View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Birkbeck Institutional Research Online

BIROn - Birkbeck Institutional Research Online

Pijl, M. and Bussu, G. and Charman, T. and Johnson, Mark and Jones, Emily J.H. and Pasco, G. and Oosterling, I. and Rommelse, N.N.J. and Buitelaar, J. (2019) Temperament as an early risk marker for Autism Spectrum Disorders? A longitudinal study of high-risk and low-risk infants. Journal of Autism and Developmental Disorders 49 (5), pp. 1825-1836. ISSN 0162-3257.

Downloaded from: http://eprints.bbk.ac.uk/25740/

Usage Guidelines:
Please refer to usage guidelines at http://eprints.bbk.ac.uk/policies.html or alternatively contact lib-eprints@bbk.ac.uk.

Temperament as an early risk marker for autism spectrum disorders? A longitudinal study of high-risk and low-risk infants

Abstract

To investigate temperament as an early risk marker for autism spectrum disorder (ASD), we examined parent-reported temperament for high-risk (HR, n=170) and low-risk (LR, n=77) siblings at 8, 14, and 24 months. Diagnostic assessment was performed at 36 months. Group-based analyses showed linear risk gradients, with more atypical temperament for HR-ASD, followed by HR-Atypical, HR-Typical, and LR siblings. Temperament differed significantly between outcome groups ($0.34 \geq \eta_p^2 \geq 0.03$). Machine learning analyses showed that, at an individual level, HR-ASD siblings could not be identified accurately, whereas HR infants without ASD could. Our results emphasize the discrepancy between group-based and individual-based predictions and suggest that while temperament does not facilitate early identification of ASD individually, it may help identify HR infants who do not develop ASD.

Key words: Autism spectrum disorder, high-risk, temperament, longitudinal, machine learning

Temperament can be defined as relatively stable individual differences in activity, affectivity, attention, and self-regulation that are shaped throughout development by complex interactions between genetic, biological, and environmental factors (Shiner et al. 2012). Given that temperament traits can be linked to neurobiological systems (White et al. 2012; Whittle et al. 2006) and are already measurable at an early age, potentially before psychopathology begins to emerge, temperament could function as a potential risk marker of later psychopathology (Nigg 2006; Fox 2004; Perez-Edgar and Fox 2005). The aim of this study was to investigate temperament as an early risk marker for autism spectrum disorders (ASD) in the high-risk (HR) younger siblings of children diagnosed with ASD and low-risk (LR) controls. Research has shown that 18.7% of HR siblings are diagnosed with ASD themselves (Ozonoff et al. 2011), and that 19% of HR siblings have some traits common to ASD, but not sufficient to warrant a clinical diagnosis (Georgiades et al. 2013). By applying a HR design, shared and unique characteristics of temperament between and within familial HR siblings (diagnosed with ASD, atypically developing, or typically developing) and LR siblings can be studied to reveal possible early predictors of later ASD or atypical development.

Most temperament frameworks encompass three traits during early childhood: (1) *surgency/approach* referring to engagement with the environment, positive emotions, and activity level; (2) *negative affect/withdrawal* including negative emotions such as anger, sadness, and fear; and (3*) effortful control* referring to regulation of attention, emotions, and behaviors (Putnam et al. 2001). In infancy, effortful control is described as orienting/regulation, focusing on soothability (pace of recovery from distress) and cuddliness (expression of enjoyment and molding of the body to the

caregiver) (Gartstein and Rothbart 2003). In the current study, we refer to this construct as *effortful control* in both infancy and toddlerhood.

Previous research has revealed that these three broader traits can differentiate children with ASD from others from 12 months onward (see Table 1). First, low levels of the trait *surgency* (i.e., approach behaviors, positive affect, and activity level) have been associated with later ASD (Del Rosario et al. 2014; Garon et al. 2009; Garon et al. 2016; Zwaigenbaum et al. 2005; Macari et al. 2017). However, findings up to 1 year are discrepant, showing that HR siblings that develop ASD have *higher* levels of surgency than high-risk siblings who do not develop ASD (Del Rosario et al. 2014; Clifford et al. 2013). This discrepancy suggests that temperamental patterns change with development, but could also reflect differences in the applied construct of surgency across age and as used in different temperament measures. In-depth examination at a dimensional level showed contrasting patterns for activity levels, with lower levels of activity being seen in infants with (or at risk of) ASD during the first year (Del Rosario et al. 2014; Bolton et al. 2012; Zwaigenbaum et al. 2005), followed by elevated levels of activity around the second year (Bolton et al. 2012; Garon et al. 2009). Second, higher levels of the temperament trait *negative affect* have been consistently associated with ASD from 12 months onward (Clifford et al. 2013; Garon et al. 2009; Garon et al. 2016; Zwaigenbaum et al. 2005; Bolton et al. 2012; Macari et al. 2017). Lastly, children with ASD have more self-regulatory difficulties (*effortful control*) from around the first birthday onward (Clifford et al. 2013; Garon et al. 2009; Garon et al. 2016; Zwaigenbaum et al. 2005; Gomez and Baird 2005; Bolton et al. 2012; Macari et al. 2017). However, Del Rosario et al. (2014) did not find any differences in negative affect or effortful control between HR-ASD and LR siblings during early childhood, which could be due to the

use of different instruments to assess temperament. See Table 1 for a detailed overview of the abovementioned studies focusing on temperament traits in ASD.

[Table 1 about here]

Most of the abovementioned studies focused on differences in distinct temperament traits at separate time points (e.g., the level of surgency at 12 months) and did not integrate findings across various traits and time. To the best of our knowledge, only two studies investigated the time course of temperament in young children at risk of ASD (Del Rosario et al. 2014; Garon et al. 2016). The investigation of trajectories of temperament across multiple time points is potentially more informative than measures of temperament at single time points, because it provides information about the structure of change across early childhood. In addition, investigating the integration of different temperament traits at different time points could help to combine complementary information across traits. Furthermore, while previous studies investigated temperamental differences between groups, they did not examine whether temperament provides information about individual outcomes. Although findings on group differences are valuable in terms of finding relevant biomarkers for ASD, there is often substantial overlap between groups in individual variation, making prediction for individual infants difficult. To fully judge whether temperament is useful in the early prediction of ASD, analyses at an individual level are needed.

