THE JOURNAL OF CHILD PSYCHOLOGY AND PSYCHIATRY

ACAMH THE ASSOCIATION FOR CHILD AND ADOLESCENT MENTAL HEALTH

Journal of Child Psychology and Psychiatry **:* (2015), pp **–** doi:10.1111/jcpp.12421

Diagnostic stability in young children at risk for autism spectrum disorder: a baby siblings research consortium study

Sally Ozonoff,[1] Gregory S. Young,[1] Rebecca J. Landa,[2] Jessica Brian,[3] Susan Bryson,[4] Tony Charman,[5] Katarzyna Chawarska,[6] Suzanne L. Macari,[6] Daniel Messinger,[7] Wendy L. Stone,[8] Lonnie Zwaigenbaum,[9] and Ana-Maria Iosif[10]

[1]MIND Institute, University of California Davis, Sacramento, CA, USA; [2]Kennedy Krieger Institute, Baltimore, MD, USA; [3]Bloorview Research Institute, University of Toronto, ON, Canada; [4]Department of Pediatrics, Dalhousie University, Halifax, NS, Canada; [5]Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, Kings College London, London, UK; [6] Child Study Center, Yale University School of Medicine, New Haven, CT, USA; [7]Department of Psychology, University of Miami, Coral Gables, FL, USA; [8]Department of Psychology, University of Washington, Seattle, WA, USA; [9]Department of Pediatrics, University of Alberta, Edmonton, AB, Canada; [10]Department of Public Health Sciences, University of California Davis, Davis, CA, USA

Background: The diagnosis of autism spectrum disorder (ASD) made before age 3 has been found to be remarkably stable in clinic- and community-ascertained samples. The stability of an ASD diagnosis in prospectively ascertained samples of infants at risk for ASD due to familial factors has not yet been studied, however. The American Academy of Pediatrics recommends intensive surveillance and screening for this high-risk group, which may afford earlier identification. Therefore, it is critical to understand the stability of an ASD diagnosis made before age 3 in young children at familial risk. **Methods:** Data were pooled across seven sites of the Baby Siblings Research Consortium. Evaluations of 418 later-born siblings of children with ASD were conducted at 18, 24, and 36 months of age and a clinical diagnosis of ASD or Not ASD was made at each age. **Results:** The stability of an ASD diagnosis at 18 months was 93% and at 24 months was 82%. There were relatively few children diagnosed with ASD at 18 or 24 months whose diagnosis was not confirmed at 36 months. There were, however, many children with ASD outcomes at 36 months who had not yet been diagnosed at 18 months (63%) or 24 months (41%). **Conclusions:** The stability of an ASD diagnosis in this familial-risk sample was high at both 18 and 24 months of age and comparable with previous data from clinic- and community-ascertained samples. However, almost half of the children with ASD outcomes were not identified as being on the spectrum at 24 months and did not receive an ASD diagnosis until 36 months. Thus, longitudinal follow-up is critical for children with early signs of social-communication difficulties, even if they do not meet diagnostic criteria at initial assessment. A public health implication of these data is that screening for ASD may need to be repeated multiple times in the first years of life. These data also suggest that there is a period of early development in which ASD features unfold and emerge but have not yet reached levels supportive of a diagnosis. **Keywords:** Preschool children, autism spectrum disorders, diagnosis.

Introduction

The stability of an autism spectrum disorder (ASD) diagnosis made at a young age is of high interest, given the impact of early intervention, the provision of which requires early identification. While studies performed over the past two decades robustly demonstrated a high degree of stability in children aged 3 years or older at first diagnosis (Woolfenden, Sarkozy, Ridley, & Williams, 2012), there was initial concern about the stability of diagnosis for children identified before age 3. Both clinicians and researchers raised important questions, given the costs of early autism treatment, about the youngest age at which a reliable diagnosis could be made. In many communities, there was a general reluctance to diagnose children before age three. Questions about the permanence of diagnosis have been highlighted by recent empirical reports of children who, in middle or later childhood, no longer meet criteria for ASD (Anderson, Liang, & Lord, 2014; Fein et al., 2013; Orinstein et al., 2014). However, in recent years, multiple studies have demonstrated impressive stability in children diagnosed before three years as well, with a meta-analysis reporting an overall stability rate of 86.3% for maintaining an ASD diagnosis over time (Rondeau et al., 2011). Similar findings were reported by Woolfenden et al. (2012) in a systematic review of 10 studies of toddlers diagnosed before their third birthday.

A review of stability and other classification indices from all previous studies of children younger than 36 months at first diagnosis was conducted and can be seen in Table 1. The table aggregates across ASD subtypes and uses dichotomous classifications of ASD and Not ASD. The first set of studies reported in Table 1 followed only children with a diagnosis of ASD and did not include a comparable sample of children without ASD. Positive predictive value, which reflects stability of the diagnosis, is high, with

Conflict of interest statement: No conflicts declared.

Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA 02148, USA

Table 1 Previously published stability studies of children diagnosed with autism spectrum disorder (ASD) before age 3

	ASD *n*	Not ASD *n*	Time 1 age	Time 2 age	True positives	False positives	False negatives	True negatives	Sensitivity (%)	Specificity (%)	Positive predictive value (Stability) (%)	Negative predictive value (%)
All ASD, no non-spectrum												
Stone 1999	37	0	31 m	43 m	31	6					84	
Takeda 2005	57	0	31 m	67 m	57	0					100	
Turner 2007	48	0	2 y	4 y	30	18					63	
Paul 2008	37	0	22 m	37 m	37	0					100	
Itzchak 2009	68	0	25 m	37 m	66	2					97	
Both ASD & non-spectrum – clinically ascertained												
Eaves 2004	43	6	33 m	57 m	40	3	0	6	100	67	93	100
Lord 2006	130	42	2 y	9 y	124	6	11	31	92	84	95	74
Chawarska 2007	27	4	14–25 m	-	27	0	1	3	96	100	100	75
Sutera 2007	73	17	16–30 m	-	60	13	0	17	100	57	82	100
Kleinman 2008	61	16	27 m	53 m	46	15	0	16	100	52	75	100
Chawarska 2009	61	28	22 m	47 m	61	0	3	25	95	100	100	89
Worley 2011	53	61	23 m	31 m	38	15	12	49	76	77	72	80
Corsello 2013	26	6	30 m	3–8 y	20	5	2	4	91	44	80	67
Both ASD & non-spectrum – community ascertained												
Cox 1999	12	38	20 m	42 m	12	0	9	29	57	100	100	76
Ventola 2007	46	17	27 m	-	38	8	0	17	100	68	83	100
van Daalen 2009	53	78	26 m	45 m	46	7	2	76	96	92	87	97
Guthrie 2013	56	26	19 m	37 m	56	0	3	23	95	100	100	88

