Candidate diagnostic biomarkers for neurodevelopmental disorders in children and adolescents: a systematic review

Samuele Cortese[1-5], Marco Solmi[1,6-9], Giorgia Michelini[10,11], Alessio Bellato[12], Christina Blanner[13], Andrea Canozzi[14], Luis Eudave[15], Luis C. Farhat[16], Mikkel Højlund[17,18], Ole Köhler-Forsberg[19,20], Douglas Teixeira Leffa[21,22], Christopher Rohde[20,23], Gonzalo Salazar de Pablo[24-27], Giovanni Vita[14], Rikke Wesselhoeft[18,28], Joanna Martin[29], Sarah Baumeister[30], Natali S. Bozhilova[24,31], Christina O. Carlisi[32], Virginia Carter Leno[33], Dorothea L. Floris[34,35], Nathalie E. Holz[30,35-37], Eline J. Kraaijenvanger[30], Seda Sacu[30], Isabella Vainieri[38], Giovanni Ostuzzi[14], Corrado Barbui[14], Christoph U. Correll[9,39-41]

[1]Centre for Innovation in Mental Health, School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, UK; [2]Clinical and Experimental Sciences (CNS and Psychiatry), Faculty of Medicine, University of Southampton, Southampton, UK; [3]Solent NHS Trust, Southampton, UK; [4]Hassenfeld Children's Hospital at NYU Langone, New York University Child Study Center, New York, NY, USA; [5]Division of Psychiatry and Applied Psychology, School of Medicine, University of Nottingham, Nottingham, UK; [6]Department of Psychiatry, University of Ottawa, Ottawa, ON, Canada; [7]Department of Mental Health, Ottawa Hospital, Ottawa, ON, Canada; [8]Ottawa Hospital Research Institute (OHRI) Clinical Epidemiology Program, University of Ottawa, Ottawa, ON, Canada; [9]Department of Child and Adolescent Psychiatry, Charité Universitätsmedizin, Berlin, Germany; [10]Department of Biological & Experimental Psychology, School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK; [11]Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles (UCLA), Los Angeles, CA, USA; [12]School of Psychology, University of Nottingham, Semenyih, Malaysia; [13]Mental Health Center, Glostrup, Copenhagen University Hospital - Mental Health Services CPH, Copenhagen, Denmark; [14]Department of Neuroscience, Biomedicine, and Movement Sciences, Section of Psychiatry, University of Verona, Verona, Italy; [15]Faculty of Education and Psychology, University of Navarra, Pamplona, Spain; [16]Department of Psychiatry, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil; [17]Department of Psychiatry Aabenraa, Mental Health Services in the Region of Southern Denmark, Aabenraa, Denmark; [18]Clinical Pharmacology, Pharmacy, and Environmental Medicine, Department of Public Health, University of Southern Denmark, Odense, Denmark; [19]Psychosis Research Unit, Aarhus University Hospital - Psychiatry, Aarhus, Denmark; [20]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; [21]ADHD Outpatient Program & Development Psychiatry Program, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil; [22]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA; [23]Department of Affective Disorders, Aarhus University Hospital - Psychiatry, Aarhus, Denmark; [24]Department of Child and Adolescent Psychiatry, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [25]Child and Adolescent Mental Health Services, South London and Maudsley NHS Foundation Trust, London, UK; [26]Early Psychosis: Interventions and Clinical-detection (EPIC) Lab, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [27]Institute of Psychiatry and Mental Health, Department of Child and Adolescent Psychiatry, Hospital General Universitario Gregorio Marañón School of Medicine, Universidad Complutense, CIBERSAM, Madrid, Spain; [28]Child and Adolescent Mental Health Odense, Mental Health Services in the Region of Southern Denmark, Odense, Denmark; [29]MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK; [30]Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany; [31]School of Psychology, University of Surrey, Guilford, UK; [32]Division of Psychology and Language Sciences, University College London, London, UK; [33]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [34]Department of Psychology, University of Zurich, Zurich, Switzerland; [35]Donders Institute for Brain, Cognition, and Behavior, Radboud University Nijmegen, Nijmegen, The Netherlands; [36]Department for Cognitive Neuroscience, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands; [37]Institute of Medical Psychology and Medical Sociology, University Medical Center Schleswig Holstein, Kiel University, Kiel, Germany; [38]Department of Clinical, Educational and Health Psychology, University College London, London, UK; [39]Psychiatry Research, Northwell Health, Zucker Hillside Hospital, New York, NY, USA; [40]Department of Psychiatry and Molecular Medicine, Zucker School of Medicine, Hempstead, NY, USA; [41]Center for Neuroscience, Feinstein Institute for Medical Research, Manhasset, NY, USA

Neurodevelopmental disorders – including attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder, communication disorders, intellectual disability, motor disorders, specific learning disorders, and tic disorders – manifest themselves early in development. Valid, reliable and broadly usable biomarkers supporting a timely diagnosis of these disorders would be highly relevant from a clinical and public health standpoint. We conducted the first systematic review of studies on candidate diagnostic biomarkers for these disorders in children and adolescents. We searched Medline and Embase + Embase Classic with terms relating to biomarkers until April 6, 2022, and conducted additional targeted searches for genome-wide association studies (GWAS) and neuroimaging or neurophysiological studies carried out by international consortia. We considered a candidate biomarker as promising if it was reported in at least two independent studies providing evidence of sensitivity and specificity of at least 80%. After screening 10,625 references, we retained 780 studies (374 biochemical, 203 neuroimaging, 133 neurophysiological and 65 neuropsychological studies, and five GWAS), including a total of approximately 120,000 cases and 176,000 controls. While the majority of the studies focused simply on associations, we could not find any biomarker for which there was evidence – from two or more studies from independent research groups, with results going into the same direction – of specificity and sensitivity of at least 80%. Other important metrics to assess the validity of a candidate biomarker, such as positive predictive value and negative predictive value, were infrequently reported. Limitations of the currently available studies include mostly small sample size, heterogeneous approaches and candidate biomarker targets, undue focus on single instead of joint biomarker signatures, and incomplete accounting for potential confounding factors. Future multivariable and multi-level approaches may be best suited to find valid candidate biomarkers, which will then need to be validated in external, independent samples and then, importantly, tested in terms of feasibility and cost-effectiveness, before they can be implemented in daily clinical practice.

Key words: Biological markers, neurodevelopmental disorders, ADHD, autism spectrum disorder, communication disorders, intellectual disability, motor disorders, specific learning disorders, tic disorders, genome-wide association studies, neuroimaging, neurophysiology

(World Psychiatry 2023;22:129–149)

Limitations related to the subjective nature of psychiatric diagnoses have prompted, in the past decades, several lines of investigation aimed at identifying valid biomarkers that can assist in the diagnosis, prediction, prognosis and management of mental health conditions.

According to the US Food and Drug Administration (FDA) - National Institute of Health (NIH) Biomarker Working Group, a *biomarker* is defined as "a characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention"[1]. Based on their main clinical application, biomarkers can be grouped as: a) *diagnostic*, used to detect or confirm the presence of a disease or medical condition or to identify homogeneous subtypes of the disease; b) *monitoring*, to monitor the status of a disease and the

response to a treatment; c) *pharmacodynamic*, to evaluate the response to a clinical intervention; d) *predictive*, to predict the probability to develop any effect following a clinical intervention; e) *prognostic*, to identify the probability of developing a clinical event in individuals with a disease or a clinical condition; f) *safety*, to evaluate the probability of developing an adverse event following an intervention; and g) *susceptibility/risk*, to quantify the risk of an individual to develop a disease or medical condition[2].

Valid and usable at scale biomarkers, if identified, promise to allow the clinical implementation of precision medicine in psychiatry[2-7], whereby: a) individual patients would receive the proper diagnosis, and therefore proper treatment, more quickly; b) they would be matched more accurately to the treatments they are most likely to respond to; c) treatment could be started before symptoms reach a severe level and/or lead to dysfunction, increasing the likelihood of expedited recovery; d) clinicians could more easily identify who is most at risk for relapse and recurrence.

However, the path for the identification of a biological characteristic as a valid biomarker in real-world clinical settings is a long one, and needs to follow rigorous steps. The biomarker needs first to be *sensitive*, i.e., accurately identify as positive those individuals who have the outcome of interest, and *specific*, namely, accurately label as negative those individuals who do not have the outcome of interest. Although there are no established benchmarks for these metrics, quantitative measures that allow diagnostic accuracy with at least 80% sensitivity and 80% specificity are often considered as clinically useful[8].

The consensus report by the American Psychiatric Association (APA) Work Group on Neuroimaging Markers of Psychiatric Disorders suggested that a promising biomarker should have two or more independent well-powered studies providing evidence of sensitivity and specificity at least of 80%[9]. In addition, a biomarker would need to: a) have good *positive predictive value* (PPV), which refers to the proportion of individuals who have the outcome of interest among those who tested positive; b) have good *negative predictive value* (NPV), indicating the proportion of individuals who do not have the outcome of interest among those who tested negative; c) have good *internal validity*, i.e., measure the intended feature in an unbiased way, without relevant influence of confounding factors; d) be *externally valid*, so that the results of the studies assessing the candidate biomarker are generalizable to the population of interest in real-life clinical settings; and e) be reliable, in terms of *test-retest reliability* (i.e., being consistent with itself when measured on several occasions) and *inter-rater reliability* (i.e., being consistent when measured across different raters)[10]. Furthermore, a biomarker should change in a dynamic and reliable way in relation to the progress/change of the clinical condition[2].

Steps for biomarker discovery should therefore include an initial phase where a clinically relevant question is identified; a phase testing internal validity, ruling out the possible role of confounding factors; a subsequent phase where external validity is tested, assessing PPV and NPV in independent, targeted samples; and a last phase where the biomarker is tested to assess whether it brings a significant benefit in relation to standard clinical practice, with acceptable number needed to assess (NNA) and number needed to treat (NNT), i.e. the number of individuals that should be assessed or treated in order to benefit one additional individual compared to those who are not assessed or treated. Crucially, this last phase should also assess if the biomarker is cost-effective in relation to standard practice[10].

Based on the pathophysiological overlap across disorders, it has been suggested that at least some of the candidate biomarkers may have a transdiagnostic nature across mental health conditions[11]. However, for at least some peripheral biomarkers, it is possible that their transdiagnostic nature be related to the chronic stress or allostatic load associated with a variety of psychiatric conditions[12]. The notion of *transdiagnosticity* of peripheral biomarkers has been supported by a systematic review showing that, out of the six molecules most commonly referred to as "biomarkers" in studies of schizophrenia, major depressive disorder and bipolar disorder, five – brain-derived neurotrophic factor (BDNF), tumor necrosis factor (TNF)-alpha, interleukin (IL)-6, C-reactive protein (CRP), and cortisol – were proposed across these disorders[12], even though without a rigorous transdiagnostic framework. Furthermore, a systematic review and meta-analysis of electrophysiological correlates of performance monitoring in four common childhood disorders – attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), Tourette's syndrome, and obsessive-compulsive disorder – found a significant overlap in electrophysiological correlates across these disorders[13,14].

Recent umbrella reviews have shown that, in the case of many putative biomarkers for ASD and ADHD, most meta-analyses claiming significant associations were likely inflated by high risk of bias, including excess of significance bias[15-17]. By pooling different studies and increasing power, meta-analyses frequently find significant results. However, in this specific field, what determines the credibility of a diagnostic biomarker is replication of findings in terms of specificity, sensitivity, accuracy, and predictive value[9], rather than a pooled effect size of association. Hence, a systematic review accounting for these variables is needed. In the present systematic review, we focus on diagnostic biomarkers of neurodevelopmental disorders, alongside oppositional defiant disorder (ODD) and conduct disorders (CD), in children and adolescents.

