Title: The N170 face-sensitive brain response: toward a stratification biomarker for ASD

Mason, L.[1], Moessnang. C[2]., Chatham, C[3]., Ham., L.[3], Tillmann, J.[4], Dumas, G.[5,6], Ellis, C.[4], Leblond C.S.[5], Cliquet F.[5], Bourgeron, T.[5], Beckmann, C.[7], Charman, T.[4], Oakley, B.,[4] Banaschewski, T.[2], Meyer-Lindenberg, A.[2], Baron-Cohen, S.[8], Bölte, S.[9], Buitelaar, J.K.[7], Durston, S.[10], Loth, E.[4], Oranje, B.[10], Persico, A.[11], Dell'Acqua, F.,[4] Ecker, C.[12], Johnson, M.H.[8,1], Murphy, D.[4], Jones, E.J.H.[1] & the EU-AIMS LEAP Team*

[1] Centre for Brain and Cognitive Development, Birkbeck, University of London

[2] Department of Child and Adolescent Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany

[3] F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland.

[4] Roche Pharma Research and Early Development, Neuroscience and Rare Diseases, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd., Basel, Switzerland

[5] Human Genetics and Cognitive Functions, Institut Pasteur, UMR3571 CNRS, Université de Paris, Paris, (75015) France

[6] Precision Psychiatry and Social Physiology laboratory, CHU Sainte-Justine Research Center, Department of Psychiatry, University of Montreal, Quebec, Canada

[7] Department of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behavior, Radboudumc, Nijmegen, The Netherlands

[8] Department of Psychology, University of Cambridge

[9] Center of Neurodevelopmental Disorders (KIND), Centre for Psychiatry Research Child and Adolescent Psychiatry, Stockholm Health Care Services, Region Stockholm, Stockholm, Sweden.

[10] NICHE-lab, Dept. of Psychiatry, Brain Center of University Medical Center Utrecht, Utrecht, the Netherlands

[11] Universita Campus Bio-Medico, Rome, Italy

[12] Curtin Autism Research Group, School of Occupational Therapy, Social Work and Speech Pathology, Curtin University, Perth, Western Australia.

*Correspondence to: Emily Jones, e.jones@bbk.ac.uk

One sentence summary: We show that an early-stage neural response to faces (the N170) is slower in ASD; is associated with quantitative variation in common genes linked to autism; and may be useful in predicting change in real-world social functioning in a clinical trial context.

1. Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterised by difficulties in social communication, but also great heterogeneity. To offer individualised medicine approaches, we need to better target interventions by stratifying autistic people into subgroups with different biological profiles and/or prognoses. We sought to validate neural responses to faces as a potential stratification biomarker in ASD by measuring neural (electroencephalography/EEG) responses to faces (critical in social interaction) in N=436 children and adults with and without ASD. The speed of early-stage face processing (N170 latency) was on average slower in ASD than age-matched controls. In addition, N170 latency was associated with responses to faces in the fusiform gyrus during an fMRI task and polygenic scores for ASD, triangulating links to social biology. Critically, within the ASD group N170 latency predicted change in adaptive socialisation skills over an 18-month follow-up period; data-driven clustering identified a subgroup with slower brain responses and poor social prognosis. Use of a distributional data-driven cut-off was associated with predicted improvements of power in simulated clinical trials targeting social functioning. Taken together, this provides converging evidence for the utility of the N170 as a stratification biomarker to identify biologically and prognostically defined subgroups in ASD, and may provide a blueprint for similar endeavours in other psychiatric conditions.

One sentence summary: N170 latency to faces relates to fusiform activity and ASD genetics, predicts social prognosis, and could improve power in clinical trials.

2. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition associated with difficulties in social interaction and communication and the presence of restricted or repetitive behaviors or interests and sensory sensitivities[1]. The causes of ASD are highly heterogenous, with multiple identified genetic factors[2] and several possible environmental factors that likely interact with genetic background[3]. Symptom presentation is also highly variable, both in core symptomatology but also the presence or absence of a range of associated conditions like anxiety, depression, intellectual disability or language delay[4]. To move towards individualised interventions/support strategies, we need to stratify this heterogeneous population into more biologically and prognostically homogenous subgroups. This will allow particular support strategies to be better targeted to those most likely to benefit, and allow individuals with different prognoses to make better-informed decisions about intervention choices. To do this, it may be fruitful to focus separately on the social and non-social domains, given substantial evidence that they are both genetically[5] and phenotypically[6] separable.

Difficulties in engaging with the social world are central to the symptom profile of ASD. Social difficulties in ASD can compromise everyday adaptive functioning[7] and social withdrawal and associated loneliness are major risk factors for conditions experienced at higher-than-expected rates in autistic people, including depression and anxiety[8–10]. Thus, developing new support strategies for social functioning is critical to boosting mental wellbeing, adaptive functioning and independence skills, and quality of life for autistic

people[11]. However, social difficulties may be associated with many different neurobiological alterations; identifying "stratification biomarkers" -- objective measures used to identify more biologically or prognostically homogenous subgroups -- will be crucial to developing individualised support strategies. This could improve the ethics of support strategy provision through allowing participants greater personalisation of the risk-benefit ratio of particular approaches, and should improve statistical power and generalisability of clinical trials. Despite decades of research on ASD there remain no validated stratification biomarkers[12]. Key steps include identifying metrics that are individually reliable, mechanistically sensitive, relevant to a known biological system, predictive of prognosis, and that have a clear potential context of use for clinical settings [13].

Here, we focus on the N170 neural response to faces as a candidate stratification biomarker for social functioning in ASD[14]. Social interactions are complex and fast-moving, and as such expertise with faces is central to mature social interaction[15]. The N170[16] is a face-sensitive event-related potential of negative polarity peaking around 170ms after stimulus onset over occipito-temporal electrodes (representing the coordinated firing of groups of neurons in simultaneously active lateral occipital areas[17], including the fusiform gyrus[18]); it likely reflects face expertise[17] built through experience[19 20 21] and is sensitive to configural processing measured via stimulus inversion[22]. Critically, its amplitude and latency have moderate to good intra-individual reliability (0.6-0.8 [23–25]). In a meta-analysis, autistic people showed on average a longer latency (slower) N170 to faces (Hedges g=0.36) but not to non-face control stimuli, relative to neurotypical controls[26]. Within ASD, faster N170 latency relates to better holistic face processing[27]; stronger adaptive socialisation[28]; and fewer social

difficulties. Prognostically, N170 latencies have been related to trajectories of social symptoms from childhood to adolescence[29].

Although this early work is promising, there are several shortcomings. Some investigations have not observed consistent associations between variation in the N170 and aspects of face recognition or social behavior[25,30–34]. Sample sizes for most studies are small (n<100) and so power to detect relatively modest associations is low, particularly for subgrouping or stratification approaches; associations are also almost all cross-sectional and thus lack clinical and prognostic utility. Finally, there are no identified potential "contexts of use" (clinically useful applications) for the N170 in a clinical trial setting (such as use as an inclusion or exclusion criteria, as a surrogate endpoint, or as a baseline covariate). To move from strong preliminary evidence to validate the N170 as a stratification biomarker with clinical utility we need large and robust studies to show a) a replicable and robust signal of differences between autistic people and controls; b) that the N170 is mechanistically sensitive to alterations in expertise-based processing; c) that the N170 relates to variation in related biologically meaningful signals (e.g. derived from genetics, brain imaging); d) that the N170 correlates with a clinically-relevant behavioral measure and shows robust prognostic utility for later variation in social functioning in either all people with ASD, or a particular subgroup; and e) to define a clear context of use for the N170 within a clinical setting.

1.1. Present study

We examined whether the N170 meets criteria for a stratification biomarker for social functioning in a large heterogeneous population of individuals with ASD and controls tested

in a multisite European longitudinal study (the Longitudinal European Autism Project, LEAP[35]). Our final sample size of 436 is approximately the same size as the combined (independent) sample size included in a recent meta-analysis[26], and several times larger than any previous individual study. First, we examined whether differences in N170 latency relate to clinical phenotype (categorical ASD diagnostic status). Second, we examined mechanistic sensitivity by testing whether N170 latency associates with measures of expertise-sensitive processing (the magnitude of the face inversion effect in EEG[16]). Third, we tested whether N170 latency associates with ASD-relevant biological pathways by examining its association with the magnitude of brain responses to faces in core social brain areas measured through fMRI[36]; and to variation in ASD polygenic scores (PGS) - the aggregated effect of many common variants previously associated with ASD. Fourth, we examined prognostic utility by testing whether N170 latency predicts later social functioning within the ASD group. For this purpose we selected the Vineland-II Socialisation scale and its constituent subdomains (Play and Leisure Time, Coping Skills and Interpersonal Relationships) because a) this provides a measure of 'real world' adaptive social functioning suited to measuring the kinds of social activities that may be associated with improved quality of life and thus potentially valuable to target with support or intervention[37]; b) the face validity of a link between the N170 and social function and c) previous evidence of associations between the N170 and the Vineland Socialisation domain[28,29]. We examined both dimensional associations across the full cohort, and within data-derived clusters (defined on the basis of averaged EEG data). To further probe clinical significance, we then examined whether the subdomains of the Vineland socialisation scale at follow-up that associated with the N170 were themselves related with

broader measures of clinical change (Clinical Global Impressions[38]) and quality of life (Child Health and Illness Profile[39]). Finally, to explore putative context-of-use we used Monte-Carlo simulations to model the potential gain in power from using an exemplar distributional-defined N170 cut-off (derived by creating a normative model of N170 latency change with age, then computing the position of each individual with autism within that distribution[40]) to enrich a clinical trial for participants who are more likely to show no spontaneous improvements in their social functioning over time.

2. Results

2.1. Relation between N170 and Clinical Phenotype

To ensure relevance to autism as a categorical diagnosis, we first sought to establish that we replicated key diagnostic effects reported in previous work within our large heterogenous sample[26]. As expected, N170 latency (N170L) to upright faces decreased with age ($F(2,430)=49.78$, $p<.001$, $\eta_p^2= 0.188$; Figure S1), and was on average slower in the ASD than the control group ($F(1,430)=9.43$, $p=0.002$, $\eta_p^2= 0.021$; Figure 1). N170L did not differ across hemispheres ($F(1,430)=.68$, $p=0.41$, $\eta_p^2= 0.002$) and diagnosis, age and hemisphere did not interact (all $Fs<1.7$, all $ps>0.2$, $\eta_p^2s< 0.004$).

Adding site and sex and their interactions with diagnostic group did not alter the statistical pattern (main effect of group on latency ($F(1,426)=7.03$, $p=0.02$, $\eta_p^2= 0.013$); diagnostic group effects did not significantly vary by site ($F(1,426)=1.36$, $p=0.25$, $\eta_p^2=0.003$), or sex ($F(1,426)=0.16$, $p=0.69$, $\eta_p^2=0.000$) and sex had no overall effect on N170L ($F(1,426)=.07$, $p = 0.80$, $\eta_p^2<0.001$). The main effect of diagnostic group also survived controlling for medication use in those who provided reports (SM2.1; $F(1,415)=5.35$, p 0.02, $\eta_p^2=0.013$). There were no significant diagnostic group differences for N170 amplitude or P1 amplitude and latency (see SM2.3, 2.4) and results were robust to variation in attention to the stimuli during the task (SM 2.2; Figure S2). Consistent with our expectations, N170L did not show strong promise as a diagnostic biomarker. Using leave-one-out cross validation, a logistic regression predicting ASD status from N170L was significant (est = 9.57, SE=3.30, z=2.9, p = 0.0037). However, the area under the curve was 0.56 indicating relatively poor prediction,

with a sensitivity of 0.19 and a specificity of 0.81. Separating the group by sex suggested slightly better performance in males (est = 11.92, SE=4.1, z=2.91, p = 0.0036, AUC =0.57, n=301, sens= 0.13, spec=0.88) than females (est = 5.31, SE=5.7, z=0.93, p = 0.35, AUC= 0.46, n=135, sens = 0.61, spec=0.33).

2.2. Relation between N170 and Core and Associated Symptoms:

Correlations with symptomatology were computed within the ASD group and controlled for age (for controls see SM Table S6). Faster N170L related to fewer examiner-rated social symptoms (Autism Diagnostic Observational Scale/ADOS Social Affect) (r(209)=.20, p=0.003) but not restricted and repetitive behaviors (r(209) = .064, p=0.36; see Table S6). No associations were observed with associated symptoms (internalising, externalising or IQ), indicating specificity (Table S6). If sex was added to the model, there were no significant sex differences in the magnitude of the association between N170L and ADOS Social Affect (ANOVA including sex, age, N170L and sex*N170L as predictors and ADOS Social Affect as the dependent variable; interaction between sex and N170L F(1,207) =.66, p = 0.42, η_p^2= 0.003).

2.3. Mechanistic Relevance: The Inversion Effect

Inverting a face disrupts processing of its configuration and extraction of identity; thus, if case-control differences in the N170L represent altered face processing we should see corresponding differences in the modulation of the N170 by face inversion. The ASD group indeed showed diminished inversion effects (response to upright vs inverted face stimuli) for

N170L relative to the control group (interaction: F(1,430) = 7.67, p = 0.006, η_p^2=0.018); ASD F(1,243) = 0.02, p = 0.9, η_p^2=0.000; controls F(1,187) = 15.86, p <0.001, η_p^2=0.078; SM2.6). If sex was added to the model, the magnitude of the group difference in the inversion effect did not vary by sex (F(1,427) = .27, p = .61, η_p^2=0.001). Consistent with N170 sensitivity to configural processing, within the ASD group N170 latency to upright faces correlated with the magnitude of the effect of inversion on N170 amplitude and latency (age controlled, latency inversion (r(246)=0.49, p<0.001; amplitude inversion r(246)=-0.23, p<0.001).

2.4. Triangulation with other biological measures

2.4.1. Relation to face-sensitive fMRI responses

To establish whether our measured N170 signal was related to core face-sensitive brain regions, we used functional magnetic resonance imaging (fMRI) to assess the blood oxygenation level dependent (BOLD) response within a bilateral, a-priori defined anatomical mask of the fusiform face area (FFA; 2318 voxel, small volume correction; Figure S3 [41]). FFA is a brain region considered as one of the primary sources of the N170 response[18]. Face-sensitive responses of the FFA were assessed as differential BOLD response to a face matching condition compared to a control condition (see Methods 4.6). Across the ASD (n=99) and control groups with data available (n=100), N170 latency associated with the face-sensitive response in the right fusiform gyrus (peak voxel at Montreal Neurological Institute/MNI [30 -64 -10] (t=3.93, P_{SVC}=.032; R^2=.131, N170 β=0.285, p < 0.001; Figure 2)

with fair regional specificity (see Figure S4); this effect was not modulated by diagnostic group ($F(1,190) \leq 7.38$, $P_{SVC} \geq .860$) or age ($F(1,188) \leq 9.09$, $P_{SVC} \geq .619$).

2.4.2. Relation to common genetic variation

To establish whether the N170L related to core genetic variation related to ASD, after quality-control (see SM1.5) we computed PGS scores for 350 individuals with autism and 242 controls from European descents using the PRSice-2 tool[42]. Within this sample 198 ASD and 133 controls had EEG data. For the ASD PGS and a range of other comparisons (see SM Table S5), the genome wide association study (GWAS) summary statistics were used as a reference (all with an additive model). Results showed that longer N170 latency associated with higher PGS for ASD (Figure 3); Spearman's $r2$=0.026; p=0.0031; participants with ASD $r2$=0.022; p=0.039; controls $r2$=0.024; p=0.074) but not comparison traits such as schizophrenia, brain volume, intelligence or body mass index (Figure S5). Of note, the correlation between N170L and ASD-PGS remains significant when the latter was computed with the new SBayesR method[43] (p= 0.035). Interestingly, there was also a significant positive correlation with the PGS for Attention Deficit Hyperactivity Disorder (ADHD)[44] ($r2$=0.01; p=0.039) and a negative correlation with a recent PGS computed for scores on the "Reading the Mind in the Eyes" test[45] ($r2$=0.01; p=0.03), a measure of cognitive empathy that can be more challenging for autistic participants[46].

2.5. Prognostic Utility

2.5.1. Dimensional Relation to Vineland Socialisation

To determine whether the N170L may have prognostic utility, we examined relations between N170 at the first assessment wave with changes in Vineland Socialisation and its subdomains at the second wave. Within the ASD group, and controlling for age and baseline scores, simple partial correlation showed that faster N170 latency at baseline associated with greater improvement in the Vineland socialisation domain's Play and Leisure Time subdomain v-scale scores between baseline and the follow-up visit ($r(141)=-0.235$, $p=0.005$; Bonferroni-corrected for four comparisons $\alpha=0.02$). If sex was included, the association between N170L and change in Play and Leisure V-scores did not significantly vary in strength by sex (ANCOVA; $F(1,139) = 0.055$, $p = 0.82$, $\eta_p^2 = 0.00$). Other scales did not show this relation (*ps*>0.5, SM Table S7) and this was not confounded by variable time delays between baseline and follow-up (SM2.8.1). Using leave-one-out cross validation and controlling for baseline score and age, a linear regression confirmed a significant predictive relation between N170L and change in Vineland Play and Leisure V-scores (est.=-24.39, SE=8.48, $t(143)=-2.88$, $p = 0.0046$). Controlling for the effect of site ($t(142)=1.59$, $p = 0.11$) did not remove the effect (est.=-21.80, SE=8.58, $t(143)=-2.54$, $p = 0.01$). Confirming its clinical relevance, Play and Leisure V-scores at follow-up significantly varied across the five outcome categories of the Clinical Global Impressions scale (caregiver judgment of change between baseline and follow-up expressed as "a lot/ a little worse", "about the same", "a little/ lot better"; $F(4,180)=3.78$, $p = 0.006$, $\eta_p^2 = 0.079$; controlling for age $F(4,180)=3.49$, $p = 0.009$, $\eta_p^2 = 0.074$) and higher scores at follow-up were cross-sectionally associated with higher scores for Achievement ($r(162)=0.36$, $p < 0.001$; controlling for age $r(158)=0.38$, $p < 0.001$), Satisfaction ($r(176)=0.21$, $p = 0.004$; controlling for age $r(173)=0.23$, $p =0.002$),

Comfort ($r(176)=0.17$, $p = 0.023$; controlling for age $r(173)=0.17$, $p =0.028$), and Resilience ($r(176)=0.20$, $p = 0.007$; controlling for age $r(173)=0.23$, $p =0.001$) as measured by the Child Health and Illness Profile.