The current study prospectively followed familial HR and LR siblings during their first 3 years of life, with the aim of observing differences in temperament between outcome groups. For these outcome groups, the HR group was divided into

HR-ASD (HR siblings subsequently diagnosed with ASD at 36 months), HR-Atypical (HR siblings not diagnosed with ASD, but with some evidence of atypical development) and HR-Typical siblings (HR siblings with typical development). The objectives were 1) to investigate *group* differences in early temperament at and across multiple time points between HR-ASD, HR-Atypical, HR-Typical, and LR siblings, and 2) to examine whether temperament (both single traits and profiles) during the first 2 years of life (both separate time points and trajectories) can help to predict ASD and atypical development at 36 months at an *individual* level. For the latter objective we extended previous work by using machine learning algorithms to combine complementary information about different temperament factors in order to identify the best predictive combination of factors. We expected that trajectories of temperament would differentiate between outcome groups and that the integration of different domains of temperament measured at different time points would improve the prediction of ASD in an individual as compared to prediction based on a single domain and/or time point. Further, based on their risk status, we hypothesized that HR-ASD would show the most 'atypical' temperament (i.e., low surgency, high negative affect, low effortful control), followed by HR-Atypical siblings, HR-Typical siblings, and LR siblings.

Methods

Participants and procedure

As part of the British Autism Study on Infant Siblings (BASIS: www.basisnetwork.org), 247 infants (170 HR and 77 LR) were assessed at four time points during their first three years of life. Data for 104 infants were collected during the first phase of the longitudinal study, which were also reported by Clifford et al. (2013). Ethical approval was given by NHS NRES London RC (06/MRE02/73, 08/H0718/76), and one or both parents gave informed consent. Most of the infants were born full-term (i.e., N=236 were born ≥36 weeks, N=11 were born between 32 and 36 weeks) and none of them had known medical or developmental conditions at the time they were enrolled. The HR infants had at least one older sibling with a clinical diagnosis of ASD (hereafter: 'proband'), confirmed in most cases by expert clinicians using information from the Development and Wellbeing Assessment (DAWBA - Goodman et al. 2000) and the Social Communication Questionnaire (SCQ - Rutter et al. 2003a). No known other significant conditions were present in the proband or extended family members (e.g., Fragile X syndrome, tuberous sclerosis). LR siblings were recruited from a volunteer database at the Birkbeck Centre for Brain and Cognitive Development. There was no ASD in first-degree family members of LR siblings (as confirmed through a parent interview regarding family medical history).

Of 247 siblings recruited, data for 33 HR and 9 LR siblings were excluded from the current study because of a substantial amount of missing data. Further information about this exclusion criterion is presented in the Measures section. We also excluded infants with no information about outcome status (N=4 HR, N=2 LR). The final sample comprised 133 HR infants (65 male; 48.9%) and 66 LR infants (28

male; 42.4%). All infants were examined at approximately 8 months (mean=8.4, SD=1.3, hereafter 8 months), 14 months (mean=14.8, SD=1.4, hereafter 14 months), around their second birthday (mean=25.4, SD=1.9, hereafter 24 months), and around their third birthday (mean=38.6, SD=2.2, hereafter 36 months).

Measures

Infant and toddler temperament. Two measures of temperament, appropriate to the child's age, were administered. Parents completed the Infant Behavior Questionnaire-Revised (IBQ-R - Gartstein and Rothbart 2003) at the 8- and 14-month visits, and the Early Childhood Behavior Questionnaire (ECBQ - Putnam et al. 2006) at the 24-month visit. Both measures are reliable and well-validated parent-reported questionnaires that are scored on a Likert scale ranging from 1 (never) to 7 (always). The IBQ-R was designed to assess temperament in the first year of life and contains 14 dimensions based on 184 items. The ECBQ was developed for children aged 18 to 36 months and consists of 18 dimensions based on 201 items. Three broad factors can be identified with both the IBQ-R and the ECBQ: Surgency, Negative Affect, and Effortful Control (labeled 'Orienting' on the IBQ-R). Of note, although both the IBQ-R and ECBQ provide a similar 3-factor model, the loading on the factors is different. See Putnam et al. (2001) for a discussion of this structure of temperament.

To ensure the validity of the temperament measures, dimensions were only calculated if data on ≥70% of items were available. Similarly, factors were only computed if ≥70% of dimension scores were available. Given that this study focused on longitudinal trajectories of temperament at 8, 14, and 24 months, participants were

only included if data on ≥70% of the factors were available across the three time points.

Outcome characterization. At the 36-month visit, various clinical research measures were used to characterize the outcome of the HR siblings. The Autism Diagnostic Observation Schedule (ADOS-2 - Lord et al. 2012), the Autism Diagnostic Interview (ADI-R - Rutter et al. 2003b), and the SCQ (Rutter et al. 2003a) were used to obtain information about ASD symptomatology. In addition, the Mullen Scales of Early Learning (Mullen 1995) and the Vineland Adaptive Behavior Scale-II (Sparrow et al. 2005) were assessed to gather information about the child's development and adaptive functioning, respectively. Experienced clinical researchers (TC, GP) reviewed the outcomes of each HR sibling. Consensus ICD-10 or DSM-5 criteria were used to ascertain ASD diagnostic outcome. Among the 133 HR siblings enrolled in this study, 24 HR siblings met criteria for ASD (hereafter: 'HR-ASD') and 34 HR siblings did not meet criteria for ASD, but scored above the ASD threshold on the ADOS and/or ADI-R and/or scored >1.5 SD below the population mean on the MSEL receptive language, expressive language, and/or early learning composite score [hereafter: 'HR-Atypical']). The remaining 75 HR siblings were considered to be developing typically (hereafter: 'HR-Typical'). No formal research diagnoses were assigned to the LR group, but none of the LR infants had a community clinical ASD diagnosis. See Table 2 for detailed demographics of the included participants.

[Table 2 about here]

Statistical analyses

Multiple imputation with the expectation maximization algorithm was used to account for missing data (Tabachnik and Fidell 2001). In addition, a Van der Waerden transformation was applied to data for temperament factors, which transforms raw scores into z-scores corresponding to the estimated cumulative proportion of the distribution analogous to a particular rank (using Statistical Package for the Social Sciences [SPSS] version 22).

