

# 1 Prematurity and Genetic Liability for Autism Spectrum Disorder

2 Yali Zhang<sup>1,2</sup>, Ashraf Yahia<sup>1,2</sup>, Sven Sandin<sup>3,4,5</sup>, Ulrika Åden<sup>6,7,8</sup> and Kristiina Tammimies<sup>1,2\*</sup>

3 <sup>1</sup>Center of Neurodevelopmental Disorders (KIND), Centre for Psychiatry Research, Department of  
4 Women's and Children's Health, Karolinska Institutet

5 <sup>2</sup>Astrid Lindgren Children's Hospital, Karolinska University Hospital, Region Stockholm, Stockholm,  
6 Sweden

7 <sup>3</sup>Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

8 <sup>4</sup>Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA

9 <sup>5</sup>Seaver Center for Research and Treatment at Mount Sinai, New York, USA

10 <sup>6</sup>Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

11 <sup>7</sup>Department of Neonatology, Division of Neonatal Medicine, Karolinska University Hospital,  
12 Stockholm, Sweden

13 <sup>8</sup>Department of Bioclinical sciences, Linköping University, Linköping, Sweden

14

## 15 Abstract

16 **Background:** Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by  
17 diverse presentations and a strong genetic component. Environmental factors, such as prematurity,  
18 have also been linked to increased liability for ASD, though the interaction between genetic  
19 predisposition and prematurity remains unclear. This study aims to investigate the impact of genetic  
20 liability and preterm birth on ASD conditions.

21 **Methods:** We analyzed phenotype and genetic data from two large ASD cohorts, the Simons  
22 Foundation Powering Autism Research for Knowledge (SPARK) and Simons Simplex Collection (SSC),  
23 encompassing 78,559 individuals for phenotype analysis, 12,519 individuals with genome  
24 sequencing data, and 8,104 individuals with exome sequencing data. Statistical significance of  
25 differences in clinical measures were evaluated between individuals with different ASD and preterm  
26 status. We assessed the rare variants burden using generalized estimating equations (GEE) models  
27 and polygenic load using ASD-associated polygenic risk score (PRS). Furthermore, we developed a

**NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

28 machine learning model to predict ASD in preterm children using phenotype and genetic features  
29 available at birth.

30 **Results:** Individuals with both preterm birth and ASD exhibit more severe phenotypic outcomes  
31 despite similar levels of genetic liability for ASD across the term and preterm groups. Notable,  
32 preterm ASD individuals showed an elevated rate of de novo variants identified in exome sequencing  
33 (GEE model with Poisson family, p-value = 0.005) in comparison to the non-ASD preterm group.  
34 Additionally, a GEE model showed that a higher ASD PRS, preterm birth, and male sex were  
35 positively associated with a higher predicted probability for ASD, reaching a probability close to 90%.  
36 Lastly, we developed a machine learning model using phenotype and genetic features available at  
37 birth with limited predictive power (AUROC = 0.65).

38 **Conclusions:** Preterm birth may exacerbate the multimorbidity present in ASD, which was not due to  
39 the ASD genetic factors. However, increased genetic factors may elevate the likelihood of a preterm  
40 child being diagnosed with ASD. Additionally, a polygenic load of ASD-associated variants had an  
41 additive role with preterm birth in the predicted probability for ASD, especially for boys. We propose  
42 that incorporating genetic assessment into neonatal care could benefit early ASD identification and  
43 intervention for preterm infants.

44

45 **Keywords:** Prematurity, Autism Spectrum Disorder, Genetics, Polygenic risk score, Machine learning,  
46 Generalized estimating equations model.

47

48

## 49 **Introduction**

50

51 Autism Spectrum Disorder (ASD) is an early-onset neurodevelopmental condition characterized by  
52 challenges in social interaction, communication, and restrictive and repetitive behaviors and interests

53 [1]. In addition to these core symptoms, individuals with ASD have multiple co-occurring  
54 neurodevelopmental, psychiatric, and physical conditions, which contribute to clinical heterogeneity  
55 [2].

56  
57 The etiology of ASD is multifaceted and not yet fully elucidated [3,4]. However, genetic factors  
58 account for up to 80-90% of the liability for ASD [4–6]. Rare de novo variants (DNV), especially those  
59 affecting the gene function of constraint genes, are shown to be enriched in ASD [7,8]. Rare inherited  
60 variants in ASD-related genes are also shown to be overtransmitted from parents to their children  
61 with ASD [5,9]. In addition to rare variants, genome-wide association studies (GWAS) have identified  
62 a few common variants associated with ASD, and the polygenic load calculated using polygenic risk  
63 score (PRS) has demonstrated predictive ability for ASD and ASD traits [10,11]. Furthermore, ASD PRS  
64 can uniquely predict variability in cognitive performance [11].

65  
66 In addition to genetic factors, there are several environmental factors associated with ASD [3]. The  
67 most robustly associated environmental stressor is prematurity, with ASD likelihood in preterm about  
68 two to four folds higher than in term, and ASD likelihood increasing as gestational age at birth  
69 decreases [12,13]. Although preterm birth involves both genetic and environmental components  
70 [14], it is typically discussed as an environmental factor in ASD studies [3,15]. Preterm birth, defined  
71 as delivery with gestational age before 37 weeks, can be further categorized into four preterm sub-  
72 categories: extremely preterm (<28 weeks), very preterm (28-31 weeks), moderate preterm (32-33  
73 weeks), and late preterm (34-36 weeks). Prematurity is not only associated with ASD but also with  
74 neurocognitive development and other health outcomes [16–18]. Earlier studies investigating  
75 phenotypes in children with ASD suggested that extremely or very preterm ASD children have more  
76 language deficits and developmental delays compared to term ASD children [19,20]. However, others  
77 reported that no significant differences in development were found when studying the entire  
78 preterm group [21]. Moreover, preterm birth as an exposure is associated with various comorbidities

79 in ASD, including attention and behavioral problems, neurological disorders, and growth deficiency.  
80 However, more investigations are needed to understand ASD phenotypic spectrum in preterm and  
81 term birth as well as how different sub-groups of prematurity contribute to specific medical  
82 outcomes and traits.

83

84 Interestingly, preterm infants have been found to have increased DNV rates compared to term [22],  
85 and various de novo CNVs in preterm were found related to neurodevelopmental disorders genes  
86 (NDD genes) [23], but it remains uncertain whether DNV burden is further elevated when both ASD  
87 and preterm birth are present. Moreover, the relationship between ASD polygenic load and  
88 prematurity has only been evaluated by Cullen et al based on cognition, but no interaction was found  
89 between ASD PRS and gestational age at birth [15]. Furthermore, there are indications that a small  
90 fraction of preterm individuals would have recognizable genetic disorders [24]. However, there have  
91 not been specific studies focusing on genetic factors within preterm individuals and ASD.

92

93 While genetic and phenotypic studies on the population level are informative, these analyses may  
94 miss interactions and non-linear relationships within or between factors on an individual level. To  
95 address this complexity, machine learning (ML) has the potential to identify patterns in high-  
96 dimensional data that traditional statistical methods may overlook, aiding in prediction. To date, ML  
97 prediction models have emerged to predict ASD using different data sources, such as routine medical  
98 assessments and electronic records [25,26], genetic data [27], and integrative models [28]. However,  
99 none of the published ML models currently predict ASD in preterm children. In the existing ML  
100 models for ASD prediction, included features are typically collected when the child is at least 1-2  
101 years of age or older [25,29]. It remains unclear whether integrating phenotype and genetic  
102 information available at birth could enable earlier ASD identification in preterm infants.

103

104 In this study, we aimed to enhance our understanding of ASD in preterm children by analyzing both  
105 clinical and genetic data in two large ASD cohorts, the Simons Foundation Powering Autism Research  
106 for Knowledge (SPARK) [8,30], and the Simons Simplex Collection (SSC) [31]. Across individuals with  
107 different ASD and prematurity sub-groups, we first examined their phenotype severity through the  
108 prevalence and multimorbidity of other medical diagnoses. Thereafter, we assessed the burden of  
109 rare and common sequence-level variants. Finally, we built an ML model using both phenotype and  
110 genetic features that could be obtained at birth to predict ASD in preterm individuals.

111

112

## 113 **Methods**

114

### 115 **Study cohorts**

116

117 The Simons Foundation Powering Autism Research for Knowledge (SPARK) database, initiated by the  
118 Simons Foundation Autism Research Initiative (SFARI), recruited families in the USA with one or  
119 more children diagnosed with autism spectrum disorder (ASD) [30]. We utilized demographic and  
120 phenotype data from the SPARK collection version 9 with a release date 2022-12-12. We considered  
121 medical and psychiatric diagnosis history from the basic medical screen dataset, grouping specific  
122 diagnoses into nine diagnostic categories: behavior, development, mood, growth, birth, eating  
123 habits (Eat), neurological conditions (Neuro), visual and auditory impairments (Visaud), and sleep.  
124 This dataset includes 9196 individuals with ASD and born preterm with gestational age less than 36  
125 weeks (ASD-preterm), 65021 individuals with ASD and born term (ASD-term), and 2706 individuals  
126 without ASD and born preterm (non-ASD-preterm). We also stratified preterm individuals into four  
127 sub-groups based on gestational age: extremely preterm (<28 weeks), very preterm (28-31 weeks),  
128 moderate preterm (32-33 weeks), and late preterm (34-36 weeks). Additionally, we compared

129 quantitative measures using the Child Behavior Checklist (CBCL) t-score for 1 to 5 and 6 to 18 years  
130 of age, Developmental Coordination Disorder Questionnaire (DCDQ) final score, Repetitive Behavior  
131 Scale-Revised (RBS-R) total final score, Social Communication Questionnaire (SCQ) final score and  
132 Full-Scale Intelligence quotient (FSIQ) score. [Additional file 1: Tables S1 and S2](#) provide detailed  
133 specific diagnoses and quantitative measures descriptions.

134

135 For the genetic part, variant calling dataset SPARK genome sequencing (GS) version 1.1 was used,  
136 including 12519 individuals from 3394 families, with 315 ASD-preterm, 2788 ASD-term, and 155 non-  
137 ASD-preterm individuals. We also utilized earlier published DNV data from exome sequencing (ES) to  
138 calculate the event rate [8]. Among the 6444 ASD individuals with DNV information, 5747 were born  
139 full-term, and 697 were born preterm. Furthermore, DNV information was accessible for 210  
140 preterm children without ASD.

