (3) were missing data for three or more MSEL subscales. Using this criterion, data from 43 participants was removed resulting in a final pool of 90 IL-siblings. Of these 90 participants, missingness of AOSI data (age of administration, items 1-18) now ranged between 0 and 3.3% for all participants. Missingness of MSEL data was considerably reduced, with between 0 and 3.3% of MSEL standard score data missing from the trimmed participant data.

Randomness of Missing Data for CISS-2 Trimmed Participants

Of the 90 IL-siblings with 12-month AOSI/MSEL data and 36-month diagnostic outcomes, 23 were classified with ASD (IL-ASD) and 67 were not (IL-N). To test the hypothesis that data were missing completely at random for AOSI and MSEL variables, Little's MCAR test was conducted. Data were split according to 36-month diagnostic outcomes for data imputation purposes as these two groups are not homogenous; IL-ASD infants score higher on the AOSI and have greater range of impairments on the MSEL relative to IL-N infants.

For the 23 IL-ASD infants, Little's MCAR test were non-significant for AOSI item-level (χ^2 = 18.259, DF = 16, p = 0.309) and MSEL standard score data (χ^2 = 11.171, DF = 8, p = 0.192) indicating that data in the trimmed dataset was missing completely at random. While no item-level AOSI data was missing for IL-ASD participants, minor amounts of MSEL data were missing (between 0 and 8.7% for MSEL ELC, VR, FM, RL, and EL subscales).

For the 67 IL-N infants, Little's MCAR test were similarly non-significant for AOSI item-level (χ^2 = 7.083, DF = 21, p = 0.998) and MSEL standard score data (χ^2 = 4.549, DF = 9, p = 0.872) indicating that data in the trimmed dataset was missing completely at random. Minimal amounts

of item-level AOSI (between 0 and 1.5% across all AOSI items) and MSEL data were missing (between 0 and 1.5% of MSEL ELC, VR, FM, RL, and EL subscales).

Since IL-ASD and IL-N AOSI and MSEL missing data were randomly distributed, missing data were eligible for data imputation. Like the IL-sibling data used to generate the training and test sets, all AOSI items scoring '8' for IL-ASD and IL-N participants were replaced with '0's prior to expectation maximization imputation. Missing data was imputed via EM in SPSS GradPack Version 28 for IL-ASD and IL-N participants separately.

Assessing for Differences in Raw vs Cleaned/Imputed Dataset Statistics

Overall, EM data imputation had minimal impact on data statistics for AOSI and MSEL data of IL-ASD and IL-N participants from the CISS-2.

For IL-ASD AOSI data, EM had either no impact (signifying no missing items being replaced) or resulted in minimal differences in mean, standard deviation, and standard error between the raw data and cleaned/EM-imputed AOSI data (max absolute difference in means, standard deviation, and standard error between the raw and EM-imputed data was 0.019, 0.003, and 0.001 respectively). For IL-ASD MSEL data, EM had minimal impact on mean, standard deviation, and standard error between raw data and cleaned/EM-imputed MSEL data (max absolute difference in mean, standard deviation, and standard error between the raw and EM-imputed data was 1.042, 0.325, and 0.068 respectively).

For IL-N AOSI data, EM data had either no impact (signifying no missing items being replaced) or resulted in minimal difference in mean, standard deviation, and standard error between the raw data and cleaned/EM-imputed AOSI data (max absolute difference in means, standard deviation, and standard error between the raw and EM-imputed data was 0.053, 0.007, and 0.007 respectively). For IL-N MSEL data, EM had minimal impact on mean, standard deviation, and standard error between raw data and cleaned/EM-imputed MSEL data (max absolute difference in mean, standard deviation, and standard error between the raw and EM-imputed data was 0.191, 0.083, and 0.010 respectively).