Group-based analyses. MANCOVAs were used to investigate whether a risk gradient was present in polynomial group contrasts at separate time points. The outcome groups were ranked as follows: 1=HR-ASD, 2=HR-Atypical, 3=HR-Typical, and 4=LR, assuming that polynomial group contrasts would indicate linear risk gradients for atypical temperament (HR-ASD > HR- Atypical > HR-Typical > LR). Analyses were performed for each temperament trait separately, including group as independent variable and temperament at three time points as dependent variables (e.g., surgency at 8, 14, and 24 months). Sex was differently distributed across groups (with more males than females in the HR-ASD group), and age at intake was variable (between 5 and 11 months), introducing potential noise in results due to different starting ages. Therefore, sex and age at the first visit were included as covariates.

In post-hoc analyses, pair wise group contrasts were examined across time by performing two-way mixed ANCOVAs and paired sample t-tests, resulting in six pair wise comparisons (i.e., HR-ASD vs. HR-Atypical, HR-ASD vs. HR-Typical, HR-ASD vs. LR, HR-Atypical vs. HR-Typical, HR-Atypical vs. LR, HR-Typical vs. LR). The effect of group (e.g., HR-ASD, LR), time (8, 14, 24 months), and the interaction effect group x time on trajectories of a temperament trait was investigated, while controlling for sex and age at the first visit. A correction for multiple comparisons

was applied for the post-hoc analyses, using the false discovery rate controlling procedure with a q-value of 0.05 (Benjamini and Hochberg 1995). If Mauchly's test indicated that the assumption of sphericity had been violated, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. Following Cohen's guidelines (Cohen, 1988), effect sizes were defined in terms of the percentage of variance explained: 1, 9 and 25% were used to define small, medium, and large effects (these percentages translate into ηp2-values of 0.01, 0.06 and 0.14). Analyses contrasting the HR group (without a differentiation based on 36-month outcome) and the LR controls are described in *Supplemental Material*.

Classifier Analyses – from group-based to individual analysis. As a next step, we investigated how temperament factors at 8, 14, and 24 months related to atypical development, and more specifically ASD, at an individual level among infants in the HR group. To this end, we performed confounder-corrected support vector machine classification with 40% holdout cross-validation repeated 10 times using custom made scripts implemented in Matlab R2016b (MATLAB 9.1, The MathWorks Inc., Natick, MA, 2016). We addressed two distinct binary classification problems: distinguishing HR-ASD from HR-Atypical and HR-Typical; and distinguishing HR-ASD and HR-Atypical together from HR-Typical. In fact, while the most clinically relevant question is to distinguish HR-ASD siblings from their peers at an early age, distinguishing HR-ASD and HR-Atypical together from HR-Typical is also clinically relevant and potentially useful for early intervention. Sex and age at the first visit were included as covariates, and findings were corrected for inverse probability weighting. Features for the classifiers consisted of temperament factors (surgency, negative affect, effortful control, and all their combinations) from different time

points (8 months, 14 months, and 24 months). To exploit the longitudinal information on developmental dynamics, the intercept and slope of the developmental trajectories on single measures between 8 and 24 months were also used as features for the classifiers. Trajectories were computed for single individuals by linear regression modeling using the *lme4* software package on R (Bates et al. 2015). A total of 28 classifiers were compared to find the best predictor of HR-ASD and HR-ASD+HR-Atypical at 36 months (see *Supplemental Material* for details). For each classifier, the area under the curve (AUC) was computed to determine the best classifier, and we evaluated the classifier performance via sensitivity, specificity, accuracy, negative predictive value (NPV – i.e., true negative over negative predicted cases), and positive predictive value (PPV – i.e., true positive over positive predicted cases). 95% confidence intervals (CI) for each metric were computed using bootstrap with n=1000 repetitions for each cross-validation fold, then averaging over folds (n=10000 in total). The *p*-value of AUC was computed for each classifier through a shuffle test (n=10000 total repetitions; n=1000 repetitions for each classification fold) to test the significance of classification performance. Performance metrics are reported only when the performance was significantly different from chance level.

For both classifications (HR-ASD vs. HR-Atypical + HR-Typical | HR-ASD + HR-Atypical vs. HR-Typical), the best predicting classifier at each time point was selected based on the AUC. A nonparametric Friedman test was performed on classifier performance metrics (i.e., AUC) at each time point separately to test for significant differences in performance between distinct classifiers. If the Friedman test was significant, post-hoc paired Wilcoxon rank sum tests were performed between the classifier of interest (i.e., the one with highest AUC) and all other classifiers. Bonferroni correction was applied to avoid biasing effects due to multiple

comparisons. In addition, differences in performance of the best classifiers across time points were tested by a two-sided Wilcoxon rank sum test.

Results

Temperament differences between groups

Surgency. A polynomial group contrast indicated a linear risk gradient to be present at 14 months of age (Contrast Estimate [CE]=0.40, p=0.02), implying that LR siblings had the highest levels of surgency, followed by HR-Typical siblings, HR-Atypical siblings, and HR-ASD siblings. No significant gradient was present at 8 or 24 months (CE=-0.08, p=0.64; CE=0.27, p=0.10, respectively).

Two-way mixed ANCOVAs examining pair wise group contrasts revealed a significant group x time effect for the comparison between HR-ASD and HR-Typical siblings (F(1.77, 168.07)=3.67, p<0.05, η_p^2=0.04; see Figure 1), as well as between HR-ASD and LR siblings (F(1.86, 160.06)=3.98, p<0.05, η_p^2=0.04). Post-hoc tests revealed that both interaction effects were driven by a group x time effect between 8 and 14 months of age (F(1, 95)=6.69, p<0.05, η_p^2=0.07; F(1, 86)=9.79, p<0.01, η_p^2=0.10, respectively), with HR-ASD siblings showing diverging levels of surgency (i.e. approach behaviors, positive affect, activity level) from 8 to 14 months compared with HR-Typical and LR siblings (paired sample t-tests for each group were non-significant). In addition, for the comparison between HR-ASD and LR siblings a significant main effect of group was found between the 14- and 24-month time point (F(1, 86)=4.89, p<0.05, η_p^2=0.05), indicating stable lower levels of surgency in the HR-ASD group than in the LR group between 14 and 24 months of age.

[Figure 1 about here]

Negative Affect. A polynomial group contrast indicated a linear risk gradient to be present at 8, 14, and 24 months (CE=-0.46, p=0.004; CE=-0.38, p=0.02; CE=-0.69, p<0.001, respectively), suggesting that HR-ASD siblings showed the highest levels of negative affect, followed by HR-Atypical siblings, HR-Typical siblings and LR siblings.