Note: m, months; y, years; -, age not reported; TN, true negatives; TP, true positives; FN, false negatives; FP, false positives.
Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP); Positive predictive value = TP/(TP + FP); Negative predictive value = TN/(TN + FN).

a range of 63% to 100% across five investigations. Although stability rates and numbers of false positives can be calculated from these studies, they cannot address another important aspect of classification accuracy, false negative rates. False negatives may reflect missed diagnoses, later onset of symptoms, and/or borderline phenotypes that result in initial clinical uncertainty and caution in making early diagnoses. Therefore, longitudinal follow-up of children without autism spectrum diagnoses at the initial evaluation is critical to understanding clinical decision-making, although it is not formally needed for calculation of stability.

The next group of studies reported in Table 1 includes children with and without ASD at Time 1 so that additional classification parameters can be calculated (sensitivity, specificity, etc.). Across 8 studies with clinically ascertained samples, the positive predictive value ranged from 72% to 100% (with half of the studies over 93%) and the negative predictive value ranged from 67% to 100% (half the studies over 89%). These classification indices are highly influenced by the base rate of the condition in the samples studied (Altman & Bland, 1994). Using samples ascertained from clinics, where there has already been a degree of concern raised that was sufficient to bring the child to clinical attention, is likely to increase the base rate of ASD, in turn biasing rates of false positives and negatives, and increasing stability estimates.

Community-ascertained samples have the potential to provide less biased psychometric indices of classification accuracy. Four such studies are summarized last in Table 1. The positive predictive value ranged from 83% to 100%, comparable to the estimates for clinically ascertained samples. For practical reasons, many community-based studies employ a pre-screening design, in which only those who screen positive at Time 1 are followed longitudinally. For example, Van Daalen et al. (2009) screened 31,724 children through primary care visits at 14 months of age and then followed 131 of the screen-positives for 12 months to calculate stability indices. Similarly, Guthrie, Swineford, Nottke, and Wetherby (2013) performed a two-step screening of 5,419 children in primary care and then followed 82 children who screened positive for two years to provide their estimates of stability. Thus, even in these community-based studies, the base rates of ASD, and thus the stability estimates, may have been overestimated by the screening process and sampling frame.

Another type of sample that may contribute to understanding the stability of early ASD diagnoses is a familial-risk sample. In such studies, participants at familial risk for ASD by virtue of having an older affected sibling are generally enrolled in longitudinal studies in early infancy, before the initial behavioral signs are usually evident (Ozonoff et al., 2010) and prior to when parents begin to report concerns (Hess

& Landa, 2012; Ozonoff et al., 2009). They have not been 'pre-screened' based on symptoms before the initial evaluation, potentially reducing such sampling biases that may influence stability. In addition to identifying young children with ASD outcomes to follow, such samples also identify children with typical development and those with a wide range of clinical presentations, including subclinical difficulties in the core areas associated with ASD (Messinger et al., 2013; Ozonoff et al., 2014). Given the potential for much earlier detection, diagnosis, and treatment of children with a positive family history (Johnson & Myers, 2007; Ozonoff et al., 2011), it is critical to examine the stability of early classification in young children at familial risk of ASD. This study had two aims: (1) to examine the stability at 36 months of a clinical diagnosis of ASD made at 18 and 24 months of age in infants at familial risk for ASD and (2) to explore phenotypic differences among children who were correctly and incorrectly classified at 18 and 24 months. Addressing these aims required a large sample and thus this study utilized information from a multisite cohort of infants whose data were collected as part of an international collaboration to study infants with an older sibling with ASD.

Method

Participants

The Baby Siblings Research Consortium (BSRC) is an international network that, with support from Autism Speaks, pools data from individually funded research sites to study the development of infants at familial risk for ASD. The present analyses were carried out using data contributed from seven sites (University of Alberta, Dalhousie University, Kennedy Krieger Institute, McMaster University, University of California – Davis, University of Toronto, Yale University) whose procedures and common measures permitted data pooling. Informed consent was obtained at each site prior to data collection, as well as Institutional Review Board approval to collect and analyze de-identified data from all sites.

Infant participants were later-born biological siblings of a child with ASD (99% were full siblings). Diverse community enrollment strategies were employed across sites, including recruitment from clinics and agencies serving individuals with ASD, community events (conferences, health fairs) targeted at families affected by ASD, other ASD studies at respective sites' universities, websites targeted to ASD, word of mouth (parents referring other parents), fliers posted in the community, mailings, and media announcements. Inclusion required a documented diagnosis of DSM-IV Autistic Disorder, Asperger Disorder, or Pervasive Developmental Disorder Not Otherwise Specified in the affected older sibling and no identified neurological or genetic condition in the infant or older sibling that could account for an ASD diagnosis (e.g., fragile X syndrome). Additional inclusion criteria were maximum enrollment age of 18 months, outcome assessment age of 36 months, and availability of both a clinical diagnosis (ASD or not ASD) and scores on the Autism Diagnostic Observation Schedule (ADOS) at 18, 24, and 36 months of age. For families with multiple enrolled infants, only the infant recruited at the youngest age was included. All BSRC sites meeting these inclusion criteria were included in the present analyses, resulting in a total sample size of 418 participants across seven sites.

Measures

Clinical best estimate (CBE) diagnosis: Each site had established procedures for making clinical diagnoses at 18, 24, and 36 months, including: (1) ADOS administration by a research-reliable examiner, (2) clinical diagnosis using DSM-IV criteria, (3) diagnosis made or verified by licensed clinicians, and (4) 36-month outcome assessments performed by examiners unaware of risk group and previous diagnostic decisions. Although this study was initiated prior to the publication of DSM-5 and diagnoses were made initially using DSM-IV criteria, in order to be consistent with current practice, and given the inconsistent application of the DSM-IV subcategories (Lord et al., 2012) that may be especially the case in younger children, all clinical diagnoses were dichotomized as ASD or Not ASD for analyses.

Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore, & Risi, 2002): The ADOS is a standardized protocol that measures symptoms of ASD and provides an empirically derived cutoff for ASD that has high inter-rater reliability and construct validity. The 2002 communication + social interaction algorithm score was used because item-level data, necessary for calculation of newer algorithms, was not available from all sites.