2.5.2. Cluster analysis of grand-average EEG responses

We examined whether a cluster analysis computed on individual EEG averages at four key electrodes (P7, P8, O1 and O2) would reveal underlying 'subgroups' of participants. BIC indicated that three clusters was the most parsimonious model (BIC=33483, AIC=33294), converging after 26 iterations with a negative log-likelihood of 16603. Table S8 (SM section 2.8.2) shows diagnostic and clinical profiles of the three clusters within the ASD group (Cluster 1, n=118, 48%; Cluster 2, n=27, 11%; Cluster 3, n=101, 41%); briefly, clusters did not differ in symptom severity, IQ or sex but were significantly different in age ($F(2,245) =57.29$, $p < 0.001$, $\eta_p^2=0.320$; Cluster 1 M=20.6y, SD=4.3y; Cluster 2 M=15.1y, SD=4.8y; Cluster 3 M=12.1y, SD=4.7y). A significant difference was observed between the three clusters in N170 latency ($F(2,245) =64.32$, $p < 0.001$, $\eta_p^2= 0.975$; Figure 4; controlling for age ($F(2,245) =31.991$, $p < 0.001$, $\eta_p^2=0.209$). Bonferroni-corrected post hoc tests confirmed that Cluster 2 had significantly longer latencies than Cluster 1 ($p=0.021$) or 3 ($p<0.001$); Cluster 3 had shorter latencies than Cluster 1 ($p<0.001$). This analysis confirms that the N170 latency captures a meaningful proportion of variance in the multidimensional EEG waveform.

Clusters differed in Play and Leisure Time scores ($F(2,144) =4.41$, $p =0.014$, $\eta_p^2=0.06$; Figure 4; controlling for age ($F(2,144) =2.21$, $p =0.11$, $\eta_p^2=0.03$). Cluster 2 (with the slower

N170 latency) had significantly smaller changes in Play and Leisure Time v-scores between baseline and follow-up than Cluster 3 (p=0.019). Within Cluster 2, the association between N170 latency to upright faces at P7 and P8 explained over 25% of the variance in the change in the same subdomain Vineland Socialisation (Play and Leisure Time) scores between baseline and 18 month follow-up visit (r(19)=-0.517, p=0.023; controlling for age r(16)=-0.56, p=0.015).

2.6. Towards clinical utility

We sought to explore the potential for selecting a cut-off latency that would indicate an individual with a lower likelihood of a spontaneous improvement in social skills who thus may wish to be enrolled in a relevant intervention trial. As N170 varies with age, we used a normative modelling approach to generate age-corrected z scores (zN170L) for each participant with ASD relative to the control group[40], see SM 2.8.3, Figure S6). Play and Leisure Time v-scores were dichotomised into those who improved (scores showed an absolute increase) vs those who did not improve (scores remained the same) or declined (scores became worse). We selected an absolute threshold for v-scores given the moderate to good psychometric properties of the Vineland (see SM2.8.3.4), and the absence of any clinically established cut-off for clinically meaningful improvement. By cluster, 68% (n=13/19) of individuals in Cluster 2 did not improve/declined, compared to 58% (n=42/73) in Cluster 1 and 42% (n=22/53) in Cluster 3 ($\chi^2(2)$= 5.23, p =0.07). Using leave-one-out cross-validation and logistic regression to predict improvement vs declining v-scores from zN170L (est. -0.56, SE=0.18, z=-3.1, p = 0.002) yielded an area under the curve (AUC) of

0.65, at a sensitivity of 0.66 and specificity of 0.59 (males only: est. -0.54, SE=0.21, z=-2.61, p = 0.009, AUC=0.66, sens=0.67, spec = 0.60; females only: est. -0.65, SE=0.38, z=-1.73, p = 0.08, AUC=0.57, sens=0.61, spec = 0.5). Of note, the selection of a specific cut-off will depend on context of use. For example, a bootstrapped optimisation analysis (R, package cutpointr, 1000 iterations) indicated that in the present sample a median optimal cut-point of +0.58SD may be optimal to maximise specificity whilst keeping sensitivity reasonable (median sens=0.51, spec=0.78); conversely, to maximise sensitivity (for inclusion of as many non-responders as possible) a median cut-point of -0.02SD may be optimal (median sens. = 0.71, spec. = 0.53; see (SM2.8.3.2, Figures S7 and S8).

To provide a simplified worked example of the potential utility of the zN170L in a clinical trial, we used Monte Carlo-based clinical trial simulations to compare the statistical power by sample size in trials with and without N170 latency enrichment based on an exemplar +0.5SD cut-off (SM 2.8.3.3). A receiver operating characteristic curve (SM2.8.3.2) indicated that this exemplar N170L (normative) cut-off of +0.5SD in the present sample yields a moderate Sensitivity of 0.55 (the proportion of individuals falling above the cut-off who won't improve/ will decline) and Specificity (the proportion of individuals who fall below the cut-off who will improve spontaneously) of 0.76. In those with ASD above the +0.5SD cut-off 72% (n=42/58) did not improve/declined, compared to 40% (n=35/87) of those below the +0.5SD cut-off (χ^2 (1)= 14.5, $p < 0.001$). Briefly, 2500 randomized (1:1), placebo-controlled, 12-week clinical trials with and without enrichment were simulated using an estimated fixed effect size of intervention of Cohen's D=.45. Based on interpolation across the simulations, approximately 78 subjects per arm would be required in a non-enriched placebo-controlled

clinical trial to detect a beneficial drug effect of equivalent magnitude with an 80% probability (type II error or $\beta = 0.20$) at $\alpha = 0.025$ (one-sided, or [equivalently] $\alpha = 0.05$ two-sided). Conversely, the same 80% probability of detecting an analogous drug effect at the same α is achieved with approximately 48 subjects per arm in an enriched clinical trial. This represents a reduction in sample size of approximately 38% (Figure S7).

3. Discussion

We provide evidence that the N170 is a promising stratification biomarker that may have utility in clinical trials. Specifically, this would mean that individuals with a relatively longer N170 latency could be enrolled in a trial because they would be (probabilistically) less likely to show spontaneous improvement in their social adaptive functioning over time than those with a shorter N170. First, we show *sensitivity to clinical phenotype*: that as a group, individuals with ASD show slower N170 responses to upright faces, replicating a recent meta-analysis[26]. This effect does not vary with age, sex, attention, collection site or medication and was not confounded by IQ or associated conditions such as ADHD or anxiety. Second, variation in N170 latency associates with a marker of configural processing (the face inversion effect). Third, we show relation to other biological variables: variation in N170 latency across the cohort was associated with higher polygenic liability for ASD and with the fMRI response of a core brain region involved in face processing, the fusiform gyrus[47]. Fourth, we demonstrate *potential prognostic utility.* Variation in N170 latency associates with social clinical prognosis (change in Vineland Socialisation Play and Leisure Time scores, a subdomain that at follow-up associated with overall global impressions of change between baseline and follow-up and concurrently relates to key measures of quality of life) over an 18-month period in both dimensional and subtype analysis. Finally, we further define a potential *context of use:* we show that data-derived cut-off scores could provide substantial efficiencies in clinical trials, reducing the magnitude of potential placebo effects. Taken together, we contend that the N170 meets core criteria for consideration as a stratification biomarker for ASD.

3.1. Case/control differences

Our work replicates and extends previous demonstrations that groups of individuals with ASD show slower latency N170 responses than controls (of whom a proportion had mild intellectual disability of varied etiology)[26]. In our cohort, this was not confounded by associated internalising or externalising symptoms, IQ, or medication; the only baseline association was with observed social symptoms, a core aspect of the ASD phenotype. Delays in N170 latency are not specific to ASD – groups with conditions like schizophrenia also show alterations in N170 amplitude that relate to general face recognition ability[48]. However, schizophrenia shows substantial genetic overlap with ASD[49] and this is associated with common molecular brain-based phenotypes[50]. Thus, markers that carve heterogeneity within ASD are likely to operate transdiagnostically[51]. This observation might affect utility as (for example) a putative diagnostic biomarker (see below), but does not reduce the utility of the N170 as a potential *stratification* biomarker that may help us parse heterogeneity *within* cohorts with ASD. However, our ability to draw inferences about the degree to which the current findings regarding stratification are specific to ASD or would generalise to other conditions is limited because we did not include a control group with another developmental condition, and this will be an important step for future work.

We did not observe sex differences in N170L or in the magnitude of group differences in N170L, nor in the relation between N170L and concurrent or future measures of social behaviors. Previous observations of faster N170L in neurotypical females than males[52] (N=152M, 141F) and slower N170L in autistic females than males[37] (N=12M, 12F)

may have led to the expectation of greater group differences in females than males, but this was not borne out in the present sample. However, the predictive relations between N170 latencies and both diagnosis and change in socialisation scores were numerically stronger and only statistically significant (p<0.05) in males. These analyses should be considered in the light of our 3:1 sex ratio, which may have affected our ability to detect meaningful differences in the profile of autistic females and males or to detect effects within autistic females analysed separately. Future investigation in more balanced samples is warranted to establish whether biomarkers need to be sex stratified. Further, we did not measure gender identity, which may have a different influence on social processing than sex. Caution should thus be exercised when generalising our results to populations less well represented in our dataset.

3.2. Relation to Genetics

We show that polygenic ASD scores computed from a GWAS including over 18,000 people with ASD and 27,000 controls[53] correlate with variation in N170 latency. A previous twin study reported a genetic contribution to the N170[54]. Our study suggests that this N170L heritability is positively correlated with the heritability of ASD (and to other psychiatric heritability since ASD is genetically correlated to other conditions). It is interesting (and somewhat expected) that the strongest observed correlation is with the ASD PGS, since individuals with ASD display replicable differences in the N170 response to faces relative to controls[26]. Of note, one limitation is that the proportion of variance explained by the current polygenic ASD score is relatively low (c. 2%)[53]. Clearly, there are many other processes

implicated in autism, and many genes that remain to be identified. Nonetheless, this provides an important first step to showing that delays in N170 latency and ASD may share genetic variance. Further work could complement this approach by examining the N170 in participants carrying large-effect genetic mutations conferring liability to autism that act on putatively more subscribed neural pathways. Importantly, a recent study showed an association between PGS for ASD and an infant precursor of the N170 at 8-months that also relates to later diagnosis[55], suggesting that effects of genes linked to ASD on the neural correlates of face processing may emerge very early in development and could play a role in causal paths to symptom emergence. Of note, the correlations between PGS-ASD and N170L we observed were present in both autistic people and controls (albeit at trend level). This may be consistent with other evidence that the genetic etiology of dimensional variation in autistic traits is similar to the etiology of autism diagnoses[56] and the proposal that dimensional variation in autistic traits is underpinned by the combined effects of multiple dimensional developmental alterations[57], one of which may be indexed by a longer N170L.

Interestingly, N170L also associated with polygenic scores derived from a GWAS from 89,553 people who completed a measure of cognitive empathy called the "Reading the Mind in the Eyes" test[45]. The test requires interpretation of emotional expressions from viewing isolated pictures of eyes, and is something with which some autistic people have difficulty[46]. Although replication is needed before strong conclusions can be drawn, given the slowed N170L in autistic people vs controls is strongest when attention is directed to the eye region of faces[58], this may point to shared neurobiological pathways via the role of eye gaze on advanced emotion recognition and early-stage face processing. Further, N170L associated

with the PGS for ADHD[11] (but not with behavioral measures of externalizing). Although ASD and ADHD show substantial familial co-aggregation[12], the overlap in variant weighting in the ASD and ADHD polygenic scores tends to be more limited[2]. In part, this is likely because of the relatively small proportion of phenotypic variance currently explained by these scores. Additionally, whilst the ASD PGS score captures variance that appears relatively specific to ASD[49,59], the ADHD PGS relates to a broader range of conditions that include bipolar disorder, schizophrenia and depression[49] and with age begins to explain independent aspects of both general psychopathology and condition-specific externalising behaviour[59,60] The lower specificity of the ADHD PGS may be because the phenotypic definition of ADHD varies more between included cohorts (including some that are solely questionnaire based[44]) than for the ASD PGS (which requires a consensus community clinical diagnosis[53]). Although again replication is necessary, our results may thus suggest there are concurrent genetic contributions to N170L from both a relatively ASD-specific mechanism and a more general liability to broad psychopathology, which may begin in early development[55].

3.3. Mechanistic Utility

How does the N170 inform the mechanisms underlying social difficulties in some individuals with autism? The rich neurotypical data on the N170 provides us with many avenues for investigation. First, N170 response to faces may index the action of a dedicated face processing system that is innately programmed and selectively and exclusively engaged in the processing of faces[61]. Notably, temporally earlier components (e.g. P1) related to general

attention did not vary between groups (SM2.4). If so, our results may indicate some very early-stage alteration in face processing systems that could compromise subsequent social development[62,63]. When taking an individual differences approach, face-selective responses in the temporal lobe (e.g., fusiform gyrus) are highly correlated with the N170 component[64], as is the case in the present study. Alternatively, "it may not be the years that matter, but the mileage" – that is, the N170 may be more influenced by experience than maturation[36]. In the present sample faster N170 latencies associated with the magnitude of face inversion effects over both the latency and amplitude of the N170, supporting its relation to configural processing[65,66]. Configural processing develops more gradually than featural processing[67]. The face-sensitivity of the N170 may thus reflect the outcome of an expertise-based process of learning about faces[68]. Distinguishing between these possibilities is an important step for future work.

3.4. Prognostic Utility

We provide evidence that variation in N170 latency predicts change in social adaptive behaviour over an 18- to 24-month period. This is consistent with reports of concurrent relations between the toddler precursor of the N170 (the N290) and social adaptive behaviors[28], and predictive relations between N290 latency and trajectories of observation-measured social symptoms on the ADOS[29]. We observed dimensional relations between N170 latency and less progress in the Vineland Socialisation Play and Leisure Time subscale; relations were stronger within a data-driven subset of individuals who had particularly slow N170 latencies and no or negative change in Vineland scores. Importantly, Play and Leisure

Time scores associated with measures of Quality of Life, supporting their clinical relevance. Further, we provide a worked example of how such insights could be used to yield benefits within a clinical trial context.

This result requires replication. We did not predict that such a relation would be specific to the Play and Leisure Time subscale. This scale asks about turn-taking, understanding of rules, and independent social activity, which have face validity for activities that may reflect expertise in processing information from faces and people more broadly. Unlike for the broader Vineland scales, no estimates of minimal clinically-meaningful changes are available for subscales[69], and this is an important task for future work (in addition to establishing whether the items included are relevant and meaningful as endpoints for autistic people). Further, although we used leave-one-out validation to verify the predictive relation between the N170L and the Vineland subscale, an external replication dataset remains important. Although this was a multisite study and site did not explain significant variance (p>0.05) in prognosis or N170L, each site did not recruit sufficiently large or representative samples with prognostic data to test the generalisability of predictive models at individual sites. We must also explain why associations with this Vineland scale were solely prognostic, and not concurrent. This pattern was also observed in a longitudinal study from childhood to adolescence that found associations between the latency of the developmental precursor to the N170 and the slope of change in ADOS social symptoms over development, but not the intercept (i.e. concurrent symptoms)[29]. Changes in the brain may precede the emergence of changes in behaviour if changes in perception or attention affect learning from the environment, which over time has cumulative effects that subsequently manifest in

behaviour[70]. Predicting future trajectories may also prove more powerful than relating brain measures to concurrent behavioral measures, in part because measuring change in a single variable within a participant can add information if the baseline and follow-up measures are strongly correlated, as in this case[71]. However, large-scale rigorous tests of underpinning models will be required to make progress in this area.

3.5 Limitations

We did not include groups with other diagnoses, which could have probed specificity of prognostic validity to autistic people; no estimates of minimal clinically-meaningful changes are available for Vineland subscales, making it difficult to identify an appropriate cut-off for change over time; we did not have access to an external dataset in which to replicate our prognostic associations; although our sample was large, the sample size at individual sites was insufficient to test formal replication of findings across locations; we did not consider both sex and gender, which will be important in future work; we did not include an assessment of the meaningfulness of the Play and Leisure Time subscale to autistic people, which will be critical to judging its value as a putative intervention target.

3.6. The N170 as a Stratification Biomarker

The utility of the N170 as a putative biomarker has been widely debated (e.g.[25,72–74]). Importantly, individual differences in N170L are moderately reliable in test-retest assessments[23–25], and were strongly split-half reliable in the present cohort (SM2.11). Utility as a diagnostic biomarker is clearly limited by the substantial population overlap between

individuals with ASD and controls illustrated in the present study, and the presence of N170 delays in other conditions like schizophrenia[48]. Use as a proxy endpoint for clinical trials would require more rigorous data on phenotypic association than is available to date[25]. However, the N170 may be more appropriate for consideration as a trial enrichment marker. This would entail the use of the N170 to select a subset of a population of individuals with ASD for entry into trials targeted towards social functioning. Such 'trial enrichment' biomarkers[75] are used at the discretion of those designing support strategies. In this context, perfect sensitivity and specificity to diagnostic category would not be expected. If the N170 in part reflects an index of social expertise, individuals with ASD who have a slower N170 latency may be statistically more likely to have a poorer prognosis in their social functioning and benefit more from targeted social support strategies. Our study provides evidence for prognostic value on a subdomain of the Vineland through both dimensional and categorical analysis approaches. We also show clear proof of principle that data-driven cut-offs can identify inclusion criteria that could be used to target clinical trials to those less likely to spontaneously improve, improving power and efficiency. This is important not only in reducing the magnitude of expected placebo effects[76], but also in improving the risk benefit ratio and ability to make informed choices for individuals, although it is also important to note that restricting trial inclusion based on an N170 criteria would make recruitment even more challenging. Of note, the cut-off we chose to model was arbitrary, and investigators may choose a range of cut-offs depending on their goals. To be fully validated for clinical use, particular cut-offs would need to be replicated in an independent sample. An alternative approach that does not require an arbitrary selection may be to use the N170L as a baseline

covariate in a clinical trial to improve the precision of statistical estimates of effects. Future work should test whether the subgroups we identify may also be more likely to benefit from particular support strategies targeted to relevant biological or social systems. In summary, the promise of stratification biomarkers in psychiatry has long been recognised but not yet realised. Our work may provide a blueprint for the next generation of research studies to move from biomarker discovery to validation in order to deliver optimal outcomes for autistic people.