A two-way mixed ANCOVA revealed significant main group effects for HR-ASD vs. HR-Typical (F(1, 95)=7.47, p<0.01, η_p^2=0.07; see Figure 2), HR-ASD vs. LR (F(1, 86)=15.57, p<0.001, η_p^2=0.15), and HR-Typical vs. LR siblings (F(1, 137)=6.49, p<0.05, η_p^2=0.05). These effects indicate that, independent of age, HR-ASD siblings had developmentally stable higher levels of negative affect than HR-Typical and LR siblings, and that HR-Typical siblings had stable higher levels of negative affect than LR siblings.

[Figure 2 about here]

Effortful Control. A polynomial group contrast indicated a linear risk gradient to be present at 14 and 24 months (CE=0.69, p<0.001; CE=0.84, p<0.001, respectively), showing that LR siblings had the highest levels of effortful control, followed by HR-Typical siblings, HR-Atypical siblings and HR-ASD siblings. No significant gradient was present at 8 months of age (CE=0.21, p=0.20).

A two-way mixed ANCOVA showed significant group x time interaction effects for the comparisons between HR-ASD and HR-Typical siblings (F(1.85, 175.79)=6.95, p<0.01, η_p^2=0.07; see Figure 3), and between HR-ASD and LR siblings (F(2, 172)=8.41, p<0.001, η_p^2=0.09). Post-hoc tests revealed that the interaction effects were driven by the 8- to 14-month trajectory (F(1, 95)=8.53, p<0.01, η_p^2=0.08;

F(1, 86)=12.69, p<0.01, η_p^2=0.13, respectively), showing that the level of effortful control decreased in HR-ASD siblings from 8 to 14 months (t(23)=2.85, p=0.009) relative to the static levels of effortful control seen in HR-Typical (t(74)=-1.08, p=0.28) and LR (t(65)=-1.03, p=0.31) siblings. Between 14 and 24 months of age, significant main effects of group were found (F(1, 86)=18.90, p<0.001, η_p^2=0.17; F(1, 86)=44.22, p<0.001, η_p^2=0.34, respectively), suggesting that HR-ASD siblings had stable lower levels of effortful control than HR-Typical and LR siblings. Furthermore, significant main group effects were found between HR-ASD vs. HR-Atypical (F(1, 54)=6.28, p<0.05, η_p^2=0.10), HR-Typical vs. LR (F(1, 137)=4.31, p<0.05, η_p^2=0.03), and HR-Atypical vs. LR (F(1, 96)=5.19, p<0.05, η_p^2=0.05) siblings. These results showed that HR-ASD siblings had developmentally stable lower levels of effortful control than HR-Atypical siblings, and that LR controls had higher levels of effortful control than both HR-Typical and HR-Atypical siblings.

[Figure 3 about here]

Individual prediction of HR clinical outcome

Classification of HR-ASD among HR siblings was significantly different from chance level using measures from 14 months onward. In contrast, classification of HR-ASD and HR-Atypical together from HR-Typical was not significantly different from chance level at any of the time points, with only marginal significance at 24 months. See Table 3 and 4 for an overview of the performance metrics of classifiers that were significantly different from chance level for the two classifications (i.e., HR-ASD vs. HR-Atypical + HR-Typical, and HR-ASD+HR-Atypical vs. HR-Typical). Detailed statistics can be found in the *Supplemental Material*.

[Table 3 and 4 about here]

To evaluate which combination of temperament factors best predicted ASD at different time points, we compared the performance of the different classifiers at the separate time points, based on the AUC. The combination of all factors at 24 months provided the most promising classifier to predict ASD among HR siblings (p=0.02; mean [CI]: AUC=72% [57% to 83%]; sensitivity=85% [61% to 99%], specificity=58% [43% to 73%], accuracy=63% [49% to 75%], PPV=30% [13% to 49%], NPV=95% [86% to 100%]). However, the predictive performance was not significantly different from that of effortful control (z=-0.51, p=0.61) and its combination with other factors at 24 months (surgency + effortful control: z=-1.58, p=0.11; effortful control + negative affect: z=-0.98, p=0.33). Furthermore, effortful control had the highest predictive power at 14 months (AUC=64%), and when using the developmental trajectory between 8 and 24 months as feature for the classifiers, the integration of scores from effortful control and negative affect provided the classifier with the highest AUC (AUC=68%). After Bonferroni correction for multiple comparisons (leading to $\alpha_{Bonferroni}$ =0.017), the difference in classification performance between the combined factors at 24 months and effortful control at 14 months was not significant (Wilcoxon z=-2.14, p=0.032), and the same applies to the difference in classification performance between the combined factors at 24 months and the combined longitudinal trajectories of effortful control and negative affect (Wilcoxon z=-1.86, p=0.063).

For classification of HR-ASD plus HR-Atypical from HR-Typical, the integration of effortful control and negative affect at 24 months provided the highest

AUC (p=0.056; mean [CI]: AUC=61% [48% to 74%]; sensitivity=60% [39% to 79%], specificity=62% [45% to 79%], accuracy=61% [48% to 74%], PPV=55% [35% to 75%], NPV=68% [50% to 85%]). Since performance was not significantly different from chance level, classifier comparison was not performed.

Overall, even though effortful control and a combination of the temperament factors at 24 months predicted ASD outcome at a moderate level (AUC=71%; AUC=72%, respectively), its positive predictive value for ASD was low and none of the classifiers adequately predicted broader atypical development at 36 months.

Discussion

The current study is the first to examine differences in temperament at and across three time points in early childhood between outcome groups (i.e. HR-ASD, HR-Atypical, HR-Typical and LR siblings), and to investigate temperament at an individual level. At a group level, our findings revealed positive linear risk gradients for surgency at 14 months, and effortful control at 14 and 24 months, and negative linear risk gradients for negative affect at 8, 14, and 24 months. This indicates that temperament in early childhood was more atypical in HR-ASD siblings, followed by HR-Atypical siblings, HR-Typical siblings, and LR controls. Post-hoc pair wise comparisons indicated differences in early temperament between the outcome groups. However, the effect sizes were generally medium, especially regarding differences within the HR group. Machine learning analyses using temperament traits during infancy (i.e., 8 months) did not accurately predict ASD at 36 months at an individual level. From 14 months onward, effortful control (or its combination with other traits) had the highest predictive power for ASD as compared to other temperament traits and combinations, with a high negative predictive value, but with a positive predictive value that was far from being clinically useful. Neither the separate temperament traits nor a combination of traits was able to accurately predict broader atypical development (i.e., HR-ASD and HR-Atypical). Thus, although differences in temperament traits can be detected in infancy at a group level, this difference does not necessarily translate into an acceptably accurate prediction of ASD in the individual infant.

