Data-driven characterization of individuals with delayed autism diagnosis

Dan Aizenberg, MS[1], Ido Shalev, MA[2], Florina Uzefovsky, PhD[2]*, Alal Eran, PhD[3]*

[1] Department of Life Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel

[2] Psychology Department, Ben-Gurion University of the Negev, Beer Sheva, Israel

[3] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

*** Corresponding authors:**

Alal Eran, PhD, Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States. E-mail address: Alal_Eran@hms.harvard.edu

Florina Uzefovsky, PhD, Department of Psychology, Ben-Gurion University of the Negev, Beer Sheva, Israel. E-mail address: florina@bgu.ac.il

Word count: 3,027

Key Points

Question: Are there specific subgroups of individuals diagnosed with autism after school age?

Findings: In this data-driven analysis of a large cohort of autistic individuals, two distinct subgroups of individuals diagnosed with autism after the age of six were identified. The first included individuals requiring low levels of support, with modest comorbidity burdens; The second included individuals requiring high levels of support, with extremely high comorbidity burdens.

Meaning: The identification of opposite subgroups of individuals with delayed autism diagnosis improves our understanding of autism heterogeneity and moves us closer towards precision diagnosis of autism.

Abstract

Importance: Despite tremendous improvement in early identification of autism, ~25% of children receive their diagnosis after the age of six. Since evidence-based practices are more effective when started early, delayed diagnosis prevents many children from receiving optimal support.

Objective: To identify and comparatively characterize groups of individuals diagnosed with Autism Spectrum Disorder (ASD) after the age of six.

Design: This cross-sectional study used various machine learning approaches to classify, characterize, and compare individuals from the Simons Foundation Powering Autism Research for Knowledge (SPARK) cohort, recruited between 2015-2020.

Setting: Analyses of medical histories and behavioral instruments.

Participants: 23,632 SPARK participants.

Exposure: ASD diagnosis upon registration to SPARK.

Main Outcomes and Measures: Clusters of individuals diagnosed after the age of six (*delayed ASD diagnosis*) and their defining characteristics, as compared to individuals diagnosed before the age of six (*timely ASD diagnosis*). Odds and mean ratios were used for feature comparisons. Shapley values were used to assess the predictive value of these features, and correlation-based cliques were used to understand their interconnectedness.

Results: Two robust subgroups of individuals with delayed ASD diagnosis were detected. The first, *D1*, included 3,612 individuals with lower support needs as compared to 17,992 individuals with a timely diagnosis. The second subgroup, *D2*, included 2,028 individuals with higher support needs, as consistently reflected by all commonly-used behavioral instruments, the greatest being repetitive and restrictive behaviors measured by the Repetitive Behavior Scale – Revised (RBS-R; D1: MR =

0.6854, 95% CI = 0.6848 – 0.686; D2: MR = 1.4223, 95% CI = 1.4210-1.4238, P = 3.54×10^{-134}). Moreover, individuals belonging to D1 had fewer comorbidities as compared to individuals with a timely ASD diagnosis, while D2 individuals had more (D1: mean = 3.47, t = 15.21; D2: mean = 8.12, t = 48.26, p< 2.23×10^{-308}). A Random Forest classifier trained on the groups' characteristics achieved an AUC of 0.94. Further connectivity analysis of the groups' most informative characteristics demonstrated their distinct topological differences.

Conclusions and Relevance: This analysis identified two opposite groups of individuals with delayed ASD diagnosis, thereby providing valuable insights for the development of targeted diagnostic strategies.

Introduction

One in 36 children in the United States is diagnosed with Autism Spectrum Disorder (ASD), a heterogeneous spectrum of neurodevelopmental conditions characterized by difficulties in social cognition and communication, and restrictive and repetitive behavior (RRBs)[1]. Early intervention, starting as early as 12 months of age, can substantially affect the development and long-term outcomes of autistic children[2-6]. However, despite dramatic improvements in the proportion of children who receive a developmental screening by age 36 months, the median age of ASD diagnosis in the US is currently 49 months. Thus, many autistic individuals remain undiagnosed until and throughout school-age[1,7,8], preventing them from receiving the support they need.

Delayed diagnosis can be partially explained by demographic factors, including sex, race[9-11], socio-economic status[12,13], and autism awareness at the time of diagnosis[13]. Delayed diagnosis may also be attributed to the severity of autistic traits and one's level of required support[9-11]. However, these factors explain delayed diagnosis only partially, and most of the variance in the age of ASD diagnosis remains unexplained[12-14]. Previous studies investigating delayed ASD diagnosis have typically focused on a limited set of characteristics[12,13,15,16]. Moreover, these studies have treated individuals with delayed diagnosis as a single group, while autism is highly heterogeneous[17,18].

Autistic children often suffer from co-occurring conditions, including seizures, gastrointestinal dysfunction, and growth problems[19-30]. Some of these conditions have been shown to share genetic components with autism[31-33], correlate with other common autism comorbidities[34], or be associated with autistic traits[35-37]. An association between some comorbidities and delayed diagnosis might arise from

several reasons. The first is overshadowing of autism by a co-occurring disorder, or in other words, attributing autistic traits to another disorder that may or may not co-occur with ASD[22,38,39]. For example, ASD may be overshadowed by attention deficit hyperactivity disorder (ADHD), and it has been shown that a presentation of hyperactivity or inattention often leads to a primary diagnosis of ADHD, consequently delaying one's autism diagnosis[33,40]. Moreover, a set of comorbidities may present a unique profile that deflects the diagnostic course from ASD.

Another potential reason for certain comorbidities to be associated with delayed ASD diagnosis is that such comorbidities might share an underlying genetic basis with autistic traits appearing later in life[34]. For example, that might be the case of phenylketonuria[41] or Smith-Lemli-Opitz syndrome (SLOS)[42]. Therefore, a comprehensive characterization or profiling of individuals with delayed ASD diagnosis using their phenotypic and comorbidity profiles may shed light on the underlying processes associated with delayed diagnosis, ultimately enabling targeted screening options and improved outcomes.

Recently, attempts have been made to exploit machine learning (ML) and big data to dissect the heterogeneous autism spectrum, based, for example, on behavioral phenotypes[43], neuroanatomy and neuro functioning[44,45], information processing ability[45], and comorbidity profiles[35-37,46]. However, to date, no data-driven approach has been used to dissect delayed diagnosis. Such a large-scale systematic approach could be used to explain diagnostic delays and gain a better understanding of the heterogeneity of autism. This, in turn, could enable targeted research and diagnosis options.

Here we analyzed a large cohort of autistic individuals, focusing on those diagnosed after the age of six, considered as having a delayed diagnosis. Using various supervised and unsupervised ML algorithms, we identified and characterized two distinct groups of individuals with delayed diagnosis (denoted D1 and D2). We also leveraged this big data to quantify the relationships between the characteristics of each group.

Methods

The Boston Children's Hospital Institutional Review Board has determined that this research qualifies as exempt from the requirements of human subjects protection regulations.

