Development and Validation of a Streamlined Autism Case Confirmation Approach for Use in Epidemiologic Risk Factor Research in Prospective Cohorts

Craig J. Newschaffer, Emily Schriver, Lindsay Berrigan, Rebecca Landa, Wendy L. Stone, Somer Bishop, Diane Burkom, Anne Golden, Lisa Ibanez, Alice Kuo, Kimberly D. Lakes, Daniel S. Messinger, Sarah Paterson, and Zachary E. Warren

The cost associated with incorporating standardized observational assessments and diagnostic interviews in large-scale epidemiologic studies of autism spectrum disorders (ASD) risk factors can be substantial. Streamlined approaches for confirming ASD case status would benefit these studies. We conducted a multi-site, cross-sectional criterion validity study in a convenience sample of 382 three-year olds scheduled for neurodevelopmental evaluation. ASD case classification as determined by three novel assessment instruments (the Early Video-guided Autism Screener E-VAS; the Autism Symptom Interview, ASI; the Screening Tool for Autism in Toddlers Expanded, STAT-E) each designed to be administered in less than 30 minutes by lay staff, was compared to ADOS scores and DSM-based diagnostic assessment from a qualified clinician. Sensitivity and specificity of each instrument alone and in combination were estimated. Alternative cutpoints were identified under different criteria and two-stage cross validation was used to avoid overfitting. Findings were interpreted in the context of a large, prospective pregnancy cohort study utilizing a two-stage approach to case identification. Under initial cutpoints, sensitivity ranged from 0.63 to 0.92 and specificity from 0.35 to 0.70. Cutpoints giving equal weight to sensitivity and specificity resulted in sensitivity estimates ranging from 0.45 to 0.83 and specificity ranging from 0.49 to 0.86. Several strategies were well-suited for application as a second-stage case-confirmation. These included the STAT-E alone and the parallel administration of both the E-VAS and the ASI. Use of more streamlined methods of case-confirmation in large-scale prospective cohort epidemiologic investigations of ASD risk factors appears feasible. ***Autism Res** 2016, 00: 000–000*. © 2016 International Society for Autism Research, Wiley Periodicals, Inc.

Keywords: autism; ASD; case-confirmation; epidemiology; diagnosis; novel assessments

Introduction

The epidemiologic evidence base around potentially modifiable risk factors and ASD is still quite sparse and although leads have emerged regarding dietary factors, air pollution, and other environmental chemical exposures [Lyall, Schmidt, & Hertz-Picciotto, 2014], considerably more research will be needed to ultimately inform recommendations around personal behavior and/or policy changes. Consequently, the US Interagency Autism Coordinating Committee Strategic Plan for Autism Research has retained epidemiologic research of potentially modifiable risk factors as a continuing ASD research priority [Interagency Autism Coordinating Committee (IACC), 2012].

While it is possible to study certain risk factors using existing secondary data captured in research registries or health care providers or payers' administrative databases, these sources involve certain inherent limitations and challenges [Burke et al., 2013; Schendel et al., 2013]. Therefore, epidemiologic studies built on primary data collection are still needed to advance the field. With

From the A.J. Drexel Autism Institute, Drexel University, 3020 Market Street Suite 560, Philadelphia, PA 19104 (C.J.N., E.S., L.B.); Center for Autism and Related Disorders, Kennedy Krieger Institute, 3901 Greenspring Avenue, Baltimore, MD 21211 (R.L.); Center on Human Development and Disability, University of Washington, Box 357920, Seattle, WA 98195-7920 (W.L.S., L.I.); University of California San Francisco School of Medicine, 401 Parnassus Avenue, San Francisco, CA 94143 (S.B.); Battelle Centers for Public Health Research and Evaluation, 6115 Falls Road, Baltimore, MD 21209 (D.B.); Ichan School of Medicine at Mount Sinai, 17 East 102 Street Floor 2nd Floor - West Tower Room D2-127, New York, NY 10029 (A.G.); University of California Los Angeles School of Medicine, UCLA Center for Health of Children, Families and Communities, 1100 Glendon Avenue, Suite 850, Los Angeles, CA 90024 (A.K.); University of California Irvine School of Medicine, 252 Irvine Hall, Irvine, CA 92697 (K.D.L.); University of Miami, P.O. Box 249229, Coral Gables, FL 33101-0721 (D.S.M.); The Center for Autism Research, Department of Pediatrics, Children's Hospital of Philadelphia, University of Pennsylvania, 3535 Market Street, Suite 860, Philadelphia, PA 19104; Department of Psychology, Temple University, 1701 N. 13th Street, Philadelphia, PA 19122 (S.P.); Vanderbilt Kennedy Center Treatment and Research Institute for Autism Spectrum Disorders (TRIAD), Vanderbilt University, 110 Magnolia Circle, Nashville, TN 37203 (Z.E.W.)

Received December 22, 2015; accepted for publication June 02, 2016

Address for correspondence and reprints: Craig J. Newschaffer, A.J. Drexel Autism Institute, 3020 Market Street Suite 560, Philadelphia, PA 19104. E-mail: cjn32@drexel.edu

Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/aur.1659

respect to primary data collection on ASD outcome, the most widely used standardized, validated instruments for confirming and characterizing ASD diagnosis remain the Autism Diagnostic Observation Schedule (ADOS) [Gotham, Risi, & Lord, 2005] and the Autism Diagnostic Interview-Revised (ADI-R) [Lord, Rutter, & Le Couteur, 1994]. The ADOS consists of a series of semi-structured activities to measure core behavioral domains of ASD and includes four modules, each suitable for individuals at different ages with different verbal abilities. The time to administer and score the ADOS typically ranges from 60 to 90 minutes and, for research purposes, the tool was intended to be administered in a clinical setting by trained professionals who have received special research training and established reliability on tool administration (a process typically taking three to six months). The ADI-R is administered to a knowledgeable caregiver by a trained interviewer and comprises 93 items taking up to 2½ hours to complete. The developers of these tools have recently recommended a staged algorithm for use in children under age 4 where the ADOS is first administered and the ADI-R is used only to follow-up on those who have less decisive ADOS scores only [Kim & Lord, 2012].

However, in implementing large-scale epidemiologic studies, the costs and time associated with case-confirmation based on these approaches can be substantial. Population-based prospective pregnancy or birth cohort studies, such as the Danish National Birth Cohort [Olsen et al., 2001] or the Norwegian Autism Birth Cohort study [Stoltenberg et al., 2010], have used a less-resource intensive initial screener to flag subjects at higher risk of being cases but, even with screeners, considerable numbers of children identified as at-risk still need to be further assessed. The latest sensitivity and specificity data on the Modified Checklist for Autism in Toddlers (M-CHAT), the most widely used ASD parent-self report screener, suggest that, even with revised scoring and a brief follow-up interview (the M-CHAT-R/F), at least two percent of a general population sample of children will screen positive [Robins et al., 2014]. Therefore, in a birth cohort of 100,000 children, like that in the Danish or Norwegian studies mentioned above, more than 2,000 children would require follow-up diagnostic assessment. Recent population-based ASD case–control studies [Hertz-Picciotto et al., 2006; Schendel et al., 2012] have performed fewer diagnostic assessments but still need sample sizes well into the hundreds to be adequately powered.

In this study we sought to determine whether the employment of alternative, less time- and resource-intensive approaches to ASD case confirmation held promise for application in the setting of the epidemiologic study of ASD risk factors.

Methods

Overview

We implemented a multi-site, cross-sectional criterion validity study in a convenience sample of three-year olds. We compared ASD case classification as determined by a battery of three novel assessment instruments, described below, each of which was designed to be administered by non-expert staff in an average of 20 minutes, to gold standard ASD case confirmation (an ADOS by a reliable assessor and a DSM-based diagnostic assessment by a qualified clinician) occurring during an independent neurodevelopmental evaluation.

The Early Video-guided Autism Screener (E-VAS)

The E-VAS is a computer assisted self-interview (CASI) that elicits caregiver-reported responses to a series of voice-over narrated video vignettes. Each vignette contrasts neurotypical development (TD) and ASD social, communication, and restricted, repetitive behaviors (per DSM-5) in analogous play and social interaction contexts and are then followed by questions about the caregiver's child's behavior in domains related to the preceding vignettes. The ASD-indicators described and depicted in the E-VAS were designed to increase caregivers' awareness of qualitatively atypical ASD features, which are difficult for caregivers to detect in young children even when they have an older child with ASD. The E-VAS was developed for children between 18 and 48 months of age and uses a four-point rating scale, permitting caregivers to report on a continuum of perceived frequency or severity of behavioral atypicality in their child. The E-VAS includes a total of 29 items.

The Autism Symptom Interview (ASI)

Development of the preschool ASI (for use in children age 24–59 months) was guided by past research on the ADI-R, as well as by new analyses of previously collected ADI-R data. Because the ASI was designed to be administered quickly by interviewers with minimal training, and because it was intended to identify *current* behavior that is consistent with a diagnosis of ASD, it targets behaviors observed during the previous three-month period. Focusing ASI items on current behaviors also allowed for developing different algorithms for children of different language levels, potentially important given accumulating evidence that ASD symptom sensitivity and specificity (including symptoms measured by the ADI-R) varies substantially according to developmental characteristics of the individual [Gray, Tonge, & Sweeney, 2008]. Questions were developed in which the quality/type of behavior was clearly defined by the question itself and, when appropriate, also reflected features embedded in the parent ADI-R question's response codes at the levels showing most discrimination between ASD and non-ASD. A Likert scale of response options was used to elicit information about whether and to what extent that behavior was present. The preschool ASI includes a total of 30–40 items depending on the language level of the child. A computer-assisted personal interviewing (CAPI) tool was developed and used by study staff to administer the ASI.