Neurodevelopmental disorders is an umbrella term encompassing a broad range of conditions characterized by impaired development of cognitive, social or motor functions, or atypical functioning, usually manifesting themselves from early childhood, and having a steady course without marked remissions or relapses[18,19]. The conceptualization and grouping of these disorders have changed over time and are still a matter of debate. Currently, the ICD-11[20] includes ADHD, ASD, communication disorders, intellectual disability, motor disorders, specific learning disorders (involving reading, writing and arithmetic), and tic disorders.

Neurodevelopmental disorders are highly heterogeneous in terms of their epidemiology[21], clinical characteristics, causes[22], burden, treatment responses and tolerability[23,24], and outcomes[25]. Notably, ODD and CD are often comorbid with neurodevelopmental disorders, in particular ADHD[26].

The level of overlap between neurodevelopmental disorders

and their symptom dimensions is substantial. This is accounted for by shared or correlated risk factors, and common or overlapping molecular and neuronal mechanisms. While this co-occurrence supports the rationale for grouping these disorders together, from a clinical standpoint it is also relevant to recognize them as individual entities. Indeed, specific, distinct diagnostic categories allow clinicians to communicate about patients' characteristics with each other and with the patients and their family members/caregivers. Furthermore, patients with different categorical diagnoses respond to different treatments. For instance, psychostimulants are effective for ADHD, and so-called antipsychotics can decrease the severity of tics, but psychostimulants are not effective for tics, and antipsychotics do not improve attention regulation difficulties of ADHD.

While previous systematic reviews, meta-analyses or umbrella reviews have provided a synthesis of the evidence on specific biomarkers in specific disorders, for example on peripheral biomarkers in ADHD[16,27] or ASD[15], no systematic review has been conducted so far covering a broad range of biomarkers across neurodevelopmental disorders.

We aimed to fill this gap by conducting a systematic review of studies on promising candidate diagnostic biomarkers in children and/or adolescents with any neurodevelopmental disorder or with ODD or CD. We aimed to assess: a) which are the candidate biochemical, genetic, neuroimaging, neurophysiological and neuropsychological biomarkers that have been replicated across studies as being significantly associated with the diagnosis of specific neurodevelopmental disorders; b) how many of these biomarkers could be defined as *promising*, based on specificity and sensitivity at least of 80% in two or more independent studies; and c) for how many of these candidate biomarkers, internal as well as external validation – assessing sensitivity, specificity, PPV and NPV – have been implemented, alongside an evaluation of the cost-effectiveness of the biomarker; and d) to what extent biomarkers are disorder-specific or transdiagnostic.

METHODS

This systematic review was based on a pre-registered protocol (available at https://osf.io/wp4je/?view_only=8c349f45a9ac441490981acf946c8d9a) and was conducted in accordance with the 2020 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement[28].

Search

We searched Medline and Embase + Embase Classic, from inception until April 6, 2022. We did not apply any limit in terms of language or type of document. We used terms related to neurodevelopmental disorders (alongside ODD and CD) and "biomarker" or equivalent ("marker", "diagnostic test", and "endophenotype"), in order to retrieve studies assessing what the study authors deemed to be a potential biomarker. The exact search syntax is reported in the supplementary information.

Additionally, we searched for the largest genome-wide association studies (GWAS), as GWAS are typically based on meta-analyses of increasing numbers of samples and, as such, many previous smaller studies are sub-samples of the largest available GWAS, which will be best powered and use the latest methodologies and best practices. We also searched for neuroimaging or neurophysiological studies conducted by international consortia.

Inclusion/exclusion criteria

We included any observational study with a comparison group, assessing children or adolescents (mean age: 18 years or less) presenting with any (one or more) of the following disorders (reported here according to the ICD-11), provided that they were diagnosed using the ICD (9, 10 or 11) or the DSM (III, III-R, IV, IV-TR or 5): 6A00 Disorders of Intellectual Development; 6A01 Developmental Speech or Language Disorders; 6A01.0 Developmental Speech Sound Disorder; 6A01.1 Developmental Speech Fluency Disorder; 6A01.2 Developmental Language Disorder; 6A02 Autism Spectrum Disorder; 6A03 Developmental Learning Disorder; 6A04 Developmental Motor Coordination Disorder; 6A05 Attention Deficit Hyperactivity Disorder; 6A06 Stereotyped Movement Disorder; 6A0Y Other Specified Neurodevelopmental Disorder; 8A05.00 Tourette Syndrome; 8A05.01 Chronic Motor Tic Disorder; 8A05.02 Chronic Phonic Tic Disorder; 6C90 Oppositional Defiant Disorder; 6C91 Conduct-Dissocial Disorder.

For ASD, we also included studies with a diagnosis based on the Autism Diagnostic Observation Schedule (ADOS), that has shown acceptable diagnostic accuracy in research settings[29].

Study selection and data extraction

Two authors independently screened titles and abstracts, and any conflicts were resolved by a third senior author. All selected articles underwent full text screening by two authors independently, with conflicts resolved by consultation with a third senior author.

For each retained study, we extracted the following variables: first author, year of publication, design (cross-sectional or longitudinal), specific disorder(s) included, diagnostic criteria, number and age of cases and controls, percentage of males, percentage of White ethnicity individuals, type of biomarker(s), most adjusted effect size or p value, and inclusion of any of the following, when available: sensitivity, specificity, PPV, NPV, and receiver operating characteristic area under the curve (ROC AUC).

Study quality appraisal

We rated the quality of cross-sectional studies using BIOCROSS, an appraisal tool for cross-sectional studies using biomarker data (no tools for longitudinal studies of biomarkers are available)[30]. The following items were selected as the most appropriate for

the appraisal of studies of biochemical biomarkers: item 3 (3.1: "Was the sampling frame reported (study population source)?"; 3.2: "Was the participation rate reported (i.e., eligible persons at least 50%)?"; 3.3: "Was sample size justification or power description provided?"); item 4 (4.1: "Were the study population characteristics (i.e., demographic, clinical and social) presented?"; 4.2: "Were the exposures and potential confounders described?"; 4.3: "Were any missing values and strategies to deal with missing data reported?"); item 5 (5.1: "Did the authors clearly report statistical methods used to calculate estimates (e.g., Spearman, Pearson, linear regression)?"; 5.2: "Were key potential confounding variables measured and adjusted statistically in reported analyses?"; 5.3: "Was the raw effect size estimate (correlation coefficient, beta coefficient) or measure of study precision provided (e.g., confidence intervals, precise p value)?"); item 8 (8.1: "Were the measurement methods described (assay methods, preservation and storage, detailed protocol, including specific reagents or kits used)?"; 8.2: "Were the reproducibility assessments performed for evaluating biomarker stability?"; 8.3: "Were the quantitation methods well described?"); item 9 (9.1: "Was the laboratory/place of measurement mentioned?"; 9.2: "Were any quality control procedures and results reported (e.g., reported coefficient of variation)?"; 9.3: "Were the analyses blinded for laboratory staff?"). We selected items 3, 4, 5, 8 and 9, with exclusion of sub-items 4.2, 8.2, 9.1 and 9.3, for neuroimaging, neurophysiological and neuropsychological studies. We selected items 3, 4, 5 and 8, with exclusion of sub-item 8.3, for GWAS.

Synthesis of the evidence

We provided a qualitative synthesis of the included studies and of the level of transdiagnosticity. To assess promising biomarkers, we indicated first, when possible, the number and frequency of positive and negative replications (with the direction of the association, i.e. *increased* or *decreased*) for each biomarker assessed in at least two studies, with at least one positive finding in terms of significant associations. We then identified the biomarkers for which at least two studies reported on sensitivity, specificity, PPV, NPV and/or ROC AUC, and the biomarkers with a sensitivity and specificity of at least 80% replicated in at least two studies.

RESULTS

From an initial pool of 10,625 references, we retained 780 studies (see Figure 1, reporting the PRISMA 2020 flow chart[31]). The lists of included references and of those excluded, with reasons for exclusion after checking the full text, are reported in the supplementary information.

We present the findings in relation to each type of candidate biomarker (now onwards, for simplicity, referred to as "biomarker"), based on the primary outcome of the study (for instance, a study assessing a neurophysiological biomarker as primary outcome but including also biochemical biomarkers is reported under the section "Neurophysiology").

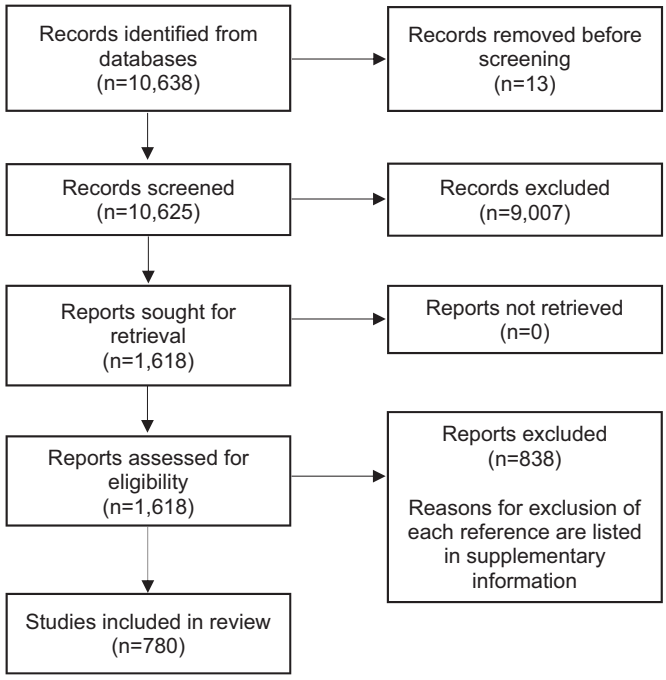

Figure 1 PRISMA 2020 flow diagram showing selection of studies for inclusion

Biochemical biomarkers

We included a total of 374 studies (359 cross-sectional and 15 longitudinal), 370 of which conducted in 58 individual countries and four in multiple countries, encompassing a total of 26,715 cases and 41,903 controls, and investigating 1,427 biomarkers (see supplementary information).

The average total BIOCROSS score (for cross-sectional studies) was 5.1 (out of 10). The average scores were 0.7 for item 3; 1.1 for item 4; 1.5 for item 5; 1.4 for item 8; and 0.5 for item 9. Therefore, the most concerning methodological issues of the included studies were related to the lack of reporting of sampling frame, participation rate and power calculation, as well as of quality procedures and blinding of the laboratory staff.

The included studies focused on a variety of biochemical biomarkers, including neurotransmitters (e.g., dopamine), hormones (e.g., oxytocin), inflammatory markers (e.g., IL-6), heavy metals (e.g., iron), antioxidants (e.g., vitamin E), and detoxifying agents (e.g., cytochrome P450 oxidase). We summarize below the findings for each neurodevelopmental disorder.

ADHD

We retained 53 studies (51 cross-sectional and two longitudinal), reported in 54 papers, from 19 countries, including a total of 4,164 participants with ADHD and 7,363 controls.

The average total BIOCROSS score was 4.9 (out of 10). The average scores were 0.8 for item 3; 1.0 for item 4; 1.3 for item 5; 1.2

for item 8; and 0.4 for item 9. Therefore, in line with the ratings across all studies of biochemical markers, the most concerning aspects were in relation to the lack of reporting of sampling frame, participation rate and power calculation, as well as of quality procedures and blinding of the laboratory staff.

The included studies assessed, collectively, 229 biomarkers (see supplementary information). Of these, 24 biomarkers were investigated in at least two studies, with at least one positive finding (see Table 1). Biomarkers with positive replications only, without negative findings, in the same direction (i.e., increased in ADHD vs. controls, or decreased in ADHD vs. controls) included: copper (two studies, increased in ADHD compared to neurotypical participants); malondialdehyde, one of the final products of polyunsaturated fatty acids peroxidation in the cells (two studies, increased); mean platelet volume (three studies, increased); and zinc (two studies, decreased).