Data preprocessing

We mined the SPARK phenotype dataset (version 5)[45], containing rich phenotypic data from 99,447 autistic individuals, aged 0-92 years. Tables populated with 25% or more of the data were used for our analysis. These include basic medical screening, Social Communication Questionnaire-Lifetime (SCQ)[47], Background history questionnaire, Repetitive Behavior Scale-Revised (RBS-R)[48], and Developmental Coordination Disorder Questionnaire (DCDQ)[49]. Of these tables, only relevant features were considered (**eTable 1**). Of note, IQ and familial medical and mental health history were not collected on a sufficiently large population and were therefore excluded from the analysis. Continuous variables were Z-standardized, and categorical variables were represented as one-hot vectors. All data was processed using Python's pandas v1.3.4[50] and NumPy v1.20.3[51] packages.

A total of 23,632 individuals who had complete information in all selected tables and whose autism diagnosis was never refuted or remitted (N = 1727) were included in the analysis. Their median age of diagnosis was 3.67 years (IQR = 2.58 – 5.75 years).

Definition of delayed diagnosis

The age cutoff for delayed diagnosis was six years, chosen for being higher than the 75th percentile of diagnosis ages in the studied cohort (5.75 years). Moreover, six years of age was previously reported to be a key turning point in autism, in which the course of autistic traits and behavior sometimes dramatically changes[52]. Autistic traits and behavior appear much earlier than this age, and by then, the relevant time window for early intervention has long since passed[5,53]. Accordingly, participants were first divided into one of two groups: those diagnosed before the age of six (median = 3.08 years, IQR = 2.42 – 4.08 years), and those diagnosed at or after the age of six (median = 8.00 years, IQR = 6.83 – 9.67 years). **eFigure 1** shows the number of individuals in each group and the distribution of their ages at diagnosis, birth year, and year of diagnosis.

Identification of groups of individuals with delayed ASD diagnosis

K-means clustering of individuals diagnosed at or after the age of six was performed on all standardized variables listed in **eTable 1**, using Python's scikit-learn package v0.23.2[54]. The number of clusters was identified using the elbow method, ensuring that every cluster contained at least 10% of individuals.

Group characterization

To characterize the delayed diagnosis groups, each variable listed in **eTable 1** was tested for differences between its distribution in each group and 1,000 bootstrapped samples from the timely diagnosis group, with the bootstrap sample size equal to the group size. For binary variables, Fisher's exact tests were performed between these samples. For continuous variables, Mann–Whitney U tests were performed and the ratio of the groups' means was saved as an interpretable statistic comparable to the odds-ratio statistic from the Fisher's exact test. Benjamini–Hochberg correction was then used to account for multiple testing, ensuring a false discovery rate of 0.05. Python's pandas v1.3.4[50] was used for bootstrapping, SciPy stats[55] v1.7.1 was used for Fisher's and Mann-Whitney tests, and Statmodels[56] v0.12.1 was used for their corrections.

Analyzing relations between co-occurring psychiatric disorders and age of ASD diagnosis

Pearson correlations were calculated between the age of ASD diagnosis and the prevalence of reported psychiatric disorders, using SciPy[55] stats v1.7.1. Benjamini–Hochberg corrections were used to account for multiple testing ($\alpha = 0.05$, Statmodels[56] v0.12.1).

Comorbidity burden comparison

For each individual, we summed the number of reported comorbid conditions. Pairwise two-sample t-tests were used to compare these between identified using SciPy[55] v1.7.1.

Group classification and model inference

A random forest classifier was trained to classify individuals into one of identified groups using Python's scikit-learn[54] package v0.23.2, on a training set of 75% of the data. Underrepresented clusters were synthetically over-sampled in the training set using the SMOTE-NC algorithm[57], implemented in Python's imbalanced-learn[58] package v0.7.0. The classifier's hyper-parameters were tuned using a grid-search with 10-fold cross-validation and with micro-averaged F1-score as the scoring method. To understand the contribution of features to the classification, Shapley values were calculated using Python's SHAP package, v0.37.0[59].

Feature interconnectivity and network analysis

To understand the relations between features of each group, we calculated Pearson's correlations between each pair of features in each group. We represented these relations in a graph whose feature vertices were joined by an edge if their correlation coefficient was > 0.1 and their Benjamini-Hochberg adjusted correlation P value was $< 1\times10^{-4}$. We then identified correlation-based cliques using Python's NetworkX v2.6.3[60]. We specifically focused on 4-cliques, defined as groups of four or more vertices that are all connected to each other. Normalized clique membership was determined by dividing the number of 4-cliques in which each (vertex) variable appeared within a group by the total number of cliques in that group.

Results

Two distinct groups of individuals with delayed ASD diagnosis differ by their co-occurring conditions and levels of autistic traits

In all, 5,640 of 23,632 (23.87%) SPARK participants received their ASD diagnosis after the age of six, consistent with recent US-wide Autism and Developmental Disabilities Monitoring (ADDM) network reports[61]. K-means clustering revealed two distinct groups of autistic individuals diagnosed after the age of six (**eFigure 2**). The first, D1, included 3,612 individuals (64%), and the second, D2, included 2,028 individuals (36%). D1 individuals were characterized by fewer autistic traits than timely diagnosed individuals, whereas D2 individuals had more autistic traits (**Figure 1**). This finding is consistent across all autism domains, as measured by multiple behavioral instruments, and spans difficulties in social communication and RRBs (**Figure 1A**), motor development (**Figure 1B**), and feeding and sleeping problems (**Figure 1C**).

Moreover, D1 individuals had lower odds of co-occurring neurodevelopmental conditions than timely diagnosed individuals, whereas D2 individuals had higher odds for such conditions (**Figure 2A-C**). These include cognitive and motor delays (**Figure 2A**), growth abnormalities (**Figure 2B**), and neural problems (**Figure 2C**). While both delayed diagnosis groups had higher rates of ADHD, anxiety, and depression or dysthymia as compared to timely diagnosed individuals, those in D2 had higher rates of neuropsychiatric disorders, including obsessive-compulsive disorder (OCD), oppositional defiant disorder (ODD), and schizophrenia (**Figure 2D**). Moreover, a positive linear trend was found between the age of diagnosis and the prevalence of neuropsychiatric disorders (**eFigure 3**).

Additionally, D1 individuals had lower odds of congenital disabilities than timely diagnosed individuals, whereas D2 individuals had higher odds for such conditions (**eFigure 4**). Similarly, prenatal and perinatal complications were less common in individuals in the D1 group as compared to timely diagnosed individuals,

whereas individuals in the D2 group had higher odds for such conditions, including fetal alcohol syndrome, premature birth, and insufficient oxygen at birth (**eFigure 5**). Finally, parents of individuals in D1 were more educated than parents of timely diagnosed individuals, while parents of individuals in D2 were less educated (**eFigure 6**).