The Screening Tool for Autism In Toddlers – Expanded (STAT-E)

The STAT-E is an interactive, play-based screening measure comprising 12 items that assess behavior in areas that represent core deficits for young children with ASD: play, imitation, requesting, and directing attention. The STAT-E expands upon the original STAT [Stone 2000, 2004], which was designed for use with children 24-36 months old, by providing alternative materials more compatible with the interests and activities of 3-year-old children extending the age range through 42 months). Consistent with the original version, the STAT-E generates a response-to-press total score that is used to indicate a child's risk for ASD and uses age-specific cutpoints (for children under 36 month of age and 36 months of age and older). Additionally, the STAT-E included two new behavioral rating metrics that index the child's level of social engagement during each individual item as well as overall, and a rating to indicate the extent to which the child exhibits atypical behaviors with regard to language, object use, body use, and sensory-seeking behaviors.

Study Population

The study population was a convenience sample of 382 children aged 24–39 months (and one of their adult caregivers) who had received, or who were scheduled to receive, a neurodevelopmental evaluation (either through a research project or for clinical purposes) at either an ASD or general neurodevelopmental disorder clinic affiliated with one of eight study sites (Children's Hospital of Philadelphia, Ichan School of Medicine at Mt Sinai, Kennedy Krieger Institute, University of California Irvine, University of California Los Angeles, University of Miami, University of Washington, and Vanderbilt University). The neurodevelopmental evaluation had to include the administration of the ADOS from an assessor who has met standards for either clinical or research reliability and the provision of a DSM-based diagnostic evaluation by qualified clinicians (either licensed clinical psychologists, board certified neurologists, or developmental pediatricians). Sites recruited two groups of children in the targeted age range, those being seen for neurodevelopmental evaluation with prior suspicion of ASD and those being seen without prior ASD suspicion. The goal of this recruitment strategy was to develop a sample that had the same general mix of developmental concerns as would a group of like-age children from the general population who screened positive on an instrument like the M-CHAT. Study enrollment began in March of 2013 and concluded in January of 2015. Of the 382 children, 300 (78.5%) had prior ASD suspicion. The Study protocol was IRB approved at all participating sites, the coordinating center (Drexel University) and the data center (Battelle Memorial Institute). Because the project was funded as a National Children's Study [Panel on the Design of the National Children's Study et al., 2014] Formative Research Project, US OMB approval was also obtained.

Data Collection

For the visit where the three above described instruments were to be administered, sites were encouraged to utilize staff that did not have extensive clinical experience working with children with ASD in order to replicate staffing conditions similar to those that were under discussion for the National Children's Study. All staff completed web-based training modules on study protocol and assessment administration prior to beginning data collection. Order of instrument administration was randomly determined. Whenever possible, study staff were kept blind as to whether a subject did or did not have prior ASD suspicion. Computer-assisted E-VAS and ASI data were captured in real time by a web-based data capture system. The STAT-E scores were entered after study visit completion. Data from the neurodevelopmental evaluation, which was completed separately by an expert research or clinical team at the same institution, were not abstracted until after the study visit was completed. The abstraction tool captured information on best-estimate clinical ASD diagnosis, any noted comorbidities, scores on general cognitive functioning tests (either previously administered or administered during the neurodevelopmental evaluation visit) as noted in the record, scores from the ADOS (and ADI-R, if also administered) given during the evaluation visit. Other DSM diagnoses that may also have been given were noted.

To assess the quality of assessor scoring and administration fidelity for the STAT-E, assessors at six of the eight sites video-recorded a small sample of their administrations and sent them for review by the STAT-E development team at University of Washington. A total of 39 videos, of which 34 were of sufficient quality for review, were received from 6 sites (University of Washington and Vanderbilt did not submit tapes because these sites had certified STAT trainers available to support staff and assure quality and fidelity of administration). To evaluate STAT-E scoring reliability for each study staff, UW reviewers scored each of the 12 STAT-E items from the videotape and then compared their scores to those of the assessors. Overall percent agreement across the 12 items was then calculated for each assessor.

Study variables

Total scores from novel assessment instruments. Each instrument generates a total score (higher scores indicating more ASD characteristics) and has suggested initial cutpoints. The ASI total score is based on distinct algorithms for children classified as verbal (based on 24 items) and nonverbal (based on 13 items). Scores can range from 0 to 72 for verbal and 0 to 39 for nonverbal. The initial cutpoints for the ASI were based on inspection of ROC curves in validation samples of 78 verbal and 30 nonverbal children with purposeful preference for sensitivity. Initial cutpoints were 27 and 14, respectively, for the verbal and nonverbal scores. Consistent with the original version of the STAT, the STAT-E total risk score was based on summing and weighting the response-to-press item score (specifically the number of fails) for each of the four subdomains (play, requesting, directing attention, and imitation), and then summing the four subdomain scores. Risk scores can range from 0 to 4 (subdomain scores are the average, as opposed to total score, in that domain). The initial cutpoints for the STAT-E score (>=2 for children less than 36 month old and >=1.75 for children 36 months old or older) are the existing published cutpoints on the STAT [Stone, Coonrod, Turner, & Pozdol, 2004; Stone & Ousley, 2008]. Because they have yet to be validated, the two new STAT-E domains, social engagement and atypical behaviors, were not factored into the scores used here. The E-VAS total score is the sum of five subdomain scores (imagination and play, flexibility with toys/routines, sharing and enjoyment, facial expressions and gestures, and unusual body movement and repetitive behaviors). Scores can range from 29 to 116. The initial cutpoint of >=53 was developed to maximize the number correctly classified in an initial validation sample ($n = 109$) recruited at the Kennedy Krieger Institute and surrounding clinics.

ASD classification. Best-estimate clinical diagnosis was abstracted from the evaluation record as ASD or no-ASD and served as the principal ASD classification measure. ADOS and ADI-R scores were also collected for consideration in the development of secondary outcomes. Because sites were collecting neurodevelopmental evaluation data under a range of circumstances (i.e., other studies, clinical evaluations), we allowed for multiple ADOS versions (ADOS-G, ADOS-2, and ADOS-T) and modules. Version and module were recorded, as were domain and summary scores (for the appropriate algorithm). It was also noted whether the ADOS evaluator was research-reliable. ADOS severity scores were calculated following the published guidelines [Gotham, Pickles, & Lord, 2009]. As there are no score conversion guidelines for the ADOS-T, these were not converted. Data on ADI-R were available for only 34 subjects, and thus were not considered further.

Covariates. Child-level covariates included gender, age, verbal status, and parent-reported ethnicity, race, education level, and household income at study visit. A child was classified as nonverbal if the parent reported he/she did not use single words, used fewer than five different words on a daily basis, or used at least five different single words on a daily basis but did not use phrases on a daily basis. We abstracted information on cognitive ability from the neurodevelopmental assessment. These data were available on 299 subjects (78.3%). Mullen Scales of Early Learning scores accounted for 99% of the available cognitive ability data. The next most commonly available test results were for the Stanford-Binet, with data on just three subjects. Consequently, we limited cognitive status data to that from the Mullen, categorizing based on whether the Early Learning Composite score was over 70.

Study staff provided information on highest level of education attained (<bachelors degree, bachelor's degree, at least some graduate training), field of study (psychology or other), years of experience in a clinical or educational setting working with children, years of experience with ASD specifically, and self-rated familiarity with ASD (low, average, high).

Analyses also considered structural covariates, including study recruitment group, study site, timing of the study visit relative to the neurodevelopmental evaluation, and order of assessment administration during the study visit.

Analytic Approach

Descriptive analysis. Descriptive analyses involved examination of the distribution of total scores from the three novel assessment instruments in ASD and non-ASD groups and across levels of covariates. Correlations between novel assessment total scores and ADOS module-specific total scores, ADOS severity scores, and Mullen ELC were also examined.

Estimating sensitivity and specificity. The sensitivity and specificity of each instrument based on its recommended cutpoint was estimated using ASD best-estimate clinical diagnosis as the gold standard. Sensitivity and specificity were also re-estimated for parallel and serial combinations of all three instruments. Parallel combinations require criteria to be met on each instrument (maximizing specificity) while serial combinations require criteria to be met on any instrument (maximizing sensitivity). Parallel combinations of each of the three possible pairings of tools were considered as well. Standard errors for sensitivity and specificity were estimated based on exact binomial, and 95% confidence intervals are presented.

Logistic regression modeling. Logistic regression models were used to explore the influence of covariates on instrument performance. For each instrument, one model including all subjects was used to predict correct classification, irrespective of ASD diagnosis. Additional separate models restricted, respectively, to subjects with and without best-estimate diagnoses of ASD were used to predict detection of true positives and true negatives. Separate models were also specified for groups of individual, staff, and structural covariates. Consequently, 27 models (three samples, three instruments, and three sets of predictors) were explored. The direction and statistical significance ($P < 0.05$) of estimated associations were examined and 95% confidence intervals calculated. Models first were run with missing values included and then rerun with missing values excluded. In addition, models were re-run using a more conservative gold-standard case confirmation definition of best-estimate clinical diagnosis plus an ADOS score indicating ASD.

Alternative cutpoint analysis. Because the evidence base supporting the initial existing cutpoints for the three novel instruments is, in all cases, limited and, more importantly, because the criteria for determining a cutpoint can vary depending on the context in which an instrument will be used, we also undertook analyses exploring alternative cutpoints using the data available here. To do so we followed the two-fold cross validation approach described by Mazumdar et al. [2003]. Under this approach the available sample is split, alternative new cutpoints are developed under the same decision rule independently in each half of the data, but the sensitivity and specificity estimates for these alternative cutpoints are calculated by applying each half's cutpoint to the opposite half's data to guard against overfitting. Because the STAT-E had initial cutpoints conditional on subject age and the ASI had initial cutpoints conditional on language stratified randomization was used when splitting the data to ensure a balance of these subject characteristics. Sensitivity and specificity are also presented for the same parallel and serial combinations of instruments examined with the initial cutpoints. Following the recommendations of Mazumadar et al. [2003], for the purpose of informing future research, another set of alternative cutpoints was also determined from the pooled data following the same decision rules used in the split-sample approach, since these are the single best-estimate values given our data and would be the cutpoints actually recommended for further consideration and study moving forward, despite the use of slightly different split-sample based cutpoints in the sensitivity and specificity estimations.