141

142 We also used the Simons Simplex Collection (SSC) cohort. SSC recruited more than 10,000 individuals  
143 from 2,000 families [31]. Due to the absence of preterm information for non-ASD individuals  
144 (siblings of ASD probands), we only conducted studies for ASD-preterm and ASD-term groups. After  
145 excluding individuals with unreliable gestational age, unknown ASD diagnosis, births occurring post-  
146 term (gestational age > 40 weeks), and missing outcomes information, we retained 1,637 probands  
147 diagnosed with ASD (157 preterm and 1479 term) for the diagnostic category analysis. Additionally,  
148 we analyzed available quantitative measures, including CBCL score, DCDQ score, SCQ score, and IQ  
149 score. Detailed descriptions of specific diagnoses and qualitative measures are provided in  
150 [Additional file 1: Tables S3 and S4](#). We incorporated the de novo variants (DNV) dataset from Ng et  
151 al in our analysis, encompassing 1450 ASD individuals after excluding post-term births [32].  
152 Additionally, the Polygenic Risk Score (PRS) dataset we used sourced from Weiner et al comprised  
153 1590 ASD individuals, excluding post-term births [33].

154

155 [De novo variant calling and analysis](#)

156

157 In SPARK, GS was conducted on the Illumina NovaSeq 6000 system. Variant calling was performed  
158 using GATK (version 3.5) with HaplotypeCaller, and all samples were jointly called by GLnexus  
159 (version 1.4.1). To find de novo variants (DNV) of children, we included all trios. For families with  
160 more than one child, each child forms a trio with their parents, resulting in multiple trios within the  
161 same family. There were 5,712 trios from 3,364 families. We used two tools to call the DNV from  
162 SPARK trios, and true DNV was selected when it was found in both tools. The DNV was called if it was  
163 labeled as “denovo” with allele balance (AB) in children higher than 0.25 in Slivar (version 0.2.8) and  
164 identified as high confidence DNV in GATK (version 4.1.4.1) [34,35]. For pseudo-autosomal regions  
165 on the sex chromosome, we separately considered the variant genotype as 1/0 in children. We did  
166 not find DNVs on chrY in pseudo-autosomal regions. Then, we did quality control to further filter the  
167 DNV by removing variants with  $GQ < 20$ ,  $DP < 10$ , gnomAD population allele frequencies  $> 0.001$ , and  
168 variants of either 10 A's or T's in a row. We filtered out DNVs on genomic centromeres and low  
169 complex regions. Then we removed DNV if 1) it can be found in other family's parents, 2) it can be  
170 found in only children but more than three families, and 3) it on positions having more than 3 multi-  
171 alleles. We identified 432,903 DNVs, including 16,155 exonic DNVs and 986 loss-of-function (LOF)  
172 variants. We filtered out 29 children with DNV counts beyond three times the standard deviation  
173 from the mean DNV count. Based on these criteria, the average number of rare DNVs per child was  
174 75.9.

175

176 We annotated DNVs by ANNOVAR and SnpEff [36,37]. DNVs with Sequence Ontology (SO) terms as  
177 "frameshift", "splice\_acceptor", "splice\_donor", "start\_lost", "stop\_gained", and "stop\_lost" in gene  
178 effect were considered as LOF DNV. Additionally, we identified variants on the neurodevelopmental  
179 disorder-related genes (NDD gene) using high-confidence ASD genes collected by the Simons  
180 Foundation Autism Research Initiative (SFARI) (2024-01-16 release) with gene scores of 1 or 2 and

181 labeled as syndromic and green gene list of Intellectual disability - microarray and sequencing  
182 (Version 5.497) on Genomics England PanelApp (2024-03-14 accessed) [38,39].

183

#### 184 [Inherited variants calling](#)

185

186 We extracted variants on the genomic protein coding NDD genes (as the NDD gene list we used for  
187 DNV analysis) using VCFtools/0.1.16 [40]. Then we annotated the inheritance mode of NDD genes  
188 using the ID gene panel app, SysNDD database (v0.1.0) and DDgenes [39,41,42]. There were 80%  
189 (1625 genes) of NDD genes annotated, including 738 dominant genes coded as monoallelic or  
190 dominant, and 944 recessive genes coded as biallelic or recessive in databases. To restrict the  
191 analysis to rare inherited variants, we used the allele frequency filter threshold of 0.001 and 0.01 for  
192 dominant and recessive genes, respectively. Variants with GQ<20, DP<10, and genotypes conflicting  
193 with the inheritance mode of located genes were filtered out. We did not find compound variants  
194 (more than one heterozygous variant on the same recessive gene for one child). For variants on  
195 dominant genes, we identified LOF following the same process in the DNV part and found damage  
196 missenses variants met at least one of the following conditions: CADD $\geq$ 20, SIFT labeled as D,  
197 POLYPHEN labeled as P and D, PHYLOP $\geq$ 2.0 or REVEL $\geq$ 0.5. From 5,712 trios in the SPARK GS  
198 database, we identified 245,671 and 4,346 inherited variants in dominant and recessive NDD genes,  
199 respectively. We identified 2,717 LOF and 39,136 damaging missense variants in genes with a  
200 dominant inheritance mode.

201

#### 202 [Polygenic risk score](#)

203

204 Based on GS jointly called variants from 12,519 SPARK participants, we performed quality control for  
205 individuals and variants using PLINK1.9 with parameters listed in [Additional file 1: Table S5](#), retaining



206 11,933 individuals and 9,558,997 variants with a total genotyping rate of 99.9% [43]. Then, genome  
207 coordinates of variants were converted from hg38 to hg19 using liftOver (Version 2017-03-14) [44],  
208 and 9,428,216 were mapped after removing duplicated variants. We calculated the posterior SNP  
209 effect size estimates using PRS-CS with ASD GWAS summary statistics from the Psychiatric Genomics  
210 Consortium (November 2017 release) (46351 individuals) and European LD reference data from  
211 1000 Genome phase III [10,45]. The default parameters used in PRS-CS also be listed in [Additional](#)  
212 [file 1: Table S5](#). The final PRS was calculated using the score function in PLINK1.9 with the estimated  
213 posterior SNP effect size. To minimize the effects on different populations, we analyzed the ancestry  
214 of individuals using Principal component analysis (PCA) by `pca` command in PLINK1.9. The ten  
215 principal components (PC1-10) were included as covariates when we calculated the association  
216 between phenotype and PRS.

217

## 218 [Statistical analysis](#)

219

220 All analyses were performed using R programming language (version 4.2.2). Commencing with an  
221 exploration of phenotypes, we investigated two pairs of groups: preterm and term birth within  
222 individuals diagnosed with ASD (ASD-preterm versus ASD-term); and individuals diagnosed with ASD  
223 to those without ASD within preterm birth (ASD-preterm versus non-ASD-preterm). An individual  
224 was considered to possess the diagnosis feature if any specific diagnosis within that diagnostic  
225 category was exhibited, no matter how many specific diagnoses there were at the same time.

226

227 In phenotypic analysis, the prevalence is reported by the frequency of individuals with the diagnosis.  
228 We examined the differences in prevalence between ASD-preterm and ASD-term, ASD-preterm and  
229 non-ASD-preterm using odds ratios with 95% confidence interval (CI) and statistical significance  
230 reported by FDR-adjusted p-values in  $\chi^2$  test. After stratifying preterm stages, we used  $\chi^2$  test to  
231 evaluate differences across preterm stages, post-hoc comparisons for each pair of preterm stages,

232 and Kendall's tau test to examine rank correlation of preterm stages with prevalences. Then,  
233 prevalences of multimorbidity (one, two, three, four, or not less than five diagnoses) were estimated  
234 by ASD and preterm status, and differences were examined by  $\chi^2$ . Additionally, we examined  
235 difference of quantitative measures in pairwise using the 2-sided Wilcoxon rank sum test and in  
236 multiple comparisons using the Kruskal Wallis rank sum test. To account for multiple testing, we  
237 applied false discovery rate correction to p-values.

238 The burden of DNV and inherited variants was evaluated by comparing rates of such variants in each  
239 subgroup categorized by ASD and preterm status. We assessed the statistical differences between  
240 groups through a generalized estimating equation (GEE) model with Poisson family and sex as a  
241 covariate. GEE model is more robust to assumptions of data following a particular data distribution  
242 and adjusts for correlations between individuals, e.g. siblings and families [46]. PRSs were z-  
243 standardized and statistical significance for PRS distribution was reported by the 2-sided Wilcoxon  
244 rank sum test. The association between targeted phenotype (y/n) and PRS in each subgroup was also  
245 evaluated in GEE logistic model with sex(m/f) and PC1-10 from ancestry checking as covariates.

246 To examine the associations between ASD diagnosis and possible variables, we modeled the  
247 probability of ASD (y/n) by fitting GEE logistic model(s) with the equation as [ASD (y/n) ~ sex (m/f) +  
248 preterm (y/n) + standardized PRS] in European population to find the correlation between ASD  
249 diagnosis and possible variables. To visualize the predicted probabilities of ASD from the GEE logistic  
250 model, we utilized `ggemmeans` function in `ggeffects` R package (version 1.5.1) [47], showing the  
251 average predicted probabilities of the ASD for specific levels of variables adjusted for other  
252 covariates in the model. After that, the variable (preterm (y/n) \* standardized PRS) was added to the  
253 GEE logistic model to check the correlation between ASD (y/n) and the interaction of preterm status  
254 and PRS. To detect the association between multimorbidity and DNV burden, GEE models (Table S4)  
255 and linear regression were used.

256

## 257 Machine learning model

258

259 This part of the analysis was performed in R (version 4.2.2). Within preterm individuals, we utilized  
260 non-ASD and ASD diagnoses as two classification flags, incorporating features obtained from data  
261 that can be collected at birth. For phenotypical variables, we included birth complications, sex, and  
262 birth-related conditions. Genetic variables encompassed the count of several types of genetic  
263 variants, the CADD score of de novo variants (DNV), and the PRS of ASD. To remove redundant  
264 variables and identify informative features, we employed the Recursive Feature Elimination (RFE)  
265 algorithm [48]. RFE is a feature selection technique that iteratively removes the least important  
266 features based on model performance, refitting the model with the remaining features until the  
267 optimal subset of features is identified. In RFE, we utilized random forest function and 10-fold cross-  
268 validation in the underlying model to assess feature importance throughout the process. For the  
269 features selected after RFE, we only retained the more general feature (e.g., retaining LOF over LOF  
270 on NDD genes) in any pair of features with a correlation coefficient above 0.7 to reduce  
271 multicollinearity. The details of selected features are listed in [Additional file 1: Table S6](#).