Groups of individuals with delayed diagnosis are characterized by extreme rates of co-occurring conditions

Next, we compared the total number of co-occurring conditions between groups. While individuals with a timely ASD diagnosis had, on average, 4.34 co-occurring conditions, D1 individuals had significantly less (mean = 3.47, t = 15.21, p = 3.01×10^{-52}) and D2 individuals had significantly more comorbidities (mean = 8.12, t = 48.26, p < 2.23×10^{-308}; **Figure 3**). Thus, individuals with delayed ASD diagnosis are characterized by extreme rates of co-occurring conditions.

Levels of autistic traits and prevalence of psychiatric disorders are the primary distinguishing features between delayed diagnosis groups

We next determined how distinguishable are the two groups of individuals with delayed diagnosis and what features contribute most to their classification. We trained a one vs. rest random forest classifier to assign an autistic individual to one of the identified groups. The model achieved a micro-averaged AUC of 0.94 and a micro-averaged average-precision of 0.89 on a test set of 25% of the data (**eFigure 7**). Thus, the identified groups could be distinguished using a relatively simple model. The most important features in this model were identified by their Shapley values,

i.e., their relative contribution to the model, considering all possible combinations of the remaining features.

Of the 20 most important features distinguishing D1 from D2 and the timely diagnosis group, 14 overlap and predict in opposite directions, reflecting the two extremes of the delayed diagnosis groups (**Figure 4**). These features include language disorders, RBS-r subscales, SCQ scores, learning disorders, sleeping problems, DCDQ scores, and DCDQ's fine motor control subscale score. Moreover, mood disorders, anxiety disorders, depression, or OCD were found to be informative characteristics of D2. Contrarily, lack of intellectual disability, speech articulation problems, or problems with eating foods contributed to the assignment of an individual to D1.

Relationships between group characteristics demonstrate unique relations between autistic traits and co-occurring conditions

Finally, we examined inter-relations between the most informative characteristics of each group. Toward that end, we calculated pairwise correlations among D1 features (**eTable 2**), D2 features (**eTable 3**), and those of the timely diagnosis group (**eTable 4**). We then assessed the strength of information flow inherent to each feature by modeling these correlations as a graph and calculating each feature's normalized clique membership, which indicates the percent of 4-cliques that contain that feature (**eTable 5**). The larger a feature's normalized clique membership, the more information flows through it. **Figure 5A-C** depicts and compares the five most inter-connected features in each group, demonstrating the difference between the role these features play in each group. For example, in D1, the two most inter-connected variables are related to RRBs, and in D2 it is motor delay and birth disabilities.

To gain a more comprehensive understanding of inter-relations between the most connected features within each group, correlation heatmaps were used to depict pairwise correlations between the most connected features and their clique members in D1 (**Figure 5D**), D2 (**Figure 5E**), and the timely diagnosed group T (**eFigure 8)**. These heatmaps demonstrate group-specific topology of feature relatedness.

Discussion

Tremendous efforts are being made worldwide to lower the age of ASD diagnosis, since early detection can greatly improve outcomes and intervention opportunities[2-6]. The present study leveraged novel machine learning approaches and big data from SPARK to identify factors contributing to delayed diagnosis. We identified two distinct groups of autistic individuals with delayed diagnosis, differing in the level of autistic traits and the degree of co-occurring conditions, and displaying distinct network structures.

Autism is highly heterogenous, and as such, a single reason for late diagnosis is unlikely. Yet, previous studies examining the differences between late and early diagnosed individuals lacked the design or analytical power to untangle such heterogeneity, resulting in conflicting evidence. While some studies suggested higher rates of co-occurring conditions[62] and no differences in autistic traits in the late-diagnosed group[15], others reported more pronounced autistic traits and higher rates of neuropsychiatric conditions, together with communication difficulties[63], language delay[13], and intellectual disability[9,13] as factors enabling earlier diagnosis. Our findings suggest the existence of two mirroring groups of individuals with delayed ASD diagnosis, which may explain the conflicting results observed in previous studies.

The first identified group, D1, was characterized by lower levels of autistic behaviors and lower odds of co-occurring neurodevelopmental and neuropsychiatric conditions, as compared to the timely diagnosed group. Individuals in this group also displayed fewer learning and intellectual disabilities and better verbal skills. These characteristics align with previous findings showing that these characteristics contribute to delayed autism diagnosis[9,12,15,64,65]. It is possible that higher verbal and intellectual skills may delay parents in seeking a diagnosis or being referred to dedicated clinics, thus delaying the time of diagnosis.

Contrarily, the second group, D2, showed higher levels of autistic traits and high rates of neurodevelopmental disorders or delays. In this group, autism is only one of many co-occurring conditions. Autism evaluations are typically focused on the specific clinical features of autism, ignoring co-occurring conditions, and could be overshadowed by them[66]. Consequently, the high rates of co-occurring conditions could make the recognition of autism more complex, delaying the age of diagnosis.

The two groups were further distinguished based on their underlying network structure and feature interconnectivity. RRBs were identified as the core characteristic of D1, showing tight intercorrelations within this group. Recent findings involving younger children with a strong indication for ASD (suggesting timely diagnosis), showed strong connectivity among RRBs, but it was considered a peripheral dimension of the network in that sample, with social communication difficulties identified as the core aspect of the network[67]. In contrast, our analysis suggests that for some individuals with delayed diagnosis, RRBs and their connectivity represent the core feature of autism.

Network analysis identified different core features in D2, such as motor delay and congenital disabilities. These conditions correlated with difficulties in social

communication, reinforcing the idea that co-occurring conditions might overshadow ASD diagnosis[66]. Nevertheless, further research is needed to explore the interplay between characteristics within these groups.

Besides offering targeted approaches for future screening and diagnosis, this study can help make sense of the heterogeneity characterizing autism. Such heterogeneity is reflected by complex genetic etiology with over one thousand known genetic variants associated with autism, most with small effect sizes and additive effects[18,68]. The two groups identified in our study may reflect differences in physiological and phenotypic bases that could be used for future targeted genetic research and better characterization of autism.

This study holds several limitations. Key among them is the absence of temporal data, such as the onset of co-occurring mental and developmental conditions, limiting our ability to clarify if and how these factors precede ASD diagnosis. Previous research suggests that preceding neurodevelopmental and neuropsychiatric conditions contribute to delayed ASD diagnosis[40,62]. For example, children with a prior ADHD diagnosis, tend to receive their ASD diagnosis an average 1.8 years later than children without ADHD[40]. Future research should employ longitudinal designs to investigate the temporal directionality of D1- and D2-informative characteristics identified in our study.

Other limitations include the exclusion of IQ measures from our data due to missing data, which could have provided valuable insights. Furthermore, SPARK participants are mostly White, limiting the generalizability of our study, and future studies should include a more diverse cohort. Finally, throughout the paper, we consistently used the term ‘ASD’ (or individuals with an ASD diagnosis), to refer to the official diagnosis given to participants according to the Diagnostic and Statistical

Manual of Mental Disorder (DSM). However, we acknowledge the controversy surrounding the use of this term, and when referring to individuals, we used identity-first language to describe autistic people, as it is the preferred choice of most autistic individuals[69].