We considered two different approaches to determining alternative cutpoints. First, alternative cutpoints were selected based on a traditional approach that weights sensitivity and specificity equally (operationalized by maximizing the Youden index (Youden & Cameron, 1950) = sensitivity+specificity-1). The second approach to choosing alternative cutpoints was informed by the risk factor relative risk bias estimation described below. Based on this a cutpoint that maximized specificity while holding sensitivity as close as possible to 50% was considered. Alternative cutpoint selection was implementing using R-code developed by Lopez-Raton et al. [2014].

Risk factor relative risk bias estimation. Because the goal of epidemiologic risk factor research is the unbiased estimation of associations between candidate risk factors and ASD, we explored the extent to which different case confirmation instruments alone or in combination using alternative cutpoints would introduce bias into such association estimates. To do so, we considered the scenario of a population-based pregnancy cohort study where a two-stage approach would be used for outcome identification and relative risks would estimate risk factor associations. We assumed the M-CHAT-R/F would be applied in the first stage of outcome assessment and that all level-one screen positives would receive case confirmation assessment at age three. As mentioned above, the latest validation data on the M-CHAT-R/F suggest that two percent of a general population would screen positive. At the M-CHAT-R/F's most recently reported 83.3% sensitivity and 99.2% specificity [Robins et al., 2014], and assuming a general population ASD prevalence of 1.47% [Baio, 2014], 60.8% of the level-one screen positives are expected to be true ASD cases. It is this level-one screen positive population that would then receive the second-stage diagnostic confirmation assessment. Given this, and assuming that ASD misclassification introduced through imperfect measurement at both the first stage (via the level-one M-CHAT screen) and at stage two (by the confirmation assessment) is non-differential with respect to the risk factor under study (a reasonable assumption for most prenatal risk factors), we calculated the expected bias in resulting relative risks (as the ratio of the estimated to true relative risk) over a range of sensitivities and specificities for candidate confirmation approaches under different assumptions regarding risk factor prevalence and true relative risk. We did this for exposure prevalences at 5%, 10% and 15%

(assuming that exposures of etiologic interest would not be extremely common) and for relative risks of 1.5 and 2.0 (assuming that the likelihood of there being true ASD risk factors with larger relative risks is probably small). We also varied true ASD prevalence around the CDC estimate, allowing that there may be variability in true prevalence across study populations, considering prevalences of 1%, 1.5% and 2%. The results were used to inform exploration of alternative cutpoints for the assessment instruments that would minimize bias in estimated risk factor relative risks.

RESULTS

Descriptive Analysis

Table 1 summarizes the characteristics of study population overall and stratified by best-estimate clinical diagnosis. Correlations between total scores on the three instruments and ADOS and ELC scores are displayed in Table 2. Scores on all three tools had positive (above 0.3), strongly statistically significant correlations with the ADOS severity score and negative, strongly significant correlations with the ELC.

Estimating Sensitivity and Specificity

Table 3 displays the sensitivity and specificity estimates for all three instruments individually, in serial and parallel combination, for the two combinations of direct observation and caregiver report, and for the combination of the two caregiver reports. Estimates are shown for the sample overall and for the sample subgroups without verbal ability and and with ELC score<=70. The direct observation STAT-E showed higher specificity and lower sensitivity (0.70 and 0.63, respectively) than the caregiver-report based ASI and E-VAS (0.91 and 0.35, 0.92 and 0.38, respectively). Specificity for all tools decreased among more impaired children (defined either as lack of verbal ability or ELC<=70). The ASI, which has different scoring and cutpoints for verbal and nonverbal children, maintained its specificity across these two groups but had reduced specificity among those children with ELC<=70. As expected, the parallel combination of all tools (which require that criteria be met on all instruments) maximized specificity (sensitivity=0.59, specificity=0.78) while serial combination (which requires criteria only be met on one instrument) maximized sensitivity (sensitivity=0.96, specificity=0.26). The parallel combination of the STAT-E with either of the caregiver report tools yielded sensitivity and specificity similar to the parallel combinations of all instruments. When sensitivity and specificity were re-estimated using the most conservative gold-standard case confirmation of best-estimate clinical diagnosis plus an ADOS score indicating ASD, sensitivity and specificity estimates were virtually unchanged (data not shown; 96.4% of cases with a clinical ASD diagnosis also had an ADOS score indicating ASD).

Logistic Regression Modeling

Table 4 summarizes results from logistic regressions examining the influence of child, assessor and study characteristics on the ability of the instruments to correctly classify subjects on ASD status. A plus sign indicates an OR point estimate above 1 and a minus sign indicates an OR below one. If the symbol is boxed, this indicates that the OR estimate was statistically significant at a $P < 0.05$. When no symbols are shown, this indicates that model fitting was not possible given the available sample size and distribution across predictors. This summary should be interpreted cautiously as even very small ORs are denoted as either positive or negative and, given the multitude of comparisons, there is elevated potential for statistically significant ORs to be false positives. Table 5 displays the OR estimates, 95% CI and *P*-values for the correct classification model in order to give a better indication of the magnitude of these effects.

Looking back at Table 4, the results for child characteristics suggest that the most consistent effect on instrument performance is impairment level of the child. Mullen ELC < =70 consistently positively influences sensitivity and negatively influences specificity, suggesting that these more impaired children are more likely to test positive for ASD on all these tools regardless of their true ASD status. The net effect on correct classification is positive, because this sample includes more true positives than true negatives. Staff effects were less consistent across tools and, interestingly, ASD experience did not have a strong or dose-consistent effect on the STAT-E. Instrument administration order did not strongly influence performance nor did the order in which the study visit occurred relative to the neurodevelopmental evaluation (although when the two occurred on the same day, we could not determine order). Membership in the ASD suspicion recruitment group was significantly associated with reduced specificity for all tools, both the parent report ASI and E-VAS as well as the direct observation STAT-E. This consistency across instruments suggests that this effect is likely due more to the presence of particular behaviors generating suspicion than to knowledge of suspicion itself influencing parent-reporting. Some site effects were also observed as site with the most prior experience with the STAT tended to have better results with the tool (OR>2 on Table 5) although other sites with less experience also had similarly sized OR point estimates. As mentioned, videotaped STAT-E assessments were received from the six sites without substantive prior STAT experience and

Table 1. Characteristics of the Study Sample

	All subjects ($n = 382$)		ASD ($n = 278$)		No ASD ($n = 104$)	
	n	%	*n*	%	*n*	%
Child characteristics						
Sex						
Male	285	74.61	212	76.26	73	70.19
Female	97	25.39	66	23.74	31	29.81
Age						
<36 months	245	64.14	178	64.03	67	64.42
>=36 months	137	35.86	100	35.97	37	35.58
Verbal ability						
Nonverbal	187	48.95	159	57.19	28	26.92
Verbal	195	51.05	119	42.81	76	73.08
Mullen ELC						
>70	130	34.03	73	26.26	57	54.81
<=70	156	40.84	135	48.56	21	20.19
Missing	96	25.13	70	25.18	26	25
Ethnicity						
Hispanic or Latino	101	26.44	77	27.7	24	23.08
Not Hispanic or Latino	281	73.56	201	72.3	80	76.92
Race						
Amer. Indian/Alaska Native/Native Hawaiian	5	1.31	4	1.44	1	0.96
Asian	36	9.42	29	10.43	7	6.73
Black/African American	35	9.16	24	8.63	11	10.58
White	267	69.9	192	69.06	75	72.12
More than one race	13	3.4	11	3.96	2	1.92
Don't Know/Refused	26	6.81	18	6.47	8	7.69
Maternal education						
<Bachelors	197	51.57	150	53.96	47	45.19
Bachelors	107	28.01	73	26.26	34	32.69
>Bachelors	64	16.75	54	19.42	23	22.12
Don't know	1	0.26	1	0.36	0	0
Family income						
<$29,999	90	23.56	64	23.02	26	25
$30,000–$49,999	55	14.4	43	15.47	12	11.54
$50,000–$74,999	66	17.28	52	18.71	14	13.46
$75,000–$99,999	56	14.66	41	14.75	15	14.42
$100,000+	100	26.18	65	23.38	35	33.66
Don't Know/Refused	15	3.92	13	4.68	2	1.92
Staff characteristics						
Education level						
<Bachelor's	125	32.72	94	33.81	31	29.81
Bachelor's	129	33.77	95	34.17	34	32.69
>=Bachelor's	127	33.25	88	31.65	39	37.5
Missing	1	0.26	1	0.36	–	–
Field of study						
Psychology	251	65.71	178	64.03	73	70.19
Other	130	34.03	99	35.61	31	29.81
Missing	1	0.26	1	0.36	–	–
0–0.5	56	14.66	36	12.95	20	19.23
1–2	122	31.94	95	34.17	27	25.96
3–6	152	39.79	113	40.65	39	37.5
6+	48	12.57	32	11.51	16	15.38
Missing	4	1.05	2	0.72	2	1.92
Yrs. ASD experience						
0–0.5	184	48.17	133	47.84	51	49.04
1–2	137	35.86	97	34.89	40	38.46
3–6	27	7.07	21	7.55	6	5.77
6+	24	6.28	18	6.47	6	5.77
Missing	10	2.62	9	3.24	1	0.96
ASD familiarity						
High	101	26.44	71	25.54	30	28.85
Average	237	62.04	173	62.23	64	61.54

