Toward Novel Tools for Autism Identification: Fusing Computational and Clinical Expertise

Laura L. Corona[1,2] · Liliana Wagner[1,2] · Joshua Wade[3] · Amy S. Weitlauf[1,2] · Jeffrey Hine[1,2] · Amy Nicholson[1,2,4] · Caitlin Stone[1,2] · Alison Vehorn[1] · Zachary Warren[1,2,4,5]

Abstract
Barriers to identifying autism spectrum disorder (ASD) in young children in a timely manner have led to calls for novel screening and assessment strategies. Combining computational methods with clinical expertise presents an opportunity for identifying patterns within large clinical datasets that can inform new assessment paradigms. The present study describes an analytic approach used to identify key features predictive of ASD in young children, drawn from large amounts of data from comprehensive diagnostic evaluations. A team of expert clinicians used these predictive features to design a set of assessment activities allowing for observation of these core behaviors. The resulting brief assessment underlies several novel approaches to the identification of ASD that are the focus of ongoing research.

Keywords Autism spectrum disorder · Assessment · Young children · Machine learning

Although many caregivers of children with autism spectrum disorder (ASD) report developmental concerns within the first two years of a child's life (Sacrey et al. 2018; Zuckerman et al. 2015), only 44% of children receive a diagnosis before 36 months of age, and the median age of first ASD diagnosis in the United States is after 4 years (Maenner et al. 2020). Unfortunately, these delays create additional stress for families (Oswald et al. 2017) and may prevent children from accessing early intervention services, which are important predictors of child outcomes (Fuller and Kaiser 2019; Landa 2018). Multiple barriers contribute to diagnostic delays, including extensive wait lists (Gordon-Lipkin et al. 2016) and limited access to qualified providers (Bishop-Fitzpatrick

and Kind 2017). These barriers are exacerbated by socioeconomic, geographic, and linguistic disparities (Antezana et al. 2017; Durkin et al. 2017; Khowaja et al. 2015). Together, these challenges highlight the need for novel approaches to the early identification of ASD that meet families' needs and connect children with essential services.

Established best practices in the diagnosis of ASD include a clinical interview with primary caregivers, comprehensive assessment of a child's cognitive or developmental functioning, and observation of a child's play and social interactions using standardized, semi-structured assessments (Huerta and Lord 2012). Though valuable, such evaluations often involve multi-hour testing sessions and/or multiple appointments that can be taxing for children and families, particularly for those with geographic or transportation barriers. Further, there is evidence that some young children can be identified as having ASD based on a briefer evaluation (Juárez et al. 2018; Swanson et al. 2014). A tiered model that streamlines risk classification and early intervention access for those children with clear phenotypic profiles of ASD may, by reducing need for comprehensive testing, simultaneously reduce wait times for those children whose complex presentations warrant additional evaluation (Zwaigenbaum and Warren 2020). At present, however, most phenotypic presentations are funneled into the same model of care, regardless of provider capacity or family preference.

✉ Laura L. Corona
  Laura.L.Corona@vumc.org

1 Vanderbilt Kennedy Center/Treatment and Research Institute for Autism Spectrum Disorders, Vanderbilt University Medical Center, Nashville, TN, USA

2 Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

3 Adaptive Technology Consulting, LLC, Nashville, TN, USA

4 Department of Psychiatry & Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN, USA

5 Department of Special Education, Vanderbilt University, Nashville, TN, USA

Several models have been developed to increase access to evaluation and early intervention services within community settings. These include providing ASD diagnostic training and consultation to pediatric providers (Hine et al. 2020; Keehn et al. 2020; Mazurek et al. 2019), embedding psychologists within pediatric clinics (Hine et al. 2018), and leveraging partnerships with early intervention providers (Juárez et al. 2018; Stainbrook et al. 2019; Yingling 2019). Such models have demonstrated positive outcomes including reduced wait times for families, reduced travel burden, and family satisfaction with alternate diagnostic processes. However, even these models are limited by reliance on multiple providers, utilization of existing tools for ASD assessment which incur training and materials costs, and scheduling or personnel burdens which limit their transportability into other community systems of care.

An alternate approach to the development of novel tools for ASD assessment has been the application of advanced computational strategies, such as machine learning, that attempt to distill extensive assessment measures into smaller sets of questions and behavioral observations that could be used for more efficient assessment (Wall et al. 2012). To date, such approaches have demonstrated limited clinical impact and have been thoughtfully critiqued (Bone et al. 2015). In particular, identifying a limited set of behavioral codes with strong predictive validity does not account for the evaluation processes, clinical judgment, and expertise of well-trained providers that ultimately result in the assignment of these codes—and the further assignment of an ASD diagnosis or risk classification. While applying advanced computational strategies may distill large amounts of data into key observations, a machine learning strategy and analysis alone does not result in a meaningful methodology for abbreviated ASD assessment.