272

273 We applied R package caret (version 6.0-94) [49] to train the ML models. Given the limited sample  
274 size and the higher proportion of ASD samples than non-ASD, we conducted nested cross-validation  
275 (NCV) and hyperparameter tuning with grid search to enhance model performance. In NCV, we  
276 partitioned the data into 10 folds in the outer loop, with nine folds used for training and the  
277 remaining fold for testing. Within the inner loop, we performed repeated 5\*5-fold cross-validation,  
278 and the model with the best performance was applied to the outer loop. We employed three  
279 algorithms—Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Linear Support Vector  
280 Machine (SVM)—to construct the models. The values of hyperparameter tuning for each model are  
281 detailed in [Additional file 1: Table S7](#). We reported evaluation metrics, including accuracy with a 95%  
282 CI, area under the receiver operating characteristic curve (AUROC), specificity, sensitivity, and F1-

283 score. Moreover, the SHapley Additive exPlanations (SHAP) values for features were computed and  
284 visualized using the R package SHAPforxgboost (version 0.1.3) [50], quantifying the contribution of  
285 each feature to individual model predictions in terms of direction and magnitude.

286

287

## 288 **Result**

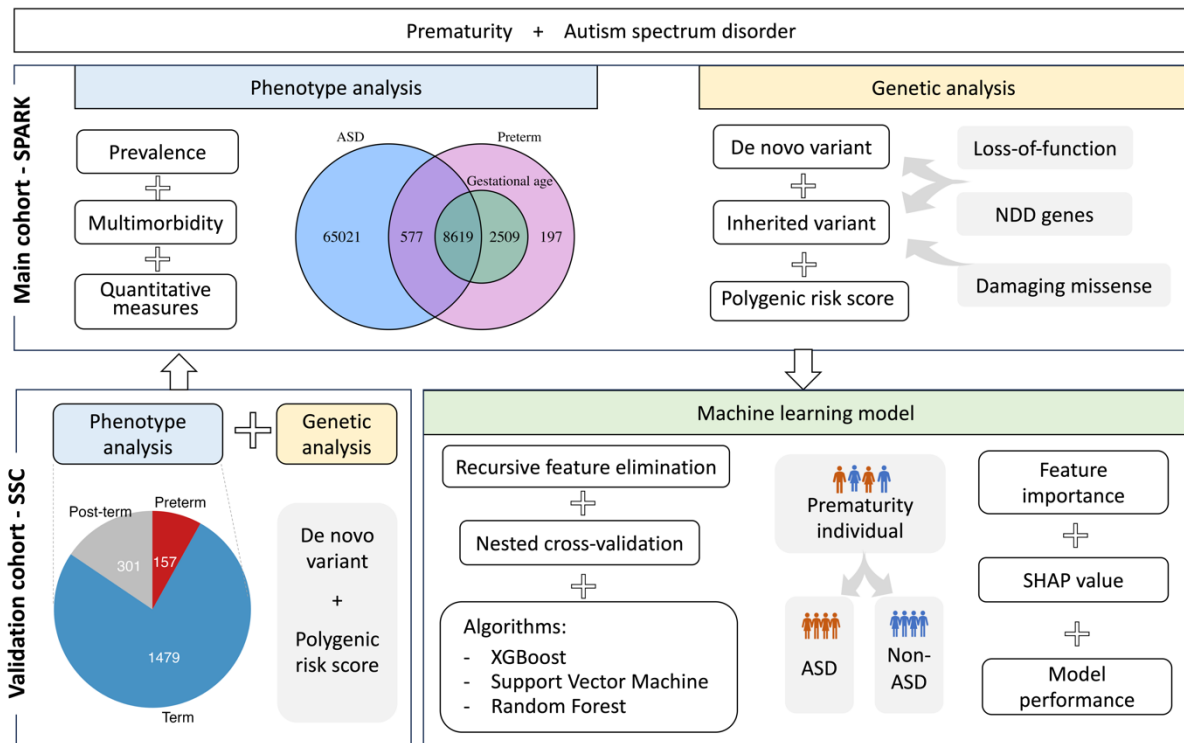
289

### 290 **Phenotype comparison across ASD and prematurity**

291

292 We utilized basic medical screening data from 181,248 individuals in the SPARK version 9 cohort  
293 (release date 2022-12-12). Among them, 74,217 (41%) were diagnosed with ASD, and 11,902 (7%)  
294 were born preterm. When performing phenotype comparison, we grouped individuals with 9,196  
295 individuals with ASD and being preterm (ASD-preterm), 65,021 individuals with ASD but being term  
296 (ASD-term), and 2,706 preterm individuals without ASD diagnosis (non-ASD-preterm) (Figure 1, Table  
297 1). In the SSC cohort, gestational age records were available only for probands with ASD, of which  
298 157 were preterm and 1,479 were term. We stratified the preterm stage based on gestational age at  
299 birth (Table 1).

300



301

302

303

304

305

306

307

308

309

310

311

312

313

**Figure 1. Overview design of the study.** Firstly, we performed phenotype analysis on diagnosis prevalence, burden of multimorbidity and quantitative measures in the SPARK cohort. The sample size of SPARK is shown in the Venn diagram, with blue indicating ASD, pink indicating preterm with unknown gestational age, and green indicating preterm with known gestational age. Secondly, we analyzed the de novo variant and inherited variant burden, separately, focusing on loss-of-function variants and damaging missense and if these affected neurodevelopmental disorder (NDD) genes. Additionally, we utilized polygenic risk scores for common variants associated with ASD. For validation, we applied similar analyses in the SSC cohort. Thirdly, we integrated phenotype and genomic data to train the machine learning models with different algorithms to predict ASD diagnosis in the preterm group. Shapley additive explanations (SHAP) values assess the effect of each feature on the model performance.

314

**Table 1. Characteristics of the analyzed samples.**

	Cohort	SPARK (version 9)		SSC
		ASD	Non-ASD	ASD
<b>Phenotypic analysis</b>	Number of samples	74217	/	1636
	Preterm	9196	2706	157
	Extremely preterm (GA <28 weeks)	699 (8%)	109 (4%)	/
	Very preterm (GA 28-31weeks)	987 (11%)	274 (10%)	2 (1%)
	Moderate preterm (GA 21-33 weeks)	1274 (14%)	372 (14%)	11 (7%)
	Late preterm (GA 34-36 weeks)	5659 (62%)	1754 (65%)	144 (92%)

		Unknown GA	577 (6%)	197 (7%)	/
		Male:Female	6860:2336	1358:1348	139:18
		Term	65021	/	1479
		Male:Female	48069:16952	/	1270:209
<b>Genetic analysis</b>	De novo variant (WGS)	Preterm	309	164	/
		Male:Female	257:52	90:74	/
		Term	2728	/	/
		Male:Female	2172:556	/	/
	De novo variant (WES)	Preterm	697	210	137
		Male:Female	563:134	119:91	121:16
		Term	5747	/	1313
		Male:Female	4557:1190	/	1129:184
	Inherited variant	Preterm	310	165	/
		Male:Female	258:52	90:75	/
		Term	2742	/	/
		Male:Female	2182:560	/	/
PRS	Preterm		305	161	155
		Male:Female	252:53	86:75	137:18
	Term		2702	6472	1435
		Male:Female	2134:568	2817:3655	1236:199
<b>Machine learning</b>		Preterm	279	150	/
		Male:Female	230:49	81:69	/

315 / Mark as data were not analyzed. GA = Gestational age at birth.

316

317 We performed analysis on nine available diagnostic categories recorded in the basic medical

318 screening dataset in SPARK phenotype database version 9, involving behavior, development, mood,

319 growth, birth, eating habits (eat), neurological conditions (neuro), visual and auditory impairments

320 (visual), and sleep (details of diagnostic categories are described in [Additional file 1: Table S1](#)). For

321 each category, we assigned a binary variable indicating the presence or absence of conditions within

322 that category, rather than counting the number of specific diagnoses. The prevalence of all diagnostic

323 categories analyzed was higher in ASD-preterm compared to ASD-term ([Figure 2A](#)). Specifically, ASD-

324 preterm had higher odds ratio (OR) for all diagnostic categories, with the highest being for birth and

325 growth diagnoses (OR=2.18 and 2.18,  $\chi^2$  tests with False Discovery Rate [FDR]-adjusted p-

326 value= $5.6 \times 10^{-55}$  and  $9.2 \times 10^{-158}$ , respectively). Additionally, preterm birth was associated with a

327 modestly increased likelihood of other behavioral diagnoses compared to term in the ASD group

328 (OR=1.2, p-value= $5 \times 10^{-15}$ ). We also observed significantly different prevalences of diagnostic  
329 categories ( $\chi^2$  test with FDR-adjusted p value < 0.001 for all categories) when we considered different  
330 sub-groups of preterm birth ([Additional file 1: Figure S1A](#)). Furthermore, we identified linear trends  
331 across different preterm stages, with groups of lower gestational age showing a higher prevalence in  
332 growth, eating, neuro, and visual diagnostic categories (Kendall's tau test, FDR-adjusted p-  
333 value=0.04). Almost all the preterm sub-groups had a higher prevalence of diagnostic categories  
334 compared to the term stage (FDR-adjusted p-values of the post-hoc comparisons of  $\chi^2$  test are in  
335 [Additional file 1: Table S8](#))