Table 1. Continued

	All subjects ($n = 382$)		ASD ($n = 278$)		No ASD ($n = 104$)	
	n	%	*n*	%	*n*	%
Low	26	6.81	17	6.12	9	8.65
Missing	18	4.71	17	6.12	1	0.96
Study characteristics						
Site						
A	38	9.95	26	9.35	12	11.54
B	17	4.45	11	3.96	6	5.77
C	14	3.66	9	3.24	5	4.81
D	52	13.61	36	12.95	16	15.38
E	23	6.02	18	6.47	5	4.81
F	18	4.71	11	3.96	7	6.73
G	84	21.99	63	22.66	21	20.19
H	136	35.6	104	37.41	32	30.77
Instrument order						
ASI/STAT-E/E-VAS	48	12.57	37	13.31	11	10.58
ASI/E-VAS/STAT-E	82	21.47	56	20.14	26	25
STAT-E/ASI/E-VAS	56	14.66	40	14.39	16	15.38
STAT-E/E-VAS/ASI	67	17.54	44	15.83	23	22.12
E-VAS/ASI/STAT-E	73	19.11	55	19.78	18	17.31
E-VAS/STAT-E/ASI	55	14.4	46	16.55	9	8.65
Missing	1	0.26	–	–	1	0.96
Visit Order						
Same day	56	14.66	42	15.11	14	13.46
Study visit/ neurodev eval	204	53.4	151	54.32	53	50.96
Neurodev eval/ study visit	70	18.32	45	16.19	25	24.04
Missing	52	13.61	40	14.39	12	11.54
Age	382	33.1 (4.1)	278	33.2 (4.1)	104	32.9 (4.2)
ADOS severity score	363	5.5 (2.6)	268	6.3 (2.2)	95	3.5 (2.4)
ADOS1 module 1 nonverbal	101	14.9 (5.1)	88	15 (4.9)	13	14.3 (6.5)
ADOS1 module 1 verbal	84	11.5 (4.8)	57	12.7 (4.1)	27	9.1 (5.4)
ADOS1 module 2	25	8.8 (5.9)	14	12.9 (4.2)	11	3.6 (3)
ADOS -T Nonverbal	30	17.2 (7.6)	22	19.7 (5.5)	8	10.3 (8.6)
ADOS-T verbal	11	10.6 (6.9)	5	16.8 (3.6)	6	5.5 (3.8)
ADOS1 module 1 nonverbal	53	18.9 (6.2)	48	19.9 (5.0)	15	9 (8.7)
ADOS-2 module 1 verbal	44	13.3 (6.9)	29	16.7 (4.7)	15	6.6 (5.3)
ADOS-2 module 2	28	9.5 (6.4)	14	14.3 (4.5)	14	4.6 (3.9)
Mullen ELC	286	72.6 (24)	208	66.5 (18.8)	78	88.8 (27.3)

were reviewed by expert STAT assessors. 39 videos were received, 34 were reviewable (in that both the child and assessor were visible), from 19 different assessors. Mean percent agreement between staff and reviewers on the scoring of the 12 STAT-E items was 86%. In the standard STAT training (which leads to certification) correct scoring of at least 10 of 12 items is needed for an administrator to be considered reliable. This level of reliability was achieved on 24 of the 34 (70%) of the reviewed administrations.

Risk factor relative risk bias estimation

As described above, we estimated anticipated bias in risk factor relative risk estimates due to outcome misclassification. We assumed a study design involving a prospective cohort with two-stage outcome assessment, where the instruments under consideration here might serve as the second-stage case confirmation tool. Among the parameters that were varied, exposure prevalence, at least over the range considered, had the smallest influence on bias. Therefore, in Figure 1 we show a heat map of degree of bias for a range of second-stage sensitivities (y-axis) and specificities (x-axis) for different assumed ASD prevalences and relative risks. Specificity has a greater influence on bias than sensitivity with the differential influence magnified as prevalence decreases. Consequently, for relative risk estimation purposes, more of a premium should be placed on specificity than sensitivity. Thus, we sought to explore an alternative cutpoint that favored specificity over sensitivity. Examining the heat maps, it can be seen that when sensitivity is lowered to 50%, if a specificity of at least 80% can be attained, the percent bias in the relative risk tends to be under 10%. While cutpoints could be selected where specificity is even higher, and sensitivity correspondingly lower, the gains in relative risk bias minimization are minimal and the total number of subjects confirmed as cases drops. Although maintaining an absolute number of identified cases is not the first research priority in this

Table 2. Correlation Coefficients (Pearson Unless Indicated) Between Instrument Total Scores and ASOS Total Scores and Mullen ELC Score

	STAT-E *r* (*P*-val)	ASI Verbal[a] *r* (*P*-val)	ASI Nonverbal[b] *r* (*P*-val)	E-VAS *r* (*P*-val)
ADOS severity score (*n* = 363)	0.49 (<0.01)	0.32 (<0.01)	0.45 (<0.01)[c]	0.39 (<0.01)
ADOS1 module 1 nonverbal (*n* = 101)	0.65 (<0.01)[c]	NA	0.41 (<0.01)[c]	0.33 (0.01)[c]
ADOS1 module 1 verbal (*n* = 84)	−0.38 (<0.01)	0.20 (0.07)	NA	0.24 (0.03)
ADOS1 module 2 (*n* = 25)	−0.15 (0.47)	0.14 (0.51)	− (*n* = 1)	0.13 (0.53)
ADOS -T nonverbal (*n* = 30)	0.54 (<0.01)[c]	NA	0.37 (0.04)	0.36 (0.05)
ADOS-T verbal (*n* = 11)	0.37 (0.27)	0.26 (0.45)[c]	NA	0.40 (0.23)
ADOS-2 module1 nonverbal (*n* = 53)	0.56 (<0.01)[c]	NA	0.39 (<0.01)[c]	0.44 (<0.01)[c]
ADOS-2 module1 verbal (*n* = 44)	0.58 (<0.01)	0.27 (0.07)[c]	NA	0.33 (0.03)[c]
ADOS-2 module 2 (*n* = 28)	0.36 (0.06)[c]	0.55 (<0.01)	− (*n* = 2)	0.61 (<0.01)
Mullen ELC (*n* = 286)	−0.54 (<0.01)[c]	−0.30 (<0.01)[c]	−0.32 (<0.01)[c]	−0.42 (<0.01)[c]

[a] 195 subjects with verbal ASI.

[b] 187 subjects with nonverbal ASI.

[c] Spearman correlation coefficient.

– correlation was not calculated because too few subjects (*N*<3).

[NA] not applicable because instrument did not have any of that type of ADOS module.

Table 3. Sensitivity and Specificity of Individual Instruments and Instrument Combinations With Best-Estimate Clinical Diagnosis as the Gold Standard Based on Initial Recommended Cutpoints

	Full Sample (*N* = 382)				Nonverbal Subgroup (*N* = 187)				Mullen ELC < = 70 Subgroup (*N* = 156)			
	Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
STAT-E	0.63	(0.57, 0.69)	0.70	(0.60, 0.79)	0.81	(0.73, 0.86)	0.54	(0.34, 0.72)	0.76	(0.67, 0.83)	0.52	(0.30, 0.74)
ASI	0.91	(0.87, 0.94)	0.35	(0.26, 0.44)	0.94	(0.90, 0.97)	0.25	(0.11, 0.45)	0.96	(0.92, 0.99)	0.05	(0.00, 0.24)
E-VAS	0.92	(0.88, 0.95)	0.38	(0.29, 0.49)	0.93	(0.88, 0.97)	0.21	(0.08, 0.41)	0.97	(0.93, 0.99)	0.10	(0.01, 0.30)
Parallel[a] (All)	0.59	(0.53, 0.65)	0.78	(0.69, 0.85)	0.77	(0.69, 0.83)	0.61	(0.41, 0.79)	0.73	(0.64, 0.80)	0.52	(0.30, 0.74)
Serial[b] (All)	0.96	(0.93, 0.98)	0.26	(0.18, 0.35)	0.97	(0.93, 0.99)	0.11	(0.02, 0.28)	1.00	(0.97, 1.00)	0.05	(0.00, 0.24)
Parallel[a] (STAT-E + ASI)	0.60	(0.54, 0.66)	0.75	(0.66, 0.83)	0.79	(0.71, 0.85)	0.61	(0.41, 0.79)	0.74	(0.66, 0.81)	0.52	(0.30, 0.74)
Parallel[a] (STAT-E + E-VAS)	0.61	(0.55, 0.67)	0.78	(0.69, 0.85)	0.78	(0.71, 0.84)	0.61	(0.41, 0.79)	0.73	(0.65, 0.81)	0.52	(0.30, 0.74)
Parallel[a] (ASI + E-VAS)	0.88	(0.84, 0.92)	0.42	(0.33, 0.52)	0.91	(0.86, 0.95)	0.29	(0.13, 0.49)	0.94	(0.89, 0.97)	0.10	(0.01, 0.30)

[a] Parallel combination requires above threshold level on ALL instruments to be considered a test positive.

[b] Serial combination requires above threshold level on ANY instrument to be considered a test positive.

context, it may be of value for ancillary projects to have a robust cohort of identified cases on which to build follow-up studies. Consequently, the second alternative cutpoint explored was one that dropped sensitivity to as close to 50% as possible. Figure 2 shows the ROC curves used to select the cutpoints under the Youden index and 50% sensitivity criteria for each instrument. In keeping with the original approach of the ASI and the STAT-E, alternative cutpoints were created conditional on age for the STAT-E and verbal ability for the ASI. The values for each of the cutpoints as well as the area under the curves (AUC) are also displayed on the figure.