Recognizing both the limitations of current tools for ASD evaluation and the shortcomings of prior work utilizing machine learning, the goal of the present work was to develop a brief assessment tool for ASD symptoms in toddlers by fusing computational and clinical expertise that can then be adapted for use across formats, settings, and providers. While a machine learning approach in isolation is not a sufficient strategy for realizing a viable, stand-alone assessment tool, machine learning represents an opportunity to elucidate patterns in clinical assessments and clinical decision-making in a way that can inform the development of novel tools (Sarkar et al. 2018). Below, we describe a computational approach to the identification of key behaviors to inform diagnostic tool development based on available clinical registry datasets, followed by the translation of these predictive models to guidelines for clinical observations of child behavior. The resulting tool, a basic framework for symptom identification, has subsequently been adapted into assessment platforms for use in telemedicine-based evaluation (Corona et al. 2020), intelligent applications for screening (Adiani et al. 2019; R43 MH115528, R44 MH115528), and enhanced training protocols for medical residents (Hine et al. 2019).

Methods

Participants

The analyses below were completed using a clinical research database housed within a university medical center. This database includes phenotypic data for individuals with and without ASD at the time of diagnosis as evaluated by a team of over 20 research-reliable psychological providers (i.e., providers with certified expertise in standardized ADOS-2 administration) across autism-focused research studies and outpatient clinics. Within the targeted age range of this work, the database included scores from 737 toddlers (77% male, 23% female) between the ages of 14–33 months (M = 25.7, SD = 3.7) whose families had provided consent for inclusion at the time of a diagnostic evaluation. Included in this database were toddlers' scores from diagnostic evaluations using the ADOS-2 (Lord et al. 2012; Luyster et al. 2009), the Mullen Scales of Early Learning (MSEL; Mullen 1995), and the Vineland Adaptive Behavior Scales, Second Edition-Interview Form (Vineland-II; Sparrow et al. 2005), as well as the clinician diagnostic impression from that visit (ASD, global developmental delay, and so on). Data span the years 2012–2016, with all participants administered the Toddler Module of the ADOS-2 (including two children ages 30–32 months who are retained here to reflect the true analyses underlying algorithmic creation). Within this sample, 70% of toddlers were classified as having ASD and 30% were classified as not meeting criteria for ASD (see Table 1). Of participants not diagnosed with ASD, approximately 62% received a diagnosis of global developmental delay, 30% received other unspecified diagnoses (such as language delay or behavioral disorders), and 8% received no diagnosis.

Data Analytic Procedure

Analytic Approach

Machine Learning (ML), a branch of Artificial Intelligence, offers a powerful means by which to infer patterns within datasets (Bishop 2006). *Feature selection* techniques in particular can be used to reveal components of a dataset that most effectively distinguish between classes of data. In the current work, ML techniques were used to carry out an exploratory analysis of behavioral assessment variables (i.e., individual items from each of the assessment instruments; henceforth *features*) with the aim of

Table 1 Participant demographics and scores from diagnostic evaluations

	ASD (n=515)	Non-ASD (n=222)
Toddlers		
Age in years (*m* [*SD*])	2.17 [0.29]	2.07 [0.33]
Male	79.0%	73.4%
Female	21.0%	26.6%
ADOS-2 Total Score (*m* [*SD*])	21.32[4.62]	6.95[5.05]
MSEL-ELC (*m* [*SD*])	56.17 [10.99]	69.94 [14.99]
Visual Reception (*T*)	27.60 [9.66]	35.77 [10.87]
Fine Motor (*T*)	27.02 [9.26]	34.19 [11.21]
Receptive Language (*T*)	22.06 [6.62]	31.64 [11.39]
Expressive Language (*T*)	22.84 [5.81]	31.78 [9.74]
VABS-2-ABC (*m* [*SD*])	71.17 [16.10]	79.99 [9.89]
Communication (SS)	66.71 [9.70]	80.44 [12.19]
Daily Living Skills (SS)	72.61 [10.83]	83.08 [12.99]
Socialization (SS)	71.34 [8.17]	81.83 [10.38]
Motor Skills (SS)	82.80 [9.10]	85.20 [11.05]

ASD autism spectrum disorder, *DD* developmental delay, *TD* typically developing, *MSEL-ELC* mullen scales of early learning, early learning composite standard score, *T* T-score, *VABS-2-ABC* vineland adaptive behavior scales-second edition, adaptive behavior composite, *SS* standard score

identifying the most discriminating classes of features (i.e., those features best differentiating ASD from non-ASD cases). A ML approach, as opposed to a statistical approach (e.g., exploratory factor analysis Furr 2017) was used (1) because our goal was not to identify specific discriminating codes of established instruments but instead to reveal broader discriminatory patterns in behavioral observations, and (2) because of the overtly ML-oriented nature of the dataset—specifically, the structured dataset was ideally suited for supervised classification methods given its binary labeling (ASD, non-ASD), rich feature set, and relatively large size.

Feature Space Exploration

Feature selection techniques, and *feature engineering* more broadly (Géron 2019), were used in the current work to identify a minimal subset of features that would yield clinically acceptable levels of model accuracy, sensitivity, and specificity. To achieve this goal, we applied both established and novel feature engineering methods with the aim of identifying the optimal feature set for use in model development. Table 2 summarizes and compares the feature engineering models applied to the data.