4. Methods

4.1 Study design

Data was taken from the Longitudinal European Autism Project (LEAP), a European multisite longitudinal observational study with two complete waves of assessment and a third ongoing; for a comprehensive clinical characterisation of the full LEAP cohort see[4]. The full LEAP sample comprises n=453 autistic and n=311 control participants, which was based on power calculations showing that this sample size was sufficient to detect small effect sizes in the full cohort, and moderate effect sizes if split into 2, 3 or 4 subgroups (see Additional file 3 in[35]). SM1.2 provides full inclusion/exclusion criteria. Participants are aged between 6 and 31 and participate in a battery of assessments that included clinical measures, EEG, eye-tracking, neuroimaging, genetics and cognitive testing; the protocol received advice from the European Medicines Agency to ensure suitability for regulatory submission[77]. At each site, an independent ethics committee approved the study; all participants (where appropriate) and their parent/legal guardian provided written informed consent. The objectives of the LEAP

study are to identify stratification biomarkers for autism; in this analysis, we predicted that the N170 derived from EEG responses to faces would predict social trajectories, based on previous literature[28,29].

4.2. Participants

We included the total sample of 436 participants with and without autism with valid EEG data, ranging in age from 6-31 years and with full-scale IQs between 50 and 148 (Table S1B; SM1.1). No outliers were excluded.

4.3. Clinical Measures

The *Autism Diagnostic Observation Schedule* (ADOS[78]), a standardised social interaction observation assessment, was used to assess current symptoms in ASD participants (Module 1: n=1; Module 2: n=1; Module 3: n=102; Module 4: n=140; missing: n=2). Calibrated Severity Scores (CSS) for Social Affect (SA), Restricted and Repetitive Behaviors (RRB) and Overall Total (range 1-10; higher – more severe) were used to provide standardised autism severity measures that account for differences in the modules administered. Internalising and externalising behaviours were measured using the *Development and Well-Being Assessment (DAWBA;*[79]*),* a semi-structured parent/carer interview designed to generate six categories of prediction scores (very unlikely (~0.1%) to probable (risk score >70%)) for ICD-10 [80] and DSM-IV-TR [81] psychiatric diagnoses.

Cognitive function (IQ) was assessed with either the *Wechsler Abbreviated Scales of Intelligence-Second Edition* (WASI-II), or if unavailable the WISC-III/IV in children and

WAIS-III/IV in adults. Adaptive skills were measured using the *Vineland Adaptive Behavior Scale-Second Edition* [82] , a semi-structured parent interview that assesses adaptive functioning. Because of our interest in the social domain, we used the standard scores from the Socialisation domain of the Vineland and the v-scale scores (*M*=15, *SD*=5, lower=fewer skills) from its three constituent subdomains (Play and Leisure Time, Interpersonal Relationships and Coping Skills). In addition, Quality of Life and everyday adaptive functioning (see[83]) were measured with the *Child Health and Illness Profile – Child Edition*, a 45-item parent-report measure consisting of five domains: physical/psychological Comfort (*How often did your son/daughter have pain that really bothered him/her?*); Satisfaction (*How often does your son/daughter feel happy?*); Resilience (*How often does your son/daughter have an adult he/she can go to for help with a real problem?*); Risk Avoidance (*How often does your son/daughter do things that are dangerous?*); and Achievement (*How did he/she do in his/her schoolwork?*).

4.4. Follow-up Assessment

Participants were invited to complete a second wave of clinical assessments 12- to 24-months after their baseline visit (see Table 1). Of the 436 participants included at baseline, 311 participants (71%) returned on average for a second visit 19.6 months after the baseline visit (*SD*=3.3 months; Min=12.2 months, Max=30.5 months). We used follow-up data from the Vineland Socialisation domain (available for *N*=153 participants), and its three subdomains (play and leisure time, interaction and coping skills). The Clinical Global Impressions Scale (CGI) was used to ask the parent or participant (adults) about their perceived overall change

between baseline and follow-up; categories are “A lot worse”, “A little worse”, “About the same”, “A little better”, “A lot better”).

4.5. EEG

EEG procedure: Five sites acquired EEG data at baseline, following international standards[84] using three different systems (SM1.3.1). Testing teams from each site attended initial training in London in 2013, followed by site visits to ensure correct set-up of equipment and that SOPs were followed. All sites then attended weekly telephone conferences to discuss data acquisition and quality.

Face task design: Using TaskEngine[85]/Presentation, participants were presented with three upright or inverted faces (Caucasian, African-American and Asian[86], subtending 12.4 degrees), repeated 168 times over four blocks (Figure S1 left). Each trial began with a randomly-selected fixation icon (2.9 degrees of visual angle positioned where the eye region of the face would subsequently appear in both upright and inverted conditions.

EEG processing: Data were uploaded from each site to a central repository in their raw, manufacturer-specific, proprietary formats and pre-processed in EEGlab[87] to harmonise data in a common format (62 channel montage, referenced to FCz with sampling rate 1kHz; SM1.3.1). Visual stimulus timing was measured and corrected at all sites except for UCBM using a photodiode (see SM1.3.2.1). In Fieldtrip[88], raw EEG data were epoched from -200ms to 800ms post stimulus-onset; bandpass filtered 0.1Hz-30Hz with 2000ms padding; and resampled to 500Hz. Artifacts were identified and removed with a custom-written automatic

algorithm (SM 1.3.2.3.) and whole-scalp artifacts (voltages > +/-100µV or a range of > 150µV or 0 µV) were detected and interpolated using a spherical spline algorithm where at least 3 neighbouring channels were artifact free. EOG artifacts were detected on frontal electrodes FP1/z/2, AF7/8 and contaminated trials removed. Grand averaged data was corrected to baseline (mean amplitude from -200 to 0ms) and average re-referenced. Finally, P1 (O1/O2) and N170 (P7, P8) peak amplitude and latency were extracted through an automatic algorithm with hand supervision (see SM1.3.2.4; SM Table S2 and S3).

4.6. fMRI Collection and Processing

Functional brain responses were acquired on 3 Tesla MRI scanners as part of the LEAP protocol using a well-established face matching task[89], with alternating blocks of faces (showing angry and fearful emotions) and control conditions. In the emotional face condition, a target face had to be matched to one of two probes (identity match) by pressing the left or right button of a response device. Analogously, in the control condition, participants were asked to match a target shape (circle or ellipses) to two test shapes. fMRI and EEG data were available for 99 individuals with ASD and 100 controls. Functional imaging data were preprocessed and statistically analysed using standard analysis routines implemented in SPM12 (http://www.fil.ion.ucl.ac.uk/spm/; see SM1.4). For each individual, we estimated one contrast image that reflected brain areas with higher sensitivity for emotional faces compared to shapes. These contrast images were subjected to a group-level analysis to assess the association of face-sensitive functional responses with the EEG derived N170 latency.

4.7. Genetics

SNP genotyping was performed at the "Centre National de Recherche en Génomique Humaine (CNRGH)" using the Infinium OmniExpress-24v1 BeadChip (> 700K markers) from Illumina. After quality-control and ancestry correction (see SM1.5) the PGS was computed for 350 individuals with autism and 242 controls from European descents using PRSice-2 tool[42]. Within this sample 198 ASD and 133 controls had EEG data. For the ASD PGS and a range of other comparisons (see SM Table S5), the genome wide association study (GWAS) summary statistics were used as a reference (all with an additive model). For the linkage disequilibrium-based single nucleotide polymorphism (SNP) pruning only SNPs with a MAF>1% and with a $R^2 < 0.1$ in windows of 500kb were selected. PGS were adjusted for principal component ancestry using PC1 to PC4 and standardized (z-scored) using the European typically developing participants as a reference.

4.8. Statistical Approach

Analyses were run in SPSS24 and R1.3.959 and corrected for multiple comparisons within each core question; statistical thresholds were set at two-sided $p<0.05$ unless otherwise noted: a) Relation to clinical phenotype: We used linear modelling of the latency of the N170 on face upright trials by diagnosis (ASD/control) x hemisphere (left [O1/P7] / right [O2/P8]) x age group (children 6-11 / adolescents 12-17 / adults 18-30). We additionally tested stability of the group effect when adding Sex as a fixed effect and Site as a covariate, or medication use as a fixed effect (number of self-reported central nervous system-relevant medications being taken). We examined specificity to the N170 latency by repeating this model with N170 amplitude, and P1 latency and amplitude. We tested efficacy as a diagnostic biomarker

using logistic regression with leave-one-out cross-validation (R package caret). We examined concurrent associations between N170 and core (ADOS Social Affect, ADOS Restricted and Repetitive Behavior Scale, Social Responsiveness Scale-2, Vineland Socialisation domain and constituent subscales) and associated (DAWBA externalising and internalising scales) symptoms and IQ, controlling for age. B) Mechanistic Utility: We used a series of partial correlations corrected for age within each diagnostic group to examine the relation between N170 latency and inversion effect magnitude (N170 amplitude to upright – inverted faces). C) Relation to biology: For genetics, Spearman's rank correlation tests were performed to study the relation between the ASD PGS[53] and N170 latency (averaged across P7 and P8). For fMRI, individual contrast images were subjected to a voxel-wise group-level analysis using a general linear model to assess the association of fMRI responses with the EEG derived N170 latency. We additionally assessed effects of diagnosis and age while controlling for effects of sex and site. Effects were evaluated at a statistical threshold of P=0.05, family-wise error corrected (FWE) at the voxel level within a bilateral mask of the fusiform gyrus (2318 voxel) based on the Anatomical Automatic Labeling Atlas (Tzourio-Mazoyer et al., 2002, Figure S7), using small volume correction (SVC). D) Prognostic utility: We used partial correlations (SPSS) and regression models with leave-one-out cross-validation (R package caret) to examine the relation between N170 latency and the domain and subdomain scores of the Vineland socialisation scale at the follow-up visit, controlling for age and score on the same measure at the baseline visit and to explore whether relations varied by sex or were affected by controlling for site. We used Pearson's correlation to examine the relation between any Vineland scores that significantly ($p<0.05$) associated with

the N170 and five domain scores of the CHIP (Quality of Life). Then, we used a data-driven decomposition approach to examine whether meaningful variance in the EEG data related to future social behaviour. To identify clusters, we took individual event-related potentials from four electrodes over which brain responses to faces are seen (O1, O2, P7 and P8) and ran a spatial principal components analysis (PCA;SM2.3) on the downsampled signal (to prevent collinearity - 167Hz) in Matlab. We took the loadings of each individual participant on the top 7 PCA components (see Figure S2) and subjected the scores to a Gaussian-mixture-model based cluster analysis (Regularisation value 0.1, diagonal covariance matrix, 10 replicates) across the whole sample. We then examined whether the N170 latency, change in Vineland Socialisation subscale scores between Baseline and Follow-up and their interrelation varied across clusters using general linear models. E) Context of Use/Potential Utility in a Clinical Context: To examine the potential utility of the N170 in a clinical trial context, we first fit a normative model of N170 latency on age (see SM2.8.3.1 for further details), and used the derived z scores in a leave-one-out cross-validated logistic regression with area under the curve calculation (R package caret) to examine predictive validity for individuals likely to improve or be stable/decline in their Vineland Socialisation subscale scores. We then identified possible cut-points optimised for sensitivity or specificity for subsequent change in Vineland Socialisation subscale scores using bootstrapping (1000 runs) in R package cutpointr. Using an exemplar zN170L cut-off selected to have good sensitivity for detecting and excluding “improvers”, we then used Monte Carlo simulations to determine the effect of restricting clinical trial entry to those predicted to have stable or decreasing Vineland Socialisation subscale scores over time.

5. List of Supplementary Materials

SM2.4. P1 to upright faces - case/control effects

SM2.4.1. Latency:

SM2.4.2. Amplitude:

SM2.5 Core and Associated symptoms

SM2.6. Inversion effects - N170 latency and amplitude

SM2.6.1. Latency

SM2.6.2 Amplitude

SM2.7. Relation to fMRI

SM2.8. Relation to dimensional socialisation

SM2.8.1. Relation to subdomains of the Vineland

SM2.8.2 Cluster analysis

SM2.8.3 Selecting an N170 cut-off

SM2.9 Genetic associations with other phenotypes

SM2.10: Additional fMRI information

SM2.11: Split-half reliability of the N170

Figure S1. Grand average ERPs to face inverted (upper) and face upright (lower) conditions, in three-year age bins.

Figure S2: Top left panel: ERPs at each hemisphere (columns) and in the ASD and NT groups (rows), elicited by subjects with >90% trials attended (blue line) and <90% attended (red line). Top right panel: Relationship between N170 latency and percentage of trials attended, at the left and right hemispheres. Bottom: Illustration of the seven principal components of the individually averaged EEG data concatenated across electrodes from P7 (red), P8 (blue), O1 (yellow), O2 (green) that were entered into the cluster analysis. Coloured lines indicate the effects of different downsampling approaches.

Figure S3: Illustration of the a-priori defined mask of the fusiform face area.

Figure S4: Illustration of fMRI brain maps reflecting the association between face-sensitive BOLD response and N170 latency at different height threshold levels.

Figure S5: Correlation analyses between N170 latency and other polygenic scores.

Figure S6: Normative modelling of the z-N170L.

Figure S7: Top: Statistical Power by Sample Size for Placebo-Controlled N170 Latency Enriched vs. Non-Enriched Clinical Trials with a Simulated Interventional Effect Equivalent to a Cohen's D of 0.45 in the Non-Enriched Population and a Simulated 12-Week Trial Duration. Bottom: Receiver Operating Characteristic Curve showing the achieved sensitivity and specificity for detecting non-improvers using different zN170L cut-offs.

Figure S8a: Effect on prognosis at different N170 latency cut-offs.

Figure S8b: Example of bootstrapped cutoffs to maximize sensitivity given a reasonable level of specificity (constructed with R package cutpointr).

Figure S8c: Example of bootstrapped cutoffs to maximize specificity given a reasonable level of sensitivity (constructed with R package cutpointr).

Table S1A: Recruitment profile of the sample with EEG data.

Table S1B: Clinical and diagnostic profile of individuals with EEG data within the LEAP sample.

Table S2: Reasons for EEG data loss separated by group.

Table S3: Clinical and diagnostic profile of individuals who did and did not provide EEG data within the LEAP sample

Table S4: Delays in milliseconds observed in stimulus presentation and corrected in analysis.

Table S5. Information of the PGS best model fit of each trait.

Table S6: Partial correlations for association between N170 latency at P7/P8 to upright faces and associated symptoms, controlled for age.

Table S7: Relation between the N170 and prognostic change in the Vineland Socialisation subdomains

Table S8: Clinical profile of the three clusters within the ASD group.

Table S9. Summary of internal reliability of the N170 ERP component by hemisphere.

6. References

1. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. (American Psychiatric Association, 2013).

2. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).

3. Modabbernia, A., Velthorst, E. & Reichenberg, A. Environmental risk factors for autism: an evidence-based review of systematic reviews and meta-analyses. *Mol. Autism* **8**, 13 (2017).

4. Charman, T. *et al.* The EU-AIMS Longitudinal European Autism Project (LEAP): clinical characterisation. *Mol. Autism* **8**, 27 (2017).

5. Warrier, V. *et al.* Social and non-social autism symptoms and trait domains are genetically dissociable. *Commun. Biol.* **2**, 328 (2019).

6. Mandy, W. P. L., Charman, T. & Skuse, D. H. Testing the Construct Validity of Proposed Criteria for DSM-5 Autism Spectrum Disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 41–50 (2012).

7. Tillmann, J. *et al.* Investigating the factors underlying adaptive functioning in autism in the EU-AIMS Longitudinal European Autism Project. *Autism Res.* **12**, 645–657 (2019).

8. Cacioppo, S., Grippo, A. J., London, S., Goossens, L. & Cacioppo, J. T. Loneliness: Clinical Import and Interventions. *Perspect. Psychol. Sci.* **10**, 238–249 (2015).

9. Katz, S. J., Conway, C. C., Hammen, C. L., Brennan, P. A. & Najman, J. M. Childhood Social Withdrawal, Interpersonal Impairment, and Young Adult Depression: A Mediational Model. *J. Abnorm. Child Psychol.* **39**, 1227 (2011).

10. White, S. W. & Roberson-Nay, R. Anxiety, Social Deficits, and Loneliness in Youth with Autism Spectrum Disorders. *J. Autism Dev. Disord.* **39**, 1006–1013 (2009).

11. Lord, C., McCauley, J. B., Pepa, L. A., Huerta, M. & Pickles, A. Work, living, and the pursuit of happiness: Vocational and psychosocial outcomes for young adults with autism. *Autism Int. J. Res. Pract.* **24**, 1691–1703 (2020).

12. Loth, E. *et al.* Identification and validation of biomarkers for autism spectrum disorders. *Nat. Rev. Drug Discov.* **15**, 70–73 (2016).

13. Abi-Dargham, A. & Horga, G. The search for imaging biomarkers in psychiatric disorders. *Nat. Med.* **22**, 1248–1255 (2016).

14. Dawson, G. *et al.* Neurocognitive and electrophysiological evidence of altered face processing in parents of children with autism: Implications for a model of abnormal development of social brain circuitry in autism. *Dev. Psychopathol.* **null**, 679–697 (2005).

15. Jack, R. E. & Schyns, P. G. The Human Face as a Dynamic Tool for Social Communication. *Curr. Biol.* **25**, R621–R634 (2015).

16. Bentin, S., Allison, T., Puce, A., Perez, E. & McCarthy, G. Electrophysiological Studies of Face Perception in Humans. *J. Cogn. Neurosci.* **8**, 551–565 (1996).

17. Eimer, M. The Face-Sensitivity of the N170 Component. *Front. Hum. Neurosci.* **5**, (2011).

18. Gao, C., Conte, S., Richards, J. E., Xie, W. & Hanayik, T. The neural sources of N170: Understanding timing of activation in face-selective areas. *Psychophysiology* **56**, e13336 (2019).

19. Le Grand, R., Mondloch, C. J., Maurer, D. & Brent, H. P. Neuroperception: Early visual experience and face processing. *Nature* **410**, 890–890 (2001).

20. Mondloch, C. J. *et al.* The effect of early visual deprivation on the development of face detection. *Dev. Sci.* **16**, 728–742 (2013).

21. Balas, B. & Saville, A. N170 face specificity and face memory depend on hometown size. *Neuropsychologia* **69**, 211–217 (2015).

22. Rossion, B. *et al.* Spatio-temporal localization of the face inversion effect: an event-related potentials study. *Biol. Psychol.* **50**, 173–189 (1999).

23. Cassidy, S. M., Robertson, I. H. & O'Connell, R. G. Retest reliability of event-related potentials: Evidence from a variety of paradigms. *Psychophysiology* **49**, 659–664 (2012).

24. Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R. A. & van IJzendoorn, M. H. Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiol. Behav.* **130**, 13–22 (2014).