Assessing for Distribution of IL-ASD/IL-N 12-Month AOSI and MSEL Data

To ascertain data distribution prior to any follow-up statistics, Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were conducted in SPSS using the *Explore* command with data factored by 36-month diagnostic outcome (IL-ASD / IL-N). AOSI age of assessment was non-normal for IL-N infants (ps < 0.05 on both the Kolmogorov-Smirnov and Shapiro-Wilks test of normality), and normal for IL-ASD infants on the Kolmogorov-Smirnov test (p = 0.060) but not the Shapiro-Wilk test of normality (p = <0.001). Item-level AOSI data (Items 1-18) was noticeably right-skewed for IL-ASD and IL-N groups (ps < 0.001 across all AOSI items for both Kolmogorov-Smirnov and Shapiro-Wilk tests of normality barring Item 10 for IL-ASD; Kolmogorov-Smirnov statistics were unable to be calculated due to all IL-ASD participants scoring '0'). The AOSI Total Score was normally distributed for IL-ASD infants (p = 0.204) but non-normal for IL-N infants (ps < 0.001). MSEL data was considered non-normal (ps < 0.04) for age of administration, ELC, FM, and RL subscales on the Kolmogorov-Smirnov test of normality for both IL-ASD and IL-N infants, and non-normal for IL-ASD infants on the VR (p = 0.012) but not IL-N (p = 0.052). MSEL age of administration and FM scores were considered non-normal for both IL-ASD and IL-N participants (ps <0.02) on the Shapiro-Wilk test of normality, with ELC and RL scores considered non-normal for IL-N infants (ps < 0.03) and VR for IL-ASD infants (p = 0.009) only. All other MSEL subscales were considered normally distributed according to both Kolmogorov-Smirnov and Shapiro-Wilk tests of normality (ps > 0.05). Results are described in Appendix 2 Table A2.7.

Assessing for Group Differences in 12-month AOSI and MSEL Data in the Independent Validation Set

Due to the non-normal distribution of all AOSI and some MSEL data in the independent testing set, Mann-Whitney U tests were conducted to assess group differences between IL-ASD and IL-N 12-month AOSI and MSEL scoring using Benjamini & Hochberg 1995 corrections. Of the five AOSI items with Mann-Whitney p-values < 0.05 (AOSI Item 1, 3, 4, 17, and Total Score), only one was retained following Benjamini & Hochberg 1995 corrected significance levels of q^* = 0.006 (AOSI Item 3). Of the four MSEL subscales with Mann-Whitney p-values < 0.05 (ELC, VR, FM, and EL), two survived the Benjamini & Hochberg 1995 corrected significance level of q* = 0.006 (ELC and EL). Results are described in Appendix 2 Table A2.8.

Table A2.01 | Participant Demographics of IL-ASD/N Infant Siblings

Characteristic	IL-ASD (n)	IL-N (n)	χ^2	*p*-value
Sex				
Boys	87	175		
Girls	38	165		
Total n	125	340	12.123	<0.001
IL-Sib birth order				
2	73	174		
3	34	129		
4	11	25		
5	4	6		
6	0	2		
7	2	0		
8	0	2		
9	0	2		
12	1	0		
Missing	0	0		
Total n	125	340	15.395	0.052
Total Number of Children in Family				
2	67	159		
3	29	132		
4	9	32		
5	6	10		
6	1	2		
9	2	5		
12	1	0		
Missing	0	0		
Total n	125	340	6.542	0.365
Site Assessed				
Toronto (Site 1)	71	189		
Hamilton (Site 2)	14	47		
Nova Scotia (Site 3)	18	51		
Edmonton (Site 4)	22	53		
Missing	0	0		
Total n	125	340	0.755	0.860
Father's Age at IL-sib's Birth				
20-24	0	1		
25-29	4	20		
30-34	28	83		
35-39	35	108		
40-44	23	56		
45-49	8	10		
50-54	3	4		
55-59	1	1		
Missing	23	57		
Total n	102	283	38.834 [a]	0.129 [a]
Father's Ethnicity				
Aboriginal	1	4		
African	4	4		
Asian	3	8		
Caucasian	88	226		
East Indian	4	15		
Mixed	2	5		
Southeast Asian	3	6		
Missing	20	72		
Total n	102	262	4.273	0.748