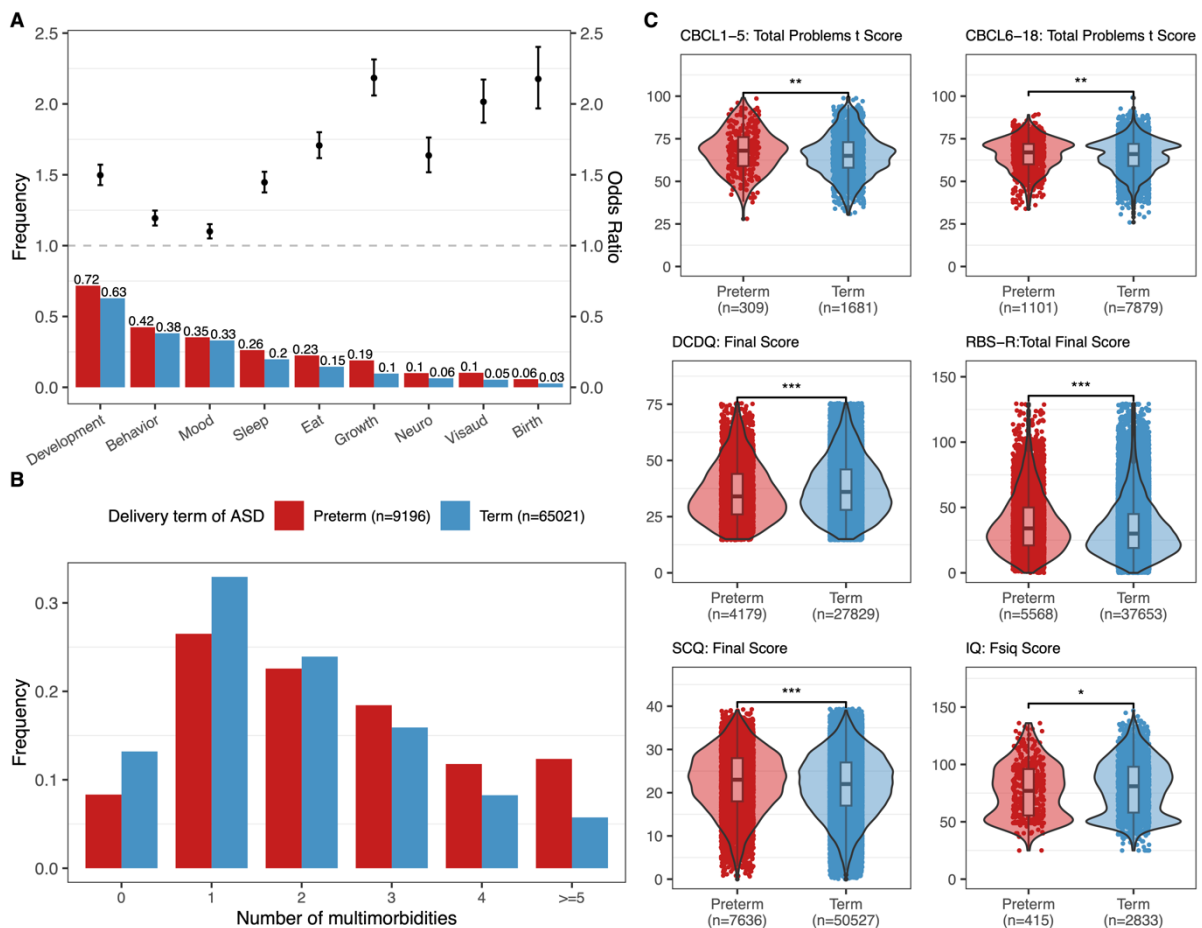
336

337 Then, we analyzed multimorbidity, indicated as the number of concurrent diagnoses among the ASD  
338 individuals, revealing that ASD-preterm exhibited a higher likelihood of the higher number of  
339 concurrent morbidities compared to ASD-term ([Figure 2B](#),  $\chi^2$  test with p-value <  $2.2 \times 10^{-16}$ ). In ASD,  
340 preterm sub-groups showed differences in the burden of multimorbidity ([Additional file 1: Figure](#)  
341 [S1B](#),  $\chi^2$  test with p-value <  $2.2 \times 10^{-16}$ ), in which extremely and moderate preterm subgroups exhibited a  
342 significantly higher burden of multimorbidity with  $\geq 5$  diagnoses compared to late preterm  
343 ([Additional file 1: Table S9](#), post-hoc test of  $\chi^2$  test with FDR-adjusted p-values =  $1.6 \times 10^{-5}$ , 0.02  
344 respectively). For diagnostic categories with more than two specific diagnoses, we found a positive  
345 linear correlation between the number of specific diagnoses, i.e., multimorbidity, and the odds ratio  
346 of preterm versus term, except for birth-related issues ([Additional file 1: Figure S1C](#)). This indicated  
347 that preterm ASD individuals tend to have more specific diagnoses within categories as well as across  
348 categories than term ASD individuals.

349

350 Next, we analyzed quantitative measures for overall behavioral challenges and specific symptom  
351 domains. We observed significant differences between ASD-preterm and ASD-term (2-sided  
352 Wilcoxon rank sum test with FDR adjustment), although the large sample size may amplify the  
353 differences ([Figure 2C](#)). ASD-preterm had increased severity of behavioral challenges (CBCL score for

354 1-5y, p-value= $1.4 \times 10^{-3}$ ; for 6-18y, p-value=0.0043), developmental coordination disorder (DCDQ final  
 355 score, p-value= $2.5 \times 10^{-21}$ ), repetitive behaviors (RBS-R score, p-value= $6.7 \times 10^{-34}$ ), and social  
 356 communication skills (SCQ score, p-value= $1.4 \times 10^{-20}$ ), as well as lower IQ scores (p-value=0.02).  
 357 Comparing different sub-groups of preterm birth, we found that extremely preterm has the lowest  
 358 DCDQ final score compared to other stages (2-sided Wilcoxon rank sum test, FDR-adjusted p-values  
 359 are 0.003, 0.002, and  $2.7 \times 10^{-5}$  when compared to very preterm, moderate preterm, and late  
 360 preterm, respectively) (Additional file 1: Figure S2).  
 361



362  
 363 **Figure 2. The phenotype comparison between preterm and term with ASD in the SPARK version 9**  
 364 **cohort.** Color bars are the same across three panels and shown at the top of panel B. A. Prevalence  
 365 and odds ratio with 95% confidence interval (CI) of diagnosis. The exact prevalence values are labeled  
 366 on the top of the bars. ORs are given among ASD individuals born preterm vs term. B. Distribution of  
 367 the number of multimorbidity. C. Differences in Child Behavior Checklist (CBCL) t-score for 1 to 5 and  
 368 6 to 18 years of age, Developmental Coordination Disorder Questionnaire (DCDQ), Repetitive  
 369 Behavior Scale-Revised (RBS-R) score, Social Communication Questionnaire (SCQ) and Full-Scale IQ  
 370 (Fsiq) among ASD individuals born preterm and term. Significance was assessed using the 2-sided



371 Wilcoxon rank sum test with the FDR-adjusted p-value marked in the plots as 0-0.001\*\*\*, 0.001-  
372 0.01\*\*, 0.01-0.05\* or NS (non-significant difference).  
373

374 To complement our analyses within the ASD individuals, we analyzed if there were any differences  
375 within preterm birth for the same phenotype measures. The ASD-preterm had more severe  
376 outcomes in comparison to non-ASD-preterm with increased severity with lower gestational age  
377 ([Additional file 1: Figure S3A-B, S4](#)). The developmental diagnostic category had the highest  
378 prevalence (72%) in the ASD group resulting in 8.8 OR (95% CI 7.9-9.7) when compared to non-ASD.  
379 For quantitative measures, the ASD group had statistically significantly higher SCQ final scores  
380 compared to the non-ASD group (2-sided Wilcoxon rank sum test,  $p\text{-value} < 2.2 \times 10^{-16}$ ) ([Additional file](#)  
381 [1: Figure S3C](#)).

382  
383 ASD-preterm and term comparisons within the SSC cohort also showed a statistically significantly  
384 higher prevalence of eating problems ( $\chi^2$  test with FDR-adjusted  $p\text{-value} = 1.4 \times 10^{-6}$ ) and a similar trend  
385 towards having more multimorbidity compared to ASD-term ([Additional file 1: Figure S5A-B](#),  $\chi^2$  test  $p\text{-}$   
386  $\text{value} = 0.045$ ). No statistically significant differences were observed in the quantitative measures  
387 ([Additional file 1: Figure S5C](#)).

388

### 389 [Genetic variants comparison across ASD and prematurity](#)

390

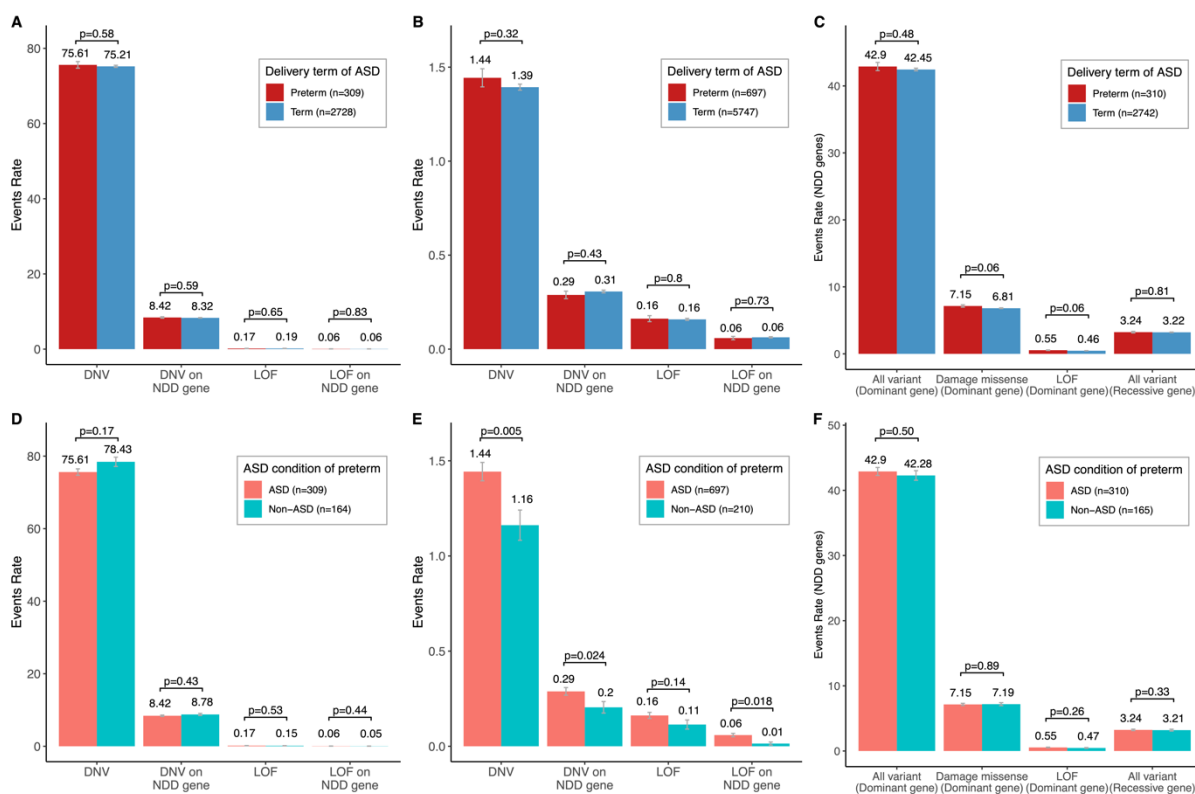
391 To investigate the burden of de novo variants, we analyzed available genome sequencing (GS) and  
392 exome sequencing (ES) data from SPARK and SSC. The population analyzed for GS included 310 ASD-  
393 preterm, 2,742 ASD-term, and 165 non-ASD-preterm individuals. The ES dataset contained 697 ASD-  
394 preterm, 5,747 ASD-term, and 210 non-ASD-preterm individuals. We did not observe any significant  
395 difference in DNV event rate or distribution of DNV numbers between ASD-preterm and ASD-term  
396 ([Figure 3A, Additional file 1: Figure S6A](#)), or between ASD-preterm and non-ASD-preterm derived

397 from GS ([Figure 3D, Additional file 1: Figure S6B](#)), even when analyzing only the exonic region  
398 ([Additional file 1: Figure S7A-B](#)). Similarly, no statistically significant differences were found  
399 comparing de novo burden from GS in 137 ASD-preterm and 1313 ASD-term in the SSC ([Additional](#)  
400 [file 1: Figure S7C](#)).