For 28 biomarkers, one or more of the following metrics were investigated: specificity, sensitivity, PPV, NPV, and ROC AUC. However, only for mean platelet volume these metrics were available from at least two studies. In both studies, specificity and sensitivity were less than 80%, and ROC AUC values were less than 0.8. Therefore, none of the biomarkers for which a significant association with ADHD was detected and replicated, without negative associations, had evidence of a specificity and sensitivity at least of 80% and ROC AUC at least of 0.8 (see also supplementary information).

Autism spectrum disorder

We included 300 studies (289 cross-sectional and 11 longitudinal), reported in 303 papers, from 55 countries, encompassing a total of 20,583 participants with ASD and 33,450 controls. The average total BIOCROSS score was 5.2 (out of 10). The average scores were 0.8 for item 3; 1.0 for item 4; 1.3 for item 5; 1.3 for item 8; and 0.7 for item 9.

The included studies evaluated, overall, 1,298 biomarkers (see supplementary information). Of these, 73 biomarkers were in-

Table 1 Candidate biochemical biomarkers investigated in at least two studies, with at least one positive finding, for attention-deficit/hyperactivity disorder (ADHD)

Biomarker	Number of studies with significant finding	Number of studies with non-significant finding	Direction	Frequency of replication (%)
Copper (urine, hair)	2	0	Increased	100
Malondialdehyde (plasma)	2	0	Increased	100
Mean platelet volume (blood)	3	0	Increased	100
Zinc (urine, hair)	2	0	Decreased	100
Cortisol (saliva, serum)	2	1	Decreased	67
Neutrophil/lymphocyte ratio (blood)	2	1	Increased	67
Oxytocin (serum)	2	1	Decreased	67
Platelet/lymphocyte ratio (blood)	2	1	Increased	67
Folate (blood)	1	1	Decreased	50
Gamma-aminobutyric acid (serum)	1	1	Increased	50
Glial cell line-derived neurotrophic factor (plasma)	1	1	Increased	50
Glutamate (serum)	1	1	Increased	50
Interleukin-6 (plasma)	1	1	Increased	50
Lymphocytes (blood)	1	1	Decreased	50
Melatonin (saliva)	1	1	Decreased	50
Monocyte/lymphocyte ratio (blood)	1	1	Increased	50
Red blood cell distribution width (blood)	1	1	Increased	50
Soluble vascular cell adhesion molecule-1 (plasma)	1	1	Increased	50
Tumor necrosis factor-alpha (plasma)	1	1	Decreased	50
Vitamin B12 (serum)	1	1	Decreased	50
Brain-derived neurotrophic factor (plasma)	2	3	Decreased	40
Neutrophils (blood)	2	1	One increased, one decreased	33
8-hydroxy-2-deoxyguanosine (serum)	1	2	Increased	33
Ferritin (serum)	1	2	Decreased	33

Table 2 Candidate biochemical biomarkers investigated in at least two studies, with at least one positive finding and more than 50% frequency of replication, for autism spectrum disorder

Biomarker	Number of studies with significant finding	Number of studies with non-significant finding	Direction	Frequency of replication (%)
2-aminobutyric acid (urine, plasma)	2	0	Increased	100
2-hydroxybutyric acid (urine)	2	0	Increased	100
8-isoprostane (urine, plasma)	3	0	Increased	100
Adrenic acid (plasma)	2	0	Decreased	100
Alanine (urine, serum)	2	0	Decreased	100
Alpha-1-antitrypsin (plasma)	2	0	Increased	100
Anandamide (serum, plasma)	2	0	Decreased	100
Arachidic acid (serum, plasma)	2	0	Increased	100
Aspartic acid (urine, plasma)	2	0	Decreased	100
Parabacteroides (gut microbiota)	2	0	Increased	100
Creatine kinase (serum, urine)	2	0	Increased	100
Coproporphyrin (urine)	4	0	Increased	100
Cysteine (serum, plasma, urine)	3	0	Decreased	100
Glutamine (blood, serum)	4	0	Decreased	100
Glutathione/oxidized glutathione ratio (serum)	3	0	Decreased	100
High-density lipoprotein (serum)	2	0	Decreased	100
Hippuric acid (urine)	2	0	Increased	100
High-sensitivity C-reactive protein (serum)	2	0	Increased	100
Heat shock protein 70 (serum, plasma)	2	0	Increased	100
Interferon-gamma-inducible protein 16 (serum)	2	0	Increased	100
Kynurenic acid (serum, urine)	2	0	Decreased	100
Lactic acid (urine)	2	0	Increased	100
Lead (urine, hair, red blood cells)	3	0	Increased	100
Neurotensin (serum)	3	0	Increased	100
Para-cresol (urine)	3	0	Increased	100
Peroxiredoxin 1 (serum, plasma)	2	0	Increased	100
Phosphatidylcholine (serum)	2	0	Decreased	100
Pregnenolone sulfate (plasma)	2	0	Decreased	100
Secreted amyloid precursor protein alpha (plasma)	3	0	Increased	100
Succinic acid (urine, plasma)	3	0	Increased	100
Transforming growth factor beta (serum, blood)	3	0	Increased	100
Thiol (serum, urine)	2	0	Decreased	100
Triglycerides (plasma)	2	0	Increased	100
Gamma-aminobutyric acid (blood, plasma, serum)	7	0	Six increased, one decreased	85
Melatonin (serum, plasma, urine)	5	0	One increased, four decreased	80
Dopamine (plasma, blood)	4	0	Three increased, one decreased	75
Glial fibrillary acidic protein (serum)	3	1	Increased	75
Glutathione (serum, plasma)	7	1	One increased, six decreased	75
Potassium (serum)	4	0	One increased, three decreased	75

Table 2 Candidate biochemical biomarkers investigated in at least two studies, with at least one positive finding and more than 50% frequency of replication, for autism spectrum disorder *(continued)*

Biomarker	Number of studies with significant finding	Number of studies with non-significant finding	Direction	Frequency of replication (%)
Leucine (serum)	3	0	Two increased, one decreased	67
Sodium (serum, plasma)	3	0	Two increased, one decreased	67
Antioxidant capacity (urine)	3	0	Two decreased, one increased	67
Arginine vasopressin (cerebrospinal fluid)	2	1	Decreased	67
Catalase (urine, plasma)	2	1	Increased	67
Citric acid (urine, plasma)	3	0	Two increased, one decreased	67
Citrulline (blood, urine)	2	1	Increased	67
Docosahexaeonic acid/arachidonic acid (plasma)	2	1	Increased	67
Epidermal growth factor (plasma)	2	1	Decreased	67
Epinephrine (plasma, blood, gut metabolites)	3	0	Two increased, one decreased	67
Glutamate (serum, blood)	2	1	Increased	67
Hexanol-lysine (urine)	2	1	Increased	67
Hypoxanthine (urine)	2	1	Increased	67
Interleukin-17-A (plasma, serum)	2	1	Increased	67
Indole-3-acetic acid (urine)	2	1	Increased	67
Oxalic acid (urine)	2	1	Increased	67
Oxidized glutathione (plasma)	2	1	Increased	67
Pentacarboxyporphyrin (urine)	2	1	Increased	67
Phosphoric acid (urine)	2	1	Decreased	67
S100 calcium-binding protein B (serum, plasma)	4	2	Increased	67
Tumor necrosis factor-alpha (saliva, serum)	2	1	Increased	67
Thyroid stimulating hormone (serum)	2	1	Decreased	67
Uric acid (serum, urine)	3	0	Two increased, one decreased	67
Vitamin E (plasma)	2	1	Decreased	67
Glutathione S-transferase (serum, plasma)	3	0	One increased, two decreased	67
Brain-derived neurotrophic factor (serum, plasma, blood)	9	1	Six increased, three decreased	67
Cortisol (saliva, plasma, gut metabolites)	3	2	Increased	60
Eicosapentaenoic acid (serum)	3	2	Increased	60
Ferritin (serum)	3	2	Decreased	60
Homocysteine (serum, urine, plasma)	9	1	Six increased, three decreased	60
Interleukin-8 (serum, plasma)	6	4	Increased	60
Creatinine (urine)	4	3	Increased	57
Mercury (blood cells, serum, urine, hair)	4	3	Increased	57
Interleukin-1-beta (plasma)	7	4	Six increased, one decreased	54

vestigated in at least two studies, with at least one positive finding and more than 50% frequency of replication (see Table 2). Biomarkers with positive replications only, without negative findings, in the same direction (i.e., increased in ASD vs. controls, or decreased in ASD vs. controls) included: 2-aminobutyric acid (two studies, increased); 2-hydroxybutyric acid (two studies, increased); 8-isoprostane, a prostaglandin isomer (three studies, increased); adrenic acid (two studies, decreased); alanine (two studies, decreased); alpha-1-antitrypsin, an enzyme inhibitor that acts as a protector against enzymes of inflammatory cells (two studies, increased); anandamide, an endocannabinoid (two studies, decreased); arachidic acid (two studies, increased); aspartic acid (two studies, decreased); parabacteroides (two studies, increased); creatine kinase, an enzyme catalyzing the conversion of creatine (two studies, increased); coproporphyrin, a product of heme synthesis (four studies, increased); cysteine (three studies, decreased); glutamine (four studies, decreased); glutathione/oxidized glutathione ratio (three studies, decreased); high-density lipoprotein (two studies, decreased); hippuric acid (two studies, increased); high-sensitivity C-reactive protein, a marker of inflammation (two studies, increased); heat shock protein 70, a molecular chaperone that stabilizes protein substrates against denaturation (two studies, increased); interferon-gamma-inducible protein 16 (two studies, increased); kynurenic acid (two studies, decreased); lactic acid (two studies, increased); lead (three studies, increased); neurotensin, a neurotransmitter/modulator (three studies, increased); para-cresol or 4-methylphenol, a phenol derivative that can be converted in an antioxidant (three studies, increased); peroxiredoxin 1, an antioxidant (two studies, increased); phosphatidylcholine, a phospholipid (two studies, decreased); pregnenolone sulfate (two studies, decreased); secreted amyloid precursor protein alpha, a neuroprotective and neurotrophic protein (three studies, increased); succinic acid (three studies, increased); human transforming growth factor beta (three studies, increased); thiol, an organosulfur protecting against oxidative stress (two studies, decreased); and triglycerides (two studies, increased).

Specificity, sensitivity, PPV, NPV and/or ROC AUC were assessed for 303 candidate biomarkers or combinations of biomarkers. When considering biomarkers reported in more than one study, with at least one study showing specificity of 80% or higher, we found 15 biomarkers. Likewise, we located 15 biomarkers reported in more than one study, with at least one study showing sensitivity of 80% or higher. Additionally, 16 biomarkers reported in more than one study had at least one study showing ROC AUC of at least 0.8 (see Table 3). There were no compounds for which PPV or NPV were reported in more than one study.

The only biomarkers showing a specificity of at least 80% in two or more studies, without studies where specificity was less than 80%, with the same direction (i.e., biomarker increased or decreased in all studies) were oxytocin (decreased, two studies) and vitamin E (decreased, two studies). Heat shock protein 70 (increased, two studies), interferon-gamma-inducible protein-16 (increased, two studies), interferon-gamma (increased, two studies), and vitamin E (decreased, two studies) showed a sensitivity of at least 80% in two or more studies, with no studies where sensitivity was less than 80%, with the same direction. Of note, the two studies on specificity and sensitivity in relation to vitamin E derived from non-independent research groups.

In relation to ROC AUC, the following candidate biomarkers showed values of at least 0.8 in two or more studies, without studies where ROC AUC was less than 0.8, with the same direction: heat shock protein 70 (increased, 2 studies), interferon-gamma (increased, two studies), mercury (increased, two studies), and vitamin E (decreased, three studies).

Therefore, similarly to ADHD, none of the biomarkers for which a significant association with ASD was detected and replicated, without negative associations, had evidence of specificity *and* sensitivity of 80% or higher, alongside ROC AUC of 0.8 or higher.