Temperament differences between HR subgroups and LR controls

At a group level, our findings showed that HR siblings with or without a subsequent ASD diagnosis could be distinguished from LR controls based on higher levels of negative affect and lower levels of effortful control (with the exception of HR-Atypical siblings regarding negative affect). These findings replicate and extend previous research (Garon et al. 2016), showing that young siblings at risk of ASD, regardless of whether they develop ASD or not, tend to use more negative emotions and have more difficulties regulating attention, emotions, and behaviors than do LR controls. Furthermore, we found that the pattern of surgency from 8 to 14 months and levels of surgency thereafter were different between HR-ASD and LR siblings, whereas levels of surgency in the HR-Typical and HR-Atypical siblings did not differ from those of the LR group. As to be expected, this suggests that, on average, low levels of approach and positive emotions are specifically associated with the development of ASD. Differences in surgency levels across time may be explained by the multi-dimensional nature of the factor surgency (Gartstein and Rothbart 2003; Putnam et al. 2006). Future research may use a dimensional or item level approach to delineate the underlying mechanisms and to enable comparison of findings between studies.

Temperament differences within at-risk siblings

Within the HR group, temperament traits distinguished HR-ASD siblings from HR siblings without a clinical diagnosis, suggesting the presence of more temperamental challenges early in life of children with subsequent ASD. Interestingly, higher levels of negative affect were already present from 8 months onward in the HR-ASD siblings, whereas effortful control started to distinguish between the groups from 14 months onward. These findings, combined with those of a recent study examining

temperament trajectories from 12 months onward (Garon et al. 2016), may indicate that early affective behaviors play an important role in the subsequent regulation of attention, emotions, and behaviors. Garon et al. (2016) found that affective components of temperament at 12 months predicted regulatory behaviors at 24 months in both HR and LR infants, and that regulatory behaviors in turn predicted ASD symptoms at 36 months in the HR sample. Future investigation of the associations between temperament traits in different outcome groups is needed, including the assessment of temperament during the first year of life.

Temperament as a potential early risk marker

The idea that temperament may be an early risk marker is in accordance with the spectrum theory (Tackett 2006), which holds that there is a shared etiology between psychopathology at the extreme negative end of a continuum of social-communicative competences and temperament traits. A study of monozygotic and dizygotic adult twins supported this idea by showing that ASD and most temperament traits share common genetic and environmental etiological factors (Picardi et al. 2015). Temperament may be a fruitful risk marker that could help differentiate between groups of children on different developmental pathways.

Nonetheless, the use of temperament traits as an early risk marker is constrained by two findings. First, identification of ASD at an individual level on the basis of temperament traits had low positive predictive value and specificity. This indicates that based on (combinations of) temperament traits a substantial number of HR siblings would be falsely classified as HR-ASD at 36 months (i.e., false positives). However, the high negative predictive values indicate that temperament traits can accurately predict which infants are *not* going to develop ASD in all likelihood. This

has still clinical value, especially for the selection of infants who might need early intervention. In other words, results at the individual level suggest that while low levels of effortful control do not predict ASD development, high levels of effortful control accurately predict typical development. The predictive value of effortful control for non-ASD development is in line with the view that effortful control, as a measure of executive function, might promote resilience, such that infants with higher levels of effortful control may be better able to compensate for atypicalities that lead to ASD outcome (Johnson 2012). However, our results highlight the difficulties of translating findings from a group to an individual level. In fact, there is often substantial overlap between groups in individual variation, making it more difficult to make predictions for individual infants. Instead of a risk marker for ASD, variation in temperament may therefore function as a *stratification marker* that allows to classify individuals with ASD into biologically more homogeneous subtypes (Loth et al. 2017). In this way, temperament may help to unravel the heterogeneous character of ASD. Importantly, the extent to which atypical temperament reflects brain alterations that predispose to ASD and/or are shared between atypical temperament and ASD need to be investigated. Additionally, future work should investigate the integration of clinical (e.g., MSEL, VABS, AOSI) and biological (e.g., eye tracking, functional imaging) measures, to improve the positive predictive value for the clinical diagnosis of ASD at an individual level (Bussu et al. 2018), and to investigate the additional value of temperament. Second, the differences found in this study mainly started to emerge around the first birthday (at both group and individual levels), which is also when behaviors related to ASD start to emerge (Ozonoff et al. 2010; Wan et al. 2013). This makes it important to ascertain whether temperament measures actually assess characteristics of temperament, or whether they just pick up the emergence of ASD

symptoms. Future research should further investigate the conceptual nature of temperament measures by examining the structure of traits in different outcome groups and in relation to ASD severity.

Limitations and Future Directions

Particular strengths of this study are its longitudinal design, which allowed the assessment of temperament trajectories across early childhood, and the differentiation between siblings based on their diagnostic status at 36 months of age. A limitation is that temperament was assessed on the basis of parent-reported measures and not on observational measures of temperament (e.g., Lab-TAB; Gagne et al. 2011). It will therefore be essential to demonstrate convergence between the parent-reported IBQ-R and ECBQ and indicators of temperament based on standardized laboratory or home assessments. Nonetheless, evidence of convergent validity between a preliminary version of the IBQ and home observations of infant temperament implies that parents' familiarity with a child's behavior may make them the best possible source of reliable information (Rothbart 1986). In addition, given that temperament is the result of complex interactions between genetic, biological, and environmental factors (Goldsmith et al., 2006; Shiner et al., 2012), the role of the environment, such as the child's family, should also be considered in temperament research. Previous research has shown that the quality of parenting interacts with individual differences in genetic variation to influence temperament traits (Voelker et al. 2009; Sheese et al. 2007).