Mullen Scales of Early Learning (Mullen, 1995): This is a standardized developmental test for children birth to 68 months that provides T scores (mean = 50, *SD* = 10) for nonverbal cognitive, receptive and expressive language, and gross and fine motor skills. The Mullen scales have excellent internal consistency and test-retest reliability.

Demographic information was collected at each site (see Table 2). Parent-reported race and ethnicity classifications of the infant were collapsed for analysis into two dichotomous variables (Caucasian/Not Caucasian and Hispanic/Not Hispanic). Another dichotomous variable was created indicating whether the infant's family was simplex (one older sibling with ASD) or multiplex (more than one older sibling with ASD).

Statistical approach

Psychometric measures of the performance of a CBE diagnostic classification at 18 and 24 months were computed. Differences in sensitivity and specificity for 18- and 24-month CBE diagnostic classification were tested using McNemar's test (Li & Fine, 2004). The positive and negative predictive values of the 18- and 24-month diagnoses were compared using Wald

Table 2 Characteristics of the sample (*n* = 418)

Age at enrollment in months, mean (*SD*)	7.1 (4.0)
Gender, *n* (%)	
Female	172 (41%)
Male	246 (59%)
Outcome (36 months), *n* (%)	
ASD	110 (26%)
Not ASD	308 (74%)
Race[a], *n* (%)	
Caucasian	308 (83%)
Non-Caucasian	61 (17%)
Hispanic[b], *n* (%)	
No	260 (95%)
Yes	14 (5%)
Multiplex Status[c], *n* (%)	
No	343 (89%)
Yes	44 (11%)

ASD, autism spectrum disorder.

[a]Frequency Missing = 49.

[b]Frequency Missing = 144.

[c]Frequency Missing = 31.

Table 3 Stability and diagnostic classification parameters at 18 and 24 months

	ASD at 36 months		Not ASD at 36 months					
	ASD (True Positives)	Not ASD (False Negatives)	ASD (False Positives)	Not ASD (True Negatives)	Sensitivity (95% CI)	Specificity (95% CI)	PPV/Stability (95% CI)	NPV (95% CI)
18 Months CBE	41	69	3	305	37.3% (28%–47%)	99.0% (97%–100%)	93.2% (81%–99%)	81.6% (77%–85%)
24 Months CBE	65	45	14	294	59.1% (49%–68%)	95.5% (92%–97%)	82.3% (72%–90%)	86.7% (83%–90%)

ASD, autism spectrum disorder; CBE, Clinical Best Estimate; CI, Confidence Interval; TP, True Positives; FN, False Negatives; FP, False Positives; TN, True Negatives. Separate analyses were conducted for 18 and 24 month CBE. Confidence intervals were computed using exact confidence limits.

Sensitivity is the percentage of those diagnosed at 36 months who were identified at the earlier visit, calculated as TP/(TP + FN) × 100.

Specificity is the percentage of those without a diagnosis at 36 months who were correctly identified as Not ASD at the earlier visit, calculated as TN/(TN + FP) × 100.

Positive predictive value (PPV) is the percentage of those identified with ASD at the earlier visit who retain the diagnosis at 36 months, calculated as TP/(TP + FP) × 100. Equivalent to stability.

Negative predictive value (NPV) is the percentage of those identified as Not ASD at the earlier visit who are verified free of diagnosis at 36 months, calculated as TN/(TN + FN) × 100.

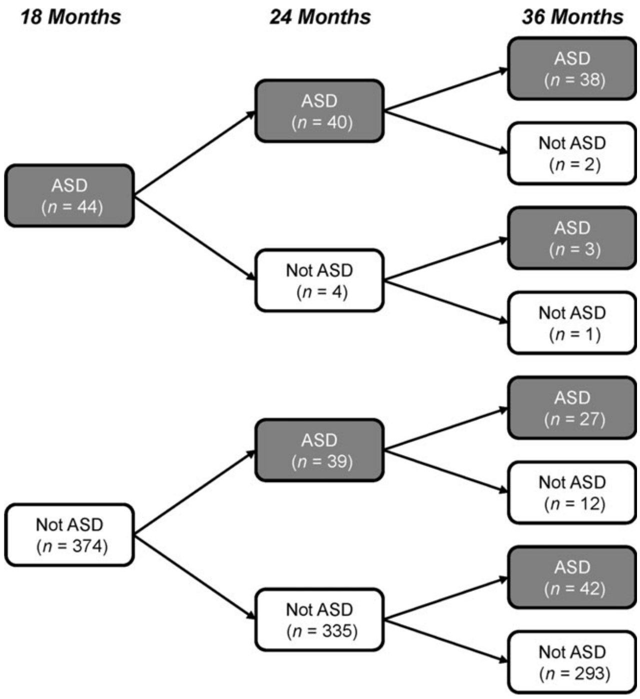

Figure 1 Stability of clinical best estimate outcome classifications across visits

test statistics derived from the weighted least square method for analyses of binary data (Wang, Davis, & Soong, 2006).

To examine group differences in ADOS and Mullen scores at the 18-, 24-, and 36-month visits, mixed-effects linear models (Laird & Ware, 1982) were employed. These models are flexible and allow for unequally spaced and missing observations. All core models included fixed effects for group membership, the linear and the quadratic effect of age (centered at 18 months), and the interaction between group and the linear age effect. To account for the correlated nature of the data, the core models included two random effects for child-specific intercepts and slopes, as well as a random effect for site. Additional fixed terms (for the interaction of the quadratic effect of age with group and for ADOS module) were also added to the core model and tested. These terms were retained in the models only if they were significant.

Residual analyses and graphical diagnostics were used to determine that model assumptions were adequately met. Positive and negative predictive values for 18- and 24-month CBE were compared using the R program SCPVTBT (www.ugr.es/~bioest/software.htm). Mixed-effect analyses were conducted using PROC MIXED in SAS Version 9.4 (SAS Institute). All tests were two-sided, with $\alpha = .05$.