25. Key, A. P. & Corbett, B. A. The unfulfilled promise of the N170 as a social biomarker. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* (2019) doi:10.1016/j.bpsc.2019.08.011.

26. Kang, E. *et al.* Atypicality of the N170 Event-Related Potential in Autism Spectrum Disorder: A Meta-analysis. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 657–666 (2018).

27. Neuhaus, E., Kresse, A., Faja, S., Bernier, R. A. & Webb, S. J. Face processing among twins with and without autism: social correlates and twin concordance. *Soc. Cogn. Affect. Neurosci.* **11**, 44–54 (2016).

28. Webb, S. J. *et al.* Developmental change in the ERP responses to familiar faces in toddlers with autism spectrum disorders versus typical development. *Child Dev.* **82**, 1868–1886 (2011).

29. Neuhaus, E. *et al.* The Relationship Between Early Neural Responses to Emotional Faces at Age 3 and Later Autism and Anxiety Symptoms in Adolescents with Autism. *J. Autism Dev. Disord.* **46**, 2450–2463 (2016).

30. Dawson, G. *et al.* Early behavioral intervention is associated with normalized brain activity in young children with autism. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 1150–1159 (2012).

31. Dawson, G., Webb, S. J. & McPartland, J. Understanding the Nature of Face Processing Impairment in Autism: Insights From Behavioral and Electrophysiological Studies. *Dev. Neuropsychol.* **27**, 403–424 (2005).

32. Garman, H. D. *et al.* Wanting it Too Much: An Inverse Relation Between Social Motivation and Facial Emotion Recognition in Autism Spectrum Disorder. *Child Psychiatry Hum. Dev.* **47**, 890–902 (2016).

33. Hileman, C. M., Henderson, H. A., Mundy, P., Newell, L. C. & Jaime, M. Developmental and Individual Differences on the P1 and N170 ERP Components in Children with and without Autism. *Dev. Neuropsychol.* **36**, 214–236 (2011).

34. Webb, S. J. *et al.* ERP responses differentiate inverted but not upright face processing in adults with ASD. *Soc. Cogn. Affect. Neurosci.* **7**, 578–587 (2012).

35. Loth, E. *et al.* The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Mol. Autism* **8**, 24 (2017).

36. Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P. & Gore, J. C. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat. Neurosci.* **2**, 568–573 (1999).

37. Coffman, M. C. *et al.* Examination of Correlates to Health-Related Quality of Life in Individuals with Fragile X Syndrome. *Brain Sci.* **10**, (2020).

38. Busner, J. & Targum, S. D. The Clinical Global Impressions Scale. *Psychiatry Edgmont* **4**, 28–37 (2007).

39. Riley, A. W., Chan, K. S., Prasad, S. & Poole, L. A global measure of child health-related quality of life: reliability and validity of the Child Health and Illness Profile - Child Edition (CHIP-CE) global score. *J. Med. Econ.* **10**, 91–106 (2007).

40. Marquand, A. F., Rezek, I., Buitelaar, J. & Beckmann, C. F. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol. Psychiatry* **80**, 552–561 (2016).

41. Tzourio-Mazoyer, N. *et al.* Neural correlates of woman face processing by 2-month-old infants. *NeuroImage* **15**, 454–461 (2002).

42. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, (2019).

43. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

44. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).

45. Warrier, V. *et al.* Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Mol. Psychiatry* **23**, 1402–1409 (2018).

46. Peñuelas-Calvo, I., Sareen, A., Sevilla-Llewellyn-Jones, J. & Fernández-Berrocal, P. The "Reading the Mind in the Eyes" Test in Autism-Spectrum Disorders Comparison with Healthy Controls: A Systematic Review and Meta-analysis. *J. Autism Dev. Disord.* **49**, 1048–1061 (2019).

47. Kanwisher, N., McDermott, J. & Chun, M. M. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *J. Neurosci.* **17**, 4302–4311 (1997).

48. Feuerriegel, D., Churches, O., Hofmann, J. & Keage, H. A. D. The N170 and face perception in psychiatric and neurological disorders: A systematic review. *Clin. Neurophysiol.* **126**, 1141–1158 (2015).

49. Brainstorm Consortium *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, (2018).

50. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).

51. Cuthbert, B. N. Research Domain Criteria: toward future psychiatric nosologies. *Dialogues Clin. Neurosci.* **17**, 89–97 (2015).

52. Nowparast Rostami, H., Hildebrandt, A. & Sommer, W. Sex-specific relationships between face memory and the N170 component in event-related potentials. *Soc. Cogn. Affect. Neurosci.* **15**, 587–597 (2020).

53. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

54. Shannon, R. W., Patrick, C. J., Venables, N. C. & He, S. 'Faceness' and Affectivity: Evidence for Genetic Contributions to Distinct Components of Electrocortical Response to Human Faces. *NeuroImage* **83**, 609–615 (2013).

55. Gui, A. *et al.* Association of Polygenic Liability for Autism With Face-Sensitive Cortical Responses From Infancy. *JAMA Pediatr.* (2021) doi:10.1001/jamapediatrics.2021.1338.

56. Lundström, S. *et al.* Autism spectrum disorders and autistic like traits: similar etiology in the extreme end and the normal variation. *Arch. Gen. Psychiatry* **69**, 46–52 (2012).

57. Constantino, J., Charman, T. & Jones, E. J. H. Clinical and translational implications of new understanding of a developmental sub structure for autism. *Annu. Rev. Clin. Psychol.* (2020).

58. Parker, T. C. *et al.* The N170 event-related potential reflects delayed neural response to faces when visual attention is directed to the eyes in youths with ASD. *Autism Res.* **14**, 1347–1356 (2021).

59. Riglin, L. *et al.* Using Genetics to Examine a General Liability to Childhood Psychopathology. *Behav. Genet.* **50**, 213–220 (2020).

60. Brikell, I. *et al.* The contribution of common genetic risk variants for ADHD to a general factor of childhood psychopathology. *Mol. Psychiatry* **25**, 1809–1821 (2020).

61. Kanwisher, N., Tong, F. & Nakayama, K. The effect of face inversion on the human fusiform face area. *Cognition* **68**, B1–B11 (1998).

62. Elsabbagh, M. *et al.* Infant Neural Sensitivity to Dynamic Eye Gaze Is Associated with Later Emerging Autism. *Curr. Biol.* **22**, 338–342 (2012).

63. Jones, E.J.H. *et al.* Reduced engagement with social stimuli in 6-month-old infants with later autism spectrum disorder: a longitudinal prospective study of infants at high familial risk. *J. Neurodev. Disord.* **8**, 7 (2016).

64. Sadeh, B., Podlipsky, I., Zhdanov, A. & Yovel, G. Event-related potential and functional MRI measures of face-selectivity are highly correlated: A simultaneous ERP-fMRI investigation. *Hum. Brain Mapp.* **31**, 1490–1501 (2010).

65. Freire, A., Lee, K. & Symons, L. A. The Face-Inversion Effect as a Deficit in the Encoding of Configural Information: Direct Evidence. *Perception* **29**, 159–170 (2000).

66. Maurer, D., Grand, R. L. & Mondloch, C. J. The many faces of configural processing. *Trends Cogn. Sci.* **6**, 255–260 (2002).

67. Mondloch, C. J., Le Grand, R. & Maurer, D. Configural Face Processing Develops more Slowly than Featural Face Processing. *Perception* **31**, 553–566 (2002).

68. Rossion, B., Curran, T. & Gauthier, I. A defense of the subordinate-level expertise account for the N170 component. *Cognition* **85**, 189–196 (2002).

69. Chatham, C. H. *et al.* Adaptive behavior in autism: Minimal clinically important differences on the Vineland-II. *Autism Res. Off. J. Int. Soc. Autism Res.* **11**, 270–283 (2018).

70. Webb, S. J., Neuhaus, E. & Faja, S. Face Perception and Learning in Autism Spectrum Disorders. *Q. J. Exp. Psychol. 2006* 1–44 (2016) doi:10.1080/17470218.2016.1151059.

71. Vickers, A. J. & Altman, D. G. Analysing controlled trials with baseline and follow up measurements. *BMJ* **323**, 1123–1124 (2001).

72. Kang, E. *et al.* Reply to: Can the N170 Be Used as an Electrophysiological Biomarker Indexing Face Processing Difficulties in Autism Spectrum Disorder? *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 324–325 (2019).

73. McPartland, J. C. Considerations in biomarker development for neurodevelopmental disorders. *Curr. Opin. Neurol.* **29**, 118–122 (2016).

74. Vettori, S., Jacques, C., Boets, B. & Rossion, B. Can the N170 Be Used as an Electrophysiological Biomarker Indexing Face Processing Difficulties in Autism Spectrum Disorder? *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 321–323 (2019).

75. Simon, R. Stratification and Partial Ascertainment of Biomarker Value in Biomarker Driven Clinical Trials. *J. Biopharm. Stat.* **24**, 1011–1021 (2014).

76. Siafis, S. *et al.* Placebo response in pharmacological and dietary supplement trials of autism spectrum disorder (ASD): systematic review and meta-regression analysis. *Mol. Autism* **11**, 66 (2020).

77. Loth, E. *et al.* Identification and validation of biomarkers for autism spectrum disorders. *Nat. Rev. Drug Discov.* **15**, 70–70 (2016).

78. Lord, C. *et al.* The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30**, 205–223 (2000).

79. Goodman, R., Ford, T., Richards, H., Gatward, R. & Meltzer, H. The Development and Well-Being Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *J. Child Psychol. Psychiatry* **41**, 645–655 (2000).

80. World Health Organization. *The ICD-10 Classification of Mental and Behavioral Disorders: Diagnostic Criteria for Research.* (1993).

81. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders - Text Revision*. (2000).

82. Sparrow, S., Cicchetti, D. & Balla, D. *Vineland adaptive behavior scales: (Vineland II), survey interview form/caregiver rating form*. (Pearson Assessments, 2005).

83. Oakley, B. F. *et al.* How do core autism traits and associated symptoms relate to quality of life? Findings from the Longitudinal European Autism Project: *Autism* (2020) doi:10.1177/1362361320959959.

84. Webb, S. J. *et al.* Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. *J. Autism Dev. Disord.* **45**, 425–443 (2015).

85. Jones, E. J. H. *et al.* Eurosibs: Towards robust measurement of infant neurocognitive predictors of autism across Europe. *Infant Behav. Dev.* **57**, 101316 (2019).

86. Tye, C. *et al.* Neurophysiological responses to faces and gaze direction differentiate children with ASD, ADHD and ASD+ADHD. *Dev. Cogn. Neurosci.* **5**, 71–85 (2013).

87. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).

88. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869 (2011).

89. Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F. & Weinberger, D. R. The amygdala response to emotional stimuli: a comparison of faces and scenes. *NeuroImage* **17**, 317–323 (2002).

90. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*. (1994).

91. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. (American Psychiatric Publishing, 2013).

92. Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **24**, 659–685 (1994).

93. Charman, T. & Gotham, K. Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders – lessons from research and practise. *Child Adolesc. Ment. Health* **18**, 52–63 (2013).

94. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

95. Baumeister, S. *et al.* Attenuated Anticipation of Social and Monetary Rewards in Autism Spectrum Disorders. *bioRxiv* 2020.07.06.186650 (2020) doi:10.1101/2020.07.06.186650.

96. Moessnang, C. *et al.* Social brain activation during mentalizing in a large autism cohort: the Longitudinal European Autism Project. *Mol. Autism* **11**, 17 (2020).

Acknowledgements

Many thanks to the participants of the LEAP study, and all members of the AIMS2 and EUAIMS consortia. **Funding:** This project has received funding from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115300, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007 - 2013) and EFPIA companies' in kind contribution; the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777394. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and AUTISM SPEAKS, Autistica, SFARI; awards from the Medical Research Council (MR/K021389/1; MR/T003057/1). **Author contributions:** LM designed, wrote paper, processed data, analysed data, EJ designed, analysed data, wrote paper; DM secured funding, supervised study overall, analysis direction and interpretation; EL secured funding, supervised study overall; MHJ designed; JT clinical phenotyping; TC clinical/design of study/interpretation; TB/CSL/FC Genetics data pre-processing/calculation of polygenic scores/ interpretation; CM fMRI processing, analysis, interpretation; CC Monte Carlo, analysis direction and interpretation; LH analysis direction and interpretation; JA,CB GD normative modelling, interpretation; TB, SBC, SB, JB, EL, DS, BO secured funding, supervised study overall; DF, CE supervised overall analytic strategy; all authors revised the manuscript for intellectual content. **Competing interests:** SB discloses that he has in the last 3 years acted as an author, consultant or lecturer for Medice and Roche. He receives royalties for text books and diagnostic tools from Hogrefe, Kohlhammer and UTB. JT is a paid

consultant to Hoffman La-Roche. TC has served as a paid consultant to F.Hoffman-La Roche and Servier and has received royalties from Sage Publications and Guilford Publications. TB served in an advisory or consultancy role for ADHS digital, Infectopharm, Lundbeck, Medice, Neurim Pharmaceuticals, Oberberg GmbH, Roche, and Takeda. He received conference support or speaker's fee by Medice and Takeda. He received royalities from Hogrefe, Kohlhammer, CIP Medien, Oxford University Press; the present work is unrelated to these relationships. AML has received consultant fees in the past two years from Boehringer Ingelheim, Elsevier, Lundbeck Int. Neuroscience Foundation, Lundbeck AS, The Wolfson Foundation, Thieme Verlag, Sage Therapeutics, von Behring Stiftung, Fondation FondaMental, Janssen-Cilag GmbH, MedinCell, Brain Mind Institute, CISSN. Furthermore he has received speaker fees from Italian Society of biological Psychiatry, Merz-Stiftung, Forum Werkstatt Karlsruhe, Lundbeck SAS France, BAG Psychiatrie Oberbayern. JB has been in the past 3 years a consultant to / member of advisory board of / and/or speaker for Takeda/Shire, Roche, Medice, Angelini, Janssen, and Servier. He is not an employee of any of these companies, and not a stock shareholder of any of these companies. He has no other financial or material support, including expert testimony, patents, royalties. DM has been paid for advisory board work by F. Hoffmann-La Roche Ltd. and Servier; and for editorial work by Springer. Other authors report no conflicts of interest. **Data and materials availability:** LEAP data is shared through an application process co-designed with autistic people that preserves security and privacy– contact the corresponding author for further details and application forms. Scripts used to implement the experimental task are covered by an MTA, which can be obtained through the corresponding author.

Figure Legends

Figure 1: Differences in N170 latency between the ASD (n=246) and control (n=190) groups across the whole cohort (general linear model; $F(1,430)=9.43$, $p=0.002$, $\eta_p^2= 0.021$). A) Grand-average event-related potential waveform with solid lines indicating the mean waveform and +/- 2 SE shaded; b) probability density function for differences in amplitude (top) and latency (bottom); c) topo-map of activation from 150ms-250ms post-face onset, electrodes P7 and P8 marked.

Figure 2: Left panel: dimensional association between face-sensitive functional responses in the fusiform gyrus and N170 latencies (ASD n=99, control n=100; general linear model; $t=3.93$, $P_{SVC}=.032$; $R^2=.131$, N170 $\beta=0.285$, $p < 0.001$); right panel: t values plotted on a brain slice.

Figure 3: Correlation between ASD polygenic scores and N170 latency responses (full sample: Spearman's $r2=0.026$; $p=0.0031$; participants with ASD n=198, $r2=0.022$; $p=0.039$; controls n=133, $r2=0.024$; $p=0.074$). The samples are 198 ASD and 133 controls from European ancestry.

Figure 4: Cluster analysis of EEG data within the ASD group; a) change in play and leisure time scores within each cluster (general linear models; Cluster 1 n=73, Cluster 2 n=19, Cluster 3 n=53; $F(2,144) =4.41$, $p =0.014$, $\eta_p^2=0.06$) ; b) N170 latency per cluster (Cluster 1 n=118, Cluster 2 n=27, Cluster 3 n=101; $F(2,245) =64.32$, $p < 0.001$, $\eta_p^2= 0.975$); c) waveforms per cluster for P7 (top) and P8 (bottom) with solid lines indicating the mean amplitude and the

shaded area depicting +/- 2 SE; d) association between N170L and change in play and leisure time scores within Cluster 2 (n=19; Pearson's $r(19)=-0.517$, $p=0.023$).

Supplementary materials

SM1. Methods

Full details of the LEAP study design [35] and clinical characteristics of the cohort [4] have been published and are briefly summarised below. The LEAP protocols are available at https://www.eu-aims.eu/fileadmin/websites/eu-aims/media/EU-AIMS_LEAP/EU-AIMS-LEAP_SOP_StudyProtocol.zip

SM1.1. Participants

Participants included in this study were recruited between January 2014 and March 2017 across five European specialist ASD centres: Institute of Psychiatry, Psychology and Neuroscience, King's College London (IoPPN/KCL, UK; *n*=152), University Medical Centre Utrecht (UMCU, Netherlands; *n=69*), Radboud University Nijmegen Medical Centre (RUNMC, Netherlands; *n*=150), Central Institute of Mental Health (CIMH, Germany; *n*=33) and the University Campus Bio-Medico (UCBM, Italy; *n*=23). Participants were recruited from existing volunteer databases, prior research cohorts, clinical referrals from local outpatient centres, special needs schools, mainstream schools and local communities.

The total sample with valid EEG comprised 436 participants, split in 246 children, adolescents, and adults with Autism Spectrum Disorder (ASD) and 190 control participants. Of the 246 participants with ASD, 205 participants had a full-scale IQ in the typical range (≥75), while 41 participants had mild intellectual disabilities (mild ID; defined by IQ between 50 and 74). The 190 control participants were split in 166 typically developing (TD) subjects and 24 individuals with mild ID and without ASD. Table S1A provides a further breakdown of sample sizes by group, IQ and age status and Table S1B provides clinical and diagnostic

data for participants included in the sample; Table S2 shows the reasons for EEG data loss and Table S3 shows clinical and diagnostic information for included and excluded participants.

Within each age band (children, adolescents, adults), participants were recruited with a similar male:female ratio (3:1) and IQ composition so that predicted cognitive/biological differences can be compared across sex and developmental stages.

Ethnicities represented in this sample include 77% Caucasian (*n*=327), 5% mixed race (*n*=21), 2% Asian (*n*=8), <1% Black (*n*=2), and 2% other (*n*=10). For 13% of participants (*n*=55), ethnicity was either not indicated (<1%) or missing (12.5%).

Annual household income was measured on an 8-point-scale ranging from <£25,000 to >£150,000, with the median annual household income being estimated at £30,000–£39,999. Highest household parental education was coded on a 5-point scale ranging from primary education to postgraduate qualifications; 67% of households had at least one parent with education beyond a high school diploma (i.e. with an undergraduate degree from university).