Of note, we also found studies exploring diagnostic classification based on models including a broad array of metabolites or microbiota, and four of these (all from China) provided a ROC AUC of at least 0.8, but none of these models was tested in additional independent studies.

Conduct disorder

We retained only five studies (three cross-sectional and two longitudinal), reported in five papers, three conducted in the US, one in Croatia and one in multiple countries, including a total of 298 participants with conduct disorder and 362 controls.

The average total BIOCROSS score was 6.3 (out of 10). The average scores were 1.0 for item 3; 1.0 for item 4; 1.7 for item 5; 1.7 for item 8; and 1.0 for item 9. So, the BIOCROSS scores were in general higher than those found for ADHD and ASD, even though deriving from a much smaller number of studies.

Overall, 13 unique biomarkers were assessed. Cortisol was the only biomarker tested in more than one study (n=2), and was found significantly associated with conduct disorder in one study but not in the other one. No values of sensitivity and specificity were reported for any biomarker in two or more independent studies.

Global developmental delay/Intellectual disability

We included only five studies (all cross-sectional), reported in six papers, one conducted in China, one in France, one in South Korea, one in Iran, and one in Turkey, encompassing a total of 954 cases of intellectual disability and 189 controls.

Our rating of the quality of the studies was lower compared to the other disorders, but this should be considered cautiously, being based on a limited number of studies. The average total BIOCROSS score was 4.0 (out of 10). The average scores were 0.7 for item 3; 0.7 for item 4; 1.3 for item 5; 1.3 for item 8; and 0.5 for item 9.

Overall, 14 unique biomarkers were assessed. BDNF was the only biomarker tested in more than one study (n=2), and was found significantly associated with intellectual disability in one study but not in the other one. No biomarkers had values of sensitivity and specificity from two or more independent studies.

Table 3 Specificity and sensitivity of at least 80% and receiver operating characteristic area under the curve (ROC AUC) of at least 0.8 in relation to diagnostic biomarkers for autism spectrum disorder investigated in at least two studies, with at least one positive finding

Biomarker	Number of studies with metrics above threshold	Number of studies with metrics below threshold	Direction of the association in studies with metrics above threshold	Frequency of replication (%)
Specificity ≥ 80%				
Oxytocin (serum, plasma)	2	0	Decreased	100
Vitamin E (plasma)	2	0	Decreased	100
Gamma-aminobutyric acid (plasma)	4	0	Three increased, one decreased	75
Brain-derived neurotrophic factor (serum)	2	1	Increased	67
Tumor necrosis factor-alpha (plasma)	3	0	Two decreased, one increased	67
Catalase (blood)	1	1	Increased	50
Glutamate (plasma)	1	1	Increased	50
Homocysteine (serum, plasma)	2	0	One increased, one decreased	50
Heat shock protein 70 (plasma)	1	1	Increased	50
Interferon-gamma (plasma)	1	1	Increased	50
Methionine (plasma)	1	1	Increased	50
Potassium (serum)	1	1	Increased	50
Interleukin-6 (plasma)	3	1	Two decreased, one increased	50
Glutathione S-transferase (plasma)	1	2	Decreased	33
Serotonin (plasma)	2	2	One increased, one decreased	25
Sensitivity ≥ 80%				
Heat shock protein 70 (plasma)	2	0	Increased	100
Interferon-gamma-inducible protein 16 (plasma)	2	0	Increased	100
Interferon-gamma (plasma)	2	0	Increased	100
Vitamin E (plasma)	2	0	Decreased	100
Sodium (plasma)	1	0	Increased	100
Gamma-aminobutyric acid (plasma)	4	0	Three increased, one decreased	75
Catalase (blood)	2	0	One increased, one decreased	50
Glutamate (plasma)	1	1	Increased	50
Potassium (serum)	1	1	Increased	50
Oxytocin (serum)	1	1	Decreased	50
Brain-derived neurotrophic factor (serum)	1	2	Increased	33
Glutathione S-transferase (plasma)	1	2	Decreased	33
Tumor necrosis factor-alpha (plasma)	1	2	Increased	33
Interleukin-6 (plasma)	3	4	Two decreased, one increased	28.5
Serotonin (plasma)	1	3	Decreased	25
ROC AUC ≥ 0.8				
Heat shock protein 70 (plasma)	2	0	Increased	100
Interferon-gamma (plasma)	2	0	Increased	100
Mercury (serum, plasma)	2	0	Increased	100
Vitamin E (plasma)	3	0	Decreased	100
Gamma-aminobutyric acid (plasma)	4	0	Three increased, one decreased	75
Glutathione S-transferase (plasma)	3	0	Two decreased, one increased	67
Interferon-gamma-inducible protein 16 (plasma)	2	1	Increased	67

Table 3 Specificity and sensitivity of at least 80% and receiver operating characteristic area under the curve (ROC AUC) of at least 0.8 in relation to diagnostic biomarkers for autism spectrum disorder investigated in at least two studies, with at least one positive finding *(continued)*

Biomarker	Number of studies with metrics above threshold	Number of studies with metrics below threshold	Direction of the association in studies with metrics above threshold	Frequency of replication (%)
Potassium (serum)	3	0	Two decreased, one increased	67
Tumor necrosis factor-alpha (plasma)	3	0	Two decreased, one increased	67
Brain-derived neurotrophic factor (serum)	1	1	Increased	50
Catalase (blood)	1	1	Increased	50
Glutamate (plasma)	1	1	Increased	50
Interleukin-6 (plasma)	4	2	Two increased, two decreased	33
Melatonin (serum)	1	2	Decreased	33
Oxytocin (serum, plasma)	2	1	Decreased	33
Serotonin (plasma, blood)	2	3	One decreased, one increased	20

Tic disorder/Tourette's syndrome

We found seven eligible studies (all cross-sectional), reported in seven papers; two conducted in China, two in the Netherlands, one in Israel, one in the US, and one in multiple countries; including a total of 569 cases of tic disorder/Tourette's syndrome and 425 controls.

The average total BIOCROSS score was 4.4 (out of 10). The average scores were 0.6 for item 3; 0.9 for item 4; 1.3 for item 5; 1.0 for item 8; and 0.7 for item 9. So, the most concerning aspects, in terms of study quality, were in relation to the lack of reporting of sampling frame, participation rate and power calculation.

Overall, 50 unique biomarkers were assessed. None was tested in more than one study.

Other or combined disorders

We found only one study for coordination developmental disorder. Only three studies included cases with more than one diagnosis, i.e., two studies assessing participants with ADHD plus ASD, reporting on non-overlapping biomarkers across the two studies, and one study including individuals with ADHD and conduct disorder/oppositional defiant disorder.

Genetics

We included five GWAS (see Table 4), covering ADHD, ASD, global developmental delay and autism, tic disorder and Tourette's syndrome, and speech/language impairment. They were conducted in the UK or US or by multinational consortia, encompassing a total of 51,083 participants with neurodevelopmental disorders and 81,918 controls.

Twelve single nucleotide polymorphisms (SNPs) were found to be significantly associated with ADHD, five with ASD, one with tic disorder/Tourette's syndrome, and none with global developmental delay or speech/language impairment. There was no overlap of significant SNPs across disorders (see Table 4).

Despite this limited number of robustly identified genetic biomarkers, several of the studies estimated the total contribution of common genetic risk factors linked to each phenotype (i.e., the "SNP-based heritability" or SNP-h^2). SNP-h^2 was estimated to be approximately 21.6% for ADHD, 11.8% for ASD, 7.7% for global developmental delay, and 21.0% for tic disorder/Tourette's syndrome.

In terms of study quality, according to the selected BIOCROSS criteria, the studies of ADHD, ASD and global developmental delay scored highly (total score: 7 out of 8), while those of tic disorder/Tourette's syndrome and speech/language impairment had moderate scores (6 out of 8, and 5 out of 8, respectively), indicating that the studies were largely well-conducted.

Of note, whereas these GWAS provided an estimate of the degree of association, none of them assessed specificity, sensitivity, PPV, NPV or ROC AUC.

We could not locate any GWAS study focusing on ODD or CD as diagnostic entities. However, there have been several GWAS related to ODD/CD which focused on a broad concept of "externalizing" problems (including, for example, substance use disorder) and consisted of primarily adult samples. The largest relevant GWAS in children[37] operationalized "aggression" and was based on symptoms in the general population, rather than disorder/diagnosis.

Neuroimaging

We included a total of 203 studies (198 cross-sectional and 5 longitudinal), 176 of which conducted in 22 individual countries and 27 in multiple countries, encompassing a total of 28,636 cases and 39,508 controls (see supplementary information).

Retained studies encompassed a variety of brain imaging techniques and measures. At the structural level, magnetic resonance imaging (MRI) morphometric measures – i.e., brain volume, sur-

Table 4 Characteristics of genome-wide association studies (GWAS) of diagnostic biomarkers in neurodevelopmental disorders

Study	Country	Design	Disorder/s	Diagnosis	N probands	N controls	Biomarker(s)	Most adjusted effect size or p value
Demontis et al[32]	Multiple	Cross-sectional	ADHD	Various (DSM/ICD)	20,183	35,191	Global SNP-h^2 rs11420276 rs1222063 rs9677504 rs4858241 rs28411770 rs4916723 rs5886709 rs74760947 rs11591402 rs1427829 rs281324 rs212178	SNP-h^2 = 0.216±0.014 All SNPs: $p<5x10^{-8}$, OR range = 0.835-0.928 and 1.079-1.124
Grove et al[33]	Multiple	Cross-sectional	ASD	Various (DSM/ICD)	18,381	27,969	Global SNP-h^2 rs910805 rs10099100 rs201910565 rs71190156 rs111931861	SNP-h^2 = 0.118±0.010 All SNPs: $p<5x10^{-8}$
Niemi et al[34]	UK and Ireland	Cross-sectional	Global developmental delay and autism	Various	6,987	9,270	Global SNP-h^2 No robust genome-wide significant SNPs	SNP-h^2 = 0.077±0.021
Yu et al[35]	Multiple	Cross-sectional	Tic disorder and Tourette's syndrome	Various	4,819	9,488	Global SNP-h^2 rs2504235	SNP-h^2 = 0.21±0.024 OR=1.16, $p=2.1\times10^{-8}$
Nudel et al[36]	UK	Cross-sectional	Speech/language impairment	Various	278	Not applicable (family based study)	No robust genome-wide significant SNPs	

ADHD – attention-deficit/hyperactivity disorder, ASD – autism spectrum disorder, SNP-h^2 – single nucleotide polymorphism-based heritability, OR – odds ratio

face area, cortical thickness (region-specific and whole-brain) – as well as structural connectivity (via diffusion tensor imaging, DTI) were included. At the functional level, different levels of functional connectivity (including effective connectivity, whole-brain connectivity, network-based connectivity, global/local efficiency, and low frequency fluctuations) were measured with task-based or resting state functional MRI. In addition, a few studies reported less commonly measured functional phenotypes, such as wavelet coherence or entropy, other measures (e.g., brain iron content in ADHD), or used imaging modalities other than MRI, e.g. functional near-infrared spectroscopy (fNIRS) (see also supplementary information).

The average total BIOCROSS score was 4.86 (out of 8). The average scores were 0.98 for item 3; 1.03 for item 4; 1.40 for item 5; and 1.44 for item 8. Therefore, the main concerns were around study population source, reporting of participation rate, and sample size justification.

Four studies included two or more neurodevelopmental disorders compared to controls; the rest focused on individual disorders. Of note, only five studies tested the candidate biomarker in an external, independent sample.

ADHD

We included 66 studies (64 cross-sectional and 2 longitudinal), 61 conducted in 17 countries and five in multiple countries, encompassing a total of 10,273 cases and 20,518 controls.

The average total BIOCROSS score was 5.14 (out of 8). The average scores were 1.00 for item 3; 1.12 for item 4; 1.50 for item 5; and 1.56 for item 8.