Conclusions

In this study, we identified and characterized two distinct groups of individuals diagnosed with ASD after the age of six. By dissecting the heterogeneous landscape of individuals with delayed ASD diagnosis, this work opens new avenues for better-powered (less heterogenous) genetic and behavioral research. Ultimately, this work enables the development of targeted diagnostic strategies and thereby improved outcomes for autistic children.

Acknowledgments

Author contributions: The original draft of the paper was written by DA, who also conducted all the main statistical analyses. DA, IS, FU, and AE were responsible for both the design of the study and the interpretation of its results. The original draft was reviewed and edited by AE, FU, and IS, who made critical contributions to the paper's content. IS performed some of the secondary analyses. The project was supervised by AE and FU, who provided funding for this study. All authors approved the submitted version of this paper. AE had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of Interest Disclosures: All authors declare no conflict of interest.

Funding/Support: This research was supported by the Israel Science Foundation (grant No. 2755/20 to AE).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: We are grateful to all of the families in SPARK, the SPARK clinical sites and SPARK staff. We appreciate obtaining access to SPARK phenotypic collection (version 5) data on SFARI Base. Approved researchers can obtain the SPARK population dataset described in this study ([SPARK Phenotype Dataset(https://www.sfari.org/resource/spark/) by applying at https://base.sfari.org. Additionally, we would like to thank the Uzefovsky and Eran lab members for engaging in valuable and fruitful discussions.

References

1. Maenner MJ, Warren Z, Williams AR, et al. Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR Surveill Summ*. 2023;72(2):1-14. doi:10.15585/mmwr.ss7202a1

2. Towle PO, Patrick PA, Ridgard T, Pham S, Marrus J. Is Earlier Better? The Relationship between Age When Starting Early Intervention and Outcomes for Children with Autism Spectrum Disorder: A Selective Review. *Autism Research and Treatment*. 2020:7605876. doi:10.1155/2020/7605876

3. Orinstein AJ, Helt M, Troyb E, et al. Intervention for Optimal Outcome in Children and Adolescents with a History of Autism. *Journal of Developmental & Behavioral Pediatrics*. 2014;35(4):247-256. doi:10.1097/DBP.0000000000000037

4. Eapen V, Črnčec R, Walter A. Clinical outcomes of an early intervention program for preschool children with Autism Spectrum Disorder in a community group setting. *BMC Pediatrics*. 2013;13(1):3. doi:10.1186/1471-2431-13-3

5. Ben Itzchak E, Zachor DA. Who benefits from early intervention in autism spectrum disorders? *Research in Autism Spectrum Disorders*. 2011;5(1):345-350. doi:10.1016/j.rasd.2010.04.018

6. Vivanti G, Dissanayake C, The Victorian AT. Outcome for Children Receiving the Early Start Denver Model Before and After 48 Months. *Journal of Autism and Developmental Disorders*. 2016;46(7):2441-2449. doi:10.1007/s10803-016-2777-6

7. Jensen CM, Steinhausen H-C, Lauritsen MB. Time Trends Over 16 Years in Incidence-Rates of Autism Spectrum Disorders Across the Lifespan Based on

Nationwide Danish Register Data. *Journal of Autism and Developmental Disorders*. 2014;44(8):1808-1818. doi:10.1007/s10803-014-2053-6

8. Avlund SH, Thomsen PH, Schendel D, Jørgensen M, Carlsen AH, Clausen L. Factors Associated with a Delayed Autism Spectrum Disorder Diagnosis in Children Previously Assessed on Suspicion of Autism. *Journal of Autism and Developmental Disorders*. 2021;51(11):3843-3856. doi:10.1007/s10803-020-04849-x

9. Shattuck PT, Durkin M, Maenner M, et al. Timing of identification among children with an autism spectrum disorder: Findings from a population-based surveillance study. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2009;48(5):474-483. doi:10.1097/CHI.0b013e31819b3848

10. Rosenberg RE, Landa R, Law JK, Stuart EA, Law PA. Factors Affecting Age at Initial Autism Spectrum Disorder Diagnosis in a National Survey. *Autism Research and Treatment*. 2011;2011:874619. doi:10.1155/2011/874619

11. Valicenti-McDermott M, Hottinger K, Seijo R, Shulman L. Age at Diagnosis of Autism Spectrum Disorders. *The Journal of Pediatrics*. 2012;161(3):554-556. doi:10.1016/j.jpeds.2012.05.012

12. Daniels AM, Mandell DS. Explaining differences in age at autism spectrum disorder diagnosis: A critical review. *Autism*. 2014;18(5):583-597. doi:10.1177/1362361313480277

13. Mandell DS, Novak MM, Zubritsky CD. Factors associated with age of diagnosis among children with autism spectrum disorders. *Pediatrics*. 2005;116(6):1480–1486. doi:10.1542/peds.2005-0185

14. Coo H, Ouellette-Kuntz H, Lam M, et al. Correlates of age at diagnosis of autism spectrum disorders in six Canadian regions. *Chronic Dis Inj Can*. 2012;32(2):90-100.

15. Jónsdóttir SL, Saemundsen E, Antonsdóttir IS, Sigurdardóttir S, Ólason D. Children diagnosed with autism spectrum disorder before or after the age of 6 years. *Research in Autism Spectrum Disorders*. 2011;5(1):175-184. doi:10.1016/j.rasd.2010.03.007

16. Larsen K. The Early Diagnosis of Preschool Children with Autism Spectrum Disorder in Norway: a Study of Diagnostic Age and Its Associated Factors. *Scandinavian Journal of Child and Adolescent Psychiatry and Psychology*. 2015;3(2):136-145. doi:doi:10.21307/sjcapp-2015-014

17. Masi A, DeMayo MM, Glozier N, Guastella AJ. An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neuroscience Bulletin*. 2017;33(2):183-193. doi:10.1007/s12264-017-0100-y

18. Wiśniowiecka-Kowalnik B, Nowakowska BA. Genetics and epigenetics of autism spectrum disorder—current evidence in the field. *Journal of Applied Genetics*. 2019;60(1):37-47. doi:10.1007/s13353-018-00480-w

19. Hansen BH, Oerbeck B, Skirbekk B, Petrovski BÉ, Kristensen H. Neurodevelopmental disorders: prevalence and comorbidity in children referred to mental health services. *Nordic Journal of Psychiatry*. 2018;72(4):285-291. doi:10.1080/08039488.2018.1444087

20. Mpaka DM, Okitundu DL, Ndjukendi AO, et al. Prevalence and comorbidities of autism among children referred to the outpatient clinics for neurodevelopmental disorders. *Pan Afr Med J*. 2016;25:82. doi:10.11604/pamj.2016.25.82.4151

21. Hirschberger RG, Kuban KCK, O'Shea TM, et al. Co-occurrence and Severity of Neurodevelopmental Burden (Cognitive Impairment, Cerebral Palsy, Autism Spectrum Disorder, and Epilepsy) at Age Ten Years in Children Born Extremely