Father's Highest Level of Education				
Some junior high school	1	1		
Some high school	4	5		
High school diploma	15	52		
Some college or specialized training	18	45		
College or university graduate	56	123		
Graduate training	20	655		
Missing	11	49		
Total n	114	291	4.526	0.476
Father's Current Occupation				
Farm workers, service works, or not employed	2	5		
Unskilled workers	7	16		
Machine operators or semi-skilled workers	16	23		
Smaller business owners, skilled manual workers	13	48		
Clerical, sales work, small farm, business owner	6	16		
Technicians, semi-professionals, small business owners	18	51		
Managers, minor professionals, farm owners	26	67		
Administrators, lesser professionals, etc.	18	44		
Major professionals	13	50		
Missing	6	20		
Total n	119	320	6.520	0.589
Mother's Age at Childbirth				
20-24	2	5		
25-29	9	34		
30-34	41	116		
35-39	40	95		
40-44	12	30		
45-49	0	2		
Missing	21	58		
Total n	104	282	17.448[a]	0.829[a]
Mother's Ethnicity				
Aboriginal	0	1		
African	5	5		
Asian	5	8		
Caucasian	85	220		
Indonesian	4	15		
Lebanese	0	1		
Middle Eastern	0	1		
Mixed	0	1		
Southeast Asian	1	5		
Missing	20	73		
Total n	105	267		
Mother's Education				
Some junior high school	0	0		
Some high school	2	6		
High school diploma	17	30		
Some college or specialized training	14	32		
College or university graduate	60	152		
Graduate training	16	63		
Missing	16	57		
Total n	109	283	4.099	0.393
Current Occupation		12		
Farm workers, service works, or not employed	28	54		
Unskilled workers	5	12		
Machine operators or semi-skilled workers	5	8		
Smaller business owners, skilled manual workers	5	12		
Clerical, sales workers, small farm, business owner	13	31		
Technicians, semi-professionals, small business owners	14	45		

Managers, minor professionals, farm owners	17	48		
Administrators, lesser professionals, etc.	12	39		
Major professionals	7	34		
Missing	19	57		
Total n	106	283	5.895	0.659
Parent relationship				
Parents are together; only father is working	48	118		
Parents are together; only mother is working	3	4		
Parents are together; both are working	63	166		
Single mother (never married or divorced)	2	2		
Single mother (mother not working; father pays support)	0	3		
Missing	9	47		
Total n	116	293	2.940	0.568
Family Socioeconomic Status				
<20	1	6		
21 to ≤ 35	24	45		
36 to ≤ 50	34	106		
51 to ≤ 66	53	129		
Missing	13	55		
Total n	112	285	33.980	0.420
Father's Age at IL-Sib's Birth	**Mean (SD)**	**Mean (SD)**	**T**	
Father	37.775 (5.697)	36.448 (5.117)	2.112	0.035
Missing	23	57		
Total n	102	283		
Father's Age at IL-Sib's Birth	**Mean (SD)**	**Mean (SD)**	**T**	
Mother	34.673 (4.223)	34.145 (4.435)	1.050	0.294
Missing	21	58		
Total n	104	282		

[a] = these values were calculated based on year-by-year ages (e.g., 25, 26, 27, etc.) and not pooled groupings (e.g., 25-29, 30-34).
T = T-statistic
SD = standard deviation

Table A2.02 | 12-Month AOSI and MSEL Data Normality Tests Factored by IL-ASD / IL-N Grouping