Conclusions

Taken together, our longitudinal study identified differences in early temperament traits between HR and LR siblings as well as between the different outcome

subgroups among HR children, as most clearly demonstrated by differences in negative affect from 8 months onward and effortful control from 14 months onward. Our results underscore the complexity of translating findings from a group to an individual level, as findings did not accurately predict ASD at an individual level. From a clinical perspective, our results indicate that temperament traits may provide useful information about which HR infants are less likely to develop ASD but are not useful in predicting which HR infants will develop ASD or an atypical outcome. Future studies should increase our understanding of the role of temperament when it comes to individualizing interventions. Knowledge about temperament traits that influence adaptive functioning might help to improve the benefit of interventions in young children at risk of ASD.

All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software, 67*, 1--48.

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

Bolton, P. F., Golding, J., Emond, A., & Steer, C. D. (2012). Autism spectrum disorder and autistic traits in the avon longitudinal study of parents and children: precursors and early signs. *J Am Acad Child Adolesc Psychiatry, 51*(3), 249-260, doi:10.1016/j.jaac.2011.12.009.

Bussu, G., Jones, E. J. H., Charman, T., Johnson, M. H., Buitelaar, J. K., & Team, B. (2018). Prediction of autism at 3 years from behavioural and developmental measures in high-risk infants: a longitudinal cross-domain classifier analysis. *J Autism Dev Disord*, doi:10.1007/s10803-018-3509-x.

Clifford, S. M., Hudry, K., Elsabbagh, M., Charman, T., Johnson, M. H., & Team, B. (2013). Temperament in the first 2 years of life in infants at high-risk for autism spectrum disorders. *J Autism Dev Disord, 43*(3), 673-686, doi:10.1007/s10803-012-1612-y.

Del Rosario, M., Gillespie-Lynch, K., Johnson, S., Sigman, M., & Hutman, T. (2014). Parent-reported temperament trajectories among infant siblings of children with autism. *J Autism Dev Disord, 44*(2), 381-393, doi:10.1007/s10803-013-1876-x.

Fox, N. A. (2004). Temperament and early experience form social behavior. *Ann N Y Acad Sci, 1038*, 171-178, doi:10.1196/annals.1315.025.

Gagne, J. R., Van Hulle, C. A., Aksan, N., Essex, M. J., & Goldsmith, H. H. (2011). Deriving childhood temperament measures from emotion-eliciting behavioral episodes: scale construction and initial validation. *Psychol Assess, 23*(2), 337-353, doi:10.1037/a0021746.

Garon, N., Bryson, S. E., Zwaigenbaum, L., Smith, I. M., Brian, J., Roberts, W., et al. (2009). Temperament and its relationship to autistic symptoms in a high-risk infant sib cohort. *J Abnorm Child Psychol, 37*(1), 59-78, doi:10.1007/s10802-008-9258-0.

Garon, N., Zwaigenbaum, L., Bryson, S., Smith, I. M., Brian, J., Roncadin, C., et al. (2016). Temperament and its association with autism symptoms in a high-risk population. [Peer Reviewed]. *Journal of Abnormal Child Psychology. Aug*(Pagination), doi:10.1007/s10802-015-0064-1.

Gartstein, M. A., & Rothbart, M. K. (2003). Studying infant temperament via the revised Infant Behavior Questionnaire. *Infant Behavior & Development, 26*(1), 64-86, doi:10.1016/S0163-6383(02)00169-8.

Georgiades, S., Szatmari, P., Zwaigenbaum, L., Bryson, S., Brian, J., Roberts, W., et al. (2013). A prospective study of autistic-like traits in unaffected siblings of probands with autism spectrum disorder. *JAMA Psychiatry, 70*(1), 42-48, doi:10.1001/2013.jamapsychiatry.1.

Gomez, C. R., & Baird, S. (2005). Identifying early indicators for autism in self-regulation difficulties. [Peer Reviewed]. *Focus on Autism and Other Developmental Disabilities, 20*(2), doi:10.1177/10883576050200020101.

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: description and initial

validation of an integrated assessment of child and adolescent psychopathology. *J Child Psychol Psychiatry, 41*(5), 645-655.

Johnson, M. H. (2012). Executive function and developmental disorders: the flip side of the coin. *Trends Cogn Sci, 16*(9), 454-457, doi:10.1016/j.tics.2012.07.001.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)*. Los Angeles, CA: Western Psychological Corporation.

Loth, E., Charman, T., Mason, L., Tillmann, J., Jones, E. J. H., Wooldridge, C., et al. (2017). The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Molecular Autism, 8*, doi:ARTN 24 10.1186/s13229-017-0146-8.

Macari, S. L., Koller, J., Campbell, D. J., & Chawarska, K. (2017). Temperamental markers in toddlers with autism spectrum disorder. *J Child Psychol Psychiatry*, doi:10.1111/jcpp.12710.

Mullen, E. (1995). *Mullen scales of early learning (AGS ed.)*. Circle Pines, MN: American Guidance Service.

Nigg, J. T. (2006). Temperament and developmental psychopathology. *J Child Psychol Psychiatry, 47*(3-4), 395-422, doi:10.1111/j.1469-7610.2006.01612.x.

Ozonoff, S., Iosif, A. M., Baguio, F., Cook, I. C., Hill, M. M., Hutman, T., et al. (2010). A prospective study of the emergence of early behavioral signs of autism. [Article]. *J Am Acad Child Adolesc Psychiatry, 49*(3), 256-266, doi:10.1016/j.jaac.2009.11.009.

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., et al. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics, 128*(3), e488-495, doi:10.1542/peds.2010-2825.

Perez-Edgar, K., & Fox, N. A. (2005). Temperament and anxiety disorders. *Child Adolesc Psychiatr Clin N Am, 14*(4), 681-706, viii, doi:10.1016/j.chc.2005.05.008.

Picardi, A., Fagnani, C., Medda, E., Toccaceli, V., Brambilla, P., & Stazi, M. A. (2015). Genetic and environmental influences underlying the relationship between autistic traits and temperament and character dimensions in adulthood [Peer Reviewed]. *Compr Psychiatry*(Pagination), doi:10.1016/j.comppsych.2014.12.018 25600422.

Putnam, S. P., Ellis, L. K., & Rothbart, M. K. (2001). The structure of temperament from infancy through adolescence. In A. E. A. Angleitner (Ed.), *Advances/proceedings in research on temperament* (pp. 165-182). Germany: Pabst Scientist Publisher.