Three established methods—χ^2 goodness of fit, information gain, and Pearson correlation—were used to rank features according to their ability to reliably predict the class label (i.e., ASD or non-ASD) using the ML toolkits scikit-learn version 0.20.3 (Pedregosa et al. 2011) and WEKA version 3.8.4 (Hall et al. 2009).

A fourth method, developed by the data analytic team, evaluated the predictive utility of aggregations of features by comparing the central tendencies of groupwise clusters within the clinical dataset. That is, rather than employing the three methods described above, we extracted new features based on the distance, in feature space, between an individual and a group (see Fig. 1). Here, "groups" refer to the children with a confirmed ASD diagnosis and the children who did not meet criteria for diagnosis of ASD. We explored the space of all possible two- and three-component feature vectors, in which the vectors were comprised of unique combinations of the ADOS-2 feature set. For example, a two-component feature vector $ASD_{a2,b1}$ represents the central tendency, or centroid, composed of features a2 and b1 within the ASD sample. Similarly, a three-component feature vector $NonASD_{a2,b1,b5}$ represents the centroid composed of features a2, b1, and b5 within the non-ASD sample. Only two- and three-component feature vector spaces were explored because distance metrics become increasingly unreliable in higher-dimensions (Aggarwal et al. 2001). Moreover, the chosen spaces were amenable to brute force search, resulting in rapid computation and evaluation.

Table 2 Comparison of model performance on holdout set

#	Model	Accuracy (%)	F-Score	Sensitivity	Specificity	UAR
1	ADOS-2 total score	94	0.959	0.987	0.833	0.910
2	Highest ranked single feature: a2	89	0.926	0.981	0.645	0.813
3	10 Highest ranked (χ^2)	86	0.902	0.902	0.729	0.815
4	10 Highest ranked (Information Gain)	85	0.901	0.920	0.661	0.791
5	10 Highest ranked (Pearson Correlation)	84	0.888	0.871	0.746	0.808
6	Top performing aggregated feature[a]	87	0.914	0.945	0.661	0.803
7	Top performing cluster + a3, d1, d2, and d5	87	0.914	0.901	0.780	0.844

UAR Unweighted Average Recall (Bone et al. 2015)

[a]Top performing aggregated feature: distance between the individual and the non-ASD centroid, which was composed of the means of ADOS-2 codes a2, b1, and b5 for the non-ASD sample
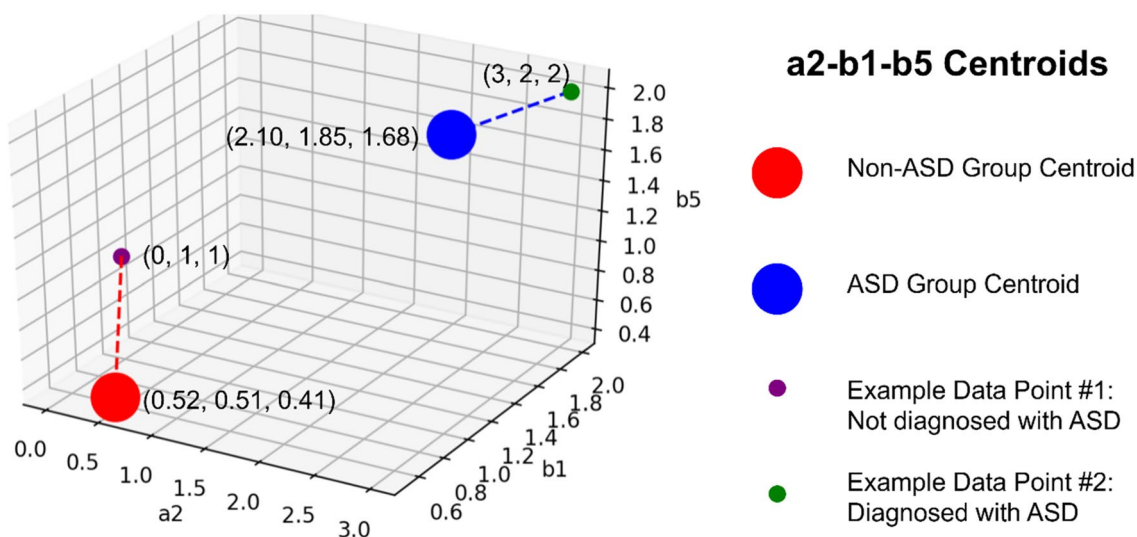
Fig. 1 The top-performing centroids resulting from the data analytic team's novel feature selection method were the three-component centroids composed of ADOS-2 codes a2, b1, and b5. The figure depicts the distance between the identified centroids and new data points, represented by the dashed lines. In much the same way that a *k*-nearest neighbors classifier predicts class association, our method defines a feature based on a measure of proximity to clusters that may reveal class association