SM1.2 Inclusion/exclusion criteria

Participant inclusion criteria for the ASD sample were an existing clinical diagnosis of ASD according to DSM-IV [89], DSM-IV-TR [80], DSM-5 [90] or ICD-10 [79] criteria and age between 6 and 30 years. ASD diagnoses were based on a comprehensive assessment of the participant's clinical history and/or current symptom profile, depending on when the participant was originally identified at that site, including the diagnostic instruments, the Autism Diagnostic Observation Schedule (ADOS; [77] and the Autism Diagnostic Interview-Revised (ADI-R; [91]. Given the better accuracy of clinical judgements [92], individuals with ASD

were not excluded if they did not reach the cut-off scores on the Autism Diagnostic Observation Schedule (ADOS) or the Autism Diagnostic Interview-Revised (ADI-R), reflecting the moderate-to-good but imperfect accuracy of individual diagnostic tools. Exclusion criteria included significant hearing or visual impairments not corrected by glasses or hearing aids, a history of alcohol and/or substance abuse or dependence in the past year and the presence of any MRI contraindications (e.g. metal implants, braces, claustrophobia) or failure to give informed written consent to MRI scanning. The presence of co-occurring psychiatric conditions was not an exclusion criterion, given their prevalence in this population. Exclusion criteria of the control group were the same, but additionally participants were excluded if they had a parent- or (where appropriate) self-report of a psychiatric disorder or scored had a *T*-score of 70 or higher on the self-report or parent-report form of the Social Responsiveness Scale-2. If on medication, all participants had to be stable (min. 8 weeks) at entrance point and over the course of the baseline visit to be included. Information on concurrent medication use was collected at the institute visit and substances were mapped to the Anatomical Therapeutic Chemical (ATC) classification system to categorise drugs as affecting/non-affecting the nervous system (ATC Level-1 code "N"; SM2.1).

SM1.3. Quality control procedures

Appropriate to a multi-centre study, quality control procedures were in place around training, and data collection/entry. Cross-site training sessions for collecting clinical data were put in place, the ADOS and ADI-R were administered and scored by qualified/certified personnel and the study was regularly monitored according to Good Clinical Practice (GCP) standards.

Of the total number of ADI-R assessments (4–5 ever/diagnostic) administered to participants in the current sample ($N = 246$), $N = 94$ were re-used from previous studies, while for the ADOS ($N = 244$), a total of $N = 30$ were re-used (all completed within the previous 12 months). For the key clinical measures completed at the institute visit (i.e. ADI-R, ADOS, IQ test), 10% of test manuals were randomly chosen to be double-entered. If a significant level of incorrect/inconsistent data was identified, all test data from all participants was checked against the original paper forms. Other procedures also included impossible values/range checks of all items, sub-scales and total scores for interview and questionnaire measures, duplicated entry detection and correction, as well as data audits and checks of scoring algorithms. When missing data was present, site coordinators were asked to secure the information if possible.

SM1.3.1. EEG procedure

Five sites acquired EEG data in LEAP: Kings College, London (KCL), The Central Institute of Mental Health, Mannheim (CIMH), University Medical Centre, Utrecht (UMCU), Radboud University Nijmegen Medical Centre (RUNMC) and University Campus Biomedico, Rome (UCBM). Three different EEG systems were used to acquire the data, Brainproducts Acticaps (KCL, CIMH, RUNMC), Biosemi Active-Two (UMCU) and Micromed (UCBM). Testing teams from each site attended initial training in London in 2013, followed by site visits to ensure correct set-up of equipment and that SOPs were followed. All sites then attended weekly telephone conferences to discuss data acquisition and quality, and to report any problems.

Data were uploaded from each site to a central repository in their raw, manufacturer-specific, proprietary formats. Preprocessing and harmonisation of this data was performed at

Birkbeck, University of London. Each dataset was first loaded into EEGLab (Delorme & Makeig, 2004). Briefly, the following steps were followed: 1) harmonisation of electrode labels to 62-channel common montage; 2) generation of horizontal electrooculogram (HEOG) channels from electrodes AF7/8 (KCL, RUNMC & UCBM only, CIMH & UMCU used external electrodes to record HEOG); 3) generation of variance-based data quality metrics and extraction of impedance values from Brainvision sites; 4) re-reference to FCz; 5) Resample to 1Khz; 6) harmonise event labels.

This process resulted in harmonised data in a common EEGLab format, upon which all subsequent task-specific analyses were performed.

Stimuli were presented using custom-written Matlab software (KCL, CIMH, UMCU, UCBM) and Presentation (UMCU).

SM1.3.2 Face ERP task processing

Of note, all analysis stages were performed blind to age, site and diagnostic status.

SM1.3.2.1 Timing correction

Visual stimulus timing was measured at all sites except for UCBM using a photodiode. The delta between stimuli being drawn on the screen and the event marker being sent was recorded over 600 trials. The average delta (summarised in Table S4) was then computed and subtracted from the event marker latencies on a per-site basis.

SM1.3.2.2 Preprocessing

All task processing was carried out in the Matlab Fieldtrip toolbox [87]. Raw EEG data were segmented into individual trials, from -200ms to 800ms post stimulus-onset. A bandpass filter of 0.1Hz-40Hz and an FFT-based DFT notch filter at 50Hz was applied, with 2000ms of padding to avoid filter edge-artefacts. After filtering, data were resampled to 500Hz.

SM1.3.2.3. Cleaning

Data were cleaned in Matlab using a custom-written automatic algorithm. Two classes of artefact were detected, 1) whole-scalp artefacts; and 2) EOG artefacts. Interpolation is only attempted on scalp artefacts, since EOG artefacts are largely blinks and eye movements, the presence of which suggests the participant was not watching the stimuli.

1. **Scalp artefacts** were detected where voltages exceeded minimum/maximum criteria of +/-100μV or a range of 150μV. Flat channels are detected as those that do not exceed a criterion of +/-0.0001μV
2. **EOG artifacts** were detected on frontal electrodes FP1/z/2, AF7/8. Blink detection involved: 1) bandpass filtering the data from 3Hz-10Hz; 2) calculating z-scores of the voltage value at each sample of each trial (relative to the distribution of samples across all trials); 3) thresholding the data with a z-score of +/-6SDs; 4) detecting contiguous runs of samples greater than 50ms in duration - these were determined to be blinks and were marked. Eye movements were detected by fitting a linear function to the data from each electrode on each trial. Trials with a linear R2 greater than .6 were determined to be eye movements and were marked.

The sequence of cleaning the data from one participant is as follows:

1. Detect scalp artefacts.
2. Find channels with greater than 80% bad trials and interpolate entire channel (on the basis that bad channels are likely caused by poor electrode contact, and are not representative of broader data quality at other channels).
3. Re-detect scalp artefacts (since presence of artefacts may change due to interpolation).
4. Interpolate on a trial x electrode basis. For each trial, any electrodes marked as having artefacts were interpolated where sufficient neighbouring electrodes were free of artefacts.
5. Re-detect scalp artefacts.
6. Identify channels with greater than 40% bad trials after interpolation. These channels are assumed to be unrepresentative of broader data quality and so are excluded from a) further artefact detection; and b) average reference. If P7/8 or O1/2 are in the excluded list, the participant is dropped (see final criteria, below).
7. Detect EOG artefacts, and re-detect scalp artefacts on non-excluded channels. Trials with a) EOG artefacts; or b) scalp artefacts remaining after all attempts at interpolation, are dropped from further analysis.

After cleaning, each dataset was inspected against four criteria. Failing to meet any of these criteria results in the participant being dropped from further analysis.

1. A minimum of 20 clean trials-per-condition (face upright/inverted).
2. A maximum of 10 channels interpolated in stage (2) above.
3. A maximum of 10 channels excluded in stage (6) above.
4. Any of P7/8 or O1/2 marked as excluded.

SM1.3.2.4. Averaging, peak detection and amplitude extraction

At this stage, clean segments were split by condition (face upright/inverted) and formed into individual averages for each participant. We extracted peak metrics (latency and amplitude) for the N170 component, and for the P1 as a test of component specificity.

P1 and N170 ERP latencies become faster with age. In order to set age-specific search windows for peak detection we formed grand average ERPs in three-year age bins (Figure S1), and manually recorded the mean peak latency at in each bin. An algorithm then searched each individual participant average for the maximum voltage value within each age-specific time window. This was done independently for each component (P1/N170), for each condition (face upright/inverted) and at each electrode (P1: O1/2, N170: P7/8), resulting in eight separate peaks for each participant.

After automatic peak detection, we performed a manual process of peak checking for each participant. The first rater (CE) visually inspected each peak and, where the algorithm had chosen incorrectly, manually corrected the search window. The algorithm then searched again within this window to find the correct peak. Regardless of whether any manual corrections were required, each peak was rated as: 1) OK, 2) presenting a double peak, 3) peak not clear, or 4) other (needs checking).

Subsequently, two expert raters (LM & EJ) reviewed CE's ratings. Where LM and EJ disagreed, a consensus was formed on the basis of (in descending priority order), 1) morphology (a clear P1 followed by an N170), and 2) expected latency. Where a consensus could not be reached, the peak was marked as "not clear".

SM1.4. fMRI processing

Acquisition parameters were harmonized across sites as closely as possible. Functional images were collected using an echo-planar imaging (EPI) sequence and structural images were acquired with a high-resolution T1-weighted magnetization-prepared rapid gradient echo sequence. fMRI data analysis followed standard processing routines in SPM12 (http://www.fil.ion.ucl.ac.uk/spm/), including slice-time correction, a two-step realignment procedure, unified segmentation and normalization to standard stereotactic space as defined by the Montreal Neurological Institute (MNI), and smoothing with an 8mm full-width-at-half-maximum Gaussian Kernel. Data were subjected to an extensive quality assessment pipeline, and datasets with excessive head motion (>20% of trials with a framewise displacement (Jenkinson et al., 2001) greater 0.5 mm) were excluded (ASD: n=28, TD: n=19). For each subject, task conditions were modeled as boxcar functions that accounted for the presentation of face blocks and shape blocks, respectively. Task regressors were convolved with the canonical hemodynamic response function (HRF) and subjected as predictors to a general linear model (GLM), along with six realignment parameters to account for head motion. During first-level model estimation, data was high-pass filtered with a cut-off of 256 s, and an autoregressive model of the first order was applied. The faces condition was subsequently contrasted to the shapes condition to identify brain responses reflecting sensitivity to emotional faces.

SM1.5 Genetics

Sample quality controls such as Sex check (based on the X chromosome homozygosity rate or the median of the Log R ratio of the X and Y chromosomes), Mendel errors (transmission errors within full trios) and Identity By State (IBS) were performed using PLINK 1.90.

Imputation of 17 million SNPs was performed using the 700k genotyped SNPs on the Michigan Imputation Server [93]. We used the HRC r1.1 2016 reference panel for a European population since a majority (>95%) of individuals in this study are from European ancestry. Principal Components Analysis (PCA) of variance standardized relationship matrix was used to evaluate the ancestry of individuals and to provide components for any covariate adjustments. Table S5 shows the best PGS model fit for each trait.

SM2. Supplementary Results

SM2.1 Effects of medication

Information on concurrent medication use was collected at the institute visit (either at baseline or if missing, retrospectively at the follow-up visit) and substances were mapped to the Anatomical Therapeutic Chemical (ATC) classification system to categorise drugs as affecting/non-affecting the nervous system (ATC Level-1 code “N”). Then, those substances categorised as affecting the nervous system were again classified in pharmacological subgroups/categories the particular medication relates to (i.e. antidepressants, antiepileptics, antimigraine preparations, antipsychotics, anxiolytics, hypnotics and sedatives, other analgesics and antipyretics, and psychostimulants and other drugs used to treat ADHD).

For the current analyses, these categories were further collapsed into those who reported taking neurodevelopmentally relevant medication (e.g. methylphenidate, anxiolytics, ASD N=90; control N=17)), those who reported taking no neurodevelopmentally relevant medication (but might be taking e.g. paracetamol; ASD n=69, control = 70); and those who

did not report on their medication use (ASD N=86 ; control N=101). Including the Medication factor in the model as a main effect and in interaction with Group, Age and Age by Group revealed a main effect of Medication ($F(2,415)=9.00$, $p < 0.001$). Posthoc Tukey tests indicated this was because mean N170 latencies were faster in the group who did not provide information about medication use relative to those who reported either using or not using medication ($ps<0.001$); those who reported using or not using medication did not differ ($p=0.68$). The effect of Medication did not interact with either Age ($F(2,415) = 1.16$, $p = 0.36$), Group ($F(2,415) = .22$, $p = 0.81$) or Age and Group ($F(4,415)=0.28$, $p = 0.89$). The main effect of Group on N170 latency remained, indicating that this is not confounded by medication status ($F(1,415) = 5.35$, $p = 0.02$).

SM2.2 Effects of visual attention

We recorded video of participants watching the task and measured visual attention to the stimuli at the two sites with the largest samples (KCL and Nijmegen).

SM2.2.1. Semi-automatic eye tracker coding

For 263 participants we recorded concurrent eye tracking data during the EEG task, and synchronised the eye tracking data to the onset and offset of each trial. For each trial we extracted the percentage of gaze to the screen. To avoid false-negatives where the eye tracker fails to track the eyes and so reports no gaze when a participant was in fact attending, we manually coded all trials in which the eye tracker reported <25% of gaze was to the screen.

SM2.2.2. Manual video coding

107 participants did not have concurrent eye tracking data and in these cases we performed manual attention coding using DataVyu (2014) software. Each coder first coded fifteen videos from the KCL site, from which we calculated a mean inter-class correlation of 0.809, indicating that coders could reliably detect and mark periods of attention and inattention. For each participant, we then recorded the video timestamps of the onset and offset of each period of attention throughout the EEG session. We then synchronised the video to each trial of the EEG data and calculated percentage attended for each trial.

SM2.2.3. Results

The ASD and NT groups did not differ on the percentage of trials attended, ASD M=93.1% (SD=17.9%), NT M=92.9%, (SD=21.1%), Mann-Whitney U=-1.10, N_{ASD}=227, N_{NT}=143, p=0.27. We also calculated a dichotomous variable (Valid Attention) for each participant, coding whether they attended for greater or less than 90% of trials. We then performed a linear effect model on N170 Latency with independent factors Diagnosis (ASD/NT), Age Group (Children, Adolescents, Adults), Valid Attention (Valid/Invalid), and repeated factor Hemisphere (P7/P8, over which the N170 was measured). Neither the main effect of Valid Attention, nor any second- or third-order interactions were significant, all F's<.719, all p's>.482 (see Figure S2 for ERPs of participants who did and did not attend to >90% of trials). We also calculated bivariate correlations between percentage of attended trials and N170 latency at each hemisphere and found not relationship, Left Hemisphere, ASD r=.02, p=.891; NT r=.03, p=.775; Right Hemisphere, ASD r=.02, p=.836; NT r=.04, p=.621 (Figure S5).

SM2.3 N170 amplitude

N170 amplitude was examined to rule out the possibility that group differences related to general factors like skull thickness or head size, which would be expected to have effects on both latency and amplitude. The sample overall showed a normative pattern of larger responses to faces in the right than the left hemisphere ($F(1,430)=28.61$, $p<.001$). Likewise, amplitudes decreased with age ($F(2,430)=64.163$, $p<.001$). The groups did not differ on N170 amplitudes across both hemispheres ($F(1,430)=0.358$, $p=0.550$), and there was no significant interaction between hemisphere and group ($F(1,430)=0.707$, $p=0.401$).

SM2.4. P1 to upright faces - case/control effects

SM2.4.1. Latency:

P1 latency decreased with age (F(2,430)=5.3, p=0.005), and did not differ between groups (F(1,430)=1.12, p=0.29). Latencies did not differ across hemisphere (F(1,430)=1.40, p=0.24), and diagnosis, age and hemisphere did not interact (all F's<.1.2, all p's>.28).

SM2.4.2. Amplitude:

The sample overall showed a normative pattern of larger responses to faces in the right than the left hemisphere (P1: F(1,430)=17.9, p<0.001). Likewise, amplitudes reduced with age, P1: F(2,430)=152.17, p<.001. The groups did not differ on **P1** amplitude across both hemispheres (P1: F(1,430)=1.82, p=0.18, and there was no significant group by laterality interaction (P1: F(1,430)=0.78, p=0.38).

SM2.5 Core and Associated symptoms

Table S6 shows the relationship between N170 latency and core and associated symptoms, controlled for age.

SM2.6. Inversion effects - N170 latency and amplitude

In addition to presenting participants with upright faces, we also included an equal number of trials with inverted (rotated 180°) faces in order to probe whether this provided supporting evidence for disruption to configural aspects of face processing. In these models, condition and hemisphere were repeated factors, and diagnosis and age group were fixed factors. A compound symmetry covariance matrix was used for both. Significant interactions were followed up with simple main effects analyses.

SM2.6.1. Latency:

As expected, latency of the N170 across the sample was slower to inverted than upright faces (N170: $F(1,430)=8.80$, $p=0.003$). There was also an interaction between diagnostic group and inversion ($F(1,430)=7.67$, $p = 0.006$) such that there was no significant effect of condition in the ASD group ($F(1,243) = 0.02$, $p = 0.9$) but there was a significant effect in the TD/ID group ($F(1,187) = 15.86$, $p <0.001$).

SM2.6.2 Amplitude:

Across the whole sample, the N170 exhibited a normative pattern of larger amplitudes to inverted than to upright faces, $F(1,430)=41.15$, $p<.001$. The inversion effect on N170 amplitude did not differ by diagnostic group ($F(1,430)=2.58$, $p=0.11$).

SM2.7. Relation to fMRI

Functional brain responses were acquired on 3 Tesla MRI scanners as part of the LEAP protocol. Acquisition parameters were harmonized across sites as closely as possible. Functional images were collected using an echo-planar imaging (EPI) sequence and structural images were acquired with a high-resolution T1-weighted magnetization-prepared rapid gradient echo sequence. Further details of the fMRI procedures are available [94,95].

fMRI data analysis followed standard processing routines in SPM12 (http://www.fil.ion.ucl.ac.uk/spm/), including slice-time correction, a two-step realignment procedure, unified segmentation and normalization to standard stereotactic space as defined by the Montreal Neurological Institute (MNI), and smoothing with an 8mm full-width-at-half-maximum Gaussian Kernel. Data were subjected to an extensive quality assessment pipeline, and datasets with excessive head motion (>20% of trials with a framewise displacement greater 0.5 mm) were excluded. For each subject, task conditions were modeled as boxcar functions that accounted for the presentation of face blocks and shape blocks, respectively. Task regressors were convolved with the canonical hemodynamic response function (HRF) and subjected as predictors to a general linear model (GLM), along with six realignment parameters to account for head motion. During first-level model estimation, data was high-pass filtered with a cut-off of 256 s, and an autoregressive model of the first order was applied. The faces condition was subsequently contrasted to the shapes condition to identify brain responses reflecting sensitivity to emotional faces.