Measure	Outcome	Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	Df	*p*-value	Statistic	Df	*p*-value
Autism Observation Scale for Infants							
Age at assessment	IL-ASD	0.198	125	<0.001	0.806	125	<0.001
	IL-N	0.172	340	<0.001	0.793	340	<0.001
Item 1	IL-ASD	0.508	125	<0.001	0.430	125	<0.001
	IL-N	0.508	340	<0.001	0.415	340	<0.001
Item 2	IL-ASD	0.525	125	<0.001	0.350	125	<0.001
	IL-N	0.529	340	<0.001	0.329	340	<0.001
Item 3	IL-ASD	0.309	125	<0.001	0.771	125	<0.001
	IL-N	0.443	340	<0.001	0.594	340	<0.001
Item 4	IL-ASD	0.390	125	<0.001	0.677	125	<0.001
	IL-N	0.448	340	<0.001	0.578	340	<0.001
Item 5	IL-ASD	0.468	125	<0.001	0.493	125	<0.001
	IL-N	0.505	340	<0.001	0.405	340	<0.001
Item 6	IL-ASD	0.459	125	<0.001	0.555	125	<0.001
	IL-N	0.495	340	<0.001	0.461	340	<0.001
Item 7	IL-ASD	0.219	125	<0.001	0.867	125	<0.001
	IL-N	0.234	340	<0.001	0.821	340	<0.001
Item 8	IL-ASD	0.404	125	<0.001	0.614	125	<0.001
	IL-N	0.493	340	<0.001	0.486	340	<0.001
Item 9	IL-ASD	0.273	125	<0.001	0.786	125	<0.001
	IL-N	0.360	340	<0.001	0.705	340	<0.001
Item 10	IL-ASD	0.523	125	<0.001	0.379	125	<0.001
	IL-N	0.540	340	<0.001	0.168	340	<0.001
Item 11	IL-ASD	0.421	125	<0.001	0.616	125	<0.001
	IL-N	0.471	340	<0.001	0.521	340	<0.001
Item 14	IL-ASD	0.352	125	<0.001	0.710	125	<0.001
	IL-N	0.433	340	<0.001	0.611	340	<0.001
Item 15	IL-ASD	0.466	125	<0.001	0.541	125	<0.001
	IL-N	0.499	340	<0.001	0.472	340	<0.001
Item 16	IL-ASD	0.523	125	<0.001	0.379	125	<0.001
	IL-N	0.524	340	<0.001	0.380	340	<0.001
Item 17	IL-ASD	0.412	125	<0.001	0.607	125	<0.001
	IL-N	0.464	340	<0.001	0.546	340	<0.001
Item 18	IL-ASD	0.498	125	<0.001	0.471	125	<0.001
	IL-N	0.541	340	<0.001	0.215	340	<0.001
Total Score	IL-ASD	0.122	125	<0.001	0.946	125	<0.001
	IL-N	0.132	340	<0.001	0.925	340	<0.001
Mullen Scales of Early Learning							
Age at assessment	IL-ASD	0.217	125	<0.001	0.798	125	<0.001
	IL-N	0.200	340	<0.001	0.778	340	<0.001
ELC	IL-ASD	0.090	125	0.015	0.986	125	0.227
	IL-N	0.087	340	<0.001	0.991	340	0.038
VR	IL-ASD	0.171	125	<0.001	0.961	125	0.001
	IL-N	0.149	340	<0.001	0.968	340	<0.001
FM	IL-ASD	0.204	125	<0.001	0.930	125	<0.001
	IL-N	0.128	340	<0.001	0.970	340	<0.001
RL	IL-ASD	0.179	125	<0.001	0.959	125	0.001
	IL-N	0.155	340	<0.001	0.956	340	<0.001
EL	IL-ASD	0.122	125	<0.001	0.971	125	0.009
	IL-N	0.112	340	<0.001	0.975	340	<0.001

Df = Degrees freedom, EL = Expressive Language, ELC = Early Learning Composite, FM = Fine Motor, RL = Receptive Language, VR = Visual Reception

Table A2.03 | 12-Month AOSI and MSEL Data Normality Tests Factored 80/20 Data Partitions