More than half of the studies (53%) reported results only as p values, which are poorly informative as significance depends on sample size. Reported effect sizes (d) were lower than 1, and frequently low (around 0.2-0.4). Of note, both specificity and sensitivity were at least 80% for four studies only. These studies were based, respectively, on a semi-supervised learning algorithm that discovers natural groupings of brains based on the spatial patterns of variation in the morphology of the cerebral cortex and other brain regions; fNIRS functional connectivity; a support vector machine (SVM) model including prefrontal cortex activity (fNIRS) during interference with inhibitory control; and cortical thickness and volume features (see supplementary information). However, importantly, there were no other studies replicating these findings. Other

measures such as PPV and NPV were reported only inconsistently.

Autism spectrum disorder

We retained 115 studies (112 cross-sectional and 3 longitudinal), 94 conducted in 14 countries and 21 in multiple countries, including a total of 17,632 cases and 18,254 controls.

The average total BIOCROSS score was 4.72 (out of 8). The average scores were 0.97 for item 3; 0.99 for item 4; 1.36 for item 5; and 1.40 for item 8.

Nearly half of the studies (47%) reported only p values. In seven studies, both specificity and sensitivity were higher than 80%: one assessing wavelet-based coherence in resting state across larger-scale functional networks; four assessing resting-state functional connectivity in different networks; and two evaluating different DTI parameters. In one study only, specificity and sensitivity were higher than 80% and ROC AUC higher than 0.8; that study used a SVM model including ten critical functional resting-state sub-networks (see supplementary information).

Conduct disorder

We found six eligible studies (including 197 cases and 194 controls), all cross-sectional, five conducted in China and one in the UK.

The average total BIOCROSS score was 4.60 (out of 8). The average scores were 1.00 for item 3; 1.00 for item 4; 1.16 for item 5; and 1.50 for item 8.

Three studies reported only p values. Sensitivity and specificity were equal to or higher than 80% in one study only, based on a convolutional neural network (CNN) model to automatically extract multi-layer high dimensional features of structural MRI (see supplementary information).

Tic disorder/Tourette's syndrome

Eight studies (196 cases and 211 controls), all cross-sectional, six conducted in China and two in the US, were retained.

The average total BIOCROSS score was 4.50 (out of 8). The average scores were 1.00 for item 3; 1.00 for item 4; 1.20 for item 5; and 1.50 for item 8.

Four of the studies (50.0%) reported p values only. Both sensitivity and specificity were at least 80% in three of the included studies. The first of these studies focused on inter-hemispheric intrinsic functional connectivity for the bilateral orbitofrontal gyrus, bilateral midbrain, and bilateral ventral striatum; the second on global functional network properties; and the third on multiscale entropy. In all these studies, ROC AUC was higher than 0.8, but no replication of the results was found.

Other disorders

We found only one eligible study on developmental delay, one on dyslexia, and one on dyslexia/learning disorders. In none of these studies, specificity and sensitivity were higher than 80%.

Neurophysiology

A total of 133 studies were retained, 121 cross-sectional, 11 longitudinal, and 1 cross-sectional plus longitudinal, 128 conducted in a total of 24 countries and five in multiple countries, including a total of 7,045 cases and 6,923 controls (see supplementary information).

The average total BIOCROSS score was 4.87 (out of 8). The average scores were 0.97 for item 3; 1.11 for item 4; 1.32 for item 5; and 1.52 for item 8. Therefore, the most critical items were related to sampling frame, participation rate, and sample size justification.

Biomarkers tested in the retained studies included electroencephalogram (EEG), magnetoencephalography (MEG), cardiovascular, acoustic startle reflex, oculomotor, actigraphy and pupillometry measures.

ADHD

N2 amplitude, contingent negative variation (CNV) amplitude, mismatch negativity (MMN) latency, gamma coherence, and activity levels had a replication rate of 100%, albeit in a small number of studies (four for N2 amplitude and two for the other measures) (see Table 5).

The average total BIOCROSS score was 4.88 (out of 8). The average scores were 0.97 for item 3; 1.12 for item 4; 1.33 for item 5; and 1.52 for item 8.

There were no biomarkers for which sensitivity, specificity, PPV, NPV and ROC AUC have been tested in more than one study per biomarker (see supplementary information).

Autism spectrum disorder

The only biomarker with a replication rate of 100% was acoustic eye-blink startle latency (see Table 6). Sensitivity, specificity, PPV, NPV or ROC AUC were not tested in more than one study per biomarker (see supplementary information).

The average total BIOCROSS score was 4.87 (out of 8). The average scores were 0.97 for item 3; 1.11 for item 4; 1.32 for item 5; and 1.51 for item 8.

Other disorders

We could not assess replication rates of biomarkers in other disorders, due to paucity of data.

Neuropsychology

We included 65 studies, 61 cross-sectional, three longitudinal, and one cross-sectional plus longitudinal, 61 conducted in a total

Table 5 Candidate neurophysiological biomarkers investigated in at least two studies, with at least one positive finding, in relation to attention-deficit/hyperactivity disorder (ADHD)

Biomarker	Number of significant effects	Number of non-significant effects	Direction	Rate of replication (%)
MEG/EEG measures				
N2 amplitude	4	0	Four increased	100
Contingent negative variation (CNV) amplitude	2	0	Two increased	100
Mismatch negativity (MMN) latency	2	0	Two increased	100
Gamma coherence	2	0	Two decreased	100
P3 amplitude	6	3	Six decreased	67
Mismatch negativity (MMN) amplitude	2	1	Two increased	67
Alpha clustering coefficient	2	1	Two decreased	67
Alpha path length	2	1	Two decreased	66
Delta power	10	2	Six increased, four decreased	50
Alpha coherence	2	0	One increased, one decreased	50
Theta/beta ratio	5	7	Five increased	42
Alpha power	13	6	Five increased, eight decreased	42
Theta power	5	9	Five increased	36
P3 latency	1	2	One increased	33
Gamma power	2	4	Two decreased	33
Alpha peak frequency	1	2	One decreased	33
Alpha asymmetry	2	4	Two increased	33
Theta coherence	3	0	One increased, two decreased	33
Beta power	9	11	Four increased, five decreased	25
Actigraphy				
Activity level	2	0	Increased	100
Oculomotor measures and visual attention				
Exploration of social information	1	2	One increased	33
Visual attention orienting	3	5	Two increased, one decreased	25
Pupillometry				
Pupil diameter changes	1	1	One decreased	50

MEG – magnetoencephalography, EEG – electroencephalography

of 24 countries and four in multiple countries, including a total of 7,335 cases and 6,341 controls (see supplementary information).

The average total BIOCROSS score was 5.09 (out of 8). The average scores were 1.04 for item 3; 1.19 for item 4; 1.69 for item 5; and 1.16 for item 8.

ADHD

Long-term and short-term memory were characterized by replication rates of 100%, but across a small number of studies (two and five, respectively) (see Table 7).

The average total BIOCROSS score was 5 (out of 8). The average scores were 0.95 for item 3; 1.14 for item 4; 1.67 for item 5; and 1.24 for item 8.

In no instance, sensitivity, specificity, PPV, NPV or ROC AUC have been tested in more than one study per biomarker (see supplementary information).

Autism spectrum disorder

Long-term and short-term memory had replication rates of 100%, but across a small number of studies (two and five, respectively) (see Table 8).

The average total BIOCROSS score was 5.17 (out of 8). The average scores were 1.09 for item 3; 1.21 for item 4; 1.74 for item 5;

Table 6 Candidate neurophysiological biomarkers investigated in at least two studies, with at least one positive finding, in relation to autism spectrum disorder

Biomarker	Number of significant effects	Number of non-significant effects	Direction	Rate of replication (%)
MEG/EEG measures				
P3 amplitude	3	1	Three increased	75
Alpha power	5	3	Five decreased	62.5
N1 amplitude	5	1	Three increased, two decreased	50
N170 amplitude	1	1	One decreased	50
N2 amplitude	2	2	Two increased	50
Mismatch negativity (MMN) amplitude	4	3	Three increased, one decreased	43
Gamma power	22	11	Thirteen increased, nine decreased	39
P1 amplitude	1	2	One increased	33
P2 amplitude	1	2	One decreased	33
Theta power	1	2	One decreased	33
Delta power	1	3	One decreased	25
Beta power	1	10	One decreased	9
Cardiovascular measures				
Heart rate	3	0	One increased, two decreased	67
Heart rate variability - high frequency	3	0	One increased, two decreased	67
Acoustic startle reflex				
Acoustic eye-blink startle latency	3	0	Three increased	100
Acoustic eye-blink startle magnitude	10	5	Ten increased	66
Acoustic eye-blink startle habituation	1	8	One decreased	11
Oculomotor measures and visual attention				
Exploration of visual stimuli	4	0	One increased, three decreased	75
Visual attention - biological motion	4	1	One increased, three decreased	60
Perseveration on visual stimuli	8	4	Six increased, two decreased	50
Visual attention - social	22	33	Eight increased, 19 decreased	34
Visual attention - non-social	11	10	Five increased, six decreased	28
Pupillometry				
Pupil light reflex - dilation	3	1	Two slower	75
Pupil light reflex - constriction	7	3	Six slower, one faster	60
Pupil diameter	4	4	Two increased, two decreased	25

MEG – magnetoencephalography, EEG – electroencephalography

and 1.12 for item 8.

We could not locate any biomarkers for which sensitivity, specificity, PPV, NPV or ROC AUC have been tested in more than one study per biomarker (see supplementary information).

Tourette's syndrome

No replication, for any biomarkers, was found in relation to Tourette's syndrome.

Are there promising biomarkers which are transdiagnostic?

As we did not find any promising biomarker according to the criteria that we set, we could not address our additional aim, i.e., to assess to what extent promising biomarkers are transdiagnostic across neurodevelopmental disorders.

However, replication rates of associations, when available, did not suggest the transdiagnostic nature of any candidate biomarkers, with the possible exception of long-term and short-term mem-

Table 7 Candidate neuropsychological biomarkers investigated in at least two studies, with at least one positive finding, in relation to attention-deficit/hyperactivity disorder (ADHD)

Biomarker	Number of significant effects	Number of non-significant effects	Direction	Rate of replication (%)
Long-term memory	2	0	Two decreased	100
Short-term memory	5	0	Five decreased	100
IQ	6	1	Six decreased	86
Other task accuracy measures	13	2	Thirteen decreased	86
Working memory	20	4	Twenty decreased	83
Sustained attention omission errors	8	2	Eight increased	80
Reaction time variability	17	5	Seventeen increased	77
Ex-Gaussian sigma	3	1	Three increased	75
Response inhibition commission errors	8	5	Eight increased	62
Interference accuracy (e.g., Stroop test)	5	3	Five decreased	62
Mean reaction time	11	7	Eleven increased	61
Ex-Gaussian tau	3	2	Three increased	60
Delay aversion	3	2	Three increased	60
Timing task variability	2	2	Two increased	50
Face/emotion recognition accuracy	1	1	One decreased	50
Face/emotion recognition speed	1	1	One decreased	50
Set shifting accuracy	3	5	Three decreased	37.5
Other memory measures	3	7	Three decreased	30
Reaction time frequency measures	4	8	Three increased, one decreased	25
Wisconsin Card Sorting Test accuracy	1	3	One decreased	25

Table 8 Candidate neuropsychological biomarkers investigated in at least two studies, with at least one positive finding, in relation to autism spectrum disorder

Biomarker	Number of significant effects	Number of non-significant effects	Direction	Rate of replication (%)
Long-term memory	2	0	Two decreased	100
Short-term memory	5	0	Five decreased	100
Working memory	4	1	Four decreased	80
Face/emotion recognition accuracy	3	1	Three decreased	75
Reaction time variability	5	2	Five increased	71
Ex-Gaussian tau	2	1	Two increased	67
Motor coordination	2	1	Two decreased	67
Other memory measures	3	2	Three decreased	60
Other task accuracy measures	3	3	Three decreased	50
Reaction time frequency measures	2	4	Two increased	33
Face/emotion recognition speed	2	1	One increased, one decreased	33
Mean reaction time	1	8	One increased	11

ory, that had 100% replication for ADHD and ASD, and of working memory, that had ~80% replication for these disorders. Similarly, there was no overlap across SNPs across neurodevelopmental disorders in the included GWAS.