Preterm. *Pediatric Neurology*. 2018;79:45-52. doi:10.1016/j.pediatrneurol.2017.11.002

22. Sacco R, Gabriele S, Persico AM. Head circumference and brain size in autism spectrum disorder: A systematic review and meta-analysis. *Psychiatry Research: Neuroimaging*. 2015;234(2):239-251. doi:10.1016/j.pscychresns.2015.08.016

23. Hill AP, Zuckerman KE, Fombonne E. Obesity and Autism. *Pediatrics*. 2015;136(6):1051-1061. doi:10.1542/peds.2015-1437

24. Curtin C, Jojic M, Bandini LG. Obesity in children with autism spectrum disorder. *Harv Rev Psychiatry*. 2014;22(2):93-103. doi:10.1097/hrp.0000000000000031

25. Viscidi EW, Triche EW, Pescosolido MF, et al. Clinical Characteristics of Children with Autism Spectrum Disorder and Co-Occurring Epilepsy. *PLOS ONE*. 2013;8(7):e67797. doi:10.1371/journal.pone.0067797

26. Sansa G, Carlson C, Doyle W, et al. Medically refractory epilepsy in autism. *Epilepsia*. 2011;52(6):1071-1075. doi:10.1111/j.1528-1167.2011.03069.x

27. Xiong J, Chen S, Pang N, et al. Neurological Diseases With Autism Spectrum Disorder: Role of ASD Risk Genes. Original Research. *Frontiers in Neuroscience*. 2019;13:349. doi:10.3389/fnins.2019.00349

28. Sigmon ER, Kelleman M, Susi A, Nylund CM, Oster ME. Congenital Heart Disease and Autism: A Case-Control Study. *Pediatrics*. 2019;144(5):e20184114. doi:10.1542/peds.2018-4114

29. Bean Jaworski JL, Flynn T, Burnham N, et al. Rates of autism and potential risk factors in children with congenital heart defects. *Congenital Heart Disease*. 2017;12(4):421-429. doi:10.1111/chd.12461

30. Rotem RS, Chodick G, Davidovitch M, Hauser R, Coull BA, Weisskopf MG. Congenital Abnormalities of the Male Reproductive System and Risk of Autism Spectrum Disorders. *American Journal of Epidemiology*. 2018;187(4):656-663. doi:10.1093/aje/kwx367

31. Mazefsky CA, Oswald DP, Day TN, Eack SM, Minshew NJ, Lainhart JE. ASD, a Psychiatric Disorder, or Both? Psychiatric Diagnoses in Adolescents with High-Functioning ASD. *Journal of Clinical Child & Adolescent Psychology*. 2012;41(4):516-523. doi:10.1080/15374416.2012.686102

32. Nelson C, Bruce SM. Children Who Are Deaf/Hard of Hearing with Disabilities: Paths to Language and Literacy. *Education Sciences*. 2019;9(2):134. doi:10.3390/educsci9020134

33. Miodovnik A, Harstad E, Sideridis G, Huntington N. Timing of the Diagnosis of Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder. *Pediatrics*. 2015;136(4):e830-e837. doi:10.1542/peds.2015-1502

34. Ozonoff S, Young GS, Brian J, et al. Diagnosis of Autism Spectrum Disorder After Age 5 in Children Evaluated Longitudinally Since Infancy. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2018;57(11):849-857.e2. doi:10.1016/j.jaac.2018.06.022

35. Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T, Linstead E. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International Journal of Medical Informatics*. 2019;129:29-36. doi:10.1016/j.ijmedinf.2019.05.006

36. Hong S-J, Valk SL, Di Martino A, Milham MP, Bernhardt BC. Multidimensional Neuroanatomical Subtyping of Autism Spectrum Disorder. *Cerebral Cortex*. 2017;28(10):3578-3588. doi:10.1093/cercor/bhx229

37. Feczko E, Balba NM, Miranda-Dominguez O, et al. Subtyping cognitive profiles in Autism Spectrum Disorder using a Functional Random Forest algorithm. *NeuroImage*. 2018;172:674-688. doi:10.1016/j.neuroimage.2017.12.044

38. Yaneva V, Ha LA, Eraslan S, Yesilada Y, Mitkov R. Detecting High-Functioning Autism in Adults Using Eye Tracking and Machine Learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2020;28(6):1254-1261. doi:10.1109/TNSRE.2020.2991675

39. Liu W, Li M, Yi L. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*. 2016;9(8):888-898. doi:10.1002/aur.1615

40. Kentrou V, de Veld DM, Mataw KJ, Begeer S. Delayed autism spectrum disorder recognition in children and adolescents previously diagnosed with attention-deficit/hyperactivity disorder. *Autism*. 2019;23(4):1065-1072. doi:10.1177/1362361318785171

41. Khemir S, Halayem S, Azzouz H, et al. Autism in Phenylketonuria Patients: From Clinical Presentation to Molecular Defects. *J Child Neurol*. 2016;31(7):843-9. doi:10.1177/0883073815623636

42. Thurm A, Tierney E, Farmer C, et al. Development, behavior, and biomarker characterization of Smith-Lemli-Opitz syndrome: an update. *Journal of Neurodevelopmental Disorders*. 2016;8(1):12. doi:10.1186/s11689-016-9145-x

43. Anwar A, Abruzzo PM, Pasha S, et al. Advanced glycation endproducts, dityrosine and arginine transporter dysfunction in autism - a source of biomarkers for clinical diagnosis. *Molecular Autism*. 2018;9(1):3. doi:10.1186/s13229-017-0183-3

44. Ardalan A, Assadi AH, Surgent OJ, Travers BG. Whole-Body Movement during Videogame Play Distinguishes Youth with Autism from Youth with Typical Development. *Sci Rep*. 2019;9(1):20094. doi:10.1038/s41598-019-56362-6

45. Feliciano P, Daniels AM, Snyder LG, et al. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron*. 2018;97(3):488-493. doi:10.1016/j.neuron.2018.01.015

46. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. *Pediatrics*. 2014;133(1):e54-e63. doi:10.1542/peds.2013-0819

47. Rutter ML, Bailey A, Lord C, Bailey A, Bailey P, Anthony BA. *The Social Communication Questionnaire Manual*. Western Psychological Services; 2003.

48. Bodfish JW, Symons FJ, Parker DE, Lewis MH. Varieties of Repetitive Behavior in Autism: Comparisons to Mental Retardation. *Journal of Autism and Developmental Disorders*. 2000;30(3):237-243. doi:10.1023/A:1005596502855

49. Wilson BN, Kaplan BJ, Crawford SG, Campbell A, Dewey D. Reliability and Validity of a Parent Questionnaire on Childhood Motor Skills. *The American Journal of Occupational Therapy*. 2000;54(5):484-493. doi:10.5014/ajot.54.5.484

50. McKinney W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* 2010:51–56.

51. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

52. Georgiades S, Tait PA, McNicholas PD, et al. Trajectories of Symptom Severity in Children with Autism: Variability and Turning Points through the Transition to School. *Journal of Autism and Developmental Disorders*. 2022;52(1):392-401. doi:10.1007/s10803-021-04949-2

53. Fuller EA, Kaiser AP. The Effects of Early Intervention on Social Communication Outcomes for Children with Autism Spectrum Disorder: A Meta-analysis. *Journal of Autism and Developmental Disorders*. 2020;50(5):1683-1700. doi:10.1007/s10803-019-03927-z

54. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.

55. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

56. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*. 2010.

57. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 2002;16:321-357. doi:10.1613/jair.953

58. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017;18(1):559–563.

59. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30.

60. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using networkx. presented at: 7th Python in Science Conf; 2008; United States. Accessed May 31, 2023. https://www.osti.gov/biblio/960616

61. Shaw KA, Bilder DA, McArthur D, et al. Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *Morbidity and*

mortality weekly report Surveillance summaries. 2023;72(1):1-15. doi:10.15585/mmwr.ss7201a1 Accessed July 26, 2024.

62. Levy SE, Giarelli E, Lee L-C, et al. Autism Spectrum Disorder and Co-occurring Developmental, Psychiatric, and Medical Conditions Among Children in Multiple Populations of the United States. *Journal of Developmental & Behavioral Pediatrics*. 2010;31(4):267-275. doi:10.1097/DBP.0b013e3181d5d03b

63. Sicherman N, Charite J, Eyal G, et al. Clinical signs associated with earlier diagnosis of children with autism Spectrum disorder. *BMC Pediatrics*. 2021;21(1):96. doi:10.1186/s12887-021-02551-0

64. Lehnhardt F-G, Falter CM, Gawronski A, et al. Sex-Related Cognitive Profile in Autism Spectrum Disorders Diagnosed Late in Life: Implications for the Female Autistic Phenotype. *Journal of Autism and Developmental Disorders*. 2016;46(1):139-154. doi:10.1007/s10803-015-2558-7

65. Brett D, Warnell F, Mcconachie H, Parr JR. Factors Affecting Age at ASD Diagnosis in UK: No Evidence that Diagnosis Age has Decreased Between 2004 and 2014. *Journal of Autism and Developmental Disorders*. 2016;46:1974–1984. doi:10.1007/s10803-016-2716-6

66. Polyak A, Kubina RM, Girirajan S. Comorbidity of intellectual disability confounds ascertainment of autism: implications for genetic diagnosis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2015;168(7):600-608. doi:10.1002/ajmg.b.32338

67. Montazeri F, Buitelaar JK, Oosterling IJ, de Bildt A, Anderson GM. Network Structure of Autism Spectrum Disorder Behaviors and Its Evolution in Preschool Children: Insights from a New Longitudinal Network Analysis Method. *Journal of*

Autism and Developmental Disorders. 2022;53:4293-4307. doi:10.1007/s10803-022-05723-8

68. Lovato DV, Herai RR, Pignatari GC, Beltrão-Braga PCB. The Relevance of Variants With Unknown Significance for Autism Spectrum Disorder Considering the Genotype–Phenotype Interrelationship. *Frontiers in Psychiatry*. 2019;10:60. doi:10.3389/fpsyt.2019.00409

69. Taboas A, Doepke K, Zimmerman C. Preferences for identity-first versus person-first language in a US sample of autism stakeholders. *Autism*. 2023;27(2):565-570. doi:10.1177/13623613221130845

Figure legends

Figure 1. Behavioral differences between groups of individuals with delayed ASD diagnosis. Differences between groups of individuals with delayed ASD diagnosis (D1, D2) and those with a timely one, in (**A**) core autism domains, (**B**) motor development, and (**C**) common symptoms. The color of each circle represents the effect size measured by the log ratio of a feature's mean in each group as compared to the feature's mean among timely diagnosed individuals. The size of the circles is proportional to $\sqrt{-log(P\ value)}$. Thus, the larger the circle, the smaller the P value. The lowest possible P value in this analysis is 6.87×10^{-278}. Circles are shown only for significant results.

Figure 2. Differences in co-occurring neurodevelopmental conditions between groups of individuals with delayed ASD diagnosis. (**A**) Differences in cognitive and motor delays; (**B**) differences in growth alterations; (**C**) Differences in sensory and neural problems; and **(D)** Differences in neuropsychiatric conditions. The color of each circle represents the effect size measured by the log ratio of a feature's mean in each group as compared to the feature's mean among timely diagnosed individuals. The size of the circles is proportional to $\sqrt{-log(P\ value)}$. Thus, the larger the circle, the smaller the P value. The lowest possible P value in this analysis is 6.87×10^{-278}. Circles are shown only for significant results.

Figure 3. Cumulative distributions of co-occurring conditions.

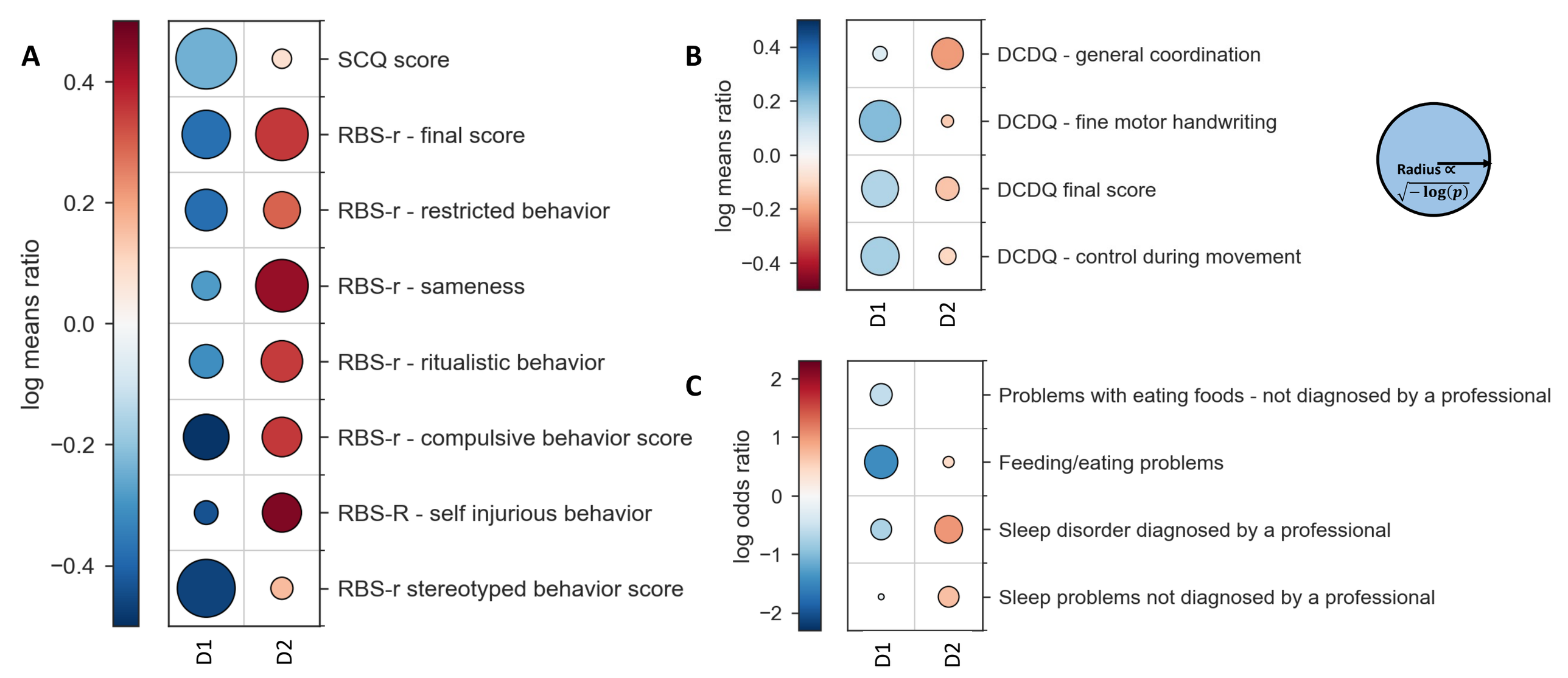
Cumulative distributions of the number of co-occurring conditions in individuals of each group. Horizontal lines depict the mean number of co-occurring conditions in each group.