Alternative cutpoint analysis

Table 6 shows the sensitivity and specificities, estimated from the two fold cross validation approach, based on the two alternative cutpoints for each instrument individually and for the same parallel and serial combinations as previously displayed in Table 3. The cutpoints recommended under Youden index maximization criteria (equally weighting sensitivity and specificity) generate similar sensitivity and specificities as the original cutpoints for the STAT-E and the ASI. The E-VAS sensitivity is lower and specificity higher in this sample for the Youden cutpoint than for the originally proposed cutpoint. Under the 50% sensitivity criteria, the STAT-E achieved 83% specificity (estimated sensitivity at 54%) while the E-VAS achieved 79% specificity (estimated sensitivity was 50%) and the ASI reached 70% specificity (estimated sensitivity was 53%). The parallel combinations of all tools elevated specificity above 90% but dropped sensitivity below 30%. The parallel pairing of the two caregiver reports, the ASI and the E-VAS, generated estimated sensitivity of 42% and specificity of 82%.

Table 4. Summary of Association Between Child, Staff, and Study Characteristics and Instrument Performance With Best-Estimate Clinical Diagnosis as the Gold Standard Based on Initial Recommended Cutpoints

	STAT-E			ASI			E-VAS		
	TP[1]	TN[2]	CC[3]	TP[1]	TN[2]	CC[3]	TP[1]	TN[2]	CC[3]
Child Characteristics									
Sex									
Male	+	+	+	-	$\boxed{-}$	-	-	$\boxed{-}$	-
Female	ref	ref	ref	ref	ref	ref	ref	ref	ref
Age									
<36 mos	ref	ref	ref	ref	ref	ref	ref	ref	ref
>36 mos	-	+	-	-	+	-	-	+	+
Verbal Ability									
Non-verbal	ref	ref	ref	ref	ref	ref	ref	ref	ref
Verbal	$\boxed{-}$	+	$\boxed{-}$	-	-	-	-	+	$\boxed{-}$
Mullen ELC									
>70	ref	ref	ref	ref	ref	ref	ref	ref	ref
<=70	$\boxed{+}$	$\boxed{-}$	+	$\boxed{+}$	$\boxed{-}$	$\boxed{+}$	$\boxed{+}$	$\boxed{-}$	$\boxed{+}$
Missing	$\boxed{+}$	-	+	+	-	+	+	$\boxed{-}$	+
Ethnicity									
Hispanic or Latino	$\boxed{-}$	-	$\boxed{-}$	+	+	+	-	$\boxed{+}$	+
Not Hispanic or Latino	ref	ref	ref	ref	ref	ref	ref	ref	ref
Race									
Asian	-	+	-	-	+	-	$\boxed{-}$	+	-
Black/African American	+	+	+	-	+	-	-	+	-
White	ref	ref	ref	ref	ref	ref	ref	ref	ref
Other	+	+	+	+	-	+	+	-	-
Maternal Education									
<Bachelors	ref	ref	ref	ref	ref	ref	ref	ref	ref
Bachelors	+	+	+	-	+	-	-	$\boxed{+}$	+
>Bachelors	+	+	+	+	$\boxed{+}$	+	+	$\boxed{+}$	$\boxed{+}$
Don't know	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Family Income									
$20,000 - $29,999	ref	ref	ref	ref	ref	ref	ref	ref	ref
$30,000 - $49,999	-	+	-	+	-	+	+	-	+
$50,000 - $74,999	$\boxed{+}$	-	+	+	-	+	+	+	+
$75,000 - $99,999	+	$\boxed{-}$	-	+	-	+	+	-	+
$100,000+	+	-	+	+	+	+	+	+	+
Don't Know/Refused	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Staff Characteristics									
Education level									
<Bachelor's	ref	ref	ref	ref	ref	ref	ref	ref	ref
Bachelor's	+	+	+	+	+	+	+	+	+
>=Bachelor's	+	-	+	-	+	-	-	+	-
Field of Study									
Psychology	+	$\boxed{+}$	$\boxed{+}$	-	+	-	-	-	-
Other	ref	ref	ref	ref	ref	ref	ref	ref	ref
Yrs. clinical/educ. experience									
<1	ref	ref	ref			ref		ref	ref
1-2	-		-			+		+	+
3-6	$\boxed{-}$		-			+		+	+
6+	-		-			-		+	-
Missing	-		-			-		+	+
Yrs. ASD experience									
<1	ref	ref	ref		ref	ref			ref
1-2	$\boxed{+}$	-	+		+	+			+
3-6	+	-	+		-	+			+
6+	-	+	-		+	$\boxed{+}$			$\boxed{+}$
Missing	+	N/A	+		N/A	+			+

Table 4. Continued

	STAT-E			ASI			E-VAS		
	TP[1]	TN[2]	CC[3]	TP[1]	TN[2]	CC[3]	TP[1]	TN[2]	CC[3]
ASD Familiarity									
High	-	+	+	-	+	-	-	+	-
Average	-	+	-	-	+	-	+	-	-
Low	ref	ref	ref	ref	ref	ref	ref	ref	ref
Missing	+	N/A	-	-	N/A	-	+	N/A	+
Study Characteristics									
Recruitment group									
No ASD suspected	ref	ref	ref	ref	ref	ref	ref	ref	ref
ASD suspected	-	$\boxed{-}$	$\boxed{-}$	+	$\boxed{-}$	-	+	$\boxed{-}$	+
Site									
A	$\boxed{+}$		$\boxed{+}$						-
B	+		+						+
C	+		+						+
D	+		+						-
E	+		+						+
F	+		-						-
G	$\boxed{+}$		$\boxed{+}$						+
H	ref		ref						ref
Instrument order									
First	ref	ref	ref	ref	ref	ref	ref	ref	ref
Second	-	$\boxed{+}$	-	+	+	+	-	-	-
Third	-	+	+	-	+	+	-	-	-
Visit Order									
Same day	ref	ref	ref	ref	ref	ref		ref	ref
Study visit/neurodev eval	+	-	-	-	$\boxed{-}$	-		-	-
Neurodev. eval/study visit	+	+	+	-	-	-		-	-
Missing	-	+	-	+	-	-		-	-

[a] TP=true positive; test positive among clinically confirmed ASD cases ($n = 278$).
[b] TN=true negative; test negative among non ASD cases ($n = 104$).
[c] CC=correctly classified; TP+TN among all subjects ($n = 382$).
Boxed symbols represent statistically significant results ($P < 0.05$).

Table 5. Association between child, staff, and study characteristics and correct classification (true positives plus true negatives) for each instrument with best-estimate clinical diagnosis as the gold standard based on initial recommended cutpoints

	STAT-E			ASI			E-VAS		
	OR	95%CI	*P*-val	OR	95%CI	*P*-val	OR	95%CI	*P*-val
Child characteristics									
Sex									
Male	1.32	(0.8, 2.2)	0.30	0.72	(0.2, 2.3)	0.58	0.63	(0.3, 1.2)	0.16
Female	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Age									
<36 months	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
>36 months	0.96	(0.6, 1.6)	0.87	0.51	(0.2, 1.4)	0.18	1.08	(0.6, 1.9)	0.78
Verbal ability									
Nonverbal	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Verbal	0.38	(0.2, 0.6)	0.00	0.33	(0.1, 1.0)	0.06	0.51	(0.3, 0.9)	0.02
Mullen ELC									
>70	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
<=70	1.40	(0.8, 2.6)	0.24	5.64	(1.5, 20.7)	0.01	2.50	(1.3, 4.8)	0.01
Missing	1.63	(0.8, 3.1)	0.14	1.76	(0.5, 5.9)	0.36	1.61	(0.8, 3.3)	0.19
Ethnicity									
Hispanic or Latino	0.46	(0.2, 0.8)	0.01	1.39	(0.4, 5.1)	0.62	1.13	(0.6, 2.3)	0.73
Not Hispanic or Latino	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Race									
Asian	0.85	(0.4, 2.0)	0.69	0.34	(0.1, 1.4)	0.13	0.44	(0.2, 1.1)	0.07
Black/African American	1.77	(0.7, 4.6)	0.24	0.88	(0.2, 5.2)	0.88	0.67	(0.3, 1.7)	0.39
White	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Other	1.70	(0.8, 3.8)	0.19	2.04	(0.2, 19.4)	0.54	0.89	(0.4, 2.2)	0.81