Model Selection

The set of 737 samples was randomly partitioned using a 70–30 split into a training set (N = 515) and a holdout set for validation (N = 222). A train-test split approach involves partitioning a dataset into two sets of different sizes—often 70–30 or 60–40 partitions—where the larger partition is typically used for model training and the smaller partition is held out exclusively for testing. This method is often used to account for the problem of overfitting in which a predictive model demonstrates excellent classification performance on a training dataset at the expense of generalizability to future or unseen data (Kuhn and Johnson 2013). In a train-test split approach, the performance of the model on the holdout set is used to determine the model's overall reliability. In the current work, the training set was used to train and compare classifiers using the four feature engineering methods described above, while the holdout set was used to evaluate the final performance of the various models. Within the scope of the training set, cross-validation was used to gauge the preliminary performance of the classifier before final evaluation on the holdout set. Cross-validation involves iteratively dividing a dataset into *k* segments, training a model on *k* − 1 segments, and then evaluating the model on the *k*th segment (Bishop 2006). The accuracy of the model on each test segment is then averaged to yield a measure of performance that is expected to be robust to variations in the data. Consistent with typical practice, the value of *k* was set to 10 in the performed analyses (Géron 2019).

Results

Multiple predictive models were trained based on the four feature selection methods described above. All features identified through the feature selection methods were individual items from the ADOS-2. A decision tree classifier was selected for model training. Decision trees are widely used in practice and perform well on certain types of data, particularly data that fit neatly into cuboid regions when viewed from the perspective of plotting data points in multidimensional space (Bishop 2006). Table 2 compares the performance of seven models evaluated on the holdout set.

Model 1, which is included as a basis of comparison for each of the other models, is simply the ADOS-2 total score, representing a trivial classifier with only one feature. Model 2, another trivial classifier, includes only one feature, frequency of spontaneous vocalizations directed to others, which emerged as one of the highest ranked features across all of the ranking methods. It is included to demonstrate the predictive accuracy of a single feature; however, the specificity of this model was inadequate (see Table 2).

Models 3–5 include the 10 highest ranked features using the χ^2, Information Gain, and Pearson Correlation ranking methods, respectively. Table 3 shows the 10 highest-ranked features for each of these three feature selection methods. Features identified as among the most predictive of ASD diagnosis across Models 3–5 include ADOS-2 codes related to: frequency of child vocalizations to others, integration of eye contact with other behaviors in the context of social overtures, showing behaviors, and the overall number of social overtures to the examiner.

Table 3 The 10 highest-ranked features for each of three feature extraction methods

χ^2	Information Gain	Pearson Correlation
Frequency of undirected vocalizations (a9)	Integration of gaze and other behaviors during social overtures (b5)*	Frequency of spontaneous vocalization directed to others (a2)*
Frequency of spontaneous vocalization directed to others (a2)*	Amount of social overtures/maintenance of attention to examiner (b16a)*	Facial expressions directed to others (b4)
Pointing (a7)	Frequency of spontaneous vocalization directed to others (a2)*	Quality of social overtures (b15)
Intonation of vocalizations or verbalizations (a3)	Unusual eye contact (b1)	Pointing (a7)
Showing (b12)*	Quality of social overtures (b15)	Showing (b12)*
Amount of social overtures/maintenance of attention to examiner (b16a)*	Intonation of vocalizations or verbalizations (a3)	Integration of gaze and other behaviors during social overtures (b5)*
Requesting (b9)	Facial expressions directed to others (b4)	Unusual eye contact (b1)
Ignore (b8)	Overall quality of rapport (b18)	Overall quality of rapport (b18)
Spontaneous initiation of joint attention (b13)	Frequency of undirected vocalizations (a9)	Requesting (b9)
Integration of gaze and other behaviors during social overtures (b5)*	Showing (b12)*	Amount of social overtures/maintenance of attention to examiner (b16a)*

*The features common to all three feature selection methods were a2, b5, b12, and b16a. Alphanumeric notations after each feature refer to ADOS-2 Toddler Module codes. (Two children over the age of 30 months were administered the Toddler Module.)

Model 6 represents the highest performing aggregated feature from the previously described clustering method. The predictive features identified by this model focused on items related to a child's use of eye contact, vocalizations directed to other people, and integration of eye contact with other forms of communication, such as sounds and gestures. Finally, Model 7 is an extension of Model 6 that includes four additional features based on a secondary feature selection analysis, conducted using an embedded feature ranking algorithm in WEKA. This method first assesses subsets of features and then selects features that are highly correlated with the class label while maintaining low correlation with one another (Hall 1999). The additional features identified in this final model included items focused on the intonation of a child's vocalization, atypical sensory interests, stereotyped hand movements, and repetitive or stereotyped interests.

The assessment of model performance focused on five performance metrics, including accuracy, sensitivity, specificity, F-score, and unweighted average recall (UAR). Accuracy, F-score, and sensitivity (also known as "recall") are commonly reported metrics in the ML literature, while specificity is often reported in the context of diagnostic testing. UAR has been suggested as a preferred performance metric for model assessment related to diagnostic testing, especially in the presence of unbalanced data (Bone et al. 2015). As such, UAR was selected as the preferred metric for comparing model performance. Based on this criterion, Model 7 was the highest performing non-trivial classifier with a UAR of 0.844. When applied to the test sample, Model 7 achieved a sensitivity of 0.90 and a specificity of 0.78.