401  
402 When analyzing DNV event rates obtained from ES data [8] from SPARK, no statistically significant  
403 differences were found between ASD-preterm and ASD-term ([Figure 3B, Additional file 1: Figure](#)  
404 [S6C](#)). However, ASD-preterm individuals had more exonic DNVs ( $p$ -value=0.005), exonic DNVs on NDD  
405 genes ( $p$ -value=0.024), and LOF affecting NDD genes ( $p$ -value=0.018) than non-ASD-preterm ([Figure](#)  
406 [3E, Additional file 1: Figure S6D](#)). We stratified the ASD-preterm and non-ASD-preterm by gestational  
407 age preterm subgroups and observed a similar trend in both the moderate and late preterm groups.  
408 However, due to the limited sample size, a statistically significant difference between ASD-preterm  
409 and non-ASD-preterm was only detected in the event rate of DNVs in NDD genes in the moderate  
410 preterm group ( $p$ -value=0.02) and in both overall DNV and DNV in NDD genes in the late preterm  
411 group ( $p$ -value=0.002 and 0.04 respectively). Interestingly, for extremely to very preterm stages, the  
412 event rates of DNV were numerically lower in ASD compared to non-ASD individuals, even though  
413 this difference was not statistically significant ([Additional file 1: Figure S8](#)).

414  
415 We also investigated the rates of inherited variants, focusing on those affecting NDD genes and  
416 protein-coding regions. From 4,974 individuals with phenotype information, we did not observe  
417 statistically significant differences between ASD-preterm and ASD-term nor between ASD-preterm  
418 and non-ASD-preterm, although ASD-preterm tend to have a numerically higher rate of rare  
419 inherited variants ([Figure 3C, 3F, Additional file 1: Figure S6E, S6F](#)).

420



421  
 422 **Figure 3. Association between genetic variant burden and subgroups with varying preterm birth**  
 423 **and ASD status in the SPARK cohort.** In ASD individuals, event rates of de novo variants (DNV)  
 424 identified through genome sequencing (A) and exome sequencing (B), and inherited variants on  
 425 dominant and recessive NDD genes identified through genome sequencing (C) were calculated. In  
 426 preterm individuals, event rates of de novo variants identified through genome sequencing (D) and  
 427 exome sequencing (E), and inherited variants in dominant and recessive NDD genes identified  
 428 through genome sequencing (F) were calculated. Data are presented as mean values  $\pm$  standard  
 429 errors as error bars. The GEE model with Poisson family and sex covariate was used to compute the  
 430 p-value to assess the differences in DNV count between groups.

431  
 432 To further test whether the multimorbidity would be a modifying factor for the differences in the  
 433 DNV burden, we computed GEE models (Additional file 1: Table S10) and found that multimorbidity  
 434 is positively correlated with GS LOF (p-value=0.037), GS LOF on NDD genes (p-value= $5.3 \times 10^{-6}$ ) and all  
 435 types of ES DNV burden (p-value= $1.1 \times 10^{-5}$ ,  $1.9 \times 10^{-9}$ ,  $6.6 \times 10^{-14}$  and  $< 2.2 \times 10^{-6}$  for DNV, LOF, DNV on  
 436 NDD genes and LOF on NDD genes respectively) across all individuals. Stratified by preterm and ASD  
 437 status, we observed this positive correlation pattern in ASD-term group for ES DNV (p-value=0.009),  
 438 ES DNV on NDD genes (p-value=0.013) and LOF on NDD genes (p-value=0.004), as well as in ASD-  
 439 preterm group for ES LOF on NDD genes (p-value=0.013) (Additional file 1: Figure S9). However,

440 except for GS DNV on NDD genes, there is no interaction between DNV burden and multimorbidity  
441 performing on ASD or preterm outcomes.

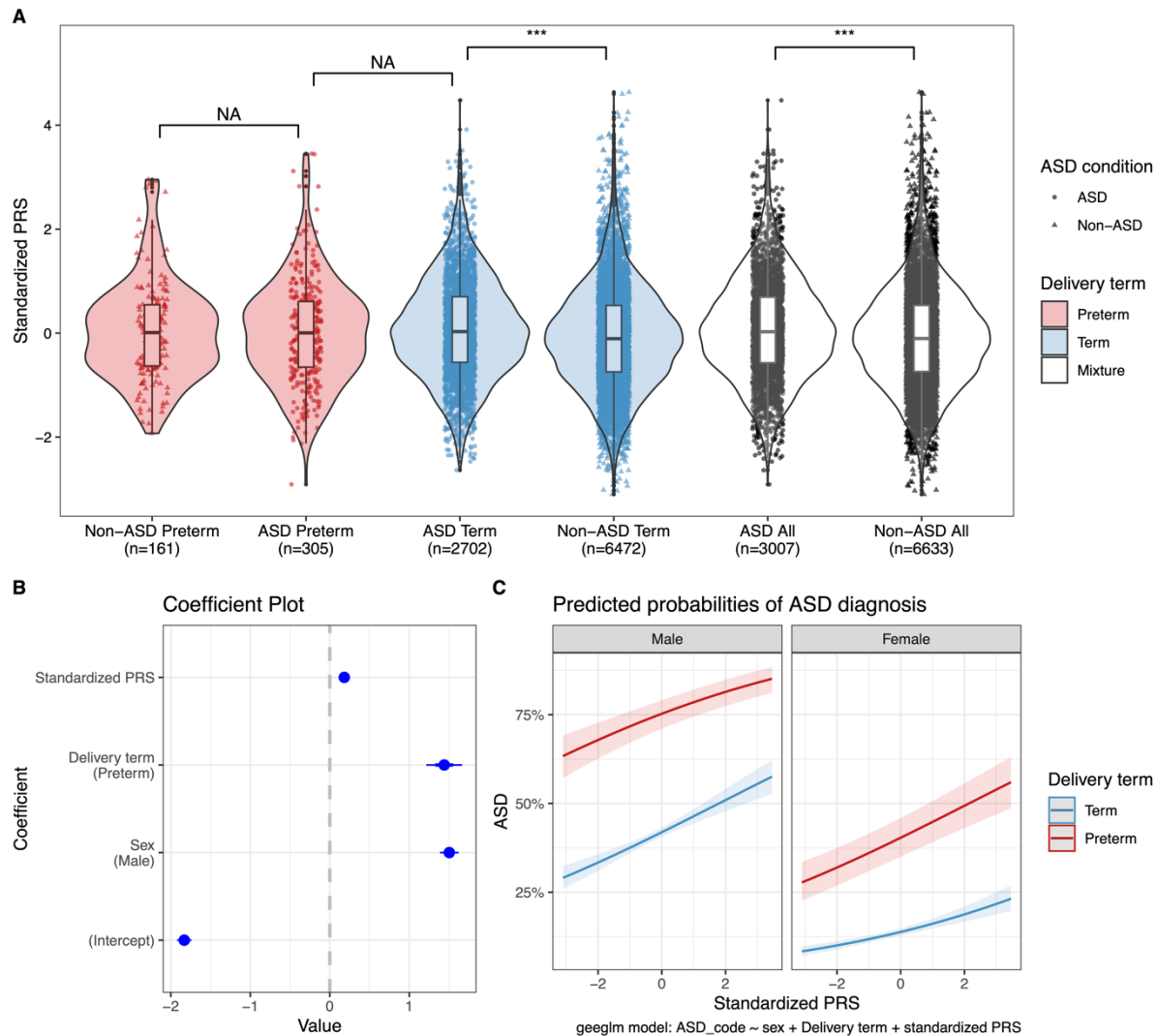
442

#### 443 [ASD Polygenic risk score and association with preterm status](#)

444

445 We calculated ASD PRS for individuals in the SPARK cohort using the most comprehensive GWAS on  
446 ASD as source data [10]. There was no significant difference in the distribution of PRS between the  
447 ASD-preterm and ASD-term groups nor between ASD-preterm and non-ASD-preterm ([Figure 4A](#)). As  
448 expected, ASD individuals had higher PRS compared to non-ASD individuals in the whole cohort  
449 displaying the usability of the PRS (2-sided Wilcoxon rank sum test,  $p\text{-value}=6.7\times 10^{-13}$ ) ([Figure 4A](#)).  
450 Additionally, after adjusting for sex and population ancestry (as indicated by principal components  
451 [PC]) in a GEE logistic model, we confirmed that there was no independent association between  
452 preterm birth and PRS (in ASD population), or between ASD diagnosis and PRS (in preterm  
453 population). The statistically non-significant association between preterm birth and PRS (in ASD  
454 population) was replicated in the SSC cohort ([Additional file 1: Figure S10](#)). Furthermore, we  
455 computed a full GEE logistic model for ASD diagnosis within European populations, showing that  
456 male sex, preterm birth, and higher PRS were all positively associated with ASD diagnosis ( $p\text{-}$   
457  $\text{value}<2\times 10^{-16}$ ,  $<2\times 10^{-16}$  and  $2.2\times 10^{-12}$ , respectively) ([Figure 4B](#)). In this model, the predicted  
458 probability of an ASD diagnosis was almost 90% for preterm-born males, with the highest PRS ([Figure](#)  
459 [4C](#)). Then, we included the interaction between preterm status and PRS in the full GEE logistic model,  
460 observing a significant association between this interaction and ASD diagnosis ( $p = 0.017$ ).

461



462

463 **Figure 4. The association of polygenic risk score (PRS) with delivery term and ASD condition in**  
 464 **SPARK cohort.** A. The distribution of standardized PRS in groups with different delivery terms and  
 465 ASD diagnosis. Statistical significance was assessed using the 2-sided Wilcoxon rank sum test with the  
 466 p-value marked in the plots as 0-0.001\*\*\* or NS (statistically non-significant difference). B.  
 467 Coefficient plot for the GEE logistic model [ASD(y/n) ~ sex(m/f) + Preterm status(y/n) + Standardized  
 468 PRS], displaying the estimated coefficients for each variable. Positive coefficients suggest an increase  
 469 in the likelihood of ASD associated with the variable, while negative coefficients indicate a  
 470 decrease. Error bars represent 95% CI. C. Visualized effect plot of GEE model, which shows average  
 471 predicted probabilities of ASD diagnosis for specific levels of variables, with color region around the  
 472 line showing 95% CI.

473

474 [Predictive model for ASD within preterm births](#)

475

476 Lastly, we investigated the potential of ML models to identify those preterm infants with a high

477 likelihood of ASD from information present at birth by combining clinical and available genetic data.