Table 5. Continued

	STAT-E			ASI			E-VAS		
	OR	95%CI	*P*-val	OR	95%CI	*P*-val	OR	95%CI	*P*-val
Maternal education									
<Bachelors	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Bachelors	1.21	(0.7, 2.5)	0.37	0.88	(0.3, 3.0)	0.84	1.03	(0.5, 2.0)	0.93
>Bachelors	1.33	(0.6, 2.5)	0.60	1.93	(0.4, 9.8)	0.43	2.50	(1.0, 6.2)	0.05
Don't know	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Family income									
$20,000–$29,999	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
$30,000–$49,999	0.97	(0.5, 2.1)	0.94	2.74	(0.5, 15.1)	0.25	1.10	(0.5, 2.5)	0.82
$50,000–$74,999	1.63	(0.7, 3.6)	0.22	2.20	(0.5, 9.6)	0.30	2.11	(0.9, 5.1)	0.10
$75,000–$99,999	0.70	(0.3, 1.6)	0.40	4.70	(0.7, 33.3)	0.12	1.29	(0.5, 3.3)	0.59
$100,000+	1.10	(0.5, 2.4)	0.80	2.24	(0.5, 10.6)	0.31	1.58	(0.7, 3.8)	0.31
Don't Know/Refused	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Staff characteristics									
Education level									
<Bachelor's	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Bachelor's	1.92	(0.9, 4.0)	0.08	1.08	(0.5, 2.4)	0.85	1.39	(0.6, 3.2)	0.44
>=Bachelor's	1.35	(0.7, 2.5)	0.36	0.88	(0.4, 1.8)	0.73	0.94	(0.5, 1.9)	0.88
Field of study									
Psychology	2.17	(1.2, 4.1)	0.01	0.86	(0.4, 1.8)	0.68	0.68	(0.3, 1.4)	0.30
Other	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Yrs. clinical/educ. experience									
<1	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
1–2	0.78	(0.3, 2.1)	0.61	2.07	(0.7, 5.9)	0.18	2.37	(0.8, 7.1)	0.13
3–6	0.50	(0.2, 1.1)	0.10	1.34	(0.6, 3.2)	0.52	1.46	(0.6, 3.5)	0.40
6+	0.58	(0.1, 2.6)	0.48	0.36	(0.1, 1.5)	0.17	0.40	(0.1, 1.6)	0.20
Missing	0.08	(0.0, 1.2)	0.07	0.23	(0.0, 3.4)	0.28	1.61	(0.1, 24.3)	0.73
Yrs. ASD experience									
<1	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
1–2	1.43	(0.8, 2.7)	0.27	1.05	(0.5, 2.1)	0.88	1.40	(0.7, 2.9)	0.37
3–6	1.11	(0.3, 3.9)	0.87	2.07	(0.5, 8.0)	0.29	1.93	(0.5, 7.1)	0.32
6+	0.30	(0.1, 1.5)	0.14	11.7	(1.7, 79.7)	0.01	11.0	(1.9, 63.8)	0.01
Missing	2.83	(0.3, 31.6)	0.40	1.63	(0.1, 41)	0.77	N/A	N/A	N/A
ASD familiarity									
High	1.94	(0.6, 6.0)	0.25	1.25	(0.4, 4.3)	0.72	0.83	(0.2, 2.8)	0.76
Average	0.90	(0.3, 2.3)	0.84	0.91	(0.3, 2.5)	0.85	1.00	(0.4, 2.8)	0.99
Low	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Missing	1.19	(0.2, 7.8)	0.86	1.44	(0.1, 17.9)	0.78	1.26	(0.1, 16.0)	0.86
Study characteristics									
Recruitment group									
No ASD suspected	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
ASD suspected	0.44	(0.2, 0.8)	0.01	0.96	(0.5, 1.8)	0.90	1.17	(0.6, 2.2)	0.64
Site									
A	2.68	(1.2, 6.0)	0.02				0.69	(0.3, 1.6)	0.40
B	3.03	(0.9, 10.1)	0.07				1.59	(0.4, 6.0)	0.49
C	3.73	(1.0, 14.5)	0.06				4.00	(0.5, 32.8)	0.20
D	1.53	(0.7, 3.1)	0.24				0.60	(0.3, 1.3)	0.18
E	2.01	(0.7, 5.9)	0.21				2.24	(0.5, 10.5)	0.31
F	0.97	(0.3, 2.8)	0.95				0.59	(0.2, 1.8)	0.36
G	4.17	(1.5, 11.8)	0.01				1.11	(0.4, 3.1)	0.84
H	Ref	Ref	Ref				Ref	Ref	Ref
Instrument order									
First	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Second	0.88	(0.5, 1.6)	0.67	1.24	(0.7, 2.2)	0.47	0.57	(0.3, 1.0)	0.07
Third	1.01	(0.6, 1.7)	0.98	1.19	(0.7, 2.1)	0.56	0.60	(0.3, 1.2)	0.13
Visit order									
Same day	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Study visit/neurodev eval	0.98	(0.3, 2.9)	0.97	0.53	(0.2, 1.2)	0.11	0.43	(0.1, 1.4)	0.17
Neurodev eval/study visit	1.89	(0.7, 5.2)	0.22	0.46	(0.2, 1.1)	0.08	0.36	(0.1, 1.1)	0.07
Missing	0.82	(0.2, 2.7)	0.74	0.95	(0.3, 2.7)	0.92	0.55	(0.1, 2.2)	0.40

DISCUSSION

The principal goal of population based epidemiologic research on ASD risk factors is unbiased estimation of measures of association. There were a number of scenarios where the candidate instruments tested here would appear to perform adequately in the context of this type of study. Table 7 lists the five instrument(s) and cutpoint criteria with sensitivity near 50% and specificity near 80%. (Note that the sensitivity and specificity estimates shown are from the full data set, rather than from the cross validation, since these are the best available estimates from our data.) Also shown is the bias introduced in a RR estimate for a risk factor with 10% prevalence and a true RR in a study cohort where the ASD prevalence is 1.5% and two-stage case finding involving the M-CHAT as the stage one screener. The last two columns show estimates of the number of identified cases at the end of the two-stage process and the estimate of the proportion of these that actually have ASD. The bias in all these situations is fairly comparable as are the number of cases and proportion correctly classified. Given this, the two approaches that would seem most feasible would be the STAT-E alone and the Parallel ASI plus E-VAS approaches (both using the cutpoints established under the sensitivity=50% criteria).

As a direct observation tool, the STAT-E could be easily integrated into a study that planned in-person visits. Moreover, there could be potential to enhance the performance of the STAT-E above what was achieved here with additional training or quality control measures. These steps come with additional cost but could be easily integrated into a comprehensive study visit-type protocol. The STAT-E requires some data entry post assessment, and thus carries additional staff time cost, but respondent time with this single instrument approach would be minimized. The appeal of the ASI plus E-VAS approach is that it could be implemented without an in-person study visit. The E-VAS could be completed online by caregivers at home and the ASI can be done by phone interview. Respondent time burden would be higher, but staff time burden would be lower.

Kim and Lord [2012] report on the sensitivity and specificity of the ADOS and ADI-R in a sample of children (a broader age range than the children here) who were participating in research studies where they received neurodevelopmental evaluations. They indicate that optimal performance (sensitivity and specificity both consistently over 80%) of these tools occurs when they are both administered and parallel criteria are used. If we used sensitivity and specificity of 85% to approximate the performance of the ADOS and ADI-R and assumed that these tools were used as the second-stage case confirmation approach in the example just discussed, the percent bias in the RR would be 5.5%, 1,180 ASD cases would be identified of which 90% would be true cases. So while this traditional approach does offer reduction in bias and increases in the numbers of identified cases and the proportion of these who are true positives, it is not apparent that these gains would justify the increased costs. At the other extreme, if the M-CHAT alone were used to identify cases without a second stage, bias would increase to over 20%, 1,250 screen positives would be considered cases and only 61% of these would actually have ASD. So with relatively little additional cost, especially if study contact at three years would be part of a general cohort follow-up plan, inclusion of the second-stage approaches highlighted above seem to offer substantive gains.

We found no evidence that performance of these instruments was strongly affected by subject demographic or assessor experience factors. Nor were there strong, consistent effects of assessment order, visit order, or site on instrument performance. However, level of impairment, as measured by subject's verbal ability and, more profoundly, their Mullen early learning composite score did influence the performance of these measures. Notably, a child with a Mullen ELC <70 was significantly more likely to be classified by all measures as a case, regardless of true case status. Therefore, ELC<70 acted to enhance sensitivity and limit specificity of all tools. This suggests that perhaps cognitive ability should be considered when developing items or cutpoints for these instruments.

We considered exploring this further here, but because 96 subjects (25%) were missing Mullen data, sample size was a limitation (only 24 non-ASD subjects had Mullen ELC scores <=70). Moreover, if an epidemiologic study is seeking to evaluate outcome in a streamlined, low-resource manner, it may not be tractable to build cognitive status assessment into the data collection scheme. Because verbal ability can be accurately assessed through a single caregiver report question, and because data on this were available for all subjects, we did go ahead and develop separate cutpoints for verbal and nonverbal children on the E-VAS and the STAT-E (recall that ASI cutpoints were already developed separately for verbal and nonverbal groups since initial ASI cutpoints were established that way). However, performance based on these cutpoints was not markedly improved over that seen for the cutpoints developed for all subjects together (see Appendix Table A1). Nonetheless, further exploration of the influence of cognitive functioning on the performance of these tools could lead to cutpoints offering performance improvements.

This study had a number of strengths. The sample size was large for an ASD validation study and the participation of multiple sites increases the diversity and potential for generalizability of the sample. All subjects underwent