Following feature selection and model comparison, a design team of six clinical experts in ASD reviewed the features identified, each of which represented a behavioral code from a standardized instrument included in the database, and then aligned each feature (and associated behavioral descriptor) with DSM-5 diagnostic criteria for ASD in young children to determine if symptoms would be captured across each core diagnostic area. Features identified by the final model (Model 7) included child behaviors related to: vocalizations directed to other people, intonation of vocalizations, overall use of eye contact, integration of eye contact with other social and communicative behaviors, and restricted or repetitive behaviors including sensory interests, repetitive motor behaviors, and repetitive play or interests.

Through a process of collaborative consensus that included independent generation of content, cross-team review for shared and discrepant material, and preliminary finalization of items deemed most clinically representative, the design team developed core behavioral descriptors reflecting those features most predictive of diagnosis in our sample (see Table 4). These descriptors were then reviewed by an internal group of 16 behavioral providers (licensed clinical psychologists, licensed senior psychological examiners, developmental-behavioral pediatricians, and postdoctoral fellows) with varying levels of ASD expertise who read the text and replied with suggested edits or clarifications in order to simplify the language for a broader audience. After the descriptors were finalized, the design team operationalized these behaviors using anchors within a Likert-style scale. For each item, a rating of 1 indicates that the ASD-related symptom is not present, a rating of 2 indicates that

Table 4 Translation of predictive features identified through computational approach to underlying constructs and behavioral ratings

Behavioral construct	Behavioral ratings		
	1 (Symptom not present)	2 (Symptom present but at subclinical levels)	3 (Symptom clearly consistent with autism spectrum)
Socially directed speech and sounds	Child often uses words or other vocalizations for a variety of social purposes (e.g. requesting, protesting, directing attention, sharing enjoyment)	Inconsistent socially directed speech	Most of the child's sounds are self-directed. May make atypical non-word noises (e.g., "digga digga")
Frequent and flexible eye contact	Child frequently makes eye contact with others and across a variety of activities	Child's eye contact seems inconsistent. Gaze seems less flexible and harder to catch than expected	Child infrequently makes eye contact. Might only make eye contact during one activity (e.g., asking for help)
Unusual vocalizations	No unusual qualities of speech/language observed. Most of child's speech is appropriate for the child's age and developmental level	Speech is not clearly unusual, but there are some differences (e.g., volume, slight repetitive quality of speech/language, unclear echoing, some occasional sounds that are unusual)	Child produces unusual jargon, sounds, or speech/language (e.g., undirected jargoning, speech of peculiar intonation, unusual sounds, repetitive vocalizations, echoing or repetitive speech/language)
Unusual or repetitive play	Child plays with toys in appropriate ways (uses toys as expected)	Child's play is not clearly unusual, but child is strongly focused on some toys, routines, or activities. May sometimes be hard to shift child's attention to something new	Child shows clearly repetitive or unusual play, such as repeatedly pushing buttons, watching how objects move, lining things up, or scrambling/dropping toys
Unusual or repetitive body movements	No unusual or repetitive body movements seen	Unclear unusual/repetitive body movements. Some repetitive jumping or very brief posturing of fingers, hands, or arms that is not clearly atypical	Child clearly shows unusual or repetitive (e.g., hand-flapping, posturing or tensing upper body, toe-walking, facial grimacing, hand/finger mannerisms) repetitive running/walking/spinning/jumping
Combines gestures, eye contact, and speech/vocalization	Child frequently points and uses other gestures to communicate. Child's gestures are usually combined with vocalizations and eye contact	Child may sometimes point or use other gestures, but less than expected. Child does not always look at you or make a sound when gesturing	Child does not usually gesture to communicate. May sometimes reach or point, but does not usually combine these with eye gaze or sounds. May move your hand or push on your body to get help
Unusual sensory exploration or reaction	No unusual sensory behavior observed	Unclear sensory exploration or reaction. May have a brief response to a sound, smell, or how something feels or moves	Child shows sensory differences. May closely inspect objects, overreact to sounds, show intense interest or dislike to textures (e.g., touching, licking, biting, refusing to touch specific toys), or clear self-injurious behavior

the symptom is present but at subclinical levels, and a rating of 3 indicates that the symptom is present and clearly consistent with ASD. Similar to the development of the behavioral descriptors, the Likert anchors were reviewed by a secondary team of non-experts to improve clarity regarding targets for behavioral observation.

After finalizing behavioral descriptions and anchors for Likert ratings, the team developed an assessment process designed to elicit observations tied to these seven key behaviors. The process was designed to (1) be administered within 20 min of interaction, in order to maximize transportability into community practice settings, (2) employ inexpensive and widely available materials to facilitate use and access to the tool with low cost burden, and (3) provide understandable instructions for community providers. Activities include opportunities for free play and partnered play, as well as social presses to provide opportunities for making requests, sharing enjoyment, and directing attention (see Table 5), all of which give children opportunities to display the communication, social interaction, and independent play skills related to the core discriminatory behaviors identified through the ML procedures described above. Together, these brief administration activities and providers' ratings of children's behaviors during the activities make up a novel tool for identification of ASD symptoms.