478 The model was developed and tested using a study population with preterm individuals classified  
479 into ASD (n=279) and non-ASD (n=150). For features used in prediction model, we also considered  
480 Combined Annotation Dependent Depletion (CADD) scores which assess the potential impact (i.e.  
481 deleterious or benign) of genetic variants on the function of genes and available for most of the  
482 DNVs. We applied Recursive Feature Elimination (RFE) and a correlation threshold of 0.7 to select 13  
483 features, including clinical features (sex, condition of birth complications, gestational age, insufficient  
484 oxygen at birth) and genomic features (number of several types of variants, CADD scores and  
485 standardized ASD-PRS) ([Additional file 1: Figure S11A, Table S6](#)). We used three algorithms to train  
486 the models ([Table 2](#)), of which the XGBoost model exhibited the highest area under the receiver  
487 operating characteristic curve (AUROC), at 0.65. The model accurately identified 69% (95% CI 0.644-  
488 0.733) ASD diagnosis in the preterm, with a sensitivity of 0.81, specificity of 0.47, and F1-score of  
489 0.77.

490  
491 The first three XGBoost models within the 10-fold training ([Additional file 1: Figure S11B](#)) were  
492 selected to visualize feature effects for this best-performing model. The feature importance varied  
493 slightly across the training XGBoost models with sex, PRS, and CADD score being the most important  
494 features ([Additional file 1: Figure S11C, S11E, S11G](#)). Using SHAP values to characterize the impact of  
495 each feature on the model's output for specific individuals, we found that sex had the highest  
496 significant impact on the model's predictions, whereas being male had a positive impact on the  
497 model's prediction. Furthermore, we demonstrate that lower gestational age, more autosomal  
498 exonic DNVs, more dominant inherited variants, more LOF variants, the presence of birth  
499 complications, and insufficient oxygen at birth drove the model towards ASD prediction result  
500 ([Additional file 1: Figure S11D, S11F, S11H](#)).

501

502 **Table 2. Performance metrics of machine learning model used to predict ASD diagnosis in preterm**  
503 **individuals.**

Algorithm	Accuracy	95% CI	AUC	Sensitivity	Specificity	F1-score
XGBoost	0.69	(0.64, 0.73)	0.65	0.81	0.47	0.77
Random forest	0.67	(0.63, 0.72)	0.63	0.86	0.33	0.77
SVM	0.70	(0.66, 0.75)	0.62	0.84	0.46	0.79

504

505

## 506 Discussion

507

508 Here, we conducted a comprehensive analysis of phenotypic differences using larger cohorts, as well  
509 as genotypic differences which have been explored in only a few studies among preterm and term-  
510 birth ASD individuals. We conclude that preterm-born ASD individuals have more diagnoses across  
511 different categories and a number of co-occurring diagnoses but similar genetic landscapes when  
512 investigating sequence-level rare DNVs and inherited variants as well as a polygenic load for ASD  
513 compared with ASD-term. Our analysis of preterm individuals with and without ASD showed similar  
514 results for the phenotype comparisons but inconsistent findings for the genetic burden. The largest  
515 de novo dataset derived from ES showed that the ASD-preterm had a higher exonic DNV event rate  
516 than the non-ASD-preterm; however, we did not validate this finding in the de novo dataset from  
517 GS. Additionally, the male with preterm status and higher polygenic load faces a higher likelihood of  
518 ASD when considering these features together. Furthermore, our ML model demonstrated potential  
519 for predicting ASD diagnosis in preterm children by integrating phenotype and genetic information.  
520 Our results provide evidence that genetic factors play a role in emerging ASD in preterm birth, but  
521 the environmental stressor of being preterm most likely contributes to the severity and  
522 multimorbidity.

523

524 Previous research has reported numerous but inconsistent findings regarding phenotypic disparities  
525 between ASD preterm and term individuals [19,21], while limited research has focused on

526 investigating the genetic link of ASD in prematurity. Unlike most phenotypic comparisons that  
527 concentrate on specific diagnostic outcomes [20,21], we first grouped the various conditions into  
528 nine broader diagnostic categories. Our results indicate that children with both preterm birth and  
529 ASD exhibit a higher prevalence of diagnoses within these categories and a higher rate of  
530 multimorbidity across different diagnostic categories. Previous studies have found that both preterm  
531 birth and ASD are associated with adverse symptoms. For instance, preterm infants are  
532 independently inherently prone to multimorbidity and severe health complications affecting  
533 multiple organs and systems [16,51,52], such as visual and auditory impairments [53], epilepsy [54],  
534 ADHD [17], and other psychiatric disorders [18]. This supports the hypothesis that environmental  
535 liability factors like preterm may influence some of the heterogeneity and higher comorbidity rates  
536 observed in ASD [55].

537

538 After stratifying preterm based on gestational age, we observed that those born with lower  
539 gestational age tend to have more severe outcomes, which is in line with the dose-effect reported in  
540 prematurity, where the likelihood of developmental issues increases with decreasing gestational age  
541 [52]. This effect is also reflected in the potentially increasing complexity of multimorbidity among  
542 groups with lower gestational age [16]. Additionally, we showed significantly more severe symptom  
543 levels, as measured by different standardized questionnaires and cognitive tests, in ASD-preterm,  
544 consistent with previous studies as well as general research comparing preterm and term birth  
545 [17,56–58]. It is important to note that with large sample sizes, even very small differences can  
546 become statistically significant. Therefore, the results of the quantitative measures should be  
547 interpreted with caution.

548

549 In idiopathic ASD, heritability is estimated to be approximately 80% [6], but in preterm born,  
550 environmental factors account for 60% of the variation in gestational age [59]. Our findings suggest  
551 that genetic factors underly, at least partly, the ASD diagnosis even in preterm but that the complex



552 phenotypic presentation, including multimorbidity, could be due to the environmental stressor of  
553 being preterm. Specifically, we did not observe significant differences in DNV numbers between  
554 preterm and term ASD individuals. We did observe suggestive evidence that DNV burden could be  
555 higher in ASD-preterm compared with non-ASD-preterm, but the finding was inconsistent, which  
556 could be due to sample size overall and within each gestational age sub-group. After stratifying by  
557 preterm stages, we observed higher point estimates for DNV event rates in ASD compared to non-  
558 ASD within the moderate to late preterm birth, while lower point estimates in ASD compared to  
559 non-ASD within extremely to very preterm birth, but most of them did not reach statistical  
560 significance. If proven statistically significant in a future study, one can speculate that this may be  
561 indicative of distinct underlying genetic mechanisms for ASD across different preterm sub-groups.  
562 Limited research indicated a higher DNV burden in overall preterm newborn genomes and primarily  
563 in genes related to embryonic brain development; however, the study did not consider ASD or  
564 another behavioral diagnosis in preterm infants [22]. The increased DNV burden could be thus due  
565 to the higher prevalence of ASD within the preterm infant group as similar findings are repeatedly  
566 shown for ASD [60].

567

568 Although GWAS studies of prematurity have identified variations in maternal and fetal genes  
569 separately [61,62], few have examined the impact of rare inherited variants. Our study did not find a  
570 difference in the burden of rare inherited variants between ASD-preterm and ASD-term individuals.  
571 This can be partially explained by the fact that the maternal genome influences prematurity more  
572 than the fetal genome [14]. Although we did not find an overall association between ASD PRS and  
573 prematurity, we show intriguing findings that those with the highest PRS could have a higher  
574 likelihood of ASD, especially in preterm infants and boys. Even after including the interaction  
575 between preterm status and PRS, these features maintained a significant association with ASD  
576 likelihood, and the interaction itself was significantly associated with ASD diagnosis. Again, these  
577 findings need validation, especially as a prior study by Cullen et al. found no evidence of an

578 interaction effect between ASD polygenic score and gestational age at birth on cognition [15].

579 However, it is important to note that the cognitive difficulty they measured is only one of the  
580 outcomes that do not imply an ASD diagnosis, and the model they used also included socio-  
581 economic status as a covariate.

582

583 Variants at *AGTR2* and *ADCY5* genes were identified as associated with gestational duration and  
584 preterm birth in the GWAS study of Zhang et al [63]. Notably, these two genes are also known as  
585 ASD-associated genes [38]. This overlap suggests that certain genetic factors may influence both  
586 preterm birth and neurodevelopment. Given that our study subjects include individuals with either  
587 preterm or ASD conditions, some of these shared genetic factors may be overlooked. In the future,  
588 the understanding of the role of these genes in both preterm birth and ASD may reveal mechanisms  
589 by which genetic susceptibility to preterm birth contributes to the increased likelihood of ASD and  
590 other NDDs observed in preterm children.

591

592 In addition to genetic factors, widespread alterations in brain development associated with preterm  
593 infants may contribute to the increase in ASD likelihood. Previous studies have indicated that  
594 reduced structural brain asymmetry and poor brain development during neonatal life may increase  
595 the liability of ASD in preterm infants [64,65]. Even in preterm children exhibiting similar ASD traits  
596 during childhood, distinct etiological trajectories have been observed involving variations in neonatal  
597 cerebellar volume and developmental delay [66].

598

599 Not all preterm infants develop ASD [67]. but we demonstrate that when genetic factors are  
600 combined with the environmental risk of preterm birth, preterm children face an elevated likelihood  
601 of ASD diagnosis. Recognizing the limitations of traditional statistical models in capturing nonlinear  
602 interactions between features, we developed an ML model to predict ASD diagnosis in preterm  
603 children at birth. Unlike previous ASD prediction models that rely on developmental trajectories or

604 typical characteristics collected as children grow [25,29,68], our ML model utilized only information  
605 available at birth, integrating phenotype and genetic information. Moreover, most previous models  
606 are based on the general population [25,68], limiting their applicability to preterm infants. However,  
607 it is necessary to build prediction models tailored specifically for preterm infants due to the  
608 heterogeneity of ASD phenotypes [1], and preterm ASD children may exhibit specific phenotypes  
609 compared to term ASD children [18]. Although our ML model did not achieve significantly higher  
610 performance, achieving 69% accuracy with a small sample size and few features demonstrates the  
611 feasibility and efficacy of integrating phenotype and genetic information for ASD prediction. There is  
612 still substantial room for improvement in model performance. Increasing the sample size would  
613 provide the model with more learning opportunities and could enhance prediction accuracy.  
614 Additionally, adding more features associated with preterm birth and ASD would benefit the  
615 prediction, such as maternal age, prenatal exposure, and fetal birth weight. Other features like  
616 intubation in the delivery room, family language, parental education, other treatment, Infection, and  
617 ventilation have also been found to have predictive ability for cognitive outcomes in very preterm  
618 infants [29]. It is important to note that we cannot identify the causal relationships between features  
619 selected by the model and ASD diagnosis. Still, we suggest that these features could potentially  
620 enhance prediction models in the future.