Measure	Partition	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	Df	*p*-value	Statistic	Df	*p*-value
Autism Observation Scale for Infants							
Age at assessment	Training	0.170	373	0.000	0.807	373	<0.001
	Testing	0.219	92	0.000	0.761	92	<0.001
Item 1	Training	0.512	373	0.000	0.399	373	<0.001
	Testing	0.486	92	0.000	0.490	92	<0.001
Item 2	Training	0.526	373	0.000	0.345	373	<0.001
	Testing	0.532	92	0.000	0.290	92	<0.001
Item 3	Training	0.418	373	0.000	0.637	373	<0.001
	Testing	0.364	92	0.000	0.714	92	<0.001
Item 4	Training	0.429	373	0.000	0.617	373	<0.001
	Testing	0.448	92	0.000	0.565	92	<0.001
Item 5	Training	0.493	373	0.000	0.432	373	<0.001
	Testing	0.499	92	0.000	0.406	92	<0.001
Item 6	Training	0.479	373	0.000	0.508	373	<0.001
	Testing	0.510	92	0.000	0.401	92	<0.001
Item 7	Training	0.232	373	0.000	0.840	373	<0.001
	Testing	0.238	92	0.000	0.826	92	<0.001
Item 8	Training	0.472	373	0.000	0.531	373	<0.001
	Testing	0.462	92	0.000	0.547	92	<0.001
Item 9	Training	0.334	373	0.000	0.737	373	<0.001
	Testing	0.349	92	0.000	0.703	92	<0.001
Item 10	Training	0.540	373	0.000	0.249	373	<0.001
	Testing	0.540	92	0.000	0.206	92	<0.001
Item 11	Training	0.463	373	0.000	0.543	373	<0.001
	Testing	0.437	92	0.000	0.574	92	<0.001
Item 14	Training	0.412	373	0.000	0.642	373	<0.001
	Testing	0.404	92	0.000	0.651	92	<0.001
Item 155	Training	0.489	373	0.000	0.494	373	<0.001
	Testing	0.487	92	0.000	0.491	92	<0.001
Item 16	Training	0.523	373	0.000	0.384	373	<0.001
	Testing	0.527	92	0.000	0.359	92	<0.001
Item 17	Training	0.447	373	0.000	0.570	373	<0.001
	Testing	0.462	92	0.000	0.547	92	<0.001
Item 18	Training	0.534	373	0.000	0.313	373	<0.001
	Testing	0.536	92	0.000	0.292	92	<0.001
Total Score	Training	0.128	373	0.000	0.922	373	<0.001
	Testing	0.173	92	0.000	0.919	92	<0.001
Mullen Scales of Early Learning							
Age at assessment	Training	0.201	373	0.000	0.789	373	<0.001
	Testing	0.233	92	0.000	0.762	92	<0.001
ELC	Training	0.072	373	0.000	0.988	373	0.004
	Testing	0.079	92	.200*	0.971	92	0.038
VR	Training	0.144	373	0.000	0.969	373	<0.001
	Testing	0.109	92	0.009	0.973	92	0.054
FM	Training	0.155	373	0.000	0.958	373	<0.001
	Testing	0.124	92	0.001	0.975	92	0.068
RL	Training	0.136	373	0.000	0.965	373	<0.001
	Testing	0.149	92	0.000	0.958	92	0.005
EL	Training	0.098	373	0.000	0.976	373	<0.001
	Testing	0.125	92	0.001	0.968	92	0.022

[a] = Lilliefors Significance Correction

* = The lower bound of the true significance level

Df = Degree's freedom, EL = Expressive Language, ELC = Early Learning Composite, FM = Fine Motor, RL = Receptive Language, VR = Visual Reception