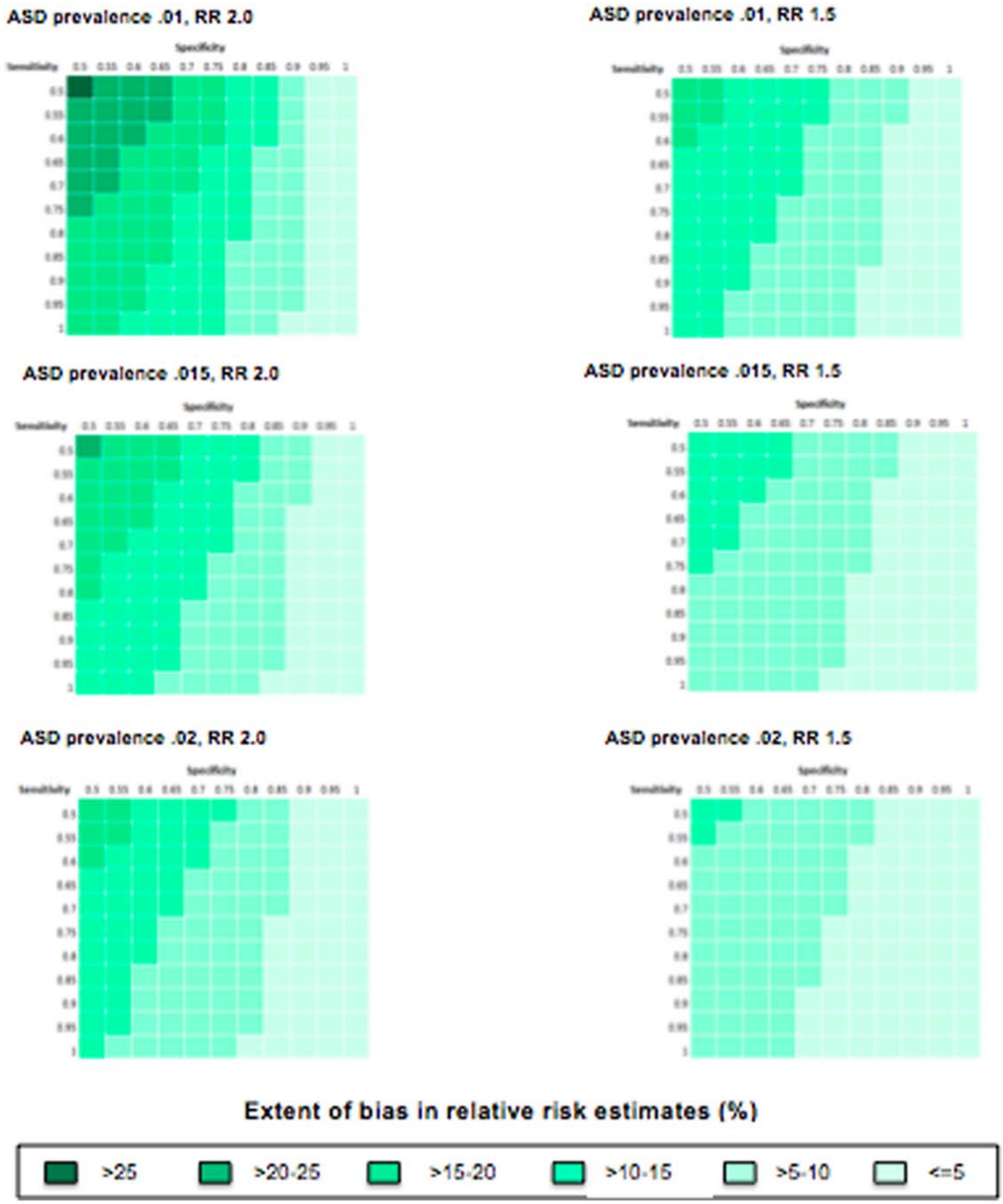

Figure 1. Extent of bias in relative risk estimates at differing stage-two ASD case confirmation sensitivities and specificities for different combinations of ASD prevalence (1%, 1.5%, 2%) and true relative risks (1.5, 2.0) with risk factor prevalence at 10%.

neurodevelopmental evaluations at established ASD research or clinical assessment centers. ADOS data were available on the majority of subjects and the use of more conservative criteria for gold standard case confirmation involving both a clinical diagnosis and ADOS administration did not affect results. Our data entry tool included built-in data quality checks that minimized the potential for data entry error.

While our sample was comparatively large, there were fewer non-cases and, consequently, specificity estimates were less precise than sensitivity estimates. There were a higher proportion of ASD cases that were confirmed on neurodevelopmental assessment than would be expected in a screen-positive general population sample. Sibling studies were among the type of research studies ongoing at the recruitment and enrollment sites that provided subjects to this project, and subjects recruited from these studies may have been at higher risk of ASD than other like-aged children with similar levels of symptom-based suspicion. Whether or not this

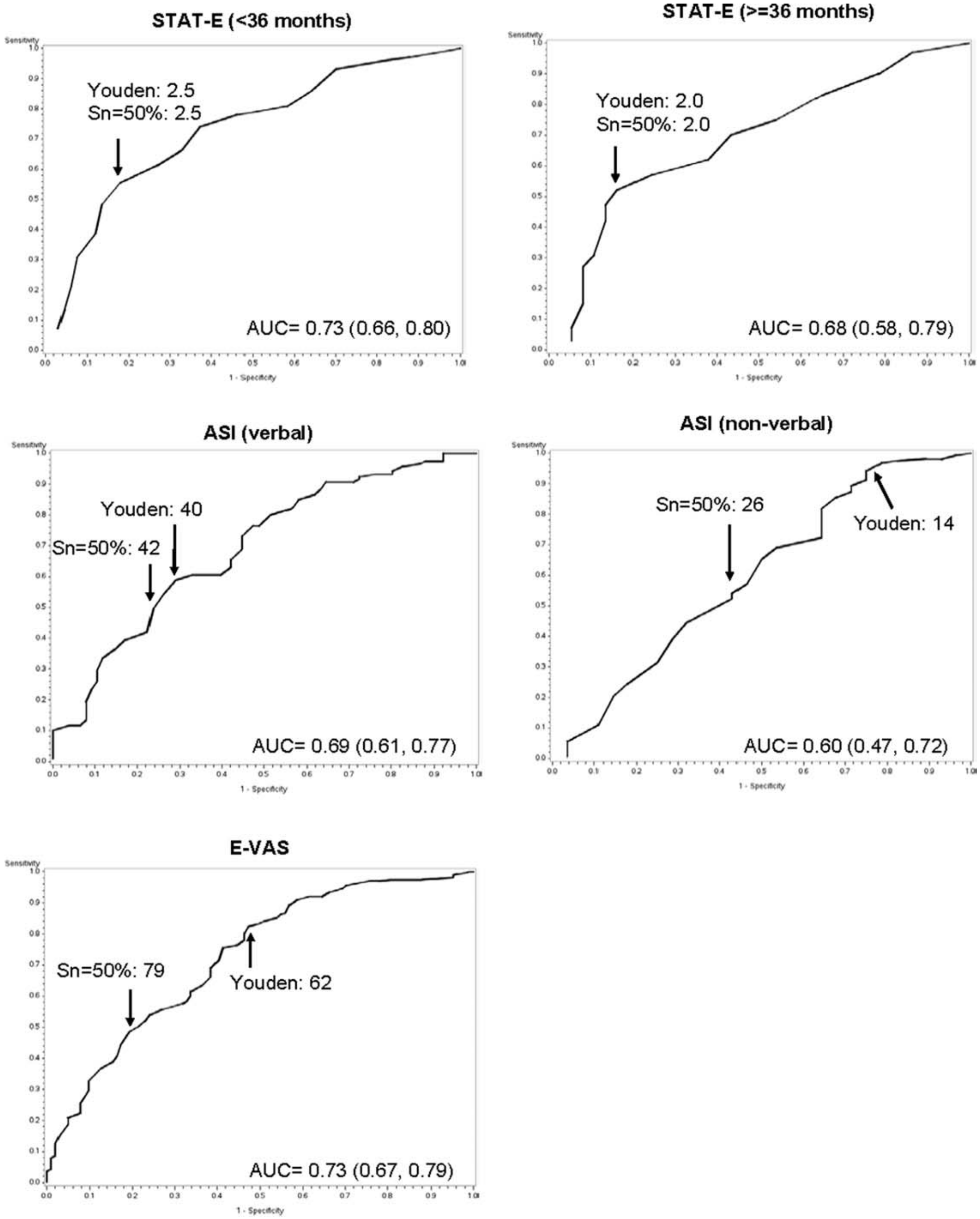

Figure 2. ROC curves by instrument with AUC (95% CI) and alternate cutpoints based on Youden index and sensitivity=50% criteria.

means that our non-cases may have been more impaired or had a higher-level of autism-like features than non-cases in a screen-positive sample is not known, but, if so, could suggest that true specificity might be under-estimated in this sample. We also had some missing data on several covariates and extremely low reporting of medical comorbidities.

The findings suggest that there is potential for less-resource-intensive case-confirmation approaches to be effectively implemented in large population-based epidemiologic research projects that use a two-stage case finding approach. The motivating example we relied on was that of a prospective general population pregnancy cohort. Other epidemiologic designs would involve different considerations. For example, case-control studies might best employ different tools, since the prevalence of true cases in a preliminary case group looking to be confirmed will be much higher than that in a stage-one

Table 6. Two-Fold Cross Validation Estimates of Sensitivity and Specificity of Individual Instruments and Instrument Combinations With Best-Estimate Clinical Diagnosis As the Gold Standard Based on Alternative (Youden Index and Sensitivity=50% Criteria) Cutpoints

	Sensitivity		Specificity	
	Estimate	95% CI	Estimate	95% CI
Youden index Cutpoint				
STAT-E	0.62	(0.56, 0.67)	0.72	(0.62, 0.80)
ASI	0.67	(0.61, 0.73)	0.53	(0.43, 0.63)
E-VAS	0.83	(0.78, 0.87)	0.49	(0.40, 0.60)
Parallel[a] (All)	0.45	(0.39, 0.51)	0.86	(0.77, 0.92)
Serial[b] (All)	0.91	(0.87, 0.94)	0.38	(0.28, 0.48)
Parallel[a] (STAT-E + ASI)	0.47	(0.41, 0.53)	0.85	(0.76, 0.91)
Parallel[a] (STAT-E + E-VAS)	0.55	(0.49, 0.61)	0.81	(0.72, 0.88)
Parallel[a] (ASI + E-VAS)	0.64	(0.58, 0.70)	0.57	(0.47, 0.66)
Sensitivity = 50% Cutpoint				
STAT-E	0.51	(0.45, 0.57)	0.83	(0.74, 0.59)
ASI	0.56	(0.50, 0.62)	0.66	(0.56, 0.75)
E-VAS	0.51	(0.45, 0.57)	0.77	(0.68, 0.85)
Parallel[a] (All)	0.27	(0.22, 0.32)	0.92	(0.85, 0.97)
Serial[b] (All)	0.78	(0.73, 0.83)	0.55	(0.45, 0.65)
Parallel[a] (STAT-E + ASI)	0.33	(0.28, 0.39)	0.91	(0.84, 0.96)
Parallel[a] (STAT-E + E-VAS)	0.31	(0.26, 0.37)	0.91	(0.84, 0.96)
Parallel[a] (ASI + E-VAS)	0.42	(0.37, 0.48)	0.81	(0.72, 0.88)

[a] Parallel combination requires above threshold level on ALL instruments to be considered a test positive.