Discussion

This work describes a computationally and clinically informed development process aimed at creating innovative measures for accurate, efficient identification of ASD in young children across a variety of clinical practice settings. This approach applied complex feature engineering to a rich phenotypic data set of toddlers with ASD and other complex developmental concerns in order to elucidate potentially valuable targets for clinical assessment and observation that could be folded into scoring systems explicitly designed for use in varied settings. The result of this translational process is a novel, brief assessment tool that has the potential to provide clinical information regarding the presence of ASD symptoms in young children.

The development process described above represents an extension of past work that has focused solely on the identification of key items within more thorough assessment measures (Wall et al. 2012). By combining a computational approach with clinical expertise, this process identified elements within a comprehensive assessment process that are predictive of ASD diagnosis and then translated these elements into key, underlying behavioral constructs. Our approach further moved beyond past work by proposing a set of assessment activities designed to elicit behaviors of clinical interest, as well as an observation and scoring system to organize and quantify clinical impressions.

It is important to recognize that the features computationally identified as most predictive of ASD diagnosis were distilled from behavioral codes assigned as part of a longer, standardized assessment and scoring procedure. These behavioral codes in isolation do not represent an appropriate, stand-alone estimate of risk for or presence of an ASD diagnosis (Bone et al. 2015). Instead, these codes represent key clinical features of concern—particularly, features related to a child's challenges with aspects of social communication (e.g., use and integration of verbal and nonverbal communication) and the presence of restricted, repetitive behaviors—that are characteristic of ASD. A brief assessment approach designed around these key clinical features, such as that described here, holds promise for identifying

Table 5 Administration activities

Activity	Administration guidelines
Child-directed play	Provider introduces 3–5 age appropriate toys and observes the child's independent play
Joint play	Provider joins child's play and initiates interactive play attempts
Calling name	Provider calls the child's name and observes response
Directing child's attention	Provider gets the child's attention, then shifts gaze and points to an object across the room while verbally directing the child to look
Familiar play routine	Provider initiates a social routine with the child (e.g., peekaboo)
Ready-set-go routine	Provider blows up a balloon, holds it in the air, and says Ready, set, go! before releasing balloon. Provider repeats several times, pausing to observe child's reaction and allow for child to respond to continue the activity
Requesting	
Bubbles	Provider blows bubbles for the child, then pauses while holding the bubbles out of reach. Provider observes whether and how child requests activity to continue
Snack	Provider offers child a snack in a container that child cannot open independently. Provider observes whether and how child requests snack or assistance
Independent play with ignoring	Provider re-presents independent play toys and removes attention from child

clear symptoms of ASD, within a short time period, in a variety of clinical settings.

The assessment activities and rating procedures developed through this work present a preliminary model and a starting point for further tool development, refinement, and investigation. Although based upon the computational selection of features most predictive of ASD within our clinical database, the activities eventually chosen to elicit these features were derived from clinical expertise with the goal of creating activities that could be easily implemented in community care settings with minimal time or financial burdens. The completion of predictive studies is an essential next step to understanding the performance of novel tools developed through this method. Translating machine learning methods into clinically meaningful assessment practices and tools is a promising approach; however, it is also critical that novel assessment tools and procedures undergo rigorous evaluation. In our ongoing work, this tool underlies four different assessment instruments under investigation. In one model, the administration instructions and behavioral ratings have been built into an interactive app designed to guide non-expert pediatric providers through the use of the tool (Adiani et al. 2019). In a second model, the tool is used to guide parents through administration activities via telehealth, while a clinician provides coaching, observes the administration, and assigns behavioral ratings (Corona et al. 2020). In a third model, the tool is integrated into an approach to meet the identified needs of pediatric residents regarding ASD evaluation (Hine et al. 2019). Finally, in a fourth model, the tool was modified slightly for use within telehealth-to-home evaluations, in direct response to disruptions in care caused by the COVID-19 pandemic (TELE-ASD-PEDS; Wagner et al. 2020). Within all of these models, evaluation of the clinical utility, user acceptability, and psychometric properties of these assessment measures is ongoing.

By definition, the preliminary nature of this work yields several limitations and essential future directions. Analysis of data regarding the performance of this tool in accurately identifying children with concern for ASD is in progress, but ongoing. The clinical dataset upon which this tool is based does not provide information regarding child race and ethnicity, medical complexity, or family variables, all of which are factors important to understanding how and for whom this assessment tool works best. Additionally, because of the point-in-time nature of the clinical assessment process, information on diagnostic stability and accuracy for children within our preliminary dataset is not available, and it is unknown how this instrument would function across a range of phenotypic profiles. Future work should also explore a range of scoring methodologies; although a Likert-style scale was chosen for ease of clinical use and behavioral anchoring, other methods may provide a more fine-grained phenotypic profile across a range of risk levels. Finally, given that the tool has been adapted for applications across provider type and setting, ongoing work is necessary to examine potential differences in scoring cut-offs and tool functionality across these groups. Addressing these limitations in ongoing work by our group and others will be essential for creating models of care that accurately identify ASD-related concerns in young children while meeting families' diverse needs (Zwaigenbaum and Warren 2020). In addition, continuing to apply machine learning approaches as our datasets grow can continue to help clinicians identify and interpret these patterns on a large scale.