Table A2.04 | Item-level AOSI Pearson correlations

	AQ1	AQ2	AQ3	AQ4	AQ5	AQ6	AQ7	AQ8	AQ9	AQ10	AQ11	AQ14	AQ15	AQ16	AQ17	AQ18	TS
AQ1	-																
AQ2	-.060	-															
AQ3	-.009	.024	-														
AQ4	.089$^{\alpha}$	.006	.140$^{\beta}$	-													
AQ5	.035	.052	.078$^{\alpha}$	-.025	-												
AQ6	.116$^{\beta}$	.015	.105$^{\alpha}$	.156$^{\beta}$	.197$^{\beta}$	-											
AQ7	.031	.089$^{\alpha}$	.094$^{\alpha}$	.059	.211$^{\beta}$	.242$^{\beta}$	-										
AQ8	.002	.105$^{\alpha}$	.335$^{\beta}$	.188$^{\beta}$	.110$^{\beta}$	.051	.147$^{\beta}$	-									
AQ9	.043	.034	.189$^{\beta}$	.159$^{\beta}$	.224$^{\beta}$	.112$^{\beta}$	.286$^{\beta}$	.408$^{\beta}$	-								
AQ10	.010	.009	.023	-.002	.044	.030	.062	.166$^{\beta}$	.098$^{\alpha}$	-							
AQ11	.127$^{\beta}$	-.082$^{\alpha}$	.170$^{\beta}$	.114$^{\beta}$	.242$^{\beta}$	.233$^{\beta}$	.159$^{\beta}$	.143$^{\beta}$	.311$^{\beta}$	.028	-						
AQ14	.009	.086$^{\alpha}$	.222$^{\beta}$	.124$^{\beta}$	.338$^{\beta}$	.231$^{\beta}$	.343$^{\beta}$	.442$^{\beta}$	.553$^{\beta}$	.184$^{\beta}$	.413$^{\beta}$	-					
AQ15	.141$^{\beta}$	.001	.084$^{\alpha}$	.058	.089$^{\alpha}$	.080$^{\alpha}$	.040	.032	.080$^{\alpha}$	-.030	.112$^{\beta}$	.112$^{\beta}$	-				
AQ16	.046	-.032	.084$^{\alpha}$	.101$^{\alpha}$	.087$^{\alpha}$	.118$^{\beta}$	.052	.112$^{\beta}$	-.004	.111$^{\beta}$	.133$^{\beta}$	.066	.074	-			
AQ17	.081$^{\alpha}$	-.017	.077$^{\alpha}$	.096$^{\alpha}$	.045	.094$^{\alpha}$	.069	.137$^{\beta}$	.099$^{\alpha}$	.115$^{\beta}$	.176$^{\beta}$	.115$^{\beta}$	.010	.073	-		
AQ18	.071	.018	.069	.004	.096$^{\alpha}$	.134$^{\beta}$	.134$^{\beta}$	.170$^{\beta}$	.096$^{\alpha}$	.197$^{\beta}$	.140$^{\beta}$	.140$^{\beta}$	.114$^{\beta}$	.079$^{\alpha}$	.169$^{\beta}$	-	
TS	.219$^{\beta}$	.184$^{\beta}$	.435$^{\beta}$	.352$^{\beta}$	.414$^{\beta}$	.434$^{\beta}$	.516$^{\beta}$	.588$^{\beta}$	.606$^{\beta}$	.224$^{\beta}$	.509$^{\beta}$	.675$^{\beta}$	.257$^{\beta}$	.240$^{\beta}$	.411$^{\beta}$	.377$^{\beta}$	-

AQ = AOSI question #, TS = Total Score

α Correlation is significant at the 0.05 level (1-tailed)

β correlation is significant at the 0.01 level (1-tailed)

Table A2.05 | Principal Parallel Axis Analysis Results of Item-Level AOSI Data

Root	Raw Data Eigenvalues	Means	95% Percentile eigenvalues
1	2.311[α]	0.375	0.481
2	0.650[α]	0.294	0.356
3	0.496[α]	0.238	0.287
4	0.323[α]	0.190	0.232
5	0.215[α]	0.147	0.185
6	0.139	0.109	0.143
7	0.088	0.072	0.103
8	0.039	0.036	0.065
9	0.005	0.002	0.029
10	-0.082	-0.031	-0.004
11	-0.093	-0.064	-0.038
12	-0.138	-0.097	-0.071
13	-0.195	-0.131	-0.104
14	-0.224	-0.166	-0.137
15	-0.241	-0.205	-0.174
16	-0.312	-0.252	-0.214

[α] = greater than the 95% percentile eigenvalues calculated from n=5000 parallel permutations of raw item-level AOSI data

Table A2.06 | Factor Analysis results for item-level AOSI data

AOSI Item		Factor				
		1	2	3	4	5
14	Social interest and affect	0.897	-0.028	0.077	0.112	-0.047
2	Disengagement of attention	0.371	0.034	-0.021	-0.002	0.021
16	Motor behaviour	-0.131	-1.003	0.091	0.022	0.027
1	Visual tracking	0.014	-0.364	-0.070	0.042	0.001
8	Eye contact	0.036	-0.030	0.812	-0.232	-0.142
3	Orients to name	-0.014	0.017	0.432	0.041	0.036
9	Reciprocal social smiling	0.160	0.039	0.376	0.184	0.023
4	Differential response to emotion	0.005	0.023	0.361	0.099	-0.009
6	Imitation	-0.046	-0.002	-0.040	0.621	-0.057
11	Reactivity	0.173	-0.034	0.045	0.444	-0.016
7	Social babbling	0.116	0.004	0.113	0.254	-0.051
15	Transitions	0.013	-0.009	-0.007	0.238	-0.047
5	Anticipatory response	0.029	-0.043	0.068	0.220	0.037
18	Atypical sensory behaviours	-0.088	0.133	-0.048	0.154	-0.738
10	Coordination of eye gaze and action	0.089	-0.154	0.048	-0.108	-0.287
17	Atypical motor behaviours	0.013	-0.047	0.075	0.067	-0.195

Note: Items 12 and 13 have been removed from the AOSI and are thusly not included or reported in any factor analysis results. Extraction Method: Maximum Likelihood. Rotation Method: Oblimin with Kaiser Normalization. [a] Cells highlighted in grey represent factor loading values >|0.250|.