Results

Table 3 provides stability and other classification indices at 18 and 24 months of age (using diagnosis at 36 months as the outcome standard) for this sample of 418 children at familial risk for ASD. More ASD diagnoses were made at 24 months ($n = 79$) than at 18 months ($n = 44$). This results in significant increases in sensitivity ($p < .001$) and decreases in the number of false negatives ($p = .003$) from 18 to 24 months of age. There is also a small but statistically significant decrease ($p = .02$) in positive predictive value from 18 months (93%) to 24 months (82%). This reflects the greater number of false

Table 4 Patterns of Clinical Best Estimate outcome classifications across visits

Clinical Best Estimate Outcome			Total	ASD at 36 months	Not ASD at 36 months	
18 months	24 months	36 months	(n = 418)	(n = 110)	(n = 308)	Classification
A	A	A	38	35%	–	True positives
A	A	N	2	–	0.7%	False positives
A	N	N	1	–	0.3%	False positives
N	A	N	12	–	4%	False positives
A	N	A	3	3%	–	False negatives
N	A	A	27	25%	–	False negatives
N	N	A	42	38%	–	False negatives
N	N	N	293	–	95%	True negatives

ASD, autism spectrum disorder; A, ASD; N, Not ASD.

positives at 24 months (n = 14) than at 18 months (n = 3). The 18- and 24-month stability rates in this familial-risk sample fall within the range of, and are consistent with, the stability rates for children under age 3 in clinic- and community-ascertained samples reviewed in Table 1.

As depicted in Figure 1, eight patterns of stability are generated when a dichotomous diagnostic decision (ASD or Not ASD) is made at three ages. Some children are consistently identified as ASD or Not ASD (i.e., AAA or NNN patterns in Table 4), others are classified in a way that evolves over time, in both directions (i.e., ANN, AAN, NAA, NNA), and still others move back and forth between ASD and Not ASD classifications at different ages (i.e., ANA, NAN). Due to the very small sample sizes in several of the subgroups and to allow for comparison with other studies that use the language of classification science (e.g., true and false positives and negatives), we consolidated the 8 patterns into four conservatively defined stability groups. Diagnosis at 36 months was used as the gold standard. A stable 'positive' early assessment was defined as meeting criteria for ASD at 18 *and* 24 months (e.g., True Positives [TP] = AAA), while a stable 'negative' early assessment was defined as not meeting criteria for ASD at both 18 *and* 24 months (e.g., True Negatives [TN] = NNN). The unstable groups were also defined conservatively, in that a classification at *either* 18 *or* 24 months that differed from the classification at 36 months led to inclusion in these groups. Thus, False Positives [FP] met ASD criteria at 18 and/or 24 months but not 36 months, while False Negatives [FN] failed to meet ASD criteria at 18 and/or 24 months but did at 36 months. The resulting classifications can be seen in Table 4.

Table 5 presents estimated means and 95% confidence intervals from the mixed-models for ADOS and Mullen scores for the four stability groups. Full details of these models are provided in Table S1. Five sets of group differences were of interest (comparisons of the FP and FN groups to the TP and TN groups, as well as to each other) and are summarized in Table 5 and Figure 2.

At 18 and 24 months, the clinical features of the FN group were intermediate between the TP and TN groups. They had higher Mullen and lower ADOS scores than the TP group, but lower Mullen and higher ADOS scores than the TN group, suggesting that, although they were not yet diagnosed with ASD, they were atypical at 18 and 24 months. By 36 months, the FN and TP groups had similar ADOS scores, but the FN group's Mullen remained higher than that of the TP group.

The patterns of group differences were quite similar for the FP group, who, like the FN group, demonstrated Mullen and ADOS scores that were intermediate between and significantly different from both the TP and TN groups at 18 and 24 months. At 36 months, the Mullen scores of the FP group remained lower and their mean ADOS score was still higher than the TN group, so they demonstrated continued atypical development. However, their 36-month ADOS scores now differed from the TP group.

We found no statistically significant differences between the FP and FN groups at either 18 or 24 months; in addition, the confidence intervals were almost completely overlapping on every measure at both ages (see Table 5) and the magnitude of the estimated differences was modest at both 18 and 24 months on all scales. At 36 months, there continued to be no differences on the Mullen scales, but the FN group now had a significantly higher ADOS score than the FP group.

Discussion

This study had two aims: (1) to examine the stability at 36 months of a clinical diagnosis of ASD made at 18 and 24 months in young children at familial risk for ASD, and (2) to explore phenotypic differences among children who were correctly and incorrectly classified at 18 and 24 months. The familial-risk design had a number of strengths. Improving upon previous studies, three longitudinal visits were conducted, the ages of which corresponded to screening ages recommended by the American Academy of Pediatrics (AAP; Johnson & Myers, 2007). In addition, the familial-risk cohort was not biased by clinical ascertainment or by the prescreening selection methods often applied to community-based samples.

Regarding Aim 1, the stability rates (i.e., positive predictive value estimates) of 93% at 18 months and 82% at 24 months in this familial-risk sample were comparable to previous studies of both clinically and community-ascertained samples younger than age three. The consistent positive predictive value across different types of samples provides some reassurance that previously published stability rates were not overly influenced by ascertainment methods. The high rates of diagnostic stability across studies and methodologies indicate that when ASD is identified at 18 or 24 months, the diagnosis is very likely to be retained, so implementation of treatment should begin as soon as possible.

The low sensitivity of an ASD diagnosis at 18 months and the decrease in stability from 18 to 24 months suggest that there may have been age-dependent differences in clinical calibration operating in this familial-risk sample. It appears that at 18 months, clinicians monitored their decision-making such that if the clinical picture was not certain, they waited to make the diagnosis until later. Indeed, the ratio of false negatives to false positives approached 5:1, suggesting that clinicians' ratings were conservative and biased towards committing as few positive identification errors as possible. But when clinicians were confident in identifying the phenotype, even at early ages (e.g., 18 months) and did make a diagnosis, they were generally correct and it was verified at subsequent visits. Another explanation for differences in clinical decision-making at the two ages may lie in the subclinical social and communication difficulties that have been documented in even very young siblings of children with ASD (Landa & Garrett-Mayer, 2006; Landa, Holman, & Garrett-Mayer, 2007; Messinger et al., 2013; Ozonoff et al., 2014). Clinicians in this study needed to differentiate between emerging signs of ASD and subclinical features more consistent with the broader autism phenotype, a much more subtle distinction than ordinarily faced in a clinic setting. This may have encouraged clinicians in the current investigation to diagnose only the most affected children at 18 months of age.