621  
622 Our study has several limitations. Firstly, the cohorts we used are specifically focused on ASD, and  
623 the control group without ASD were still siblings and parents of ASD probands, potentially  
624 underestimating genetic differences between the groups. Given the high heritability estimation in  
625 ASD, siblings with a closer relationship with ASD have a higher relative risk ratio for ASD [4]. We  
626 cannot eliminate these potential genetic influences, which may introduce biases in results and affect  
627 the prediction ability of the ML model. Secondly, we did not stratify analyses by sex due to the  
628 limited sample size, potentially overlooking sex-specific differences in ASD phenotypes and variant  
629 event rates. Thirdly, we focused here only on the sequence level variation; thus, the next step would

630 be to include more types of genetic variations. Finally, our exploration of genetic factors primarily  
631 focused on average population-level associations and NDD genes, potentially overlooking genetic  
632 effects beyond the currently known ASD-associated genes and variants that may contribute to the  
633 elevated likelihood of ASD in preterm children. Previous studies have pointed out the genetic  
634 association between preterm and ASD, such as common genetic variants linking abnormalities in the  
635 gut-brain axis with both conditions [69]. We believe that combining genetic features and more  
636 detailed phenotypic information will help to explain further why some preterm children have ASD  
637 while others do not.

638

639

## 640 **Conclusion**

641 In conclusion, we demonstrate that ASD genetic liability is similar in ASD-term and ASD-preterm,  
642 suggesting that even within preterm, genetic factors play an important role in etiology. Our study did  
643 not find evidence of a link between genetic factors and preterm birth in ASD. However, our findings  
644 suggest that preterm birth would exacerbate the severity of outcomes in ASD individuals, and this  
645 difference may be driven more by environmental factors. As we observed some differences in the  
646 rate of ES DNV in preterm individuals compared between ASD and non-ASD, we only suggest that  
647 genetic factors may increase the likelihood of a preterm child getting an ASD diagnosis and the  
648 diagnosis is not modified by the interaction between multimorbidity and DNV burden. Through the  
649 development of our ML model, we demonstrate that integrating phenotype and genetic information  
650 is feasible and holds promise for the early prediction of ASD in preterm children at birth. Our study  
651 provides insights into the phenotypic characteristics of ASD preterm individuals. We suggest that  
652 health screening for preterm birth infants should incorporate the collection of genetic data, as it  
653 better supports early clinical identification of ASD and can aid in the guidance of early intervention  
654 strategies.

655

656

657 **List of abbreviations**

658 ASD: Autism Spectrum Disorder

659 SPARK: Simons Foundation Powering Autism Research for Knowledge

660 SSC: Simons Simplex Collection

661 GS: Genome sequencing

662 ES: Exome sequencing

663 GEE: Generalized Estimating Equations

664 PRS: Polygenic Risk Score

665 DNV: De Novo Variants

666 GWAS: Genome-Wide Association Studies

667 NDD genes: Neurodevelopmental Disorders-Related Genes

668 ML: Machine Learning

669 CBCL: Child Behavior Checklist

670 DCDQ: Developmental Coordination Disorder Questionnaire

671 RBS-R: Repetitive Behavior Scale-Revised

672 SCQ: Social Communication Questionnaire

673 Fsiq: Full-Scale Intelligence Quotient

674 AB: Allele Balance

675 LOF: Loss-of-Function

676 CADD: Combined Annotation Dependent Depletion

677 PCA: Principal Component Analysis

678 PC: Principal Components

679 CI: Confidence Interval

680 RFE: Recursive Feature Elimination

681 XGBoost: Extreme Gradient Boosting

682 RF: Random Forest

683 SVM: Linear Support Vector Machine

684 AUROC: Area Under the Receiver Operating Characteristic Curve

685 SHAP: SHapley Additive exPlanations

686 GA: Gestational Age at Birth

687 OR: Odds Ratio

688 FDR: False Discovery Rate

689

## 690 **Additional files**

691 **Additional file 1:** Supplementary Figures S1–S11 and Supplementary Tables S1–S10. (PDF 3686 kb)

692

## 693 **Declarations**

694 **Ethics approval and consent to participate:** Ethical approval for the data collection and informed

695 consent were obtained from the participants within the SPARK and SSC projects. The Swedish Ethical

696 Committee approved this study and data analysis in Sweden (dnr 2020-00400).

697 **Consent for publication:** Not applicable.

698 **Availability of data and materials:** The data that support the findings of this study are available from

699 the Simons Foundation Autism Research Initiative (SFARI, <https://www.sfari.org/resource/sfaribase>)

700 but restrictions apply to the availability of these data, which were used under license for the current

701 study, and so are not publicly available. Data are however available from the authors upon

702 reasonable request and with permission of SFARI. The R scripts used to perform the main analysis

703 reported in this manuscript are available on GitHub ([https://github.com/Tammimies-](https://github.com/Tammimies-Lab/AutismPreterm_Zhang)  
704 [Lab/AutismPreterm\\_Zhang](https://github.com/Tammimies-Lab/AutismPreterm_Zhang)).

705 **Competing interests:** The authors declare that they have no competing interests.

706 **Funding:** This work was supported by grants from the Swedish Foundation for Strategic Research  
707 (FFL18-0104), the Swedish Research Council (2017-01660\_VR, 2017-03043\_VR, 2023-02111\_VR ,  
708 2023-02451\_VR), The Swedish Brain Foundation ( FO2021-0073 and F02023-0186), Strategic  
709 Research Area in Neuroscience at Karolinska Institutet (StratNeuro), the China Scholarship Council  
710 (CSC, Zhang), the Sällskapet Barnavård (SBV, Zhang). Furthermore, research reported in this  
711 publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human  
712 Development of the National Institutes of Health under Award Number R01HD113669 (NIH, Sandin).  
713 The content is solely the responsibility of the authors and does not necessarily represent the official  
714 views of the National Institutes of Health.

715 **Authors' contributions:** Y.Z. and K.T. designed the study and planned the analyses. Y.Z. performed the  
716 analyses. Y.A. provided support for the analysis. S.S., U.Å, and K.T. provided supervision and support  
717 for the analysis. Y.Z. wrote the first draft with feedback from K.T. All authors provided critical  
718 feedback and helped shape the research, analysis, and the final manuscript.

719 **Acknowledgements:** We thank all the participants of SPARK and SSC cohorts and Simons Foundation  
720 for access of the data. The computations were performed using resources provided by the National  
721 Academic Infrastructure for Supercomputing in Sweden (NAISS) through Uppsala Multidisciplinary  
722 Center for Advanced Computational Science (UPPMAX).

723

724

725 **Reference**

- 726 1. Rosen NE, Lord C, Volkmar FR. The Diagnosis of Autism: From Kanner to DSM-III to DSM-5 and  
727 Beyond. *J Autism Dev Disord.* 2021;51:4253–70.
- 728 2. Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, et al. Autism spectrum disorder. *Nat*  
729 *Rev Dis Primers.* 2020;6:5.
- 730 3. Carlsson T, Molander F, Taylor MJ, Jonsson U, Bölte S. Early environmental risk factors for  
731 neurodevelopmental disorders – a systematic review of twin and sibling studies. *Dev Psychopathol.*  
732 2021;33:1448–95.
- 733 4. Havdahl A, Niarchou M, Starnawska A, Uddin M, van der Merwe C, Warriar V. Genetic  
734 contributions to autism spectrum disorder. *Psychol Med.* 2021;51:2260–73.
- 735 5. Ruzzo EK, Pérez-Cano L, Jung J-Y, Wang L-K, Kashef-Haghighi D, Hartl C, et al. Inherited and De Novo  
736 Genetic Risk for Autism Impacts Shared Networks. *Cell.* 2019;178:850-866.e26.
- 737 6. Bai D, Yip BHK, Windham GC, Sourander A, Francis R, Yoffe R, et al. Association of Genetic and  
738 Environmental Factors With Autism in a 5-Country Cohort. *JAMA Psychiatry.* 2019;76:1035.
- 739 7. Iossifov I, O’Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo  
740 coding mutations to autism spectrum disorder. *Nature.* 2014;515:216–21.
- 741 8. Zhou X, Feliciano P, Shu C, Wang T, Astrovskaya I, Hall JB, et al. Integrating de novo and inherited  
742 variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat Genet.*  
743 2022;54:1305–19.
- 744 9. Wilfert AB, Turner TN, Murali SC, Hsieh P, Sulovari A, Wang T, et al. Recent ultra-rare inherited  
745 variants implicate new autism candidate risk genes. *Nat Genet.* 2021;53:1125–34.
- 746 10. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common  
747 genetic risk variants for autism spectrum disorder. *Nat Genet.* 2019;51:431–44.



- 748 11. Loughnan RJ, Palmer CE, Makowski C, Thompson WK, Barch DM, Jernigan TL, et al. Unique  
749 prediction of developmental psychopathology from genetic and familial risk. *Child Psychology*  
750 *Psychiatry*. 2022;63:1631–43.
- 751 12. Crump C, Sundquist J, Sundquist K. Preterm or Early Term Birth and Risk of Autism. *Pediatrics*.  
752 2021;148:e2020032300.
- 753 13. Persson M, Opdahl S, Risnes K, Gross R, Kajantie E, Reichenberg A, et al. Gestational age and the  
754 risk of autism spectrum disorder in Sweden, Finland, and Norway: A cohort study. Fasset MJ, editor.  
755 *PLoS Med*. 2020;17:e1003207.
- 756 14. Svensson AC, Sandin S, Cnattingius S, Reilly M, Pawitan Y, Hultman CM, et al. Maternal Effects for  
757 Preterm Birth: A Genetic Epidemiologic Study of 630,000 Families. *American Journal of Epidemiology*.  
758 2009;170:1365–72.
- 759 15. Cullen H, Selzam S, Dimitrakopoulou K, Plomin R, Edwards AD. Greater genetic risk for adult  
760 psychiatric diseases increases vulnerability to adverse outcome after preterm birth. *Sci Rep*.  
761 2021;11:11443.
- 762 16. Heikkilä K, Metsälä J, Pulakka A, Nilsen SM, Kivimäki M, Risnes K, et al. Preterm birth and the risk  
763 of multimorbidity in adolescence: a multiregister-based cohort study. *The Lancet Public Health*.  
764 2023;8:e680–90.
- 765 17. Allotey J, Zamora J, Cheong-See F, Kalidindi M, Arroyo-Manzano D, Asztalos E, et al. Cognitive,  
766 motor, behavioural and academic performances of children born preterm: a meta-analysis and  
767 systematic review involving 64 061 children. *BJOG*. 2018;125:16–25.
- 768 18. Johnson S, Marlow N. Preterm Birth and Childhood Psychiatric Disorders. *Pediatric Research*.  
769 2011;69:11R-18R.