[a] = Rotation converged in 10 iterations.

Table A2.07 | 12-Month AOSI and MSEL Normality Data in CISS-2 IL-Siblings

Measure	Outcome	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	Df	*p*-value	Statistic	Df	*p*-value
Autism Observation Scale for Infants							
Age at assessment	IL-ASD	0.177	23	0.060	0.766	23	<0.001
	IL-N	0.135	67	0.004	0.946	67	0.005
Item 1	IL-ASD	0.423	23	<0.001	0.621	23	<0.001
	IL-N	0.520	67	<0.001	0.353	67	<0.001
Item 2	IL-ASD	0.539	23	<0.001	0.215	23	<0.001
	IL-N	0.530	67	<0.001	0.321	67	<0.001
Item 3	IL-ASD	0.218	23	0.006	0.814	23	<0.001
	IL-N	0.410	67	<0.001	0.649	67	<0.001
Item 4	IL-ASD	0.479	23	<0.001	0.512	23	<0.001
	IL-N	0.530	67	<0.001	0.291	67	<0.001
Item 5	IL-ASD	0.462	23	<0.001	0.541	23	<0.001
	IL-N	0.452	67	<0.001	0.559	67	<0.001
Item 6	IL-ASD	0.474	23	<0.001	0.522	23	<0.001
	IL-N	0.459	67	<0.001	0.554	67	<0.001
Item 7	IL-ASD	0.249	23	<0.001	0.868	23	0.006
	IL-N	0.271	67	<0.001	0.848	67	<0.001
Item 8	IL-ASD	0.479	23	<0.001	0.512	23	<0.001
	IL-N	0.486	67	<0.001	0.500	67	<0.001
Item 9	IL-ASD	0.252	23	<0.001	0.787	23	<0.001
	IL-N	0.330	67	<0.001	0.729	67	<0.001
Item 10	IL-ASD	n/a[β]	n/a[β]	n/a[β]	n/a[β]	n/a[β]	n/a[β]
	IL-N	0.539	67	<0.001	0.163	67	<0.001
Item 11	IL-ASD	0.396	23	<0.001	0.669	23	<0.001
	IL-N	0.480	67	<0.001	0.464	67	<0.001
Item 14	IL-ASD	0.309	23	<0.001	0.733	23	<0.001
	IL-N	0.337	67	<0.001	0.684	67	<0.001
Item 155	IL-ASD	0.489	23	<0.001	0.477	23	<0.001
	IL-N	0.506	67	<0.001	0.446	67	<0.001
Item 16	IL-ASD	0.508	23	<0.001	0.412	23	<0.001
	IL-N	0.533	67	<0.001	0.322	67	<0.001
Item 17	IL-ASD	0.347	23	<0.001	0.639	23	<0.001
	IL-N	0.463	67	<0.001	0.546	67	<0.001
Item 18	IL-ASD	0.517	23	<0.001	0.402	23	<0.001
	IL-N	0.528	67	<0.001	0.352	67	<0.001
Total Score	IL-ASD	0.166	23	0.101	0.943	23	0.204
	IL-N	0.133	67	0.005	0.916	67	<0.001
Mullen Scales of Early Learning							
Age at assessment	IL-ASD	0.192	23	0.027	0.773	23	<0.001
	IL-N	0.131	67	0.006	0.956	67	0.018
ELC	IL-ASD	0.198	23	0.020	0.920	23	0.066
	IL-N	0.162	67	<0.001	0.950	67	0.009
VR	IL-ASD	0.206	23	0.012	0.884	23	0.012
	IL-N	0.108	67	0.052	0.968	67	0.085
FM	IL-ASD	0.187	23	0.036	0.874	23	0.008
	IL-N	0.188	67	<0.001	0.923	67	<0.001
RL	IL-ASD	0.203	23	0.014	0.923	23	0.076
	IL-N	0.159	67	<0.001	0.960	67	0.030
EL	IL-ASD	0.131	23	0.200*	0.958	23	0.416
	IL-N	0.090	67	0.200*	0.979	67	0.315