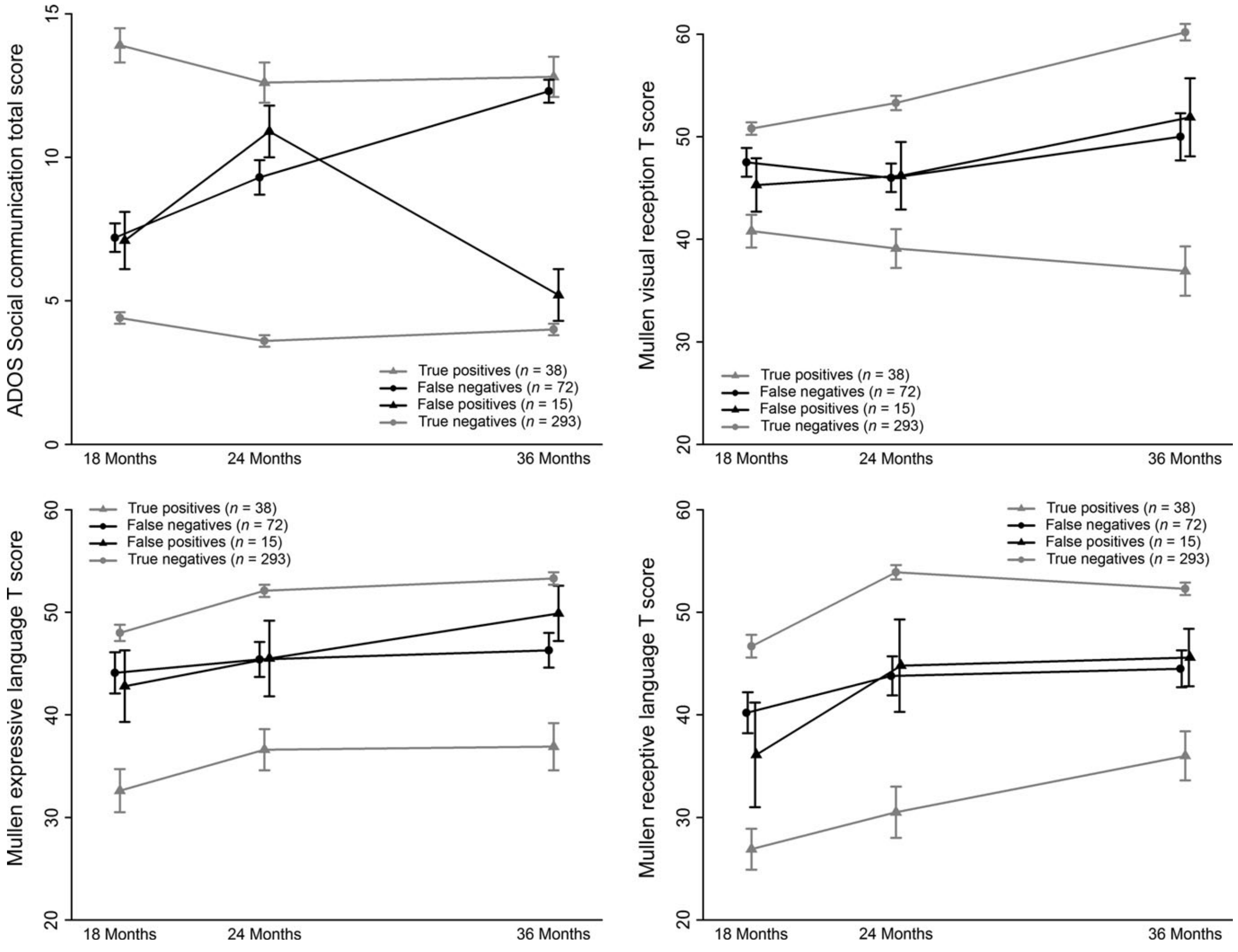

Figure 2 Means ± 1 standard errors for 18, 24, and 36 month ADOS and Mullen scores for the four stability groups

Table 5 Estimated scores and 95% confidence intervals for the four stability groups

Variable	True positives (n = 38)	False negatives (n = 72)	False positives (n = 15)	True negatives (n = 293)
ADOS Social-Communication Score				
18 months	14.0 (12.6–15.4)	7.0 (6.0–8.1)[a,b]	6.9 (4.7–9.0)[a,b]	4.5 (4.1–4.9)
24 months	12.7 (11.5–14.0)	9.2 (8.3–10.1)[a,b]	10.8 (8.8–12.7)[b]	3.5 (3.2–3.9)
36 months	12.4 (11.1–13.7)	11.6 (10.6–12.6)[b,c]	5.2 (3.2–7.2)[a,b,c]	3.0 (2.4–3.7)
Mullen Expressive Language T Score				
18 months	32.9 (28.5–37.3)	42.0 (38.2–45.8)[a,b]	40.6 (34.0–47.3)[a,b]	48.4 (46.1–50.8)
24 months	36.5 (32.7–40.4)	45.1 (41.9–48.3)[a,b]	45.2 (39.5–50.9)[a,b]	52.2 (50.0–54.4)
36 months	37.9 (33.9–41.9)	45.4 (42.1–48.8)[a,b]	48.4 (42.4–54.4)[a]	53.7 (51.6–55.9)
Mullen Receptive Language T Score				
18 months	26.6 (21.4–31.8)	37.9 (33.8–42.1)[a,b]	34.5 (25.9–43.1)[b]	48.0 (45.3–50.7)
24 months	34.1 (29.8–38.5)	44.2 (40.7–47.6)[a,b]	42.3 (35.6–49.1)[a,b]	54.2 (51.7–56.7)
36 months	36.4 (31.9–40.9)	43.8 (40.4–47.2)[a,b]	45.2 (38.6–51.7)[a,b]	53.7 (51.4–56.1)
Mullen Visual Reception T Score				
18 months	42.4 (38.7–46.1)	45.5 (42.6–48.5)[b]	44.4 (38.7–50.2)[b]	51.6 (49.5–53.6)
24 months	40.0 (36.6–43.4)	45.8 (43.1–48.5)[a,b]	45.7 (40.5–50.8)[a,b]	53.7 (51.7–55.7)
36 months	37.9 (32.8–42.9)	48.9 (45.2–52.6)[a,b]	50.8 (43.3–58.2)[a,b]	60.5 (58.2–62.8)
Mullen Fine Motor T Score				
18 months	44.3 (41.6–47.1)	50.1 (47.4–52.7)[a,b]	47.4 (43.1–51.8)[b]	52.8 (51.3–54.4)
24 months	39.4 (36.8–42.0)	45.1 (42.9–47.3)[a,b]	44.4 (40.4–48.4)[a,b]	51.0 (49.6–52.4)
36 months	34.0 (29.8–38.2)	39.7 (36.5–43.0)[a,b]	42.9 (36.5–49.2)[a,b]	52.0 (50.3–53.7)

For ADOS the estimates are for Module 1 scores; scores on Module 2 were 1.2 points higher.