SM2.8. Relation to dimensional socialisation

SM2.8.1. Relation to subdomains of the Vineland

Table S7 shows the relation between the N170L and prognostic change in the Vineland Socialisation domain score and constituent subdomains.

Effect of variability in the time between baseline and follow-up: A regression controlling for age at baseline and the time in days between baseline and follow-up assessments also showed a significant effect of N170 latency (overall model $F(3,144)= 3.05$, $p = 0.031$, $r^2=0.06$; N170L beta=-.020, $t(144)=-2.15$, $p = 0.033$); and using the rate of change of v-scale scores as the dependent variable (score difference divided by time gap) and controlling for age at baseline also showed the same effect (overall model $F(3,144)= 3.04$, $p = 0.031$, $r^2=0.06$; N170L beta=-.021, $t(144)=-2.26$, $p = 0.026$), indicating that time between baseline and follow-up did not confound results.

SM2.8.2 Cluster analysis

Table S8 shows diagnostic and clinical profiles of the three clusters within the ASD group.

SM2.8.3 Selecting an N170 cut-off

SM2.8.3.1 Normative modelling

The latency of the N170 is strongly related to age, but nonlinearly. The selection of a cutoff related to 'raw' N170 latency would thus be heavily confounded with age. Therefore, scores are to be transformed into an age-dependent space; normative modelling was selected as it does not assume an a priori distribution within each age bracket.

Normative modelling is a statistical framework for mapping between behavioural, demographic or clinical characteristics and a quantitative biological measure, providing estimates of centiles of variation across the population (Marquand et al. 2019). It is

conceptually related to the way in which height or weight norms are derived. First, we selected the neurotypical cohort from the LEAP sample (age 6 to 30 years) to build the initial model. This neurotypical cohort underwent an identical procedure to the ASD group. This allows the normative model to then be used to characterise responses within the ASD group, without any confounds from procedural differences. Second, a statistical model is estimated to model variance in a response variable (a.k.a. target or dependent variable) from a set of clinically relevant covariates (predictor or independent variables) across the reference cohort. Our dependent variable was N170 latency (average of left and right) and our clinically relevant covariate was age (continuously entered) because of demonstrated associations between age and N170 latency in our sample, and the absence of associations with other clinical variables. The Gaussian process regression was selected as the statistical model because it estimates distinct variance components and provides predictions for each participant that account for all sources of uncertainty. This normative model provides – at each age – a predicted mean latency and associated variances. Third, it is necessary to assess the accuracy of the normative model for predicting the response variable (e.g., mean-squared error, explained variance). The root mean squared error of the model was 24.4ms with 22.3% of the variance in N170 latency explained by age. For comparison, a linear fit of age on N170 latency has a RMSE of 25.4ms and explains 17.0% of the variance – an increase in variance explained by the normative model of 5.3% over a linear fit. Finally, this model can be applied to quantifying the deviations of individual samples from a target cohort (e.g. clinical cohort) with respect to this reference model. Within the present context, we applied this normative model to data from the ASD group tested within LEAP. Specifically, for each participant in the target cohort we are able to compare the measured latency to the predicted mean latency and associated predicted variance from the normative

model derived in the neurotypical cohort at the given age. The z-N170L is a simple transformation that expresses the number of standard deviations a measured N170 latency deviates from the predicted latency at the age of the participant (Figure S6).

SM2.8.3.2 Defining a z-N170L cut-off

Figure S7 shows the sensitivity and specificity of the relationship between the z-N170L and Improvers vs Non Improvers on the Vineland Socialisation Play and Leisure Time subscale.

Based on the ROC curve shown in Figure S2 (Area=0.67, standard error=0.04, 95%CI = 0.59-0.76), a cut-off of $\geq$ 0.5 for the z-N170L was selected. We here chose to maximise specificity over sensitivity in order to minimise the number of Improvers remaining in the trial sample. Figure S8a illustrates the options for different cut-offs, which may suit different trial needs (e.g. based on the risk:benefit ratio of a particular medication); Figures S8b and c illustrate bootstrapped methods (1000 iterations) for optimising cutoffs to maximise either sensitivity in the context of a reasonable level of specificity (b) or specificity in the context of a reasonable level of sensitivity (c).

SM2.8.3.3. Monte-Carlo simulation

Trials simulated as “Non-enriched” included all autistic subjects with valid time 1 and time 2 data on the Vineland Play and Leisure time subdomain, regardless of N170 latency status, from the EU-AIMS LEAP study. Trials simulated as “Enriched” included a subset of the “Non-enriched” sample, namely any subject with QC-passing N170 recordings and an N170

latency $\geq$ 0.5 SDs than the mean N170 latency expected of a typically-developing subject of equivalent age (as determined through normative modelling).

A total of 2500 randomized (1:1), placebo-controlled, 12-week clinical trials with and without enrichment were simulated using an estimated fixed effect size of intervention of Cohen's D=.45, as follows. First, the amount of change expected from a 12-week trial duration was simulated by fitting a linear model to the empirically observed change in Vineland Play and Leisure (PLT) subdomain v-scores from EU-AIMS LEAP, with baseline PLT v-score, age at time 1, and follow-up duration as predictors. Next, this fitted model was used to predict the PLT change scores expected from each subject over 12 weeks. Each subject's residual from the fitted model was then added to these predictions, as well as (for subjects simulated as randomly assigned to intervention) the Cohen's D multiplied by the residual population standard deviation (as estimated from the same model). Sample size per arm ranged from 25 to 250 subjects, simulated through sampling with replacement from either the enriched or non-enriched population (as defined above). Simulated Week 12 PLT v-scores were rounded to the nearest integer (v-scale scores are integer-valued) and truncated to fall within the permissible v-score range (0,24) before subtracting baseline PLT v-scale scores from these rounded and truncated Week 12 values to compute simulated change scores. Finally, for each simulated trial, these simulated 12-week change scores in the Vineland-II PLT were analysed using a linear model, with main effects of baseline score and intervention.

The statistical power is the probability of detecting an existent effect, in this case, the drug effect of Cohen's D of 0.45 (relative to the non-enriched population) in the hypothesized direction with a p-value <= 0.05 (two-sided). The estimated power by sample size graph for N170 latency enriched (i.e., only subjects with t2 and t1 Vineland PLT data, good quality

N170 data as determined by QC, and a delayed N170 latency, defined as > 0.5SDs from the corresponding mean age-normalized N170 latency) and non-enriched (i.e., all subjects with t2 and t1 Vineland data) is presented in the main text. Based on interpolation across the simulations, approximately 78 subjects per arm would be required in a non-enriched placebo-controlled clinical trial in order to detect a beneficial drug effect of equivalent magnitude with a 80% probability (type II error or $\beta = 0.20$) at $\alpha = 0.025$ (one-sided, or [equivalently] $\alpha = 0.05$ two-sided). Conversely, the same 80% probability of detecting an analogous drug effect at the same α is achieved with approximately 48 subjects per arm in an enriched clinical trial. This represents a reduction in sample size of approximately 38%.

2.8.3.4 Information on the psychometric properties of the Vineland

While we do not have test-retest data of the VABS-II play and leisure time sub-domain in our cohort, [81]reported moderate-to-good test-retest reliability (i.e. Intraclass Correlation Coefficients; ICC) of the VABS-II play and leisure time sub-domain ranging from .68 to .78 across different age groups as part of the development of the VABS-II (i.e. ages 7-13 (N=175): ICC=.78; ages 14-21 (N=90): ICC=.68; ages 22-71 (N=63): ICC=.78). In relation to the reliability of absolute changes in the play and leisure time sub-domain, data from the VABS-II normative sample suggests small standard error of measurements (SEMs) across our age groups studied, ranging from 0.95 (ages 22-31) to 1.62 (age 12-13[81]). Due to a lack of established cut-offs that indicate clinically meaningful improvement at the VABS-II subdomain level (see[69]) and the moderate to-good psychometric properties of VABS-II play and leisure time scores (both in terms of test-retest reliability and estimates of SEM), we have opted to used absolute changes in VABS-II.

SM2.9 Genetic associations with other phenotypes

Interestingly, individual differences in the face inversion effect on the N170 latency also tends to be associated with ASD PGS (Spearman's r2 = 0.0129 ; p = 0.039). Figure S5 shows the association between a range of different PGS and the N170L.

SM2.10: Additional fMRI information

Figure S3 shows the fusiform face area mask; Figure S4 shows the association between the BOLD response and the N170 latency at different height thresholds.

SM2.11: Split-half reliability of the N170

Method**:** We examined the internal reliability of the amplitude and latency of the N170 component to upright faces by calculating the intraclass correlation (ICC, type C-1) between odd and even-numbered trials, separately for each hemisphere (Table S9). We included only those participants with at least 40 valid trials elicited by upright faces, in order to maintain a criterion of 20 valid trials in each half of the split.

Next, we calculated individual participant average ERPs for odd and even trials, excluding participants where the N170 peak was not clear in either odd or even trial averages. After measuring the mean amplitude and peak latency of the N170 for each valid average we calculated the ICC between odd and even trials. We did this for the whole sample, by diagnosis group (ASD/NT), by age group (children/adolescents/adults) and by the presence or absence of mild intellectual disability (IQ<70).

Results: Across the whole sample, the internal reliability of the N170 component was either good or excellent in both hemispheres (see Table S9 for full details). Observed ICCs were

higher for latency (.95) than for amplitude (.84-.88). ICCs for subgroup analyses (by diagnosis, age, and presence of mild-ID) were all either good or excellent with the exception of amplitude in the left hemisphere in the mild-ID group, which was moderate, ICC=.69.

Supplementary Figures

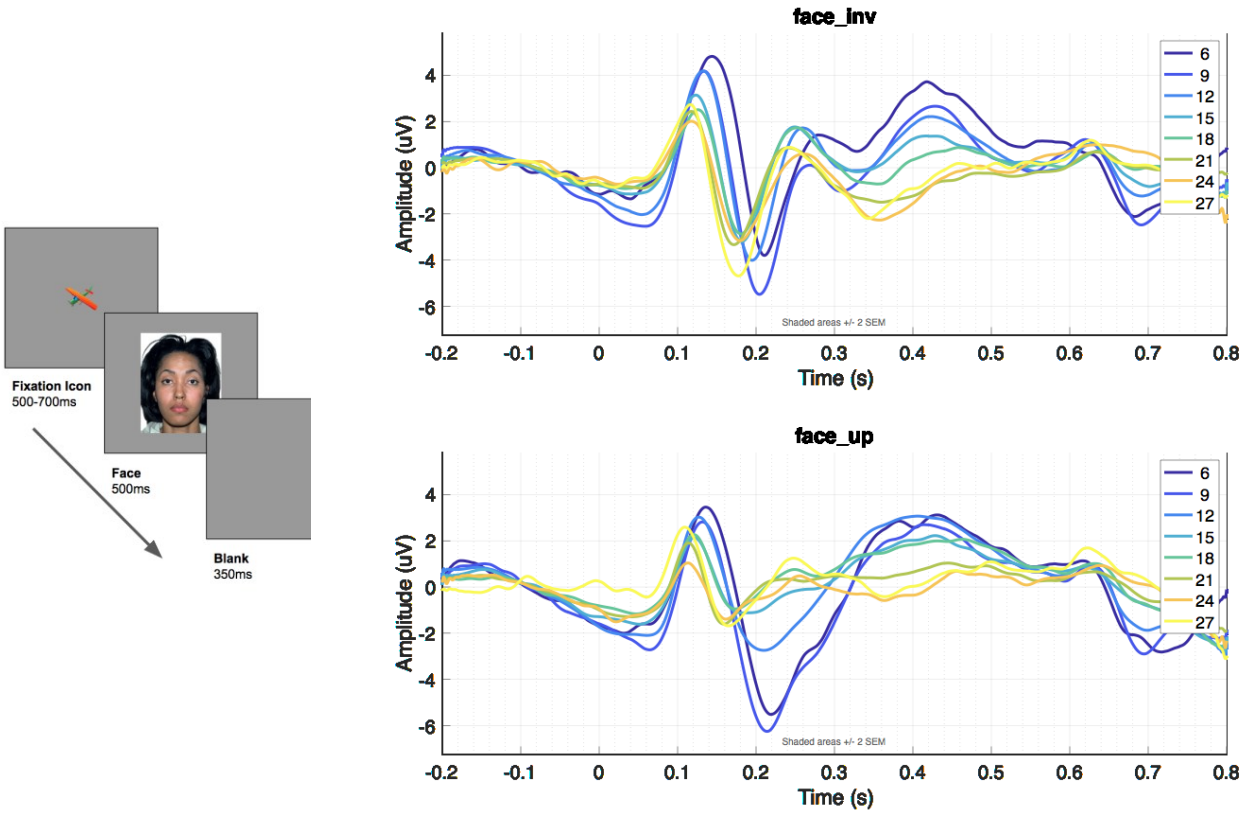


Figure S1. Grand average ERPs to face inverted (upper) and face upright (lower) conditions, in three-year age bins, with depiction of trial structure (left).

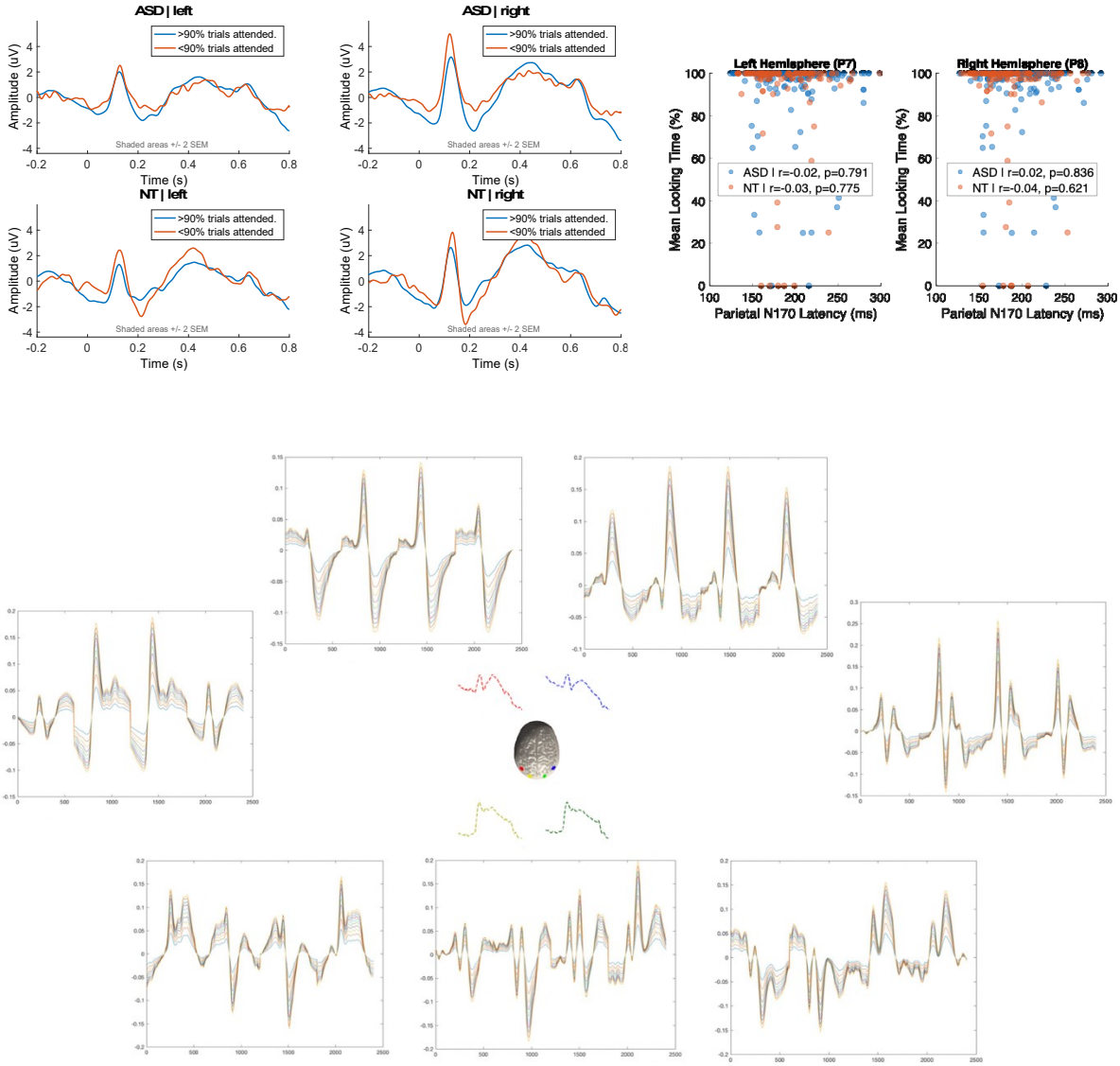


Figure S2: Top left panel: ERPs at each hemisphere (columns) and in the ASD and NT groups (rows), elicited by subjects with >90% trials attended (blue line) and <90% attended (red line). Top right panel: Relationship between N170 latency and percentage of trials attended, at the left and right hemispheres. Bottom: Illustration of the seven principal

components of the individually averaged EEG data concatenated across electrodes from P7 (red), P8 (blue), O1 (yellow), O2 (green) that were entered into the cluster analysis. Coloured lines indicate the effects of different downsampling approaches.