DISCUSSION

We conducted the first systematic review of studies on candidate diagnostic biomarkers for neurodevelopmental disorders, including 780 studies encompassing biochemical, genetic, neuroimaging, neurophysiological and neuropsychological measures.

In principle, finding valid, reliable and broadly usable biomarkers to detect or confirm the presence of any neurodevelopmental disorder would be highly valuable. Indeed, as these disorders manifest themselves early in development, an accurate and early diagnosis is crucial from a clinical and public health standpoint. However, despite decades of research and hundreds of publications, we could not find any biomarker that could be defined as promising based on evidence from two or more independent studies with specificity and sensitivity of at least 80%. Other important metrics to assess the validity of a biomarker, such as PPV and NPV, were unfrequently reported. We could not find any cost-effectiveness study.

Findings across the different areas included in this systematic review suggest that, while it is unlikely for a single candidate biomarker to become promising in terms of clinical translation, models including multiple biomarkers, converging on the same or related biological pathways, might be more successful. An additional aim of this review was to assess if promising biomarkers are transdiagnostic across neurodevelopmental disorders. We could not find evidence for this across any combination of the included disorders, but this negative finding was likely due to the absence of promising biomarkers in individual disorders in the first place.

While the body of research considered in this systematic review may seem impressive, the majority of included studies have simply focused on associations, reporting mainly p values, which are poorly informative as they are strongly affected by sample size. Whenever effect sizes were reported, these were generally in the low or moderate range, and certainly not in the range of an effect size of d=1.66 that would be needed to lead to a sensitivity and specificity of 80%[8].

Even when statistically significant associations have been reported, the way candidate biomarkers relate to the symptoms and the pathophysiology of a given disorder is unclear. Moreover, a large number of biomarkers have been significantly related with a given disorder, but in opposite directions, with equally plausible explanations, at least theoretically. For instance, a significant decrease of melatonin in ASD has been interpreted as a reflection of the genetically determined disruption of the serotonin-N-acetylserotonin-melatonin pathway[38]; by contrast, increased levels of melatonin have been explained as a consequence of a putative disruption of the blood-brain barrier in ASD[39].

Furthermore, the role of possible confounding effects when interpreting associations is crucial. Indeed, some markers may be influenced by factors such as diet, abnormal weight, stress, activity levels, smoking, or pharmacological treatment[40]. Our quality appraisal via the BIOCROSS tool indicated that controlling for confounding effects was inconsistent across studies. Importantly, the type of factors adjusted for varied substantially across studies.

Longitudinal studies may help in gaining better insight into the possible causal role of candidate biomarkers. However, only a few (n=36, 4.6%) of the included studies used a longitudinal design. This finding is consistent with evidence in relation to candidate biomarkers for other mental health conditions. For instance, a systematic review of studies on peripheral biomarkers for major psychiatric disorders found that only 34% of the included studies used a longitudinal design[12].

Beyond associations, a minority of studies focused on metrics that are crucial in order to assess to which extent a biomarker is promising, mainly including specificity, sensitivity or ROC AUC. Other important metrics, such as PPV or NPV values, were only rarely assessed. Of note, we could not find any biomarker with evidence from two or more studies with acceptable specificity and sensitivity, or evidence of acceptable PPV, NPV and ROC AUC.

Beyond the methodological issues related to small sample size, poor replicability, lack of standardization, and confounding factors, the main issue that seems to hamper the successful discovery of biomarkers is the very nature of the current psychiatric diagnoses, including the diagnosis of neurodevelopmental disorders, which are based on heterogeneous clusters of symptoms rather than underlying neurobiology. While different conceptualizations exist[41-46], clinical characterizations and delineations of psychiatric diagnoses remain problematic. Stratification of patients based on more homogeneous characteristics may move the field forward leading to more valid biomarkers. As Kapur et al[47] noted, the field of breast cancer faced a similar issue until bumps could be classified with histological tools. The Research Domain Criteria framework[48], aimed at establishing underpinning dimensions from the micro (i.e., genetic) to the macro (i.e., self-reported symptoms) levels, thus appears as a remarkable opportunity for stratification of patients with neurodevelopmental disorders and, hence, the discovery of valid diagnostic biomarkers.

Arguably, given the complexity and heterogeneity of neurodevelopmental disorders in terms of pathophysiology, it is highly unlikely that biomarker applications based on a single parameter will be meaningful in clinical practice[49-52]. Indeed, we found that models based on multiple parameters were in general associated with higher specificity, sensitivity and ROC AUC, although there was no replication of such models yet. In this regard, the scientific community focusing on neurodevelopmental disorders should be inspired by initiatives in other fields integrating several modalities in the same study, such as the Canadian Biomarker Integration Network on Depression (CAN-BIND), connecting clinical information with neuroimaging (e.g., brain structure), molecular (e.g., genetic, hormonal) and electrophysiological (e.g., response to transcranial magnetic stimulation) data[53].

However, even once biomarkers with good specificity, sensitivity and other metrics are found, they will need to be first validated in external, independent samples and then, importantly, also assessed in terms of feasibility and cost-effectiveness in daily clinical practice. Strikingly, we found only a limited number of studies with external validation, mainly limited to neuroimaging studies, and, in an additional search, no replication of studies testing the cost-effectiveness of any biomarker for neurodevel-

opmental disorders. Until this path is completed, any suggestion about the clinical relevance of candidate biomarkers would be misleading. Indeed, there have been reports of court cases where neuroimaging findings and genetic polymorphisms have been used to argue that the accused had a mitigating psychiatric disorder[40]. Our findings do not provide any evidence to support a similar approach for neurodevelopmental disorders[40].

While it is highly unlikely that diagnostic biomarkers will replace clinical assessment, they may eventually support clinical decision making. For instance, preliminary evidence from a randomized, parallel, single-blind, controlled trial showed that the diagnosis of ADHD with the support of a computerized test of attention and activity (QbTest), compared to the standard clinical diagnosis, led to an appointment length reduced by 15% (time ratio: 0.85, 95% CI: 0.77-0.93) and an increased clinicians' confidence in their diagnostic decisions (odds ratio: 1.77, 95% CI: 1.09-2.89)[54]. However, since attention is at the core of the clinical symptoms defining the diagnosis, it is debatable to what degree the measurement of attention is a candidate biomarker of ADHD or a standardized symptom assessment.

The possible future clinical implementation of diagnostic biomarkers will also need to consider important ethical aspects. Patients, lay people and some professionals are concerned that biomarkers may increase mental health stigma and discrimination. Indeed, as a reaction to the Human Genome project, fuelled by historical concerns about eugenics, national legislation has been developed in some countries to prevent genomic discrimination[55]. We argue that educational campaigns will be crucial to address issues around stigma while supporting the discovery of biomarkers.

The lack of evidence for a transdiagnostic nature of the biomarkers that have been explored in neurodevelopmental disorders so far is at odds with the conclusions of another systematic review[12], supporting a transdiagnostic nature of peripheral biomarkers across several mental health conditions (major depressive disorder, bipolar disorder and schizophrenia), as well as evidence from neurophysiological studies in children and adolescents[13]. However, the conclusions of that systematic review were based on the type of key words retrieved from relevant papers as well as on the variation (increase or decrease) of the biomarkers across disorders. By contrast, we focused on replication patterns, in line with the Report of the APA Work Group on Neuroimaging Markers of Psychiatric Disorders recommendations[9].

Moreover, the lack of evidence of transdiagnosticity from GWAS should be considered with caution, given the small sample size for neurodevelopmental disorders (particularly learning disorders) and meta-analytic evidence indicating large genetic correlations between most neurodevelopmental disorders[56]. Indeed, cross-disorder genetic correlation estimates clearly show that there are substantial shared common genetic risks (e.g., across ADHD and ASD) and therefore future studies of specific SNPs that are implicated in multiple disorders will need to be identified through multi-disorder analyses[32]. Similarly, previous large scale studies and meta-analyses of neuroimaging, neurophysiological and neuropsychological impairments have highlighted areas of overlap, particularly between ADHD and ASD[57-60].

It is worth noting that the vast majority of studies have focused on cases of one neurodevelopmental disorder in comparison to neurotypical or population controls – a design that can determine whether a measure may be a good diagnostic biomarker. Should promising diagnostic biomarkers emerge from this literature, their potential clinical utility may be to aid diagnostic decisions when it is unclear whether a child meets criteria for a given disorder. However, a much more likely scenario in clinical practice is the need for objective tools that can augment the valid differential diagnosis between different neurodevelopmental disorders or to determine whether a child should receive a diagnosis of one or more comorbid neurodevelopmental disorders. Yet, a low number of studies have conducted comparisons across different neurodevelopmental disorders.

Biochemical biomarkers

Biochemical biomarkers contributed the largest pool of studies included in the present systematic review. This fact may not be surprising, as, compared to other modalities (e.g., brain imaging), it is arguably less challenging, from a logistic and financial standpoint, to conduct studies on biochemical biomarkers. However, despite a plethora of studies in the field, replications are rare, and at times coming from the same research group.

In addition to the general issues that we have discussed above, there are issues, but also opportunities, that are specific to biochemical biomarkers. Biochemical substances analyzed in the studies retained in the present review were generally collected from blood, plasma, serum or urine samples. Collection from cerebrospinal fluid (CSF) is considered to be of particular interest, due to its proximity to the brain. However, this collection is very complex, due to the invasive procedure. Furthermore, CSF contains far less proteins than plasma, contributing to a reduction of chances to identify proteomic biomarkers[2].

An alternative approach would be the use of post-mortem brain tissues, which would boost the translational links between animal models of neurodevelopmental disorders and studies in living humans, although it should be considered that such studies are not informative on brain activity[61]. Overall, the use of post-mortem tissues for neurodevelopmental disorders is still in its infancy, and mainly limited to ASD. A recent systematic review[62] focusing on ASD and related disorders identified only three post-mortem studies assessing proteins and metabolites, without replicated findings[62]. Efforts in this field, such as the post-mortem brain tissue Autism BrainNet collection from the Simons Foundation[63], are therefore laudable and mirror a trend for other psychiatric disorders, such as the setting-up of the Douglas-Bell Canada Brain Bank[64], or the Netherlands Brain Bank for Psychiatry[65].

Another aspect relates to the type of biochemical biomarker. While a broad range of substances have been investigated, some in the field argue that metabolites ("metabolomics") should be particularly promising as, differently from genomics, they capture the dynamic nature of a disease and, in contrast to proteins ("proteomics"), they provide information on the final product of

complex interactions between proteins, signalling cascades and cellular environments[2]. However, there is usually a high degree of heterogeneity in terms of metabolite panels across studies.

Finally, the procedure to collect data is also highly relevant. Factors including time of day or length of time since last meal are known to impact the levels of certain biomarkers (e.g., cytokines, gene expression, or cortisol)[61]. Therefore, future studies should endeavour to follow standardized procedures, both within and across studies.

Genetic biomarkers

Compared to GWAS of other psychiatric disorders in adults (e.g., major depressive disorder with more than 135,000 cases[66], or schizophrenia with more than 76,000 cases[67]), the five retained GWAS of child neurodevelopmental disorders are relatively small and underpowered to detect robustly associated common genetic risk factors related to these disorders. However, the results of the available GWAS suggest that these disorders are highly polygenic, with thousands of common genetic variants that collectively contribute to an increased disorder risk.

It should be noted that GWAS of child disorders often include adults as well, and further work is needed to understand the degree to which the same genetic risk factors are implicated in childhood/remitting vs. persistent forms of disorder. This type of research has already been undertaken for some neurodevelopmental disorders, for instance ADHD[68].