[a] Significant differences ($p < .05$) from true positives.

[b] Significant differences ($p < .05$) from true negatives.

[c] Significant differences ($p < .05$) between false positives and false negatives groups.

While negative predictive value at 18 months was respectable (81.6%), the number of false negatives was quite high. For many families who already have a child with ASD, hearing that their 18-month-old does not meet criteria for a diagnosis will not be reassuring, given that the rate of missed diagnoses (18.4%) at this age is close to or higher than previously published recurrence rates for ASD (Ozonoff et al., 2011; Sandin et al., 2014). One public health implication of this study is that screening may need to be repeated *after* 24 months, since many toddlers with ASD in this sample were not identified until three years of age. While the AAP's screening guidelines (Johnson & Myers, 2007) were a step forward for clinical practice, our data suggest that they may need to go further still. For example, our results suggest that rescreening high-risk groups (e.g., siblings of children with ASD, children with developmental delays) at three years of age will identify some children whose ASD symptoms were not apparent at earlier ages.

The second aim of this study was to examine what differentiates the diagnostically stable and unstable groups. The FP and FN groups demonstrated an intermediate phenotype, with higher developmental levels and fewer ASD features than the TP group, but lower developmental functioning and more ASD symptoms than the TN group. The FP and FN groups were very similar to each other in global scores on the developmental and diagnostic tests at 18 and 24 months, so it is intriguing to speculate on the factors involved in clinical decision-making that led a clinician to diagnose one child with ASD and to classify another child with similar scores as non-ASD. There may have been particular symptom patterns that, when present, influenced clinicians to make (or not make) a clinical diagnosis. For example, a recent study identified several features at 18 months that were especially predictive of an ASD diagnosis, such as poor eye contact, lack of communicative gestures, and repetitive behaviors (Chawarska et al., 2014). It is possible that, even with similar ADOS algorithm scores, the FP and FN groups differed in individual symptoms or constellations of symptoms. Factors not measured in this study, such as medical and developmental history, level of parent or pediatrician concern, or delays in additional areas, such as motor or adaptive functioning, may also have influenced clinicians to make versus hold off on a diagnosis at 18 and 24 months.

At each age, the FN group demonstrated significantly higher developmental functioning on the Mullen than the TP group. One interpretation of these data is that the FN group was composed of higher functioning children with ASD who had a later onset of symptoms or whose symptoms were subtle at first and masked by age-appropriate language and cognitive abilities. These results are convergent with the results of a recent study that employed a data mining approach, rather than a CBE diagnostic process, to classify ASD at 18 months (Chawarska et al., 2014). In that study, a decision-tree learning algorithm identified correctly over half of the ASD cases at 18 months, but missed those who had less pronounced developmental delays and fewer symptoms of ASD. This suggests that the high rate of false negatives in this study might be linked with the developmental dynamics observed in young children developing ASD, rather than with particular classification methods.

At 36 months, the FP group continued to demonstrate significantly lower Mullen and higher ADOS scores than the TN group. Thus, they continued to experience developmental difficulties, even though they no longer met criteria for an ASD diagnosis. More differentiated clinical outcomes were assigned at 36 months at each participating site. Of the 15 children in the False Positive group, only two were considered to be typically developing or have no diagnosis at 36 months. Over half (9 of the 15) children in the FP group demonstrated atypical social-communication features consistent with the broader autism phenotype, as has been found in other familial-risk samples (Georgiades et al., 2012; Messinger et al., 2013; Ozonoff et al., 2014). Two others in the FP group were classified at 36 months with speech-language delays, one with global developmental delays, and one with other developmental concerns that did not meet criteria for another clinical classification. This suggests that a history of atypical social-communication behavior at 18 or 24 months constitutes an important clinical indicator of later problems and suggests that these children should be monitored closely after age three, even though they may no longer meet ASD criteria.

Some might wonder if the false positive cases in this study were actually children with 'optimal outcomes' (Fein et al., 2013; Sutera et al., 2007), possibly secondary to early treatment. It is challenging, however, to compare the present investigation to previous studies of optimal outcome, which followed participants much longer, into later childhood. Intervention history data were available from only a few sites in this study and the small sample size precluded formal analysis. Previous studies, however, have generally not found that number of intervention hours predicts outcome. In the meta-analysis of stability by Woolfenden et al. (2012), they note that in the subset of five studies that examined intervention hours as a predictor of outcome, none reported significant differences between the diagnostically stable and unstable groups. Anderson et al. (2014) did not find that membership in their 'very positive outcome group' was predicted by hours of treatment in early childhood. Orinstein et al. (2014) reported that children who lost their diagnosis were more likely to have received applied behavior analysis services than children who retained a diagnosis, but there were no differences between the outcome groups in number of hours of therapy. To better address the question, it is critical for future prospective studies to collect data in a systematic way on intervention history.

In this familial-risk sample, false negatives were much more common than false positives, highlighting some of the consequences of using 24 months as a final outcome age in infant sibling study designs (e.g., Shen et al., 2013; Wolff et al., 2014). While the low rate of false positives and high stability may make this a tempting strategy in terms of funding and publication timelines, it does come at some cost. In this study, over 40% of the group diagnosed with ASD at 3 years of age had not yet been identified at 24 months. While the high false negative rate in studies using 24 months as the age of final outcome may appear to present simply a conservative bias, the implications may be broader. Not only will the numbers of false negatives lead to misclassification at 24 months, potentially affecting the statistical significance of group differences, but also they may result in a nonrepresentative sample. In this study, the group diagnosed with ASD at 24 months had significantly more severe symptoms and lower developmental functioning than those who were not diagnosed until 36 months. As a result, it is possible that studies using a 24-month outcome may not be generalizable to the larger population of young children with ASD.

What are the potential lessons learned from this study in terms of clinical decision-making and diagnosis of ASD at 18 and 24 months? Could we have identified the false negatives any earlier? Is there anything that distinguishes the false positives from the true positives that would have helped clinicians realize that they would not meet criteria later and their initial diagnosis was inaccurate? There are few answers to these questions in the current dataset. The FP and FN groups are both higher functioning developmentally than the TP group, which may have clouded the clinical picture by interacting with the expression of autism symptoms. To improve early identification efforts in these clinically complex later-born siblings of children with ASD, future research could examine whether there are particular symptom patterns associated with accurate and inaccurate early classifications, as done recently by Chawarska et al. (2014) in a larger familial-risk sample.