Df = Degree's freedom, EL = Expressive Language, ELC = Early Learning Composite, FM = Fine Motor, RL = Receptive Language, VR = Visual Reception

[α] = Lilliefors Significance Correction

n/a$^{\beta}$ = statistic was not able to be calculated; all IL-ASD participants scored a '0.'
* = The lower bound of the true significance level

Table A2.08 | CISS-2 IL-ASD vs IL-N Characteristics

Measures	IL-ASD			IL-N			*U*	*Z*-score	*p*-value
	Mean	SD	Mean Rank	Mean	SD	Mean Rank			
Participant Characteristics									
n	23			67					
Gender	16M:7F			31M:36F			$\chi^2 = 3.724$		0.054
Autism Observation Scale for Infants									
Age at assessment	12.347	0.493	41.435	12.363	0.327	46.896	677.000	-0.866	0.387
Item 1 [α]	0.478	0.790	52.348	0.149	0.469	43.149	613.000	-2.313	0.021
Item 2	0.043	0.209	43.848	0.164	0.539	46.067	732.500	-0.757	0.449
Item 3 [αβ]	1.043	0.767	60.391	0.418	0.655	40.388	428.000	-3.562	<0.001[β]
Item 4 [α]	0.217	0.422	50.174	0.090	0.336	43.896	663.000	-1.826	0.068
Item 5	0.391	0.839	45.326	0.328	0.660	45.560	766.500	-0.050	0.960
Item 6	0.391	0.783	45.304	0.358	0.690	45.567	766.000	-0.056	0.955
Item 7	1.522	0.790	51.935	1.254	1.005	43.291	622.500	-1.442	0.149
Item 8	0.435	0.843	45.783	0.418	0.819	45.403	764.000	-0.085	0.932
Item 9	0.826	0.937	50.196	0.582	0.762	43.888	662.500	-1.109	0.268
Item 10	0.000	0.000	44.500	0.030	0.171	45.843	747.500	-0.833	0.405
Item 11	0.478	0.730	51.522	0.224	0.546	43.433	632.000	-1.769	0.077
Item 14	0.565	0.590	46.935	0.507	0.533	45.007	737.500	-0.349	0.727
Item 15	0.261	0.619	46.304	0.164	0.373	45.224	752.000	-0.265	0.791
Item 16	0.174	0.491	47.000	0.090	0.288	44.985	736.000	-0.614	0.539
Item 17 [α]	0.957	1.022	53.261	0.493	0.859	42.836	592.000	-2.054	0.040
Item 18	0.261	0.689	46.370	0.209	0.616	45.201	750.500	-0.340	0.734
Total Score [α]	8.043	4.940	56.152	5.478	3.735	41.843	525.500	-2.276	0.023
Mullen Scale of Early Learning									
Age at assessment	12.369	0.490	43.783	12.348	0.335	46.090	731.000	-0.366	0.715
ELC [αβ]	93.803	10.640	31.370	103.299	11.810	50.351	445.500	-3.051	0.002[β]
Visual Reception [α]	99.087	14.774	35.130	106.269	15.262	49.060	532.000	-2.224	0.026
Fine Motor [α]	88.458	13.952	36.130	96.881	11.483	48.716	555.000	-2.032	0.042
Receptive Language	92.174	16.713	38.804	98.247	15.661	47.799	616.500	-1.443	0.149
Expressive Language [αβ]	92.331	13.286	32.391	102.037	12.328	50.000	469.000	-2.791	0.005[β]

ELC = Early Learning Composite, IL-ASD = Infant siblings diagnosed with autism at 36-months, IL-N = Infant siblings not diagnosed with autism at 36-months, SD = Standard deviation, SS = standard scores

[α] = significantly different based on 2-tailed Mann-Whitney U score

[β] = survived Benjamini & Hochberg 1995 corrected significance levels for multiple comparisons ($q^* = 0.00625$)