- 770 19. Joseph RM, Lai ER, Bishop S, Yi J, Bauman ML, Frazier JA, et al. Comparing autism phenotypes in  
771 children born extremely preterm and born at term. *Autism Research*. 2023;16:653–66.
- 772 20. Leoni M, Vanes LD, Hadaya L, Kanel D, Dazzan P, Simonoff E, et al. Exploring cognitive, behavioral  
773 and autistic trait network topology in very preterm and term-born children. *Front Psychol*.  
774 2023;14:1119196.
- 775 21. Luu J, Jellett R, Yaari M, Gilbert M, Barbaro J. A Comparison of Children Born Preterm and Full-  
776 Term on the Autism Spectrum in a Prospective Community Sample. *Front Neurol*. 2020;11:597505.
- 777 22. Li J, Oehlert J, Snyder M, Stevenson DK, Shaw GM. Fetal de novo mutations and preterm birth.  
778 Williams SM, editor. *PLoS Genet*. 2017;13:e1006689.
- 779 23. Wong HS, Wadon M, Evans A, Kirov G, Modi N, O'Donovan MC, et al. Contribution of de novo and  
780 inherited rare CNVs to very preterm birth. *J Med Genet*. 2020;57:552–7.
- 781 24. Everett SS, Bomback M, Sahni R, Wapner RJ, Tolia VN, Clark RH, et al. Prevalence and Clinical  
782 Significance of Commonly Diagnosed Genetic Disorders in Preterm Infants. Preprint at  
783 <http://medrxiv.org/lookup/doi/10.1101/2023.07.14.23292662> (2023)
- 784 25. Chen J, Engelhard M, Henao R, Berchuck S, Eichner B, Perrin EM, et al. Enhancing early autism  
785 prediction based on electronic records using clinical narratives. *Journal of Biomedical Informatics*.  
786 2023;144:104390.
- 787 26. Rajagopalan SS, Zhang Y, Yahia A, Tammimies K. Machine Learning Prediction of Autism Spectrum  
788 Disorder From a Minimal Set of Medical and Background Information. *JAMA Netw Open*.  
789 2024;7:e2429229.

- 790 27. Asif M, Martiniano HFMC, Marques AR, Santos JX, Vilela J, Rasga C, et al. Identification of  
791 biological mechanisms underlying a multidimensional ASD phenotype using machine learning. *Transl*  
792 *Psychiatry*. 2020;10:43.
- 793 28. Li D, Choque Olsson N, Becker M, Arora A, Jiao H, Norgren N, et al. Rare variants in the outcome  
794 of social skills group training for autism. *Autism Res*. 2022;15:434–46.
- 795 29. Bowe AK, Lightbody G, Staines A, Murray DM, Norman M. Prediction of 2-Year Cognitive  
796 Outcomes in Very Preterm Infants Using Machine Learning Methods. *JAMA Netw Open*.  
797 2023;6:e2349111.
- 798 30. Feliciano P, Daniels AM, Green Snyder L, Beaumont A, Camba A, Esler A, et al. SPARK: A US Cohort  
799 of 50,000 Families to Accelerate Autism Research. *Neuron*. 2018;97:488–93.
- 800 31. Fischbach GD, Lord C. The Simons Simplex Collection: A Resource for Identification of Autism  
801 Genetic Risk Factors. *Neuron*. 2010;68:192–5.
- 802 32. Ng JK, Vats P, Fritz-Waters E, Sarkar S, Sams EI, Padhi EM, et al. de novo variant calling identifies  
803 cancer mutation signatures in the 1000 Genomes Project. *Human Mutation*. 2022;43:1979–93.
- 804 33. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission  
805 disequilibrium confirms that common and rare variation act additively to create risk for autism  
806 spectrum disorders. *Nature genetics*. 2017;49:978–85.
- 807 34. Pedersen BS, Brown JM, Dashnow H, Wallace AD, Velinder M, Tristani-Firouzi M, et al. Effective  
808 variant filtering and expected candidate variant yield in studies of rare human disease. *npj Genom*  
809 *Med*. 2021;6:60.

- 810 35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome  
811 Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.  
812 *Genome Res.* 2010;20:1297–303.
- 813 36. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-  
814 throughput sequencing data. *Nucleic Acids Research.* 2010;38:e164–e164.
- 815 37. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and  
816 predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*  
817 *melanogaster* strain w<sup>1118</sup>; iso-2; iso-3. *Fly.* 2012;6:80–92.
- 818 38. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, et al. SFARI Gene 2.0:  
819 a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular Autism.*  
820 2013;4:36.
- 821 39. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources  
822 expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51:1560–5.
- 823 40. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format  
824 and VCFtools. *Bioinformatics.* 2011;27:2156–8.
- 825 41. Kochinke K, Zweier C, Nijhof B, Fenckova M, Cizek P, Honti F, et al. Systematic Phenomics Analysis  
826 Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. *The*  
827 *American Journal of Human Genetics.* 2016;98:149–64.
- 828 42. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of  
829 Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal*  
830 *of Human Genetics.* 2009;84:524–33.

- 831 43. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to  
832 the challenge of larger and richer datasets. *GigaSci.* 2015;4:7.
- 833 44. Hinrichs AS. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research.*  
834 2006;34:D590–8.
- 835 45. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and  
836 continuous shrinkage priors. *Nat Commun.* 2019;10:1776.
- 837 46. Halekoh U, Højsgaard S, Yan J. The R Package geeppack for Generalized Estimating Equations. *J Stat*  
838 *Soft.* 2006;15:2.
- 839 47. Lüdtke D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of*  
840 *Open Source Software.* 2018;3:772.
- 841 48. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Stat*  
842 *Comput.* 2017;27:659–78.
- 843 49. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Soft .* 2008;28:5.
- 844 50. Liu Y, Just A. SHAPforxgboost: SHAP Plots for “XGBoost”. R package version 0.1.3.; 2020.  
845 <https://github.com/liuyanguu/SHAPforxgboost/>
- 846 51. Heikkilä K, Pulakka A, Metsälä J, Alenius S, Hovi P, Gissler M, et al. Preterm birth and the risk of  
847 chronic disease multimorbidity in adolescence and early adulthood: A population-based cohort  
848 study. Jacobsen R, editor. *PLoS ONE.* 2021;16:e0261952.
- 849 52. Johnson S, Marlow N. Growing up after extremely preterm birth: Lifespan mental health  
850 outcomes. *Seminars in Fetal and Neonatal Medicine.* 2014;19:97–104.
- 851 53. Hirvonen M, Ojala R, Korhonen P, Haataja P, Eriksson K, Gissler M, et al. Visual and Hearing  
852 Impairments After Preterm Birth. *Pediatrics.* 2018;142:e20173888.

- 853 54. Li W, Peng A, Deng S, Lai W, Qiu X, Zhang L, et al. Do premature and postterm birth increase the  
854 risk of epilepsy? An updated meta-analysis. *Epilepsy & Behavior*. 2019;97:83–91.
- 855 55. Khachadourian V, Mahjani B, Sandin S, Kolevzon A, Buxbaum JD, Reichenberg A, et al.  
856 Comorbidities in autism spectrum disorder and their etiologies. *Transl Psychiatry*. 2023;13:71.
- 857 56. Johnson S, Hollis C, Kochhar P, Hennessy E, Wolke D, Marlow N. Autism Spectrum Disorders in  
858 Extremely Preterm Children. *The Journal of Pediatrics*. 2010;156:525-531.e2.
- 859 57. Schieve LA, Baio J, Rice CE, Durkin M, Kirby RS, Drews-Botsch C, et al. Risk for cognitive deficit in a  
860 population-based sample of U.S. children with autism spectrum disorders: Variation by perinatal  
861 health factors. *Disability and Health Journal*. 2010;3:202–12.
- 862 58. Zhu JL, Olsen J, Olesen AW. Risk for Developmental Coordination Disorder Correlates with  
863 Gestational Age at Birth. *Paediatric Perinatal Epid*. 2012;26:572–7.
- 864 59. York TP, Eaves LJ, Lichtenstein P, Neale MC, Svensson A, Latendresse S, et al. Fetal and Maternal  
865 Genes' Influence on Gestational Age in a Quantitative Genetic Analysis of 244,000 Swedish Births.  
866 *American Journal of Epidemiology*. 2013;178:543–50.
- 867 60. Li C, Fleck JS, Martins-Costa C, Burkard TR, Themann J, Stuempflen M, et al. Single-cell brain  
868 organoid screening identifies developmental defects in autism. *Nature*. 2023;621:373–80.
- 869 61. Solé-Navais P, Flatley C, Steinhorsdottir V, Vaudel M, Juodakis J, Chen J, et al. Genetic effects on  
870 the timing of parturition and links to fetal birth weight. *Nat Genet*. 2023;55:559–67.
- 871 62. Liu X, Helenius D, Skotte L, Beaumont RN, Wielscher M, Geller F, et al. Variants in the fetal  
872 genome near pro-inflammatory cytokine genes on 2q13 associate with gestational duration. *Nat*  
873 *Commun*. 2019;10:3927.

- 874 63. Zhang G, Feenstra B, Bacelis J, Liu X, Muglia LM, Juodakis J, et al. Genetic Associations with  
875 Gestational Duration and Spontaneous Preterm Birth. *N Engl J Med*. 2017;377:1156–67.
- 876 64. Eklöf E, Mårtensson GE, Ådén U, Padilla N. Reduced structural brain asymmetry during neonatal  
877 life is potentially related to autism spectrum disorders in children born extremely preterm. *Autism*  
878 *Research*. 2019;12:1334–43.
- 879 65. Padilla N, Eklöf E, Mårtensson GE, Bölte S, Lagercrantz H, Ådén U. Poor Brain Growth in Extremely  
880 Preterm Neonates Long Before the Onset of Autism Spectrum Disorder Symptoms. *Cereb Cortex*.  
881 2015;bhv300.
- 882 66. Hadaya L, Vanes L, Karolis V, Kanel D, Leoni M, Happé F, et al. Distinct Neurodevelopmental  
883 Trajectories in Groups of Very Preterm Children Screening Positively for Autism Spectrum Conditions.  
884 *J Autism Dev Disord*. 2024;54:256–69.
- 885 67. Hee Chung E, Chou J, Brown KA. Neurodevelopmental outcomes of preterm infants: a recent  
886 literature review. *Transl Pediatr*. 2020;9:S3–8.
- 887 68. Amit G, Bilu Y, Sudry T, Avgil Tsadok M, Zimmerman DR, Baruch R, et al. Early Prediction of  
888 Autistic Spectrum Disorder Using Developmental Surveillance Data. *JAMA Netw Open*.  
889 2024;7:e2351052.
- 890 69. Sajdel-Sulkowska EM, Makowska-Zubrycka M, Czarzasta K, Kasarello K, Aggarwal V, Bialy M, et al.  
891 Common Genetic Variants Link the Abnormalities in the Gut-Brain Axis in Prematurity and Autism.  
892 *Cerebellum*. 2019;18:255–65.
- 893