Although the labels of false positive and false negative were used in this study in accordance with conventions in classification science, they may be misleading or even inappropriate. The way these terms are usually employed in classification science is to indicate diagnostic errors or failures of the assessment protocol to identify true underlying patterns. In this study, however, inclusion in these groups may also be due to later emerging phenotypes or symptom patterns that change with age. The pattern of ADOS scores over time clearly falls in the FP group and rises in the FN group. Since all sites maintained high standards for initial training and ongoing reliability of ADOS administration, it is unlikely that clinician error resulted in these changing patterns over time. It is more likely that shifting phenotypes in the toddlers, transient autism signs in the former group and later emerging signs in the latter, are responsible for the changes in classification. Indeed, the pattern of rising ADOS scores in the FN group is consistent with multiple previous studies demonstrating a period in which symptoms are on the increase but have not yet reached levels at which a diagnosis can be confirmed (Landa & Garrett-Mayer, 2006; Ozonoff et al., 2010, 2014). The current data suggest that the unstable diagnostic classifications may not be diagnostic errors as much as they are reflections of an unfolding, emerging picture that goes in both directions (symptoms intensifying and lessening). Finally, it is worth reiterating that the 'unstable' FP and FN groups were defined very conservatively in this study, with misclassifications at *either* 18 or 24 months leading to inclusion in these groups. While it may be alarming that such a large proportion of children with ASD went undiagnosed by expert clinicians in the second year of life, it is likely that many of these children were nonetheless eligible for early intervention services, given their lower developmental functioning and higher level of ASD symptoms than the TN group.

Infancy is characterized by rapid changes in development as well as significant behavioral variability from moment to moment, features which themselves make early diagnosis challenging. Fisch (2012) cites low test-retest correlations across multiple developmental areas in infancy and points out the psychometric and norming limitations of many measures of infant development. Yet the stability of an ASD diagnosis, both in the present investigation and in numerous previous studies (Rondeau et al., 2011; Woolfenden et al., 2012), is impressive and is substantially higher than the stability rate reported for developmental delay classifications. Moura et al. (2010) studied a population-based cohort of 3,907 infants, tested at 12 months and again at 24 months with the Batelle Developmental Screening Inventory. Of the 390 suspected of developmental delay at 12 months, only 58 continued to test positive at 24 months, yielding a stability estimate of 15% that is considerably lower than the 80% or better rates reported for ASD in the current and previous investigations.

This study had several limitations. Infant sibling study designs have inherent biases that differ from clinic- and community-based investigations. Biased enrollment of infants with higher levels of parental concern cannot be ruled out. This, or other unknown biases of the infant sibling methodology, may have contributed to a slight (and not surprising, given the restrictive inclusion criteria) elevation in recurrence rate in this sample, relative to previously reported rates (Gronberg, Schendel, & Parner, 2013; Ozonoff et al., 2011; Sandin et al., 2014).

Currently, there are no published studies comparing the clinical phenotypes of familial and non-familial cohorts and so the results of the present investigation may or may not generalize to the general population of young children with ASD. This caveat notwithstanding, it is critical that we understand the stability of early diagnoses in the familial-risk group. Such children have the potential to be identified early, since the AAP recommends performing more intensive surveillance on infants with a positive family history of ASD (Johnson & Myers, 2007). The high stability and low rate of false positive diagnoses documented in this study support the AAP guidelines for extra surveillance for this high-risk group and provide reassurance that early screening, assessment, and referral to intervention will not be wasted effort. However, the modest negative predictive value and high rate of false negatives found in this study at 18 and 24 months also suggest that, even in the context of the intensified surveillance that occurs in infant sibling studies, not all children are demonstrating clear enough clinical phenotypes to be identified prior to 36 months, particularly those with higher cognitive levels. More work is clearly needed to guide future surveillance efforts for this population.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Parameter estimates (standard errors) for the mixed-effects regression models predicting ADOS social-communication algorithm and Mullen T scores.

Acknowledgements

This original article was invited by the journal as part of a special issue; it has undergone full, external peer review.

Data collection for this manuscript was supported by National Institutes of Health Grants R01MH068398 (PI: Ozonoff), P01 HD003008, Project 1 (PI: Chawarska), R01MH087554 (PI: Chawarska), and R01MH059630 (PI: Landa), and grants from the Canadian Institutes of Health Research 62924 and 102665 (PIs: Zwaigenbaum, Bryson), NeuroDevNet (PI: Zwaigenbaum), and Autism Speaks Canada (PIs: Zwaigenbaum, Bryson). Autism Speaks provided funding for the Baby Siblings Research Consortium database. The authors gratefully acknowledge Diane Larzelere, UC Davis, for her editorial assistance with this manuscript, Alycia Halladay, for her organizational support when she was at Autism Speaks, and especially the children and families who participated at each site.

The authors have declared that they have no competing or potential conflicts of interest.

Correspondence

Sally Ozonoff, UC Davis MIND Institute, 2825 50th Street, Sacramento, CA, USA; Email: sozonoff@ucdavis.edu.

Key points

- Clinical diagnoses of ASD made before age 3 years have been shown in previous research to be quite stable in samples of children ascertained from clinics or the community.
- Stability was comparably high in a large sample of children under age 3 at heightened familial risk for ASD. Few children were classified as having ASD at 18 or 24 months who were not confirmed at 36 months.
- Sensitivity of the clinical diagnosis was relatively low at 18 and 24 months, with close to half the sample not diagnosed until 36 months of age.
- These data suggest that screening for ASD should be repeated multiple times during the first years of life.

References

Altman, D.G., & Bland, J.M. (1994). Diagnostic tests 2: Predictive values. *British Medical Journal*, *309*, 102.

Anderson, D.K., Liang, J.W., & Lord, C. (2014). Predicting young adult outcome among more and less cognitively able individuals with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *55*, 485–494.

Chawarska, K., Shic, F., Macari, S., Campbell, D.J., Brian, J., & Landa, R., … & Bryson, S. (2014). 18-month predictors of later outcomes in younger siblings of children with autism spectrum disorder: A baby siblings research consortium study. *Journal of the American Academy of Child and Adolescent Psychiatry*. *53*, 1317–1327.