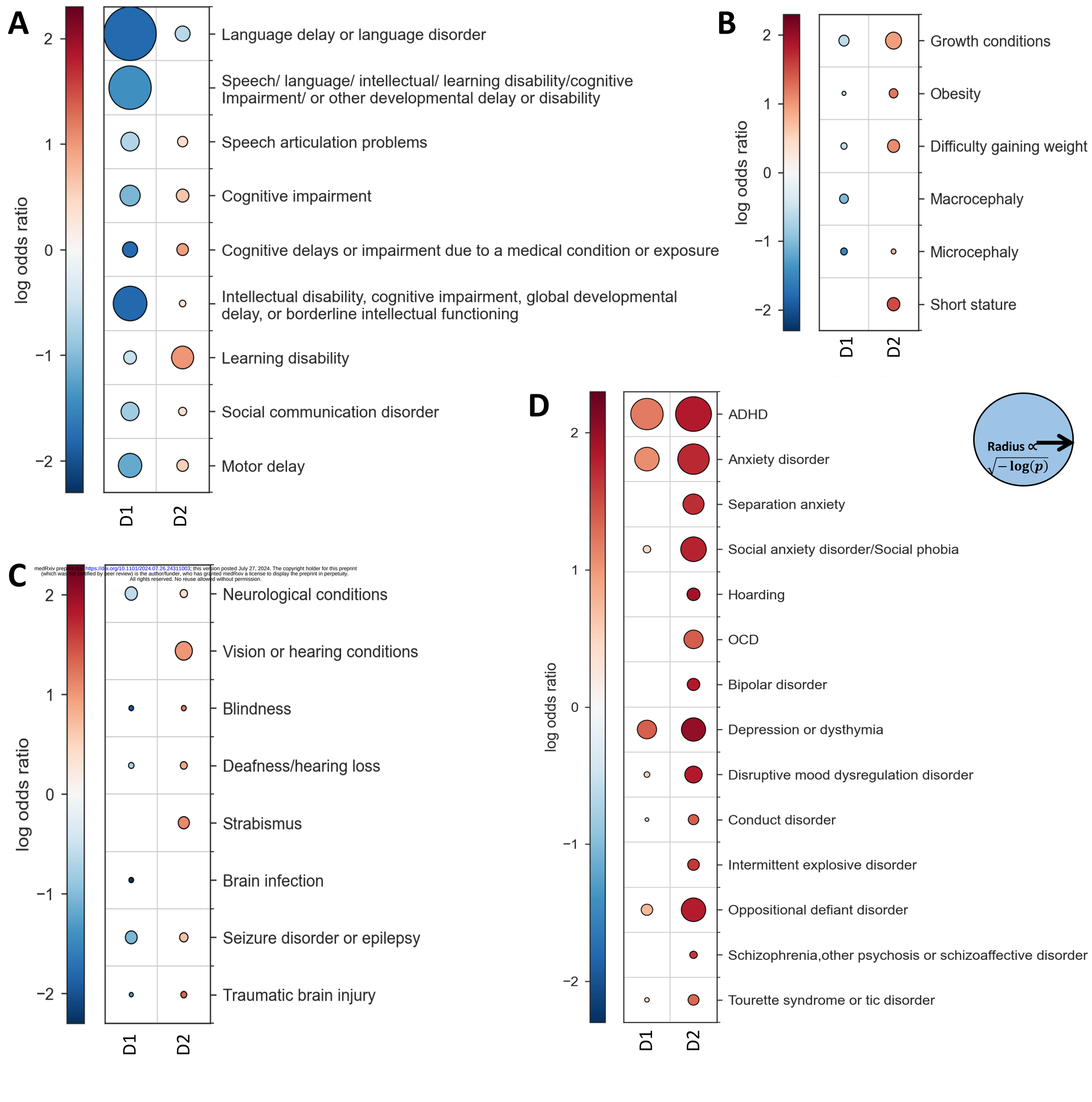
Figure 4. Feature importance using Shapley values.

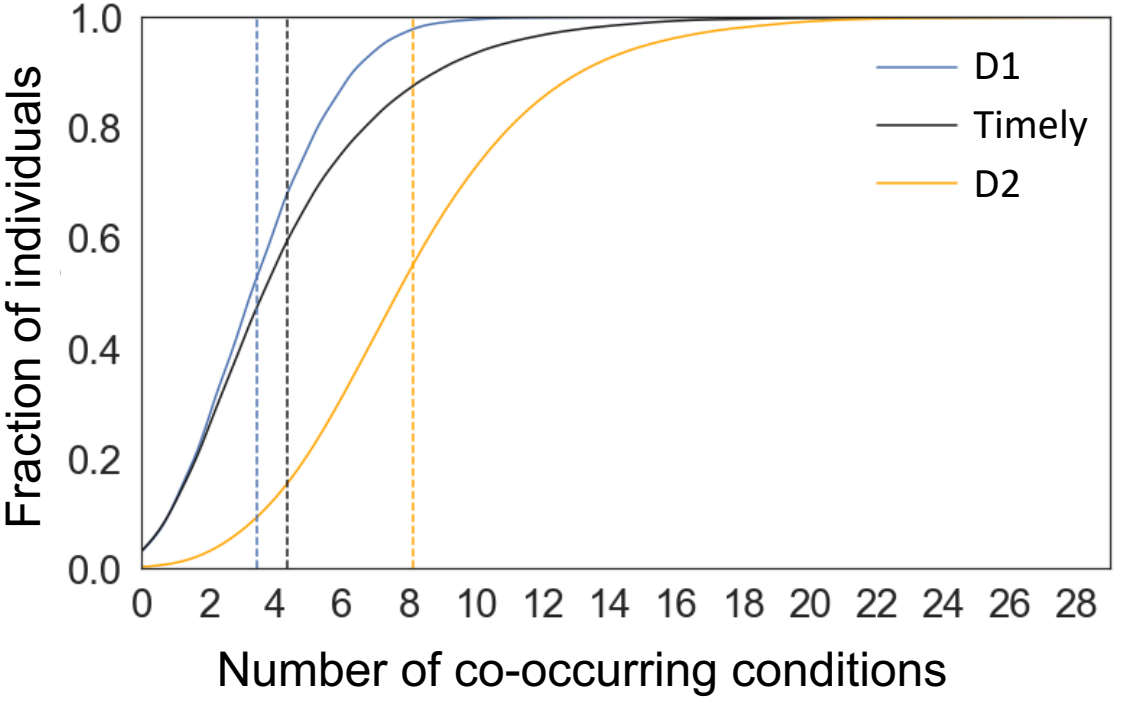
Most distinguishing features of (**A**) D1, and (**B**) D2, based on Shapley values. Each dot in the plot corresponds to a sample. The features are ordered on the y-axis by their total impact on the model output. The x-axis shows the relative impact on the model output for a given sample. The color shows the relative sample value for every feature.