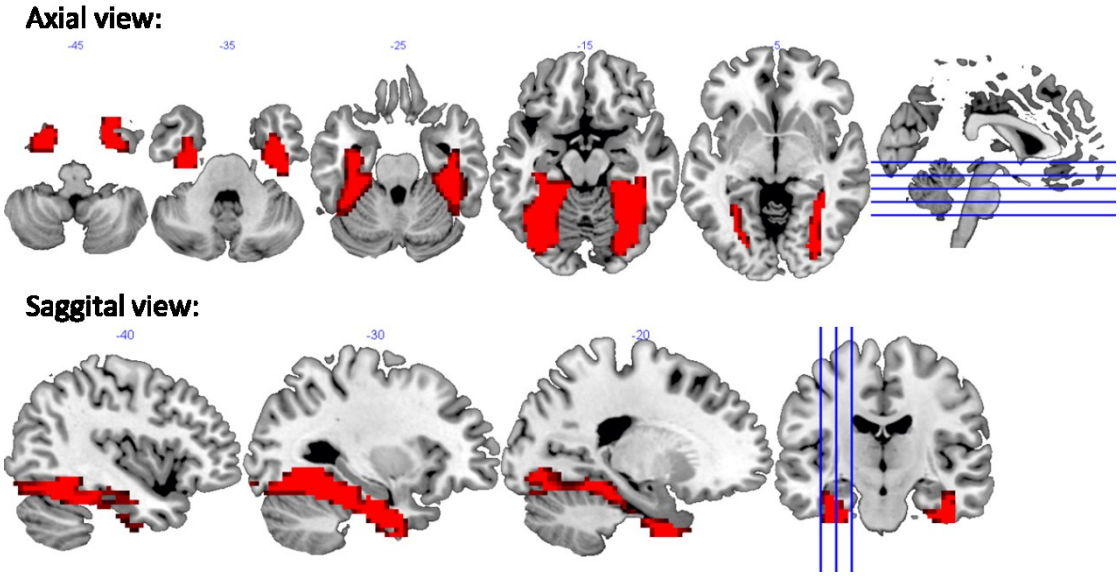


Figure S3: Illustration of the a-priori defined mask of the fusiform face area.

The mask of the fusiform face area was derived from the Anatomical Automatic Labelling Atlas (Tzourio-Mazoyer et al., 2002).

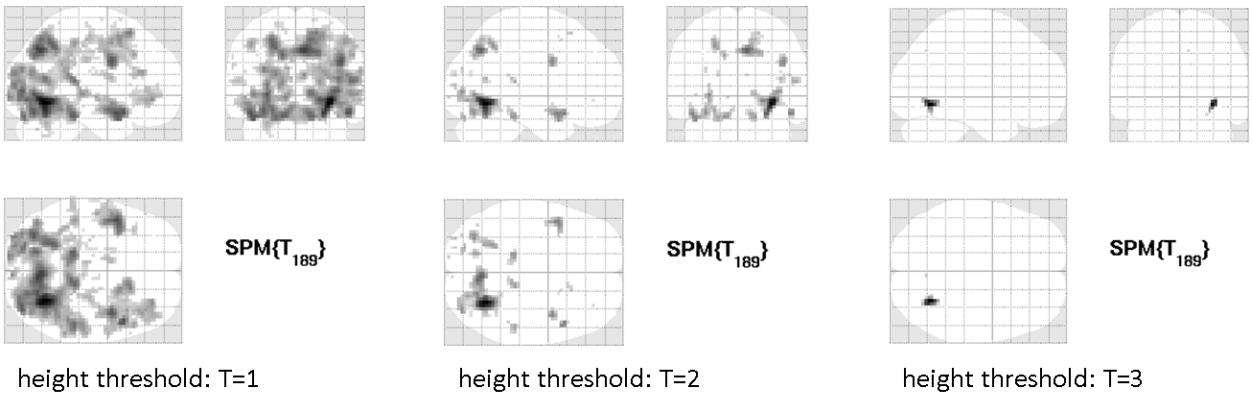


Figure S4: Illustration of fMRI brain maps reflecting the association between face-sensitive BOLD response and N170 latency at different height threshold levels.

The height threshold was defined at t-value t=1, t=2 and t=3. The peak voxel is located in the right fusiform face area MNI [30 -64 -10], t=3.93.

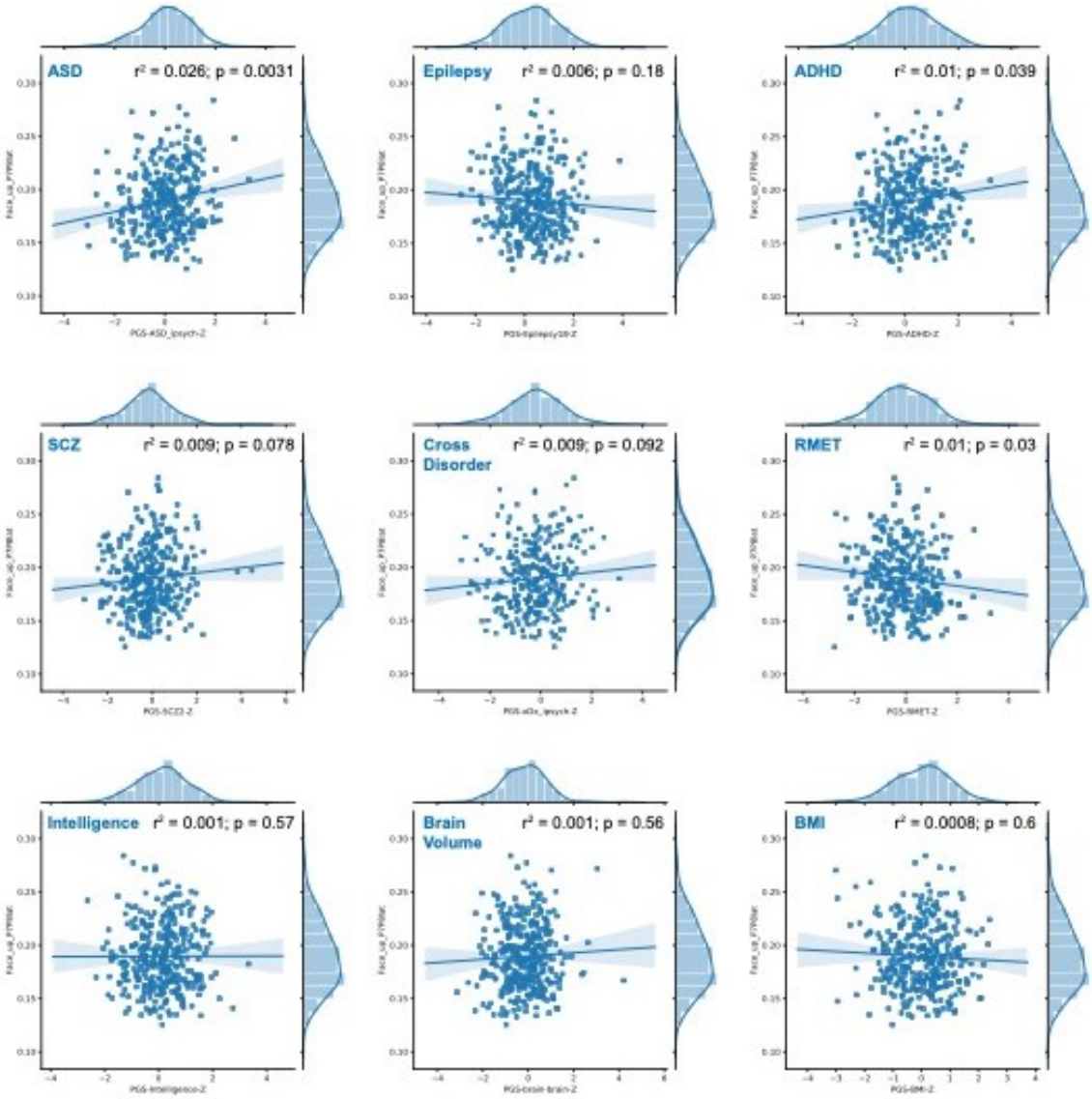


Figure S5: Correlation analyses between N170 latency and other polygenic scores.

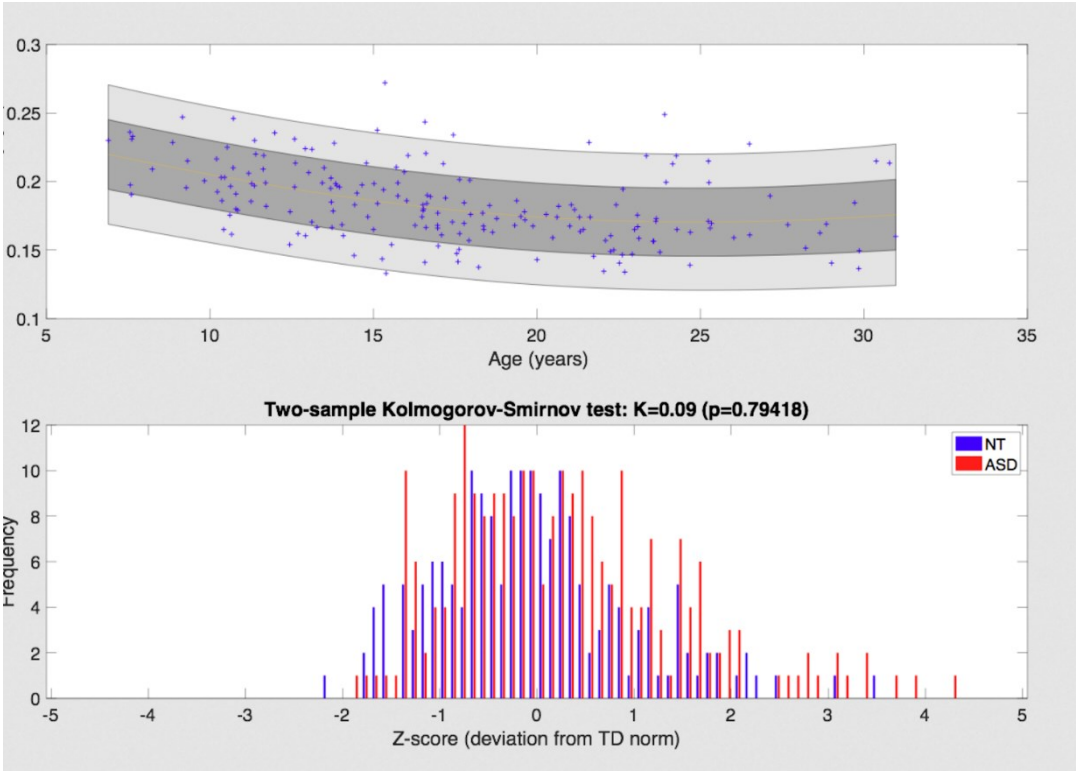
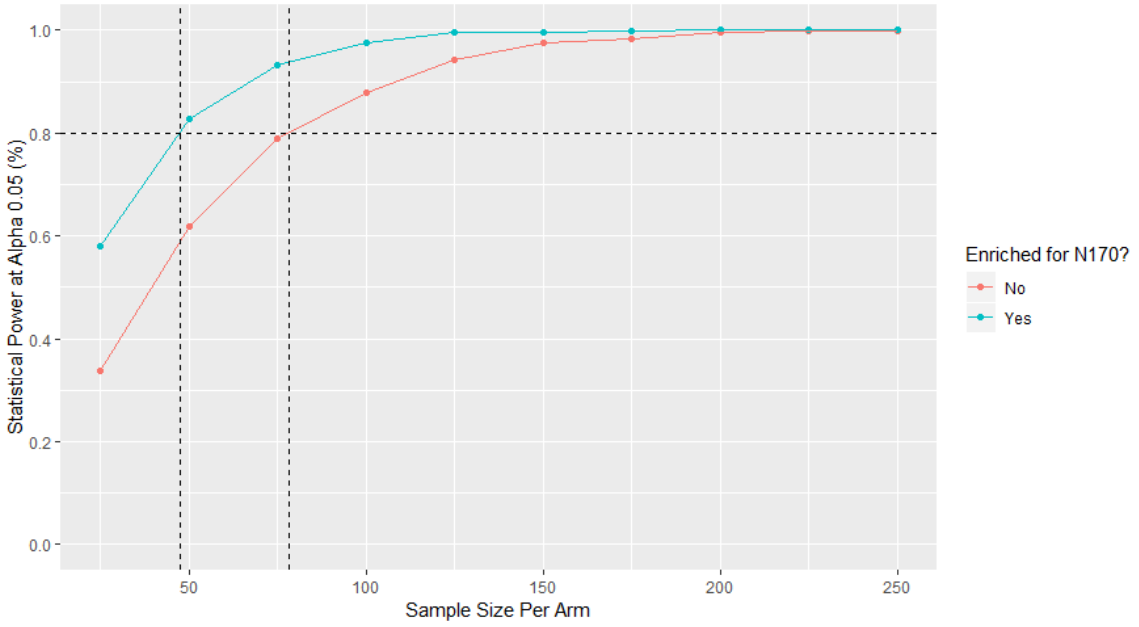


Figure S6: Normative modelling of the z-N170L.

Note: In the upper graph, the dark grey shading represents 1 standard deviation away from the age-adjusted mean and the pale grey shading represents 2 standard deviations away. The yellow line represents the age-adjusted mean.



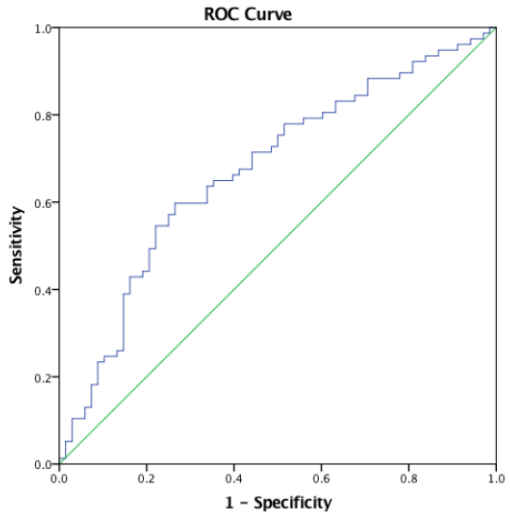


Figure S7: Top: Statistical Power by Sample Size for Placebo-Controlled N170 Latency Enriched vs. Non-Enriched Clinical Trials with a Simulated Interventional Effect Equivalent to a Cohen's D of 0.45 in the Non-Enriched Population and a Simulated 12-Week Trial Duration. Bottom: Receiver Operating Characteristic Curve showing the achieved sensitivity and specificity for detecting non-improvers using different zN170L cut-offs.

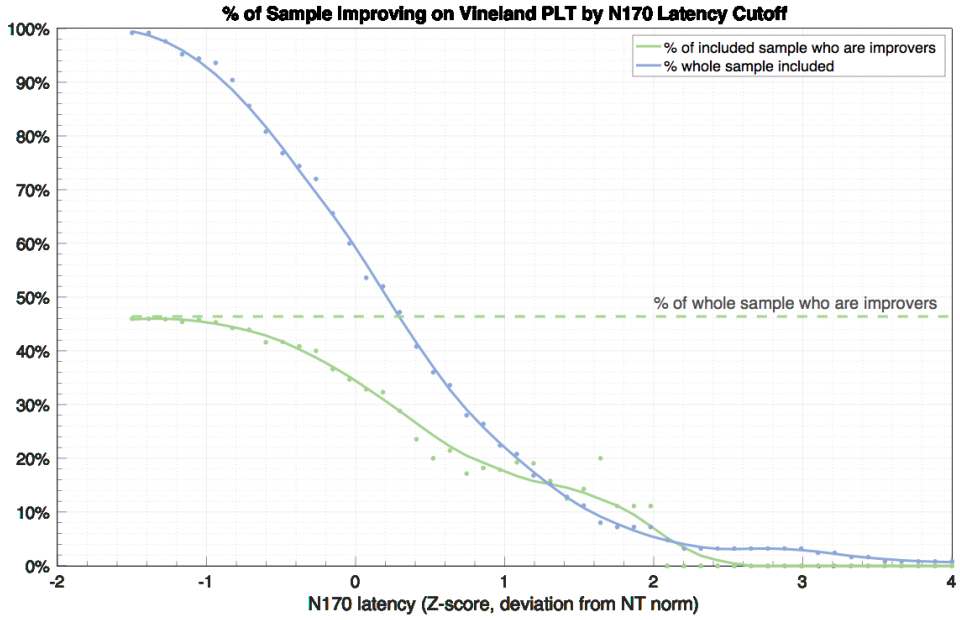


Figure S8a: Effect on prognosis at different N170 latency cut-offs.

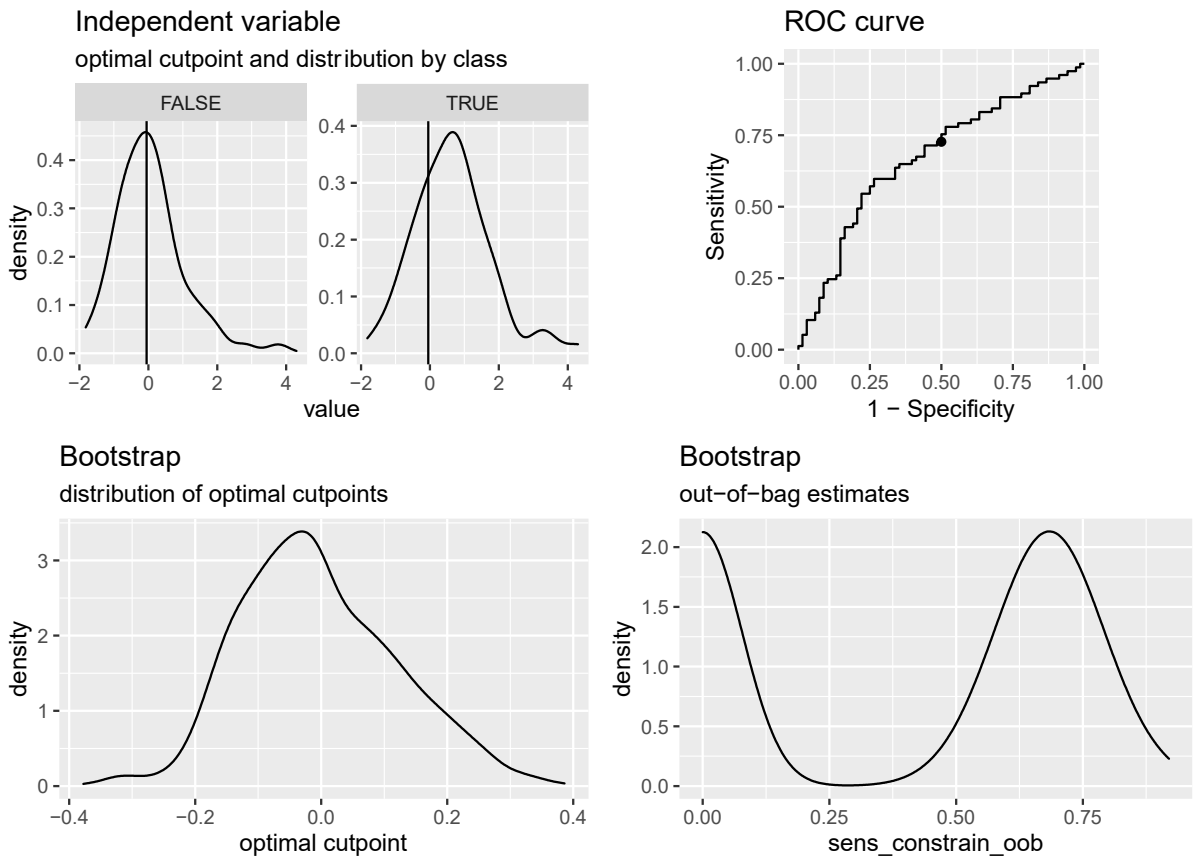


Figure S8b: Example of bootstrapped cutoffs to maximize sensitivity given a reasonable level of specificity (constructed with R package cutpointr).