Fein, D., Barton, M., Eigsti, I.M., Kelley, E., Naigles, L., Schultz, R.T., … & Tyson, K. (2013). Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry*, *54*, 195–205.

Fisch, G.S. (2012). Autism and epistemology III: Child development, behavioral stability, and reliability of measurement. *American Journal of Medical Genetics Part A*, *158*, 969–979.

Georgiades, S., Szatmari, P., Zwaigenbaum, L., Bryson, S., Brian, J., … & Garon, N. (2012). A prospective study of autistic-like traits in unaffected siblings of probands with autism spectrum disorder. *JAMA Psychiatry*, *70*, 42–48.

Gronberg, T.K., Schendel, D.E., & Parner, E.T. (2013). Recurrence of autism spectrum disorders in full- and half-siblings and trends over time: A population-based cohort study. *JAMA Pediatrics*, *167*, 947–953.

Guthrie, W., Swineford, L.B., Nottke, C., & Wetherby, A.M. (2013). Early diagnosis of autism spectrum disorder: Stability and change in clinical diagnosis and symptom presentation. *Journal of Child Psychology and Psychiatry*, *54*, 582–590.

Hess, C.R., & Landa, R.J. (2012). Predictive and concurrent validity of parent concern about young children at risk for autism. *Journal of Autism and Developmental Disorders*, *42*, 575–584.

Johnson, C.P., Myers, S.M., & the Council on Children with Disabilities (2007). Identification and evaluation of children with autism spectrum disorders. *Pediatrics*, *120*, 1183–1215.

Laird, N.M., & Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Landa, R., & Garrett-Mayer, E. (2006). Development in infants with autism spectrum disorders: A prospective study. *Journal of Child Psychology and Psychiatry*, *47*, 629–638.

Landa, R.J., Holman, K.C., & Garrett-Mayer, E. (2007). Social and communication development in toddlers with early and later diagnosis of autism spectrum disorders. *Archives of General Psychiatry*, *64*, 853–864.

Li, J., & Fine, J. (2004). On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. *Statistics in Medicine*, *23*, 2537–2550.

Lord, C., Petkova, E., Hus, V., Gan, W., Lu, F., Martin, D.M., … & Risi, S. (2012). A multisite study of the clinical diagnosis of different autism spectrum disorders. *Archives of General Psychiatry*, *69*, 306–313.

Lord, C., Rutter, M., DiLavore, P.C., & Risi, S. (2002). *Autism diagnostic observation schedule manual*. Los Angeles: WPS.

Messinger, D., Young, G.S., Ozonoff, S., Dobkins, K., Carter, A., Zwaigenbaum, L., … & Sigman, M. (2013). Beyond autism: a baby siblings research consortium study of high-risk children at three years of age. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*, 300–308.

Moura, D.R., Costa, J.C., Santos, I.S., Barros, A.J., Matijasevich, A., Halpern, R., … & Barros, F.C. (2010). Natural history of suspected developmental delay between 12 and 24 months of age in the 2004 Pelotas birth cohort. *Journal of Paediatrics & Child Health*, *46*, 329–336.

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

Orinstein, A.J., Helt, M., Troyb, E., Tyson, K.E., Barton, M.L., Eigsti, I.M., … & Fein, D.A. (2014). Intervention for optimal outcome in children and adolescents with a history of autism. *Journal of Developmental & Behavioral Pediatrics*, *35*, 247–256.

Ozonoff, S., Iosif, A.M., Baguio, F., Cook, I.C., Hill, M.M., Hutman, T., … & Young, G.S. (2010). A prospective study of the emergence of early behavioral signs of autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*, 256–266.

Ozonoff, S., Young, G.S., Belding, A., Hill, M.M., Hill, A., Hutman, T., … & Iosif, A. (2014). The broader autism phenotype in infancy: When does it emerge? *Journal of the American Academy of Child and Adolescent Psychiatry*, *53*, 398–407.

Ozonoff, S., Young, G.S., Carter, A., Messinger, D., Yirmiya, N., & Zwaigenbaum, L., … & Stone, W.L. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics*, *128*, e488–e495.

Ozonoff, S., Young, G.S., Steinfeld, M.B., Hill, M.M., Cook, I., Hutman, T., … & Sigman, M. (2009). How early do parent concerns predict later autism diagnosis? *Journal of Developmental and Behavioral Pediatrics*, *30*, 367.

Rondeau, E., Klein, L.S., Masse, A., Bodeau, N., Cohen, D., & Guilé, J.M. (2011). Is pervasive developmental disorder not otherwise specified less stable than autistic disorder? A meta-analysis. *Journal of Autism and Developmental Disorders*, *41*, 1267–1276.

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C.M., & Reichenberg, A. (2014). The familial risk of autism. *JAMA*, *311*, 1770–1777.

SAS Institute, Inc. *SAS/STAT Version 9.4.* Cary, NC: SAS Institute, Inc.

Shen, M.D., Nordahl, C.W., Young, G.S., Wootton-Gorges, S.L., Lee, A., Liston, S.E., ... & Amaral, D.G. (2013). Early brain enlargement and elevated extra-axial fluid in infants who develop autism spectrum disorder. *Brain*, *136*, 2825–2835.

Sutera, S., Pandey, J., Esser, E.L., Rosenthal, M.A., Wilson, L.B., Barton, M., ... & Fein, D. (2007). Predictors of optimal outcome in toddlers diagnosed with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*, 98–107.

Van Daalen, E., Kemner, C., Dietz, C., Swinkels, S.H., Buitelaar, J.K., & Van Engeland, H. (2009). Inter-rater reliability and stability of diagnoses of autism spectrum disorder in children identified through screening at a very young age. *European Child & Adolescent Psychiatry*, *18*, 663–674.

Wang, W., Davis, C.S., & Soong, S.J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*, *25*, 2215–2229.

Wolff, J.J., Botteron, K.N., Dager, S.R., Elison, J.T., Estes, A.M., Gu, H., ... & Piven, J. (2014). Longitudinal patterns of repetitive behavior in toddlers with autism. *Journal of Child Psychology and Psychiatry*, *55*, 945–953.

Woolfenden, S., Sarkozy, V., Ridley, G., & Williams, K. (2012). A systematic review of the diagnostic stability of autism spectrum disorder. *Research in Autism Spectrum Disorders*, *6*, 345–354.

Accepted for publication: 25 February 2015