Putnam, S. P., Gartstein, M. A., & Rothbart, M. K. (2006). Measurement of fine-grained aspects of toddler temperament: the Early Childhood Behavior Questionnaire. *Infant Behav Dev, 29*(3), 386-401, doi:10.1016/j.infbeh.2006.01.004.

Rothbart, M. K. (1986). Longitudinal Observation of Infant Temperament. *Dev Psychol, 22*(3), 356-365, doi:Doi 10.1037/0012-1649.22.3.356.

Rutter, M., Bailey, A., & Lord, C. (2003a). *SCQ. The Social Communication Questionnaire*. Los Angeles, CA: Western Psychological Services.

Rutter, M., Le Couteur, A., & Lord, C. (2003b). *ADI-R: Autism diagnostic interview—revised* Los Angeles, CA: Western Psychological Services.

Sheese, B. E., Voelker, P. M., Rothbart, M. K., & Posner, M. I. (2007). Parenting quality interacts with genetic variation in dopamine receptor D4 to influence temperament in early childhood. *Dev Psychopathol, 19*(4), 1039-1046, doi:10.1017/S0954579407000521.

Shiner, R. L., Buss, K. A., McClowry, S. G., Putnam, S. P., Saudino, K. J., & Zentner, M. (2012). What is temperament now? Assessing progress in temperament research on the twenty-fifth anniversary of Goldsmith et al. (1987). [Peer Reviewed]. *Child Development Perspectives, 6*(4), 436-444.

Sparrow, S. S., Balla., D. A., Cicchetti, D. V., & Doll, E. A. (2005). *Vineland adaptive behavior scales (Vineland-II)—2nd edition*. Mineapolis: Pearson.

Tabachnik, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Needham Heights, MA: Allyn and Bacon.

Tackett, J. L. (2006). Evaluating models of the personality-psychopathology relationship in children and adolescents. *Clin Psychol Rev, 26*(5), 584-599, doi:10.1016/j.cpr.2006.04.003.

Voelker, P., Sheese, B. E., Rothbart, M. K., & Posner, M. I. (2009). Variations in catechol-O-methyltransferase gene interact with parenting to influence attention in early development. *Neuroscience, 164*(1), 121-130, doi:10.1016/j.neuroscience.2009.05.059.

Wan, M. W., Green, J., Elsabbagh, M., Johnson, M., Charman, T., Plummer, F., et al. (2013). Quality of interaction between at-risk infants and caregiver at 12-15 months is associated with 3-year autism outcome. *J Child Psychol Psychiatry, 54*(7), 763-771, doi:10.1111/jcpp.12032.

White, L. K., Lamm, C., Helfinstein, S. M., & Fox, N. A. (2012). Neurobiology and neurochemistry of temperament in children. In S. R. L. Zentner M (Ed.), *Handbook of Temperament*. New York, Guilford: 2012.

Whittle, S., Allen, N. B., Lubman, D. I., & Yucel, M. (2006). The neurobiological basis of temperament: towards a better understanding of psychopathology. *Neurosci Biobehav Rev, 30*(4), 511-525, doi:10.1016/j.neubiorev.2005.09.003.

Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. [Article]. *International Journal of Developmental Neuroscience, 23*(2-3), 143-152, doi:10.1016/j.ijdevneu.2004.05.001.

Figure Caption Sheet

Figure 1. Estimated Means for Surgency by Diagnostic Group and Time controlled for Sex and Age at Start

Figure 2. Estimated Means for Negative Affect by Diagnostic Group and Time controlled for Sex and Age at Start

Figure 3. Estimated Means for Effortful Control by Diagnostic Group and Time controlled for Sex and Age at Start

Table 1. Summary of findings on the three temperament traits and/or dimensions related to the traits in infants and toddlers with (or at risk of) ASD.

Study	Participant description (N)	0-11 months			1-2 years			2-3 years		
		SU	NA	EC	SU	NA	EC	SU	NA	EC
Clifford et al., 2013	HR-ASD (17), HR-Atypical (12), HR-Typical (24), LR (50)	↑[1]	ns	ns	↑[1]↓[2]	ns	↓[3]	ns	↑[2]	↓[2]
Del Rosario et al., 2014	HR-ASD (10-16), HR-non ASD (7-27)	↑↓	ns	ns	↓	ns	ns	↓	ns	ns
Garon et al., 2009	HR-ASD (34), HR-non ASD (104), LR (73)							↓↑[4]	↑[2]	↓[4]
Garon et al., 2016	HR-ASD (98), HR-non ASD (285), LR (162)				↓[5]	↑[6]	ns	↓[5,6]	↑[6]	↓[5,6]
Zwaigenbaum et al., 2005	HR-ASD (19), HR non-ASD (46), LR (23)	↓[4]	ns	ns	ns	↑[4]	↓[4]	↓[4]	ns	↓[4]
Gomez & Baird, 2005	ASD (65), TD (120)						↓			
Bolton et al., 2012	ASD (85), non-ASD (13885)	↓[7]	ns	ns				↑[7]	↑[7]	↓[7]
Macari et al., 2017	ASD (165), DD (58), TD (92)							↓[8]	↑[9]	↓[8]

Notes. ADOS = Autism Diagnostic Observation Schedule; ADI-R = Autism Diagnostic Interview - Revised; ASD = infants or toddlers diagnosed with autism spectrum disorders; DD = developmentally delayed infants or toddlers; DSM-IV-TR = Diagnostic and Statistical Manual of Mental Disorders, fourth version text revision; EC = Effortful Control; HR-ASD = at-risk siblings subsequently diagnosed with ASD; HR-Atypical = at-risk siblings not diagnosed with ASD, but following an atypical development; HR-Typical = at-risk siblings following a typical development; ICD-10 = International Classification of Diseases, tenth version; LR = low-risk controls; MSEL = Mullen Scales of Early Learning; NA = Negative Affect; SU = Surgency; TD = typically developing infants or toddlers; Vineland = Vineland Adaptive Behavior Scales; Marked cells indicate findings based on the temperament trait's composite score instead of findings based on dimensions or constructs related to the broader trait. Dimensions or constructs that could not be related

to one of the three traits were not included in this table; Empty cells not investigated.
[1] HR-ASD as compared to HR-Typical; [2] HR-ASD as compared to LR; [3] HR-ASD as compared to LR and HR-Atypical; [4] HR-ASD as compared to HR non-ASD and LR; [5] HR-ASD as compared to HR non-ASD; [6] HR (HR-ASD and HR non-ASD) as compared to LR; [7] Findings reported here are controlled for gender; [8] ASD as compared to both DD and TD; [9] ASD as compared to TD.