Furthermore, for many child neurodevelopmental phenotypes, the largest available genetic analyses have focused on continuously distributed symptoms/traits in general population cohorts of children (e.g., the Avon Longitudinal Study of Parents and Children[69]), which only include a small number of diagnosed "cases". These studies were not included in this review, due to being beyond its scope, but it is plausible that biological insights which are gained from GWAS of traits/symptoms may also be relevant to diagnosed disorders, due to a large degree of shared genetic risks across disorders and traits for many neurodevelopmental conditions[70]. It should be also considered that, in addition to GWAS, studies have begun to uncover rare genetic variants, such as copy number variants or protein truncating mutations, especially in ASD[71,72], which should be assessed as possible diagnostic biomarkers.

Overall, although genetic discovery still has a long way to go to be potentially informative for neurodevelopmental disorders in children, existing GWAS can already be applied via polygenic risk score methods to gain insights into phenotypic heterogeneity, and thus inform research on diagnostic biomarkers.

Neuroimaging biomarkers

From a methodological standpoint, we highlight three important aspects that have hampered biomarker discovery and that are particularly applicable to the neuroimaging field. First, it has been noted that this field has mainly been in a mechanistic discovery phase, whereby the main focus has been on detecting alterations in brain imaging measures rather than on searching promising biomarkers[10]. Some in the field have suggested that although, ideally, biomarkers would be based on neurobiologically and mechanistically interpretable findings, this might not always be necessary, as long as biomarkers are rigorously validated. In a parallel with drug development, serendipitously discovered medications with proven clinical effectiveness were incorporated into clinical practice before their biological mechanisms were fully elucidated[10].

Second, brain development is significantly affecting case-control comparisons, and differences in developmental stage could account for greater heterogeneity during childhood and adolescence. Even if biomarkers are found, the lack of reference models of brain development renders the interpretation of certain patterns as a maturational delay or acceleration in neurodevelopmental disorders very difficult. In this context, machine learning approaches have just recently embraced advances that allow the characterization of normative trajectories and parsing of the heterogeneity at the individual level[73]. Notably, these individual-level statistics have revealed a higher predictive power of functionality when compared to unmodelled raw data[74]. Likewise, in line with the complexity of processes and mechanisms underpinning most psychiatric disorders, advanced modelling techniques[75] allow for the integration of multimodal, multivariate imaging features in neurodevelopmental disorders, which hopefully will advance biomarker discovery.

Third, neuroimaging studies included in this review, and in general across neuroimaging literature, provided effect sizes as Cohen's d. However, this metric may not be interpretable if derived out of non-normal distributions, as is often encountered in neuroimaging[8].

In terms of translation/implementation in clinical practice, it is often reported that neuroimaging biomarkers present the disadvantage of higher costs in relation to other modalities (e.g., EEG). However, it should be noted that costs may decrease over time, and the focus should be on cost-effectiveness, rather than cost *per se*. It would be worthwhile to assess to what extent neuroimaging biomarkers could avoid additional expenses, related to delayed or wrong diagnosis, to the health care system.

Neurophysiological and neuropsychological biomarkers

Several neurophysiological and neuropsychological measures have only been investigated in a small number of studies, and mainly in children with ADHD or ASD. Findings for these modalities are highly mixed and suggest very few promising biomarkers. With the exception of markers of memory performance (decreased in both ADHD and ASD), highest replication rates were generally evident for measures that have been investigated to a lesser extent.

Findings appeared more consistent for neuropsychological than for neurophysiological biomarkers. This is likely because the

ceiling/floor effects of neuropsychological measures mean that impaired profiles for a given measure are more likely to emerge consistently in the same direction (e.g., decreased working memory accuracy in children with ADHD)[76]. In contrast, atypical profiles may represent either increases or decreases relative to neurotypical controls for most neurophysiological measures (e.g., increased or decreased EEG connectivity or power).

Of note, previous studies indicate that neurophysiological profiles are highly heterogeneous in children with neurodevelopmental disorders, particularly with ADHD[77] and ASD[78], meaning that the lack of replication on these measures may not be solely attributable to methodological limitations of original studies (e.g., unrepresentative and underpowered samples). This is demonstrated by studies identifying data-driven subgroups of patients characterized by different EEG profiles, which appear associated with various clinical characteristics[79] and different rates of treatment response[80,81].

Another important consideration to make for this type of measures is that, similar to the neuroimaging literature, most of the research on neurophysiological and neuropsychological markers has focused on identifying possible mechanisms implicated in neurodevelopmental disorders (mechanistic discovery phase), rather than on developing biomarkers. Our search explicitly focused on potential biomarkers (or similar terms), and thus did not retrieve studies that investigated relevant measures, but without identifying them with these terms. The limited focus on biomarker development from this literature is also reflected in the very limited number of studies reporting diagnostic metrics (e.g., ROC AUC, sensitivity, specificity) required for establishing whether potential case-control differences at the group level can point to viable biomarkers. Future studies combining data-driven subgrouping techniques to parse heterogeneity with formal tests of biomarker properties may be particularly promising for identifying candidate biomarkers from neurophysiological and neuropsychological assessments.

Limitations

The findings of this systematic review should be considered in the light of some limitations. First, we used the term "biomarker" or equivalent terms (marker, diagnostic test, endophenotype) to retrieve studies in which the authors themselves had labeled their measure(s) as a "(bio)marker", but we could not search for all possible (bio)markers individually, which would have not been feasible. Other systematic reviews[e.g.,12] on biomarkers have used the same strategy. This limitation is particularly relevant for neuroimaging, neurophysiological and neuropsychological studies, of which only a portion used the term "biomarker" or equivalents in the article.

A meta-analytic synthesis was beyond the scope of this review. However, given the generally limited number of studies for each specific biomarker, it would have not been possible to explore sources of heterogeneity in relation to meta-analytic estimates. Therefore, our approach in terms of a narrative presentation of the data is preferable and appropriate for the current stage of the field. Moreover, we could not locate any specific tool for the quality appraisal of longitudinal studies. Rather than adapting the current BIOCROSS for cross-sectional studies, we took a more conservative and cautious approach and we did not rate the quality of longitudinal studies; however, they were only 4.6% of the total number of studies.

Even though we were careful in determining the number of positive and negative replications for each biomarker, it is possible that some studies selectively reported only positive findings, thus biasing our estimates. Furthermore, while we endeavoured to count participants from the same sample only once, the total numbers of participants reported in this systematic review are approximate, because some research groups reported results with partially overlapping samples. Finally, we focused on child-related biomarkers, but we did not include environmental biomarkers, or maternal biomarkers during pregnancy, which were beyond the scope of this work and would require an additional, specific systematic review.

CONCLUSIONS

The present work is the most comprehensive systematic review of candidate diagnostic biomarkers for neurodevelopmental disorders in children and adolescents, and should guide future research in the field. Results point to the need for well-powered studies, replication, standardization of the procedures, use of multimodal approaches in the same study, focus on metrics that are relevant for the validity of a biomarker – as opposed to assessing and reporting mere associations – and an increased focus on disorders less well investigated, such as tic disorder/Tourette's syndrome, intellectual disability, learning and language disorders, as well as a design comparing two or more neurodevelopmental disorders.

It is hoped that in the future the biomarker research in youth with neurodevelopmental disorders will benefit from larger samples, consistent methods, concerted efforts focusing on replication, building on recent consortia and other promising ongoing efforts[82,83]. This research should follow the lead of biomarker research in adults with severe mental disorders[84,85] and of other areas of medicine[86,87], that can inform appropriate assessment techniques. Future research should focus on machine learning and other advanced data analytic techniques as well as multivariable and multi-level biomarker approaches that may arguably be best suited to match the complexities of mental disorders.

ACKNOWLEDGEMENTS

S. Cortese is supported by the UK National Institute for Health and Care Research, NIHR (grant nos. NIHR203035, RP-PG-0618-20003, NIHR203684 and NIHR130077). G. Michelini is supported by a Klingenstein Third Generation Foundation fellowship; L.C. Farhat by the São Paulo Research Foundation; D. Teixeira Leffa by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation; G. Salazar de Pablo by the Alicia Koplowitz Foundation: C. Carlisi and V. Carter Leno by a Wellcome Trust Postdoctoral Fellowship; D. Floris by the European Union's Horizon 2020 research and innovation pro-

gramme (Marie Skłodowska-Curie grant agreement no. 101025785); N. Holz by the German Research Foundation and the Radboud Excellence Fellowship. S. Cortese and M. Solmi have contributed equally to this work. Supplementary information on the study is available at https://osf.io/wp4je/?view_only=8c349f45a9ac441490981acf946c8d9a.