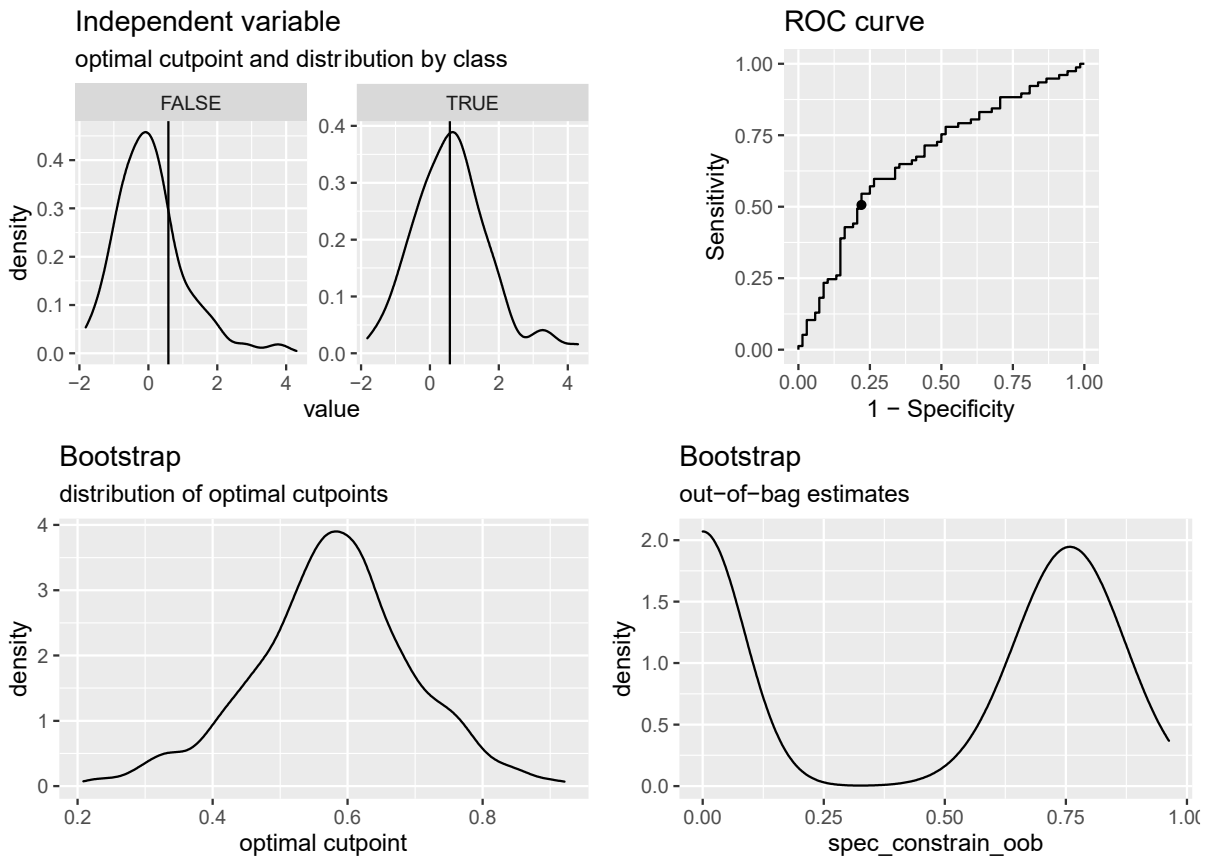


Figure S8c: Example of bootstrapped cutoffs to maximize specificity given a reasonable level of sensitivity (constructed with R package cutpointr).

Supplementary Tables

Group	**Intellectual functioning**	**Children** (6-11 years)	**Adolescents** (12-17 years)	**Adults** (18-30 years)	**Total**
ASD	ASD-no ID (IQ≥75)	50	75	80	205
	ID-ASD (IQ 40-74)	0	22	19	41
	ASD total	**50**	**97**	**99**	**246**
Control	TD (IQ≥75)	38	63	65	166
	ID-control (IQ 40-74)	0	14	10	24
	Control total	**38**	**77**	**75**	**190**
	Grand Total	**88**	**174**	**174**	**436**

Table S1A: Recruitment profile of the sample with EEG data.

Baseline visit	Autism Spectrum Disorder (*n*=246)	Control (*n*=190)
Sex (males:females); % of females	66:180 (27%)	69:121 (36%)
Age (years)	17.0 (5.6)	17.5 (5.7)
Verbal IQ (WASI/WISC)	96.0 (19.4)	102.3 (20.6)
Performance IQ (WASI/WISC)	97.1 (21.0)	102.0 (20.0)
Full-scale IQ (WASI/WISC)	96.6 (19.0)	102.3 (19.0)
VABS Communication Standard Score	74.9 (16.8)	85.7 (24.6)
VABS Daily Living Standard Score	73.7 (16.2)	86.5 (22.4)
VABS Socialisation Standard Score	70.6 (16.6)	93.2 (24.0)
VABS Play and Leisure Time V-Score	10.3 (3.4)	13.2 (3.7)
VABS Coping Skills V-Score	11.5 (3.5)	14.6 (4.4)
IVABS Interpersonal Relationships V-Score	8.5 (3.1)	13.5 (4.2)
ADI-R Social	16.0 (7.2)	NA
ADI-R Communication	12.7 (5.8)	NA

ADI-R Restricted and Repetitive Behaviors	4.2 (2.7)	NA
ADOS – CSS Total	5.3 (2.7)	NA
ADOS – CSS SA	6.0 (2.6)	NA
ADOS – CSS RRB	4.7 (2.7)	NA
DAWBA Externalizing	1.9 (1.7)	0.9 (1.1)
Follow-up visit	Autism Spectrum Disorder (*n*=223)	Control (*n*=155)
Sex (males:females); % of females	62:161 (28%)	59:96 (38%)
Age (years)	18.2 (5.6)	18.2 (5.4)
VABS Communication Standard Score	74.0 (17.9)	NA
VABS Daily Living Standard Score	75.6 (16.3)	NA
VABS Socialisation Standard Score	74.4 (17.6)	NA
VABS Play and Leisure Time V-Score	11.2 (3.3)	NA
VABS Coping Skills V-Score	12.1 (3.7)	NA
VABS Interpersonal Relationships V-Score	9.5 (3.3)	NA
ADOS – CSS Total	5.3 (2.8)	NA

ADOS – CSS SA	5.8 (2.6)	NA
ADOS – CSS RRB	5.1 (2.8)	NA

Table S1B: Clinical and diagnostic profile of individuals with EEG data within the LEAP sample.

Data is M(SD). Participants at the Cambridge site for whom EEG was not attempted were excluded. Clinical and diagnostic information at follow-up only presented for those with valid baseline and follow-up scores. VABS was not administered to controls at follow-up. IQ = Intelligence Quotient; VABS = Vineland Adaptive Behavior Scales-II; ADI-R = Autism Diagnostic Interview – Revised 4-to-5-years/ever algorithm scores; ADOS CSS Total, SA, RRB = Autism Diagnostic Observation Schedule Calibrated Severity Scores for Total, Social Affect and Restricted and Repetitive Behaviors; DAWBA = Development and Well-Being Assessment.

	Autism Spectrum Disorder (N = 453)	Control (N = 311)
Not collected due to site (UCAM did not collect EEG data)	61	34
Not collected for that individual (e.g., ran out of time in session)	46	16
Technical or upload error, or other	20	11
Did not complete EEG battery	23	9
Too few trials (<20 artifact-free)	31	25
Poor peaks	26	26
Included in final sample with EEG data	246	190

Table S2: Reasons for EEG data loss separated by group.

Variable/scale	Autism Spectrum Disorder		Control	
	Excluded M(SD) n=207	Included M(SD) n=246	Excluded M(SD) n=121	Included M(SD) n=190
Sex (males:females); % females		66:180 (27%)		69:121 (36%)
Age (years)	16.6 (6.2)	17.0 (5.6)	16.7 (6.4)	17.5 (5.7)
Vineland Communication Standard Score	74.9 (20.0)	74.9 (16.8)	89.3 (31.0)	85.7 (24.6)
Vineland Daily Living Standard Score	70.3 (18.8)	73.7 (16.2)	85.1 (25.0)	86.5 (22.4)
Vineland Socialisation Standard Score	68.7 (17.2)	70.6 (16.6)	87.6 (30.0)	93.2 (24.0)
(Play and Leisure time)	9.6 (3.8)	10.3 (3.4)	13.0 (4.7)	13.2 (3.7)
(Coping skills)	10.6 (3.4)	11.5 (3.5)	13.1 (4.9)	14.6 (4.4)
(Interaction)	8.6 (3.4)	8.5 (3.1)	13.0 (5.4)	13.5 (4.2)
ADI Social*	17.4 (6.1)	16.0 (7.2)	NA	NA
ADI Communication*	13.7 (5.5)	12.7 (5.8)	NA	NA
ADI RRB*	4.3 (2.6)	4.2 (2.7)	NA	NA
ADOS CSS*	5.6 (3.0)	5.3 (2.7)	NA	NA
ADOS Social Affect*	6.1 (2.8)	6.0 (2.6)	NA	NA
ADOS RRB *	4.9 (2.9)	4.7 (2.7)	NA	NA
Verbal IQ	94.3 (22.6)	96.0 (19.4)	103.2 (20.2)	102.3 (20.6)
Performance IQ	94.7 (22.3)	97.1 (21.0)	102.6 (21.5)	102.0 (20.0)
Full-scale IQ		96.6 (19.0)		102.3 (19.0)
DAWBA externalising	2.1 (1.7)	1.9 (1.7)	0.8 (1.2)	0.9 (1.1)
Time 2				
T2 Vineland Socialisation Standard Score	73.0 (16.2)	74.8 (17.3)	NA	NA
(T2 Play and Leisure)	10.46 (3.6)	11.27 (3.3)	NA	NA
(T2 Coping skills)	11.9 (3.5)	12.3 (3.7)	NA	NA
(T2 Interpersonal r/ships)	8.9 (3.0)	9.5 (3.2)	NA	NA

NA = not available

Table S3: Clinical and diagnostic profile of individuals who did and did not provide EEG data within the LEAP sample. Note: Data from participants from the Cambridge site, where EEG was not collected, were excluded. Data are M (SD). RRB = Restricted and repetitive behaviors; CSS = calibrated severity score.

Site	Delay (ms)
KCL	49.2
CIMH	47.6
RUNMC	25.8
UMCU	6.5
UCBM	n/a

Table S4: Delays in milliseconds observed in stimulus presentation and corrected in analysis.

Trait	PGS R2	PGS P-value	GWAS P-value threshold	Number of SNPs	MAF
BMI	0.0037	0.167	0.08	29797	0.01
ASD	0.0094	0.027	0.04	29591	0.01
SCZ	0.0050	0.11	0.01	10848	0.1
ADHD	0.012	0.013	0.17	71033	0.01
Intelligence	0.0012	0.42	0.35	172222	0.01
Cross Disorder	0.0046	0.12	0.03	8480	0.1
Brain volume	0.021	0.00097	0.2	52278	0.01
RMET	0.013	0.0084	0.01	5149	0.1
Epilepsy	0.0024	0.26	0.02	13011	0.01

Table S5. Information of the PGS best model fit of each trait. DOIs: BMI

https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_da

ta_files; ASD 10.1038/s41588-019-0344-8 ; SCZ 10.1038/nature13595 ; ADHD 10.1038/s41588-018-0269-7 ; Intelligence 10.1038/s41588-018-0152-6 ; Cross disorder 10.1038/s41593-018-0320-0 ; Brain volume https://doi.org/10.1093/cercor/bhz241; RMET 10.1038/mp.2017.122 ; Epilepsy 10.1038/s41467-018-07524-z. PGS R2, Variance explained by the PGS; PGS P-value, P-value of the model fit; GWAS P-value Threshold, Best P-value Threshold that differentiated the best the two group of European participants (autism with or without ID and typically developing from the LEAP study); Number of SNP, Number of SNPs included in the model; MAF, Minor Allele Frequency threshold; BMI, Body Mass Index; ASD, Autism Spectrum Disorder; ADHD, Attention Deficit and Hyperactivity Disorder; SCZ, Schizophrenia; Brain volume; RMET, Read the Mind in the Eyes Test.

	ASD	NT
Age (years)	r(246) = -.44, p < 0.001	r(190) = -.41, p < 0.001
Core symptoms		
ADOS-CSS SA	r(209) = .20, p = 0.003	N/A
ADOS-CSS RRB	r(209) = .064, p = 0.36	N/A
Vineland Socialisation domain	r(206) = -.017, p = .80	r(57) = .016, p = .90
Coping	r(208) = 0.01, p = 0.88	r(58) = -0.02, p = .88
Interpersonal Relationships	r(209) =-0.063, p = 0.37	r(57) = .10 p = .45
Play and Leisure Time	r(209) = 0.014, p = 0.84	r(58) = -0.02, p = .88

Social Responsiveness Scale	r(200) = 0.11, p = .13	r(104) = 0.04, p = .71
Associated symptoms		
Verbal IQ	r(239) = -0.05, p = 0.45	r(187) = -0.05, p = 0.48
Performance IQ	r(240) = -0.1, p = 0.137	r(187) = -0.06, p = 0.46
DAWBA internalising	r(211) = 0.05, p = 0.476	r(151) = .031, p = 0.70
DAWBA externalising	r(211) = 0.05, p = 0.513	r(151) = .079, p = 0.33

Table S6: Partial correlations for association between N170 latency at P7/P8 to upright faces and associated symptoms, controlled for age.

Subdomain	Follow-Up Assessment [b]
	N=145 with baseline EEG +Vineland at follow-up
Coping	r(141) = -0.05, p = 0.538
Interpersonal Relationships	r(137) =0.022, p = 0.796
Play and Leisure Time	r(141) = -0.235, p = 0.005 [c]
Vineland Socialisation	r(136) = -0.058, p = 0.500

Table S7: Relation between the N170 and prognostic change in the Vineland Socialisation subdomains

[b] Controlling for age and baseline score

[c] Corrected p-value for 4 comparisons p = 0.02

Note: 80% of the participants had their follow-up assessments between 450 and 750 days (c. 14 months to 2 years) after the baseline assessments (mean = 596 days, SD = 98.1, range = 149–914)

	Cluster 3 (n=101)	Cluster 1 (n=118)	Cluster 2 (n=27)
N170 Latency *F*(2,245) =64.32, *p* < 0.001; age covaried *F*(2,245) =31.991, *p* < 0.001	.171 (.023)*	.205 (0.03)*	.220 (0.03)*
Vineland Play and Leisure Time Change F(2,144) =4.41, p =0.014; age covaried F(2,144) = 2.21, p = 0.11	1.72 (3.34)	.52 (3.17)	-.46 (2.48)*
Age (years) *F*(2,245)=57.3, *p* <0.001	20.6 (4.28)*	15.1 (4.84)*	12.1 (4.67)*
Sex	27F (27%), 74M	32F (27%), 86M	7F (26%), 20M

$\chi^2(2)=0.017$, p=0.99.			
ADOS SA CSS *F*(2,211) = .52, p = 0.60	5.87 (2.56)	6.22 (2.68)	5.80 (2.90)
ADOS RRB CSS *F*(2,211) = 1.81, p = 0.17	4.82 (2.61)	4.79 (2.72)	3.72 (2.69)
Verbal IQ *F*(2,241)=.33, *p* =0.72	96.8 (20.2)	94.9 (18.3)	97.33 (21.5)
Performance IQ *F*(2,242)=1.77, *p* =0.17	96.8 (20.2)	94.9 (18.3)	97.3 (21.5)

Table S8: Clinical profile of the three clusters within the ASD group. Clusters have been ordered by N170 latency for ease of interpretation. Figures are mean (standard deviation).

	Left Hemisphere (P7)				Right Hemisphere (P8)			
Group	N >40 trials (% of main analysis)	N valid N170 peaks (% of reliability analysis)	ICC Amplitude (95% CI)	ICC Latency (95% CI)	N >40 trials (% of main analysis)	N valid N170 peaks (% of reliability analysis)	ICC Amplitude (95% CI)	ICC Latency (95% CI)
Whole Sample	293 (67.2%)	250 (85.3%)	.84 (.81-.87)	.95 (.93-.96)	263 (69.0%)	263 (84.8%)	.88 (.85-.91)	.95 (.94-.96)
ASD	156 (69.0%)	129 (82.7%)	.84 (.78-.89)	.94 (.92-.96)	165 (65.7%)	135 (81.8%)	.91 (.88-.93)	.96 (.95-.97)
NT	137 (78.3%)	121 (88.3%)	.83 (.77-.88)	.95 (.92-.96)	145 (73.2%)	128 (88.3%)	.84 (.78-.88)	.94 (.92-.96)

Children (6-12)	82 (63.6%)	74 (90.2%)	.81 (.71-.88)	.95 (.93-.97)	89 (60.5%)	77 (86.5%)	.87 (.81-.92)	.97 (.96-.98)
Adolescents (13-17)	86 (80.4%)	71 (82.6%)	.77 (.65-.85)	.92 (.88-.95)	90 (76.3%)	75 (83.3%)	.85 (.77-.90)	.92 (.88-.95)
Adults (18-30)	125 (75.8%)	105 (84.0%)	.80 (.72-.86)	.92 (.89-.95)	131 (71.2%)	111 (84.7%)	.83 (.76-.88)	.92 (.89-.95)
No ID (IQ>70)	250 (73.5%)	216 (86.4%)	.86 (.82-.89)	.95 (.93-.96)	265 (69.7%)	227 (85.7%)	.88 (.85-.91)	.96 (.94-.97)
Mild ID (IQ<70)	43 (70.5%)	34 (79.1%)	.69 (.46-.83)	.94 (.88-.97)	45 (65.2%)	36 (80.0%)	.87 (.76-.93)	.95 (.89-.97)

Table S9. Summary of internal reliability of the N170 ERP component by hemisphere.