Table 2. Sample characterization (means and standard deviations) for low-risk siblings and subgroups of high-risk siblings.

	HR-ASD (N=24)	HR-Atypical (N=34)	HR-Typical (N=75)	LR (N=66)
Sex (% male)	75[a]	47.1	41.3[b]	42.4[b]
Age				
8 months	8.3 (1.4)	8.6 (1.0)	8.5 (1.3)	8.3 (1.4)
14 months	14.8 (1.6)	14.7 (1.4)	14.9 (1.3)	14.7 (1.3)
24 months	25.4 (2.8)	25.4 (2.1)	26.0 (1.9)[a]	24.7 (1.0)[b]
36 months	38.0 (2.0)	38.0 (2.8)	38.5 (1.8)	38.4 (2.7)
MSEL[1]				
8 months	98.0 (15.5)[a]	100.0 (13.8)	106.3 (15.8)	107.7 (12.6)[b]
14 months	89.8 (17.3)[a]	96.5 (14.0)[a]	99.8 (14.6)	106.0 (15.0)[b]
24 months	94.5 (24.8)[a]	99.2 (21.8)[a]	104.9 (15.9)[a]	115.4 (14.2)[b]
36 months	98.0 (26.7)[a]	95.9 (24.4)[a]	115.1 (15.5)[b]	118.1 (15.0)[b]
ADOS severity[2,3]				
36 months	5.1 (3.0)[a]	5.1 (2.2)[a]	1.5 (0.9)[b]	2.5 (1.8)[c]

Superscripted letters that differ from other superscripted letters indicate significant differences across groups for the given measure ($p \leq 0.05$). Values without superscript letters indicate no significant differences from another group.

[1] Mullen Scales of Early Learning (Mullen, 1995) Early Learning Composite Standard Score

[2] Autism Diagnostic Observation Schedule-2 (ADOS-2 - Lord et al., 2012)

[3] ADOS-2 calibrated severity score (Gotham et al., 2009)

Table 3. HR-ASD vs. HR-Typical + HR-Atypical. Performance metrics of temperament factors for classifying HR siblings who later develop ASD from their peers.

	Classifier	*p*	AUC	Sensitivity	Specificity	Accuracy	PPV	NPV
14 months	Effortful control	0.047	64.4 [46.7, 80.3]	69.6 [37.4, 97.8]	59.2 [44.5, 73.4]	60.9 [47.9, 73.6]	26.6 [9.9, 45.2]	90.3 [78.3, 99.4]
24 months	Effortful control	0.006	71.4 [57.0, 82.6]	88.0 [66.2, 100]	54.8 [39.8, 69.0]	60.6 [47.0, 73.0]	29.2 [12.7, 46.5]	95.6 [87.6, 100]
	Surgency + Effortful control	0.021	66.7 [49.6, 80.8]	79.3 [49.3, 100]	54.1 [39.1, 68.8]	58.5 [44.9, 71.7]	26.7 [10.7, 44.3]	92.6 [80.8, 100]
	Effortful control + Negative affect	0.031	69.9 [55.0, 82.6]	82.6 [57.0, 98.9]	57.3 [42.4, 71.8]	61.7 [48.5, 74.7]	28.9 [12.3, 47.2]	94.0 [84.5, 99.7]
	All factors	0.020	71.5 [57.1, 83.5]	84.8 [60.5, 98.9]	58.2 [43.4, 72.6]	62.8 [49.4, 75.5]	29.9 [13.0, 48.5]	94.8 [85.7, 99.7]
Longitudinal trajectory	Effortful control	0.042	66.4 [49.2, 80.3]	79.3 [48.9, 100]	53.5 [38.9, 67.9]	57.9 [44.7, 70.6]	26.9 [11.2, 44.0]	92.3 [80.2, 100]
	Effortful control + Negative affect	0.013	67.6 [50.5, 81.7]	77.3 [47.3, 99.1]	57.8 [42.8, 72.1]	61.1 [47.9, 74.0]	28.0 [11.5, 46.2]	92.4 [81.0, 99.7]
	All factors	0.048	65.1 [48.1, 79.3]	75.2 [44.8, 97.9]	55.0 [40.2, 69.5]	58.5 [44.9, 71.7]	26.0 [9.9, 43.8]	91.5 [79.3, 99.3]

Notes. All classifiers reported in this table significantly differed from prediction at chance level (shuffle test $p<0.05$). All metrics are reported as *mean [95% confidence interval]*. 95% confidence interval was computed using bootstrap. The classifiers are divided based on the data used as features: data collected at 14 months, data collected at 24 months, intercept and slope of the longitudinal trajectory between 8 and 24 months at the individual level. The abbreviations: AUC = area under the curve; PPV = positive predictive value; NPV = negative predictive value.

Table 4. HR-ASD + HR-Atypical vs. HR-Typical. Performance metrics of temperament factors for classifying the HR atypical group as a whole (including atypically developing siblings and those who later develop ASD) from typically developing siblings.

	Classifier	*p*	AUC	Sensitivity	Specificity	Accuracy	PPV	NPV
24 months	Surgency + Effortful control	0.058	60.6 [47.0, 73.5]	63.7 [43.0, 82.9]	57.6 [39.6, 74.7]	60.2 [46.8, 72.8]	52.7 [34.2, 70.9]	68.2 [49.1, 85.4]
	Effortful control + Negative affect	0.056	61.1 [48.0, 74.1]	59.8 [39.2, 79.2]	62.4 [45.4, 78.9]	61.3 [48.3, 74.3]	55.0 [34.2, 70.9]	67.9 [49.9, 84.7]
	All factors	0.051	60.0 [46.3, 72.9]	58.8 [37.7, 78.7]	61.2 [43.7, 77.6]	60.2 [47.0, 72.6]	53.2 [33.0, 72.0]	66.7 [48.5, 83.7]

Notes. None of the classifiers performed significantly different from chance level (shuffle test $p<0.05$). Here we report classifiers performing marginally different from random. All metrics are reported as *mean [95% confidence interval]*. 95% confidence interval was computed using bootstrap. Abbreviations: AUC = area under the curve; PPV = positive predictive value; NPV = negative predictive value.