[b] Serial combination requires above threshold level on ANY instrument to be considered a test positive.

Table 7. Select Instrument(s) and Cutpoint Criteria With Sensitivity and Specificity Estimates As Well As %Bias in Relative Risk, Number ASD Identified Cases, and Proportion of Identified Cases Correctly Classified Under the Assumption the Instruments Are Applied as Second-Stage Case Confirmation Tools (MCHAT Used in Stage One) in a 10,000 Subject Cohort Study With 1.5% ASD Prevalence, a 10% Risk Factor Prevalence, and a True Relative Risk of 2.0

Second-stage case confirmation instrument(s)	Cutpoint criteria	Sen	Spec	RR bias (%)	Identified ASD cases (*n*)	Cases correctly classified (%)
Parallel[1] (STAT-E + ASI)	Initial	0.60	0.75	11.4	947	79.2
Parallel[1] (STAT-E + E-VAS)	Initial	0.61	0.78	10.0	948	81.7
STAT-E	Sn=50%	0.54	0.83	9.1	809	83.4
Parallel[1] (STAT-E + E-VAS)	Youden	0.49	0.89	7.0	686	87.4
Parallel[1] (ASI + E-VAS)	Sn = 50%	0.42	0.82	11.6	669	78.7

screen-positive sample in a birth cohort and, at the same time, an initial pool of controls would have a much lower true ASD prevalence than stage-one screen positives. In addition to study design, ethical concerns will need to be considered in determining which grouping of assessments is most appropriate to administer in the context of a research study. In the motivating example of the general population pregnancy cohort discussed here, presumably families screening positive at stage one would be advised, or referred directly, at that point to a clinical service provider for additional assessment. The extent to which the study itself decided it could or should provide additional clinically actionable assessment to the family should be a topic of discussion among research teams and IRBs.

Finally, it must be emphasized that analyses presented here are not meant to address whether these instruments should be used in the clinic. In clinical application, tradeoffs between sensitivity and specificity will be very different and there will be altered constraints regarding the cost and ease of administration of different types of instruments. Findings here should not be generalized to the clinic, although the raw data we collected on this sample could still be informative for clinical applications and analyses of these data to address questions of clinical utility will be undertaken in the future.

Acknowledgments

Funded by the *Eunice Kennedy Shriver* National Institute for Child Health and Human Development, as National Children's Study Formative Research Project LOI3-QUEX-08 through NICHD Contracts No. HHSN275201 00005I and HHSN275200800033C.

Drs. Sunitha Mathew and Alexa Fraser (Westat, Rockville MD) as well as Drs. Laura Caulfield and Lynn Goldman (Johns Hopkins Bloomberg School of Public Health, Baltimore MD) oversaw the contracts under which this study was performed. Ms. Amy Swanson (Vanderbilt University, Nashville TN) contributed to STAT adaptation. The authors thank Dr. Gitanjali Taneja (NCS Program Office, Bethesda MD) for her support of this project.

Dr. Stone receives a share of author royalties for sales of the original STAT. Dr. Bishop receives a share of author royalties for sales of the Autism Diagnostic Observation Schedule, 2nd Edition (ADOS-2). Any royalties generated as part of her own research or clinical work are donated to charity.

References

Baio, J. (2014). Prevalence of Autism Spectrum Disorder Among Childen Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2010. Morbidity & Mortality Weekly Report, 63(SS02), 1–21.

Burke, J.P., Jain, A., Yang, W., Kelly, J.P., Kaiser, M., Becker, L., Newschaffer, C.J. (2014). Does a claims diagnosis of autism mean a true case? Autism, 18, 321–330.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. Journal of autism and developmental disorders, 39(5), 693–705.

Gotham, K., Risi, S., & Lord, C. (2005). The Autism Diagnostic Observation Schedule (ADOS): revised algorithms for improved diagnostic validity. Presented at the International Meeting For Autism Research (IMFAR), Boston, Massachusetts, May 5–7, 2005.

Gray, K.M., Tonge, B.J., & Sweeney, D.J. (2008). Using the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule with young children with developmental delay: evaluating diagnostic validity. Journal of Autism and Developmental Disorders, 38, 657–667.

Hertz-Picciotto, I., Croen, L. A., Hansen, R., Jones, C. R., van de Water, J., & Pessah, I. N. (2006). The CHARGE study: An epidemiologic investigation of genetic and environmental factors contributing to autism. Environmental Health Perspectives, 1119–1125.

Interagency Autism Coordinating Committee (IACC). (2012). IACC Strategic Plan for Autism Spectrum Disorder (ASD) Research - 2012 Update. Retrieved from http://iacc.hhs.gov/strategic-plan/2012/index.shtml.

Kim, S. H. & Lord, C. (2012). Combining information from multiple sources for the diagnosis of autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months of age. Journal of Child Psychology and Psychiatry, 53, 143–151.

López-Ratón, M., Rodrıguez-Alvarez, M. X., Cadarso-Suárez, C., & Gude-Sampedro, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests.

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, 659–685.

Lyall, K., Schmidt, R. J., & Hertz-Picciotto, I. (2014). Maternal lifestyle and environmental risk factors for autism spectrum disorders. International Journal of Epidemiology, 43, 443–464.

Mazumdar, M., Smith, A., & Bacik, J. (2003). Methods for categorizing a prognostic variable in a multivariable setting. Statistics in Medicine, 22, 559–571.

Olsen, J., Melbye, M., Olsen, S. F., Sørensen, T. I., Aaby, P., Andersen, A.-M. N., Schow, T. B. (2001). The Danish National Birth Cohort-its background, structure and aim. Scandinavian Journal of Public Health, 29, 300–307.

Panel on the Design of the National Children's Study, Implications for the Generalizability of Results, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Board on Children Youth and Families, Institute of Medicine, & National Research Council. (2014). The National Children's Study 2014: An Assessment. Washington (DC): National Academies Press (US).

Robins, D.L., Casagrande, K., Barton, M., Chen, C.-M.A., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). Pediatrics, 133, 37–45.

Schendel, D.E., Bresnahan, M., Carter, K.W., Francis, R.W., Gissler, M., Grønborg, T. K., Hultman, C.M. (2013). The international collaboration for autism registry epidemiology (iCARE): multinational registry-based investigations of autism risk factors and trends. Journal of Autism and Developmental Disorders, 43, 2650–2663.

Schendel, D. E., DiGuiseppi, C., Croen, L. A., Fallin, M. D., Reed, P. L., Schieve, L. A., Levy, S. E. (2012). The Study to Explore Early Development (SEED): a multisite epidemiologic study of autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) network. Journal of Autism and Developmental Disorders, 42, 2121–2140.

Stoltenberg, C., Schjølberg, S., Bresnahan, M., Hornig, M., Hirtz, D., Dahl, C., Alsaker, E. (2010). The Autism Birth Cohort: A paradigm for gene–environment–timing research. Molecular Psychiatry, 15, 676–680.

Stone, W., Coonrod, E., Turner, L., & Pozdol, S. (2004). Psychometric properties of the STAT for early autism screening. Journal of Autism and Developmental Disorders, 34, 691–701.

Stone, W., & Ousley, O. (2008). Screening tool for Autism user's manual. Unpublished manuscript - Vanderbilt University.

Youden, W., & Cameron, J. (1950). Use of Statistics to Determine Precision of Test Methods. ASTM Special Tech. Publ (103).

Appendix

Table A1. Two-Fold Cross Validation Estimates of Sensitivity and Specificity of Individual Instruments and Instrument Combinations With Best-Estimate Clinical Diagnosis as the Gold Standard Based on Alternative (Youden Index and Sensitivity = 50% Criteria) Cutpoints That Were All Derived Separately for Verbal and Nonverbal Subjects

	Sensitivity		Specificity	
	Estimate	95% CI	Estimate	95% CI
Youden index Cutpoint				
STAT-E	0.68	(0.62, 0.73)	0.61	(0.51, 0.70)
ASI	0.67	(0.61, 0.73)	0.53	(0.43, 0.63)
E-VAS	0.53	(0.47, 0.59)	0.65	(0.55, 0.74)
Parallel[a] (All)	0.32	(0.26, 0.37)	0.84	(0.75, 0.90)
Serial[b] (All)	0.91	(0.87, 0.94)	0.34	(0.25, 0.44)
Parallel[a] (STAT-E + ASI)	0.49	(0.43, 0.55)	0.78	(0.69, 0.85)
Parallel[a] (STAT-E + E-VAS)	0.35	(0.29, 0.40)	0.83	(0.74, 0.89)
Parallel[a] (ASI + E-VAS)	0.46	(0.40, 0.52)	0.68	(0.58, 0.77)
Sensitivity = 50% Cutpoint				
STAT-E	0.63	(0.57, 0.68)	0.60	(0..50, 0.69)
ASI	0.56	(0.50, 0.62)	0.66	(0.56, 0.75)
E-VAS	0.53	(0.47, 0.59)	0.70	(0.60, 0.79)
Parallel[a] (All)	0.32	(0.27, 0.38)	0.87	(0.78, 0.92)
Serial[b] (All)	0.83	(0.79, 0.88)	0.39	(0.30, 0.49)
Parallel[a] (STAT-E + ASI)	0.40	(0.34, 0.46)	0.83	(0.74, 0.89)
Parallel[a] (STAT-E + E-VAS)	0.37	(0.31, 0.43)	0.84	(0.75, 0.90)
Parallel[a] (ASI + E-VAS)	0.44	(0.38, 0.50)	0.78	(0.68, 0.85)

[a] Parallel combination requires above threshold level on ALL instruments to be considered a test positive

[b] Serial combination requires above threshold level on ANY instrument to be considered a test positive.