In conclusion, the present work describes the development of one preliminary approach to identifying behaviors strongly indicative of ASD in young children in a brief amount of time, using readily available materials and play-based assessment activities. By bringing together computational approaches and clinical knowledge, this work has identified several key child behaviors predictive of ASD and developed procedures for eliciting and observing these behaviors. Ultimately, an ASD diagnosis comes not from a specific assessment tool, but from a provider trained to use assessment tools, observations, and other available clinical information to make clinical decisions (De Marchena and Miller 2017; Sheldrick et al. 2019). This work attempts to provide a streamlined way of helping providers make key behavioral observations and organize these observations to inform diagnostic decision-making. In doing so, this tool may contribute to our collective ability to serve patients and families in timely, informed ways.

Compliance with Ethical Standards

References

Adiani, D., Schmidt, M., Wade, J., Swanson, A. R., Weitlauf, A., Warren, Z., & Sarkar, N. (2019, July). Usability enhancement and functional extension of a digital tool for rapid assessment of risk for autism spectrum disorders in toddlers based on pilot test and interview data. In *International conference on human–computer interaction* (pp. 13–22). Cham: Springer. https://doi.org/10.1007/978-3-030-23563-5_2.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001, January). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420–434). Berlin: Springer.

Antezana, L., Scarpa, A., Valdespino, A., Albright, J., & Richey, J. A. (2017). Rural trends in diagnosis and services for autism spectrum disorder. *Frontiers in Psychology, 8,* 590. https://doi.org/10.3389/fpsyg.2017.00590.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Bishop-Fitzpatrick, L., & Kind, A. J. (2017). A scoping review of health disparities in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 47*(11), 3380–3391. https://doi.org/10.1007/s10803-017-3251-9.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., & Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders, 45*(5), 1121–1136.

Corona, L. L., Weitlauf, A. S., Hine, J., Berman, A., Miceli, A., Nicholson, A., et al. (2020). Parent perceptions of caregiver-mediated telemedicine tools for assessing autism risk in toddlers. *Journal of Autism and Developmental Disorders.* https://doi.org/10.1007/s10803-020-04554-9.

De Marchena, A., & Miller, J. (2017). "Frank" presentations as a novel research construct and element of diagnostic decision-making in autism spectrum disorder. *Autism Research, 10*(4), 653–662. https://doi.org/10.1002/aur.1706.

Durkin, M. S., Maenner, M. J., Baio, J., Christensen, D., Daniels, J., Fitzgerald, R., et al. (2017). Autism spectrum disorder among US children (2002–2010): Socioeconomic, racial, and ethnic disparities. *American Journal of Public Health, 107*(11), 1818–1826. https://doi.org/10.2105/AJPH.2017.304032.

Fuller, E. A., & Kaiser, A. P. (2019). The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis. *Journal of Autism and Developmental Disorders.* https://doi.org/10.1007/s10803-019-03927-z.

Furr, R. M. (2017). *Psychometrics: an introduction.* Thousand Oaks, CA: Sage Publications.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* Sebastopol, CA: O'Reilly Media.

Gordon-Lipkin, E., Foster, J., & Peacock, G. (2016). Whittling down the wait time: Exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. *Pediatric Clinics of North America, 63*(5), 851–859. https://doi.org/10.1016/j.pcl.2016.06.007.

Hall, M. A. (1999). Correlation-based feature selection for machine learning. Thesis, The University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Hine, J. F., Allin, J., Allman, A., Black, M., Browning, B., Ramsey, B., et al. (2020). Increasing access to autism spectrum disorder diagnostic consultation in rural and underserved communities: Streamlined evaluation within primary care. *Journal of Developmental and Behavioral Pediatrics, 41*(1), 16–22. https://doi.org/10.1097/DBP.0000000000000727.

Hine, J. F., Dubin, A., Foster, T., Goode, R., Swanson, A., & Warren, Z. E. (2019). Need for pediatric resident training in autism spectrum disorder: Preparation for primary care. *Psychological Disorders and Research.* https://doi.org/10.31487/j.pdr.2019.04.01.

Hine, J. F., Herrington, C. G., Rothman, A. M., Mace, R. L., Patterson, B. L., Carlson, K. L., & Warren, Z. E. (2018). Embedding autism spectrum disorder diagnosis within the medical home: Decreasing wait times through streamlined assessment. *Journal of Autism and Developmental Disorders, 48*(8), 2846–2853. https://doi.org/10.1007/s10803-018-3548-3.

Huerta, M., & Lord, C. (2012). Diagnostic evaluation of autism spectrum disorders. *Pediatric Clinics of North America, 59*(1), 103. https://doi.org/10.1016/j.pcl.2011.10.018.

Juárez, A. P., Weitlauf, A. S., Nicholson, A., Pasternak, A., Broderick, N., Hine, J., et al. (2018). Early identification of ASD through telemedicine: Potential value for underserved populations. *Journal of Autism and Developmental Disorders, 48*(8), 2601–2610. https://doi.org/10.1007/s10803-018-3524-y.