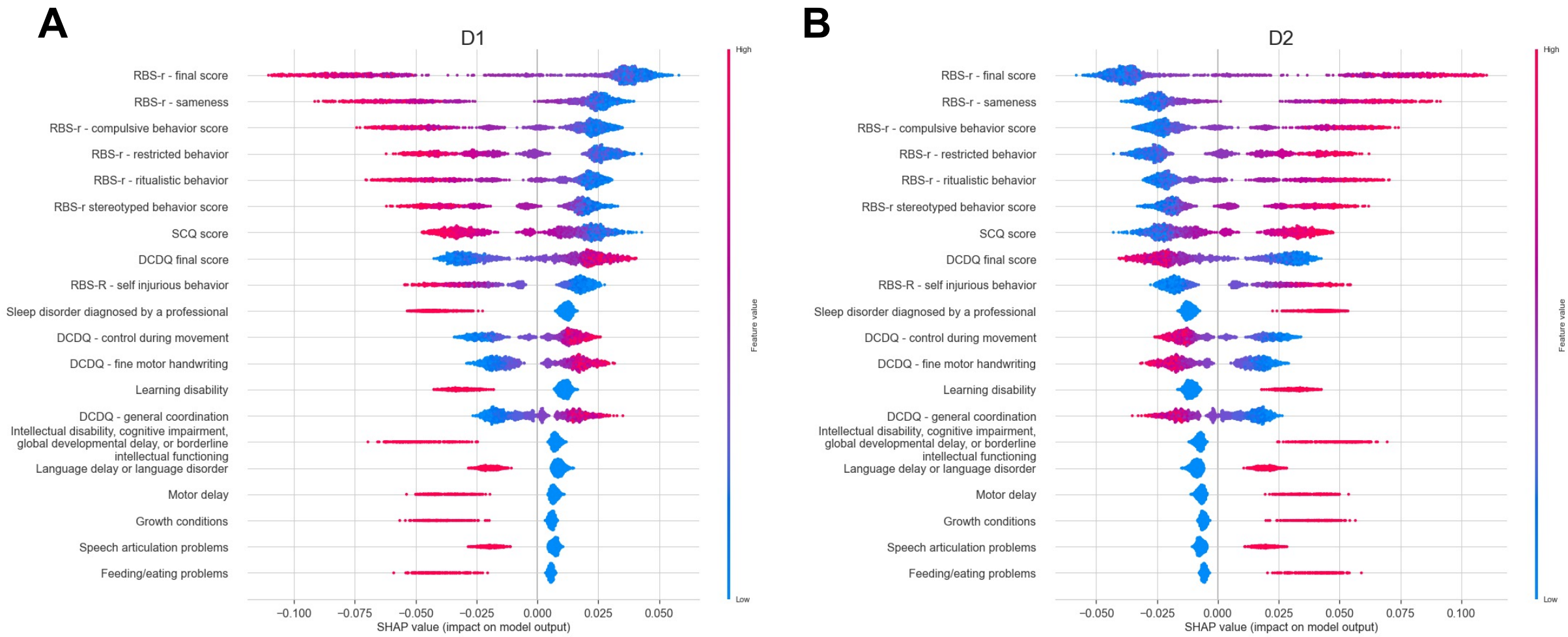
Figure 5. Feature interconnectivity and network structure of timely and delayed diagnosis groups.

Normalized clique membership is depicted for the top five most connected features in (**A**) D1, (**B**) D2, and (**C**) the timely diagnosis group, T. Correlation heatmaps of the variables most strongly correlated with the top connected features are presented for (**D**) D1, and (**E**) D2. The top connected features are denoted in bold.


A
log means ratio
0.4
0.2
0.0
−0.2
−0.4
SCQ score
RBS-r - final score
RBS-r - restricted behavior
RBS-r - sameness
RBS-r - ritualistic behavior
RBS-r - compulsive behavior score
RBS-R - self injurious behavior
RBS-r stereotyped behavior score
D1
D2
B
log means ratio
0.4
0.2
0.0
−0.2
−0.4
DCDQ - general coordination
DCDQ - fine motor handwriting
DCDQ final score
DCDQ - control during movement
D1
D2
Radius ∝ $\sqrt{-\log(p)}$
C
log odds ratio
2
1
0
−1
−2
Problems with eating foods - not diagnosed by a professional
Feeding/eating problems
Sleep disorder diagnosed by a professional
Sleep problems not diagnosed by a professional
D1
D2




1.0
0.8
0.6
0.4
0.2
0.0
Fraction of individuals
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
Number of co-occurring conditions
D1
Timely
D2


A
D1
RBS-r - final score
RBS-r - sameness
RBS-r - compulsive behavior score
RBS-r - restricted behavior
RBS-r - ritualistic behavior
RBS-r stereotyped behavior score
SCQ score
DCDQ final score
RBS-R - self injurious behavior
Sleep disorder diagnosed by a professional
DCDQ - control during movement
DCDQ - fine motor handwriting
Learning disability
DCDQ - general coordination
Intellectual disability, cognitive impairment, global developmental delay, or borderline intellectual functioning
Language delay or language disorder
Motor delay
Growth conditions
Speech articulation problems
Feeding/eating problems
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
SHAP value (impact on model output)
High
Feature value
Low
B
D2
RBS-r - final score
RBS-r - sameness
RBS-r - compulsive behavior score
RBS-r - restricted behavior
RBS-r - ritualistic behavior
RBS-r stereotyped behavior score
SCQ score
DCDQ final score
RBS-R - self injurious behavior
Sleep disorder diagnosed by a professional
DCDQ - control during movement
DCDQ - fine motor handwriting
Learning disability
DCDQ - general coordination
Intellectual disability, cognitive impairment, global developmental delay, or borderline intellectual functioning
Language delay or language disorder
Motor delay
Growth conditions
Speech articulation problems
Feeding/eating problems
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
SHAP value (impact on model output)
High
Feature value
Low