REFERENCES

1. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) resource. Bethesda: National Institutes of Health, 2016.
2. García-Gutiérrez MS, Navarrete F, Sala F et al. Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality. Front Psychiatry 2020;11:432.
3. McGorry P, Keshavan M, Goldstone S et al. Biomarkers and clinical staging in psychiatry. World Psychiatry 2014;13:211-23.
4. Lutz W, Schwartz B. Trans-theoretical clinical models and the implementation of precision mental health care. World Psychiatry 2021;20:380-1.
5. Maj M, van Os J, De Hert M et al. The clinical characterization of the patient with primary psychosis aimed at personalization of management. World Psychiatry 2021;20:4-33.
6. Stein DJ, Craske MG, Rothbaum BO et al. The clinical characterization of the adult patient with an anxiety or related disorder aimed at personalization of management. World Psychiatry 2021;L20:336-56.
7. McIntyre RS, Alda M, Baldessarini RJ et al. The clinical characterization of the adult patient with bipolar disorder aimed at personalization of management. World Psychiatry 2022;21:364-87.
8. Loth E, Ahmad J, Chatham C et al. The meaning of significant mean group differences for biomarker discovery. PLoS Comput Biol 2021;17:e1009477.
9. First M, Botteron K, Carter C et al. Consensus Report of the APA Work Group on Neuroimaging Markers of Psychiatric Disorders. https://www.psychiatry.org.
10. Abi-Dargham A, Horga G. The search for imaging biomarkers in psychiatric disorders. Nat Med 2016;22:1248-55.
11. Kapczinski F, Vieta E, Andreazza AC et al. Allostatic load in bipolar disorder: implications for pathophysiology and treatment. Neurosci Biobehav Rev 2008; 32:675-92.
12. Pinto JV, Moulin TC, Amaral OB. On the transdiagnostic nature of peripheral biomarkers in major psychiatric disorders: a systematic review. Neurosci Biobehav Rev 2017;83:97-108.
13. Bellato A, Norman L, Idrees I et al. A systematic review and meta-analysis of altered electrophysiological markers of performance monitoring in obsessive-compulsive disorder (OCD), Gilles de la Tourette syndrome (GTS), attention-deficit/hyperactivity disorder (ADHD) and autism. Neurosci Biobehav Rev 2021;131:964-87.
14. Fusar-Poli P, Solmi M, Brondino N et al. Transdiagnostic psychiatry: a systematic review. World Psychiatry 2019;18:192-207.
15. Kim JY, Son MJ, Son CY et al. Environmental risk factors and biomarkers for autism spectrum disorder: an umbrella review of the evidence. Lancet Psychiatry 2019;6:590-600.
16. Kim JH, Kim JY, Lee J et al. Environmental risk factors, protective factors, and peripheral biomarkers for ADHD: an umbrella review. Lancet Psychiatry 2020;7:955-70.
17. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. Clin Trials 2007;4:245-53.
18. Thapar A, Cooper M, Rutter M. Neurodevelopmental disorders. Lancet Psychiatry 2017;4:339-46.
19. Solmi M, Radua J, Olivola M et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022;27:281-95.
20. World Health Organization. International classification of diseases, 11th revision (ICD-11). https://icd.who.int/browse11.
21. Solmi M, Song M, Yon DK et al. Incidence, prevalence, and global burden of autism spectrum disorder from 1990 to 2019 across 204 countries. Mol Psychiatry 2022; doi: 10.1038/s41380-022-01630-7.
22. Uher R, Zwicker A. Etiology in psychiatry: embracing the reality of poly-gene-environmental causation of mental illness. World Psychiatry 2017;16:121-9.
23. Solmi M, Fornaro M, Ostinelli EG et al. Safety of 80 antidepressants, antipsychotics, anti-attention-deficit/hyperactivity medications and mood stabilizers in children and adolescents with psychiatric disorders: a large scale systematic meta-review of 78 adverse effects. World Psychiatry 2020;19:214-32.
24. Correll CU, Cortese S, Croatto G et al. Efficacy and acceptability of pharmacological, psychosocial, and brain stimulation interventions in children and adolescents with mental disorders: an umbrella review. World Psychiatry 2021; 20:244-75.
25. Godoy PBG, Sumiya FM, Seda L et al. A systematic review of observational, naturalistic, and neurophysiological outcome measures of nonpharmacological interventions for autism. Braz J Psychiatry 2022; doi: 10.47626/1516-4446-2021-2222.
26. Fairchild G, Hawes DJ, Frick PJ et al. Conduct disorder. Nat Rev Dis Primers 2019;5:43.
27. Scassellati C, Bonvicini C, Faraone SV et al. Biomarkers and attention-deficit/hyperactivity disorder: a systematic review and meta-analyses. J Am Acad Child Adolesc Psychiatry 2012;51:1003-19.e20.
28. Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021;18:e1003583.
29. Kamp-Becker I, Albertowski K, Becker J et al. Diagnostic accuracy of the ADOS and ADOS-2 in clinical practice. Eur Child Adolesc Psychiatry 2018;27: 1193-207.
30. Wirsching J, Graßmann S, Eichelmann F et al. Development and reliability assessment of a new quality appraisal tool for cross-sectional studies using biomarker data (BIOCROSS). BMC Med Res Methodol 2018;18:122.
31. Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
32. Demontis D, Walters RK, Martin J et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet 2019;51:63-75.
33. Grove J, Ripke S, Als TD et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet 2019;51:431-44.
34. Niemi MEK, Martin HC, Rice DL et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature 2018;562:268-71.
35. Yu D, Sul JH, Tsetsos F et al. Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through genome-wide association studies. Am J Psychiatry 2019;176:217-27.
36. Nudel R, Simpson NH, Baird G et al. Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. Genes Brain Behav 2014;13:418-29.
37. Ip HF, van der Laan CM, Krapohl EML et al. Genetic association study of childhood aggression across raters, instruments, and age. Transl Psychiatry 2021;11:413.
38. Pagan C, Delorme R, Callebert J et al. The serotonin-N-acetylserotonin-melatonin pathway as a biomarker for autism spectrum disorders. Transl Psychiatry 2014;4:e479.
39. El-Ansary A, Hassan WM, Daghestani M et al. Preliminary evaluation of a novel nine-biomarker profile for the prediction of autism spectrum disorder. PLoS One 2020;15:e0227626.
40. Boksa P. A way forward for research on biomarkers for psychiatric disorders. J Psychiatry Neurosci 2013;38:75-7.
41. Watson D, Levin-Aspenson HF, Waszczuk MA et al. Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): III. Emotional dysfunction superspectrum. World Psychiatry 2022;21:26-54.
42. Krueger RF, Hobbs KA, Conway CC et al. Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): II. Externalizing superspectrum. World Psychiatry 2021;20:171-93.
43. Kotov R, Jonas KG, Carpenter WT et al. Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. World Psychiatry 2020;19:151-72.
44. Waszczuk MA. The utility of hierarchical models of psychopathology in genetics and biomarker research. World Psychiatry 2021;20:65-6.
45. Astle DE, Holmes J, Kievit R et al. Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. J Child Psychol Psychiatry 2022;63:397-417.
46. Michelini G, Barch DM, Tian Y et al. Delineating and validating higher-order dimensions of psychopathology in the Adolescent Brain Cognitive Development (ABCD) study. Transl Psychiatry 2019;9:261.
47. Kapur S, Phillips AG, Insel TR. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatry 2012;17: 1174-9.
48. Sanislow CA. RDoC at 10: changing the discourse for psychopathology. World Psychiatry 2020;19:311-2.
49. Paulus MP, Thompson WK. The challenges and opportunities of small effects: the new normal in academic psychiatry. JAMA Psychiatry 2019;76:353-4.
50. Loo SK, McGough JJ, McCracken JT et al. Parsing heterogeneity in attention-deficit hyperactivity disorder using EEG-based subgroups. J Child Psychol Psychiatry 2018;59:223-31.
51. Pulini AA, Kerr WT, Loo SK et al. Classification accuracy of neuroimaging bio-

markers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. Biol Psychiatry Cogn Neurosci Neuroimaging 2019;4:108-20.
52. Michelini G, Cheung CHM, Kitsune V et al. The etiological structure of cognitive-neurophysiological impairments in ADHD in adolescence and young adulthood. J Atten Disord 2021;25:91-104.
53. MacQueen GM, Hassel S, Arnott SR et al. The Canadian Biomarker Integration Network in Depression (CAN-BIND): magnetic resonance imaging protocols. J Psychiatry Neurosci 2019;44:223-36.
54. Hollis C, Hall CL, Guo B et al. The impact of a computerised test of attention and activity (QbTest) on diagnostic decision-making in children and young people with suspected attention deficit hyperactivity disorder: single-blind randomised controlled trial. J Child Psychol Psychiatry 2018;59:1298-308.
55. Brannan C, Foulkes AL, Lázaro-Muñoz G. Preventing discrimination based on psychiatric risk biomarkers. Am J Med Genet B Neuropsychiatr Genet 2019;180:159-71.
56. Gidziela A, Ahmadzadeh Y, Micheline G et al. Genetic influences on neurodevelopmental disorders and their overlap with co-occurring conditions in childhood and adolescence: a meta-analysis. medRxiv 2022; doi: 10.1101/2022.02.17.22271089.
57. Karcher NR, Michelini G, Kotov R et al. Associations between resting-state functional connectivity and a hierarchical dimensional structure of psychopathology in middle childhood. Biol Psychiatry Cogn Neurosci Neuroimaging 2021;6:508-17.
58. Lukito S, Norman L, Carlisi C et al. Comparative meta-analyses of brain structural and functional abnormalities during cognitive control in attention-deficit/hyperactivity disorder and autism spectrum disorder. Psychol Med 2020;50:894-919.
59. Kushki A, Cardy RE, Panahandeh S et al. Cross-diagnosis structural correlates of autistic-like social communication differences. Cereb Cortex 2021;31:5067-76.
60. Tung YH, Lin HY, Chen CL et al. Whole brain white matter tract deviation and idiosyncrasy from normative development in autism and ADHD and unaffected siblings link with dimensions of psychopathology and cognition. Am J Psychiatry 2021;178:730-43.
61. Kirkpatrick RH, Munoz DP, Khalid-Khan S et al. Methodological and clinical challenges associated with biomarkers for psychiatric disease: a scoping review. J Psychiatr Res 2021;143:572-9.
62. Fetit R, Hillary RF, Price DJ et al. The neuropathology of autism: a systematic review of post-mortem studies of autism and related disorders. Neurosci Biobehav Rev 2021;129:35-62.
63. Anderson MP, Quinton R, Kelly K et al. Autism BrainNet: a collaboration between medical examiners, pathologists, researchers, and families to advance the understanding and treatment of autism spectrum disorder. Arch Pathol Lab Med 2021;145:494-501.
64. Almeida D, Turecki G. A slice of the suicidal brain: what have postmortem molecular studies taught us? Curr Psychiatry Rep 2016;18:98.
65. Rademaker MC, de Lange GM, Palmen S. The Netherlands Brain Bank for Psychiatry. Handb Clin Neurol 2018;150:3-16.
66. Wray NR, Ripke S, Mattheisen M et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet 2018;50:668-81.
67. Trubetskoy V, Pardiñas AF, Qi T et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 2022;604:502-8.
68. Rovira P, Demontis D, Sánchez-Mora C et al. Shared genetic background between children and adults with attention deficit/hyperactivity disorder. Neuropsychopharmacology 2020;45:1617-26.
69. Ruisch IH, Dietrich A, Glennon JC et al. Interplay between genome-wide implicated genetic variants and environmental factors related to childhood antisocial behavior in the UK ALSPAC cohort. Eur Arch Psychiatry Clin Neurosci 2019;269:741-52.
70. Taylor MJ, Martin J, Lu Y et al. Association of genetic risk factors for psychiatric disorders and traits of these disorders in a Swedish population twin sample. JAMA Psychiatry 2019;76:280-9.
71. Sanders SJ, He X, Willsey AJ et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron 2015;87:1215-33.
72. Satterstrom FK, Kosmicki JA, Wang J et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell 2020;180:568-84.e23.
73. Marquand AF, Kia SM, Zabihi M et al. Conceptualizing mental disorders as deviations from normative functioning. Mol Psychiatry 2019;24:1415-24.
74. Holz NE, Floris DL, Llera A et al. Age-related brain deviations and aggression. Psychol Med 2022; doi: 10.1017/S003329172200068X.
75. Groves AR, Beckmann CF, Smith SM et al. Linked independent component analysis for multimodal data fusion. Neuroimage 2011;54:2198-217.
76. Kuntsi J, Rogers H, Swinard G et al. Reaction time, inhibition, working memory and 'delay aversion' performance: genetic influences and their interpretation. Psychol Med 2006;36:1613-24.
77. Karalunas SL, Nigg JT. Heterogeneity and subtyping in attention-deficit/hyperactivity disorder – considerations for emerging research using person-centered computational approaches. Biol Psychiatry 2020;88:103-10.
78. Hobson H, Petty S. Moving forwards not backwards: heterogeneity in autism spectrum disorders. Mol Psychiatry 2021;26:7100-1.
79. Dwyer P, Wang X, De Meo-Monteil R et al. Defining clusters of young autistic and typically developing children based on loudness-dependent auditory electrophysiological responses. Mol Autism 2020;11:48.
80. Voetterl H, van Wingen G, Michelini G et al. Brainmarker-I differentially predicts remission to various attention-deficit/hyperactivity disorder treatments: a discovery, transfer, and blinded validation study. Biol Psychiatry Cogn Neurosci Neuroimaging 2022; doi: 10.1016/j.bpsc.2022.02.007.
81. Juarez-Martinez EL, Sprengers JJ, Cristian G et al. Prediction of behavioral improvement through resting-state electroencephalography and clinical severity in a randomized controlled trial testing bumetanide in autism spectrum disorder. Biol Psychiatry Cogn Neurosci Neuroimaging 2021; doi: 10.1016/j.bpsc.2021.08.009.
82. Shic F, Naples AJ, Barney EC et al. The autism biomarkers consortium for clinical trials: evaluation of a battery of candidate eye-tracking biomarkers for use in autism clinical trials. Mol Autism 2022;13:15.
83. Charman T, Loth E, Tillmann J et al. The EU-AIMS Longitudinal European Autism Project (LEAP): clinical characterisation. Mol Autism 2017;8:27.
84. Onitsuka T, Hirano Y, Nemoto K et al. Trends in big data analyses by multicenter collaborative translational research in psychiatry. Psychiatry Clin Neurosci 2022;76:1-14.
85. Kato H, Kimura H, Kushima I et al. The genetic architecture of schizophrenia: review of large-scale genetic studies. J Hum Genet 2022; doi: 10.1038/s10038-022-01059-4.
86. Veyssière H, Bidet Y, Penault-Llorca F et al. Circulating proteins as predictive and prognostic biomarkers in breast cancer. Clin Proteomics 2022;19:25.
87. Hu X, Li G, Wu S. Advances in diagnosis and therapy for bladder cancer. Cancers 2022;14:3181.

DOI:10.1002/wps.21037