Keehn, R. M., Ciccarelli, M., Szczepaniak, D., Tomlin, A., Lock, T., & Swigonski, N. (2020). A statewide tiered system for screening and diagnosis of autism spectrum disorder. *Pediatrics.* https://doi.org/10.1542/peds.2019-3876.

Khowaja, M. K., Hazzard, A. P., & Robins, D. L. (2015). Sociodemographic barriers to early detection of autism: Screening and evaluation using the M-CHAT, M-CHAT-R, and follow-up. *Journal of Autism and Developmental Disorders, 45*(6), 1797–1808. https://doi.org/10.1007/s10803-014-2339-8.

Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. In *Applied predictive modeling* (pp. 61–92). New York: Springer.

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. https://doi.org/10.1080/09540261.2018.1432574.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule, second edition (ADOS-2) manual (part 1): Modules 1–4.* Torrance, CA: Western Psychological Services.

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., et al. (2009). The Autism Diagnostic Observation Schedule—Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(9), 1305–1320. https://doi.org/10.1007/s10803-009-0746-z.

Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., et al. (2020). Prevalence of autism spectrum disorder among children aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016.

Morbidity and Mortality Weekly Report: Surveillance Summaries, 69(4), 1–12. https://doi.org/10.15585/mmwr.ss6904a1.

Mazurek, M. O., Curran, A., Burnette, C., & Sohl, K. (2019). ECHO Autism STAT: Accelerating early access to autism diagnosis. *Journal of Autism and Developmental Disorders, 49*(1), 127–137. https://doi.org/10.1007/s10803-018-3696-5.

Mullen, E. M. (1995). *Mullen Scales of Early Learning*. Circle Pines, MN: American Guidance Service.

Oswald, D. P., Haworth, S. M., Mackenzie, B. K., & Willis, J. H. (2017). Parental report of the diagnostic process and outcome: ASD compared with other developmental disabilities. *Focus on Autism and Other Developmental Disabilities, 32*(2), 152–160. https://doi.org/10.1177/1088357615587500.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12,* 2825–2830.

Sacrey, L. A. R., Zwaigenbaum, L., Bryson, S., Brian, J., Smith, I. M., Roberts, W., et al. (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. https://doi.org/10.1002/aur.1920.

Sarkar, A., Wade, J., Swanson, A., Weitlauf, A., Warren, Z., & Sarkar, N. (2018, July). A data-driven mobile application for efficient, engaging, and accurate screening of ASD in toddlers. In *International conference on universal access in human–computer interaction* (pp. 560–570). Cham: Springer. https://doi.org/10.1007/978-3-319-92049-8_41.

Sheldrick, R. C., Frenette, E., Vera, J. D., Mackie, T. I., Martinez-Pedraza, F., Hoch, N., et al. (2019). What drives detection and diagnosis of autism spectrum disorder? Looking under the hood of a multi-stage screening process in early intervention. *Journal of Autism and Developmental Disorders, 49*(6), 2304–2319. https://doi.org/10.1007/s10803-019-03913-5.

Sparrow, S. D., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland-II Adaptive Behavior Scales: Survey Forms Manual*. Circle Pines, MN: American Guidance Service.

Stainbrook, J. A., Weitlauf, A. S., Juárez, A. P., Taylor, J. L., Hine, J., Broderick, N., et al. (2019). Measuring the service system impact

of a novel telediagnostic service program for young children with autism spectrum disorder. *Autism, 23*(4), 1051–1056. https://doi.org/10.1177/1362361318787797.

Swanson, A. R., Warren, Z. E., Stone, W. L., Vehorn, A. C., Dohrmann, E., & Humberd, Q. (2014). The diagnosis of autism in community pediatric settings: Does advanced training facilitate practice change? *Autism, 18*(5), 555–561. https://doi.org/10.1177/1362361313481507.

Wagner, L., Corona, L. L., Weitlauf, A. S., Marsh, K. L., Bermn, A. F., Broderick, N. A., et al. (2020). Use of the TELE-ASD-PEDS for autism evaluations in response to COVID-19: Preliminary outcomes and clinician acceptability. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-020-04767-y.

Wall, D. P., Kosmicki, J., Deluca, T. F., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry, 2*(4), e100. https://doi.org/10.1038/tp.2012.10.

Yingling, M. E. (2019). Participation in Part C early intervention: One key to an earlier diagnosis of autism spectrum disorder? *The Journal of Pediatrics, 215,* 238–243. https://doi.org/10.1016/j.jpeds.2019.06.034.

Zuckerman, K. E., Lindly, O. J., & Sinche, B. K. (2015). Parental concerns, provider response, and timeliness of autism spectrum disorder diagnosis. *The Journal of Pediatrics, 166*(6), 1431–1439. https://doi.org/10.1016/j.jpeds.2015.03.007.

Zwaigenbaum, L., & Warren, Z. (2020). Commentary: Embracing innovation is necessary to improve assessment and care for individuals with ASD: A reflection on Kanne and Bishop (2020). *Journal of Child Psychology and Psychiatry*. https://doi.org/10.1111/jcpp.13271.