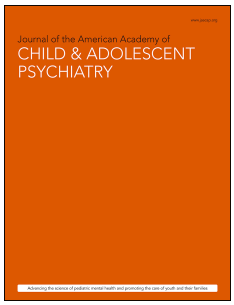

Systematic Review and Meta-analysis: Screening Tools for Attention-Deficit/ Hyperactivity Disorder in Children and Adolescents

Melissa Mulraney, PhD, Gonzalo Arrondo, PhD, Hande Musullulu, MPsych, Iciar Iturmendi-Sabater, MPsych, Samuele Cortese, MD, Samuel J. Westwood, PhD, Federica Donno, PhD, Tobias Banaschewski, MD, Emily Simonoff, MD, Alessandro Zuddas, MD, Manfred Döpfner, PhD, Stephen P. Hinshaw, PhD, David Coghill, MD

PII: S0890-8567(21)02084-0

DOI: https://doi.org/10.1016/j.jaac.2021.11.031

Reference: JAAC 3758

To appear in: *Journal of the American Academy of Child & Adolescent Psychiatry*

Received Date: 28 March 2021

Revised Date: 2 November 2021

Accepted Date: 18 November 2021

Please cite this article as: Mulraney M, Arrondo G, Musullulu H, Iturmendi-Sabater I, Cortese S, Westwood SJ, Donno F, Banaschewski T, Simonoff E, Zuddas A, Döpfner M, Hinshaw SP, Coghill D, Systematic Review and Meta-analysis: Screening Tools for Attention-Deficit/Hyperactivity Disorder in Children and Adolescents, *Journal of the American Academy of Child & Adolescent Psychiatry* (2022), doi: https://doi.org/10.1016/j.jaac.2021.11.031.

Systematic Review and Meta-analysis: Screening Tools for Attention-Deficit/Hyperactivity Disorder in Children and Adolescents

RH = Screening Tools for ADHD

Melissa Mulraney, PhD, Gonzalo Arrondo, PhD, Hande Musullulu, MPsych, Iciar Iturmendi-Sabater, MPsych, Samuele Cortese, MD, Samuel J. Westwood, PhD, Federica Donno, PhD, Tobias Banaschewski, MD, Emily Simonoff, MD, Alessandro Zuddas, MD, Manfred Döpfner, PhD, Stephen P. Hinshaw, PhD, David Coghill, MD

Drs. Mulraney and Arrondo shared first authorship of this article.

Editorial
Supplemental Material

Accepted December 17, 2021

This article was received under and accepted by Ad Hoc Editor Jonathan Posner, MD.

Dr. Mulraney and Prof. Coghill are with Murdoch Children's Research Institute, Melbourne, Australia, and the University of Melbourne, Australia. Dr. Mulraney is also with the Institute for Social Neuroscience, Ivanhoe, Australia. Dr. Arrondo and Mss. Musullulu and Iturmendi-Sabater are with the University of Navarra, Pamplona, Spain. Dr. Arrondo, Mrs. Musullulu, and Prof. Cortese are with the University of Southampton, Southampton, United Kingdom. Ms. Iturmendi-Sabater is also with University College London, United Kingdom, and the University of Toronto, Canada. Prof. Cortese is also with Solent NHS Trust, Southampton, United Kingdom; New York University, New York; and the University of Nottingham, United Kingdom. Dr. Westwood and Prof. Simonoff are with the University of Westminster, London United Kingdom. Dr. Westwood is also with King's College London, United Kingdom, and the University of Wolverhampton, United Kingdom. Dr. Simonoff is also with NIHR South London and Maudsley Biomedical Research Centre for Mental Health, London, United Kingdom. Dr. Donno and Prof Zuddas are with University of Cagliari, Italy, and "A. Cao" Pediatric Hospital, "G. Brotzu" Hospital Trust, Cagliari, Italy. Dr. Banaschewski is with the University of Heidelberg, Mannheim, Germany. Prof. Döpfner PhD is with the University Cologne (AKiP), Germany, and University of Cologne, Germany. Prof. Hinshaw PhD is with the University of California, Berkeley, and the University of California, San Francisco.

The authors have reported no funding for this work.

This article is part of a special series devoted to the subject of child and adolescent attention-deficit/hyperactivity disorder (ADHD). The series covers a range of topics in the area including genetics, neuroimaging, treatment, and others. The series was edited by Guest Editor Jonathan Posner, MD along with Deputy Editor Samuele Cortese, MD, PhD.

This work has been prospectively registered:
https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020168091.

Author Contributions
Conceptualization: Mulraney, Cortese, Banaschewski, Simonoff, Zuddas, Döpfner, Hinshaw, Coghill
Data curation: Mulraney, Arrondo, Musullulu, Iturmendi-Sabater, Westwood, Donno

Formal analysis: Mulraney
Investigation: Mulraney, Arrondo, Musullulu
Methodology: Mulraney, Arrondo, Musullulu, Iturmendi-Sabater, Cortese, Westwood, Simonoff, Coghill
Project administration: Mulraney, Arrondo
Resources: Banaschewski
Supervision: Cortese, Coghill
Writing – original draft: Mulraney
Writing – review and editing: Mulraney, Arrondo, Musullulu, Iturmendi-Sabater, Cortese, Westwood, Donno, Banaschewski, Simonoff, Zuddas, Döpfner, Hinshaw, Coghill

ORCID
Melissa Mulraney, PhD: https://orcid.org/0000-0003-1953-6481
Gonzalo Arrondo, PhD: https://orcid.org/0000-0003-3085-8959
Hande Musullulu, MPsych: https://orcid.org/0000-0002-1658-0199
Iciar Iturmendi-Sabater, MPsych: https://orcid.org/0000-0002-7239-3435
Samuele Cortese, MD: https://orcid.org/0000-0001-5877-8075
Samuel J. Westwood, PhD: https://orcid.org/0000-0002-0107-6651
Federica Donno, PhD: https://orcid.org/0000-0002-2718-0174
Tobias Banaschewski, MD: https://orcid.org/0000-0003-4595-1144
Emily Simonoff, MD: https://orcid.org/0000-0002-5450-0823
Alessandro Zuddas, MD: https://orcid.org/0000-0002-4409-0680
Manfred Döpfner, PhD: https://orcid.org/0000-0002-7929-0463
Stephen P. Hinshaw, PhD: https://orcid.org/0000-0001-6497-1082
David Coghill, MD: https://orcid.org/0000-0003-3017-9737

The authors thank the following research assistants from the University of Navarra who collaborated during the screening and risk of bias assessment: Patricia Diaz-Sanchez, MS, Irati García-Arbizu, MS, Teodora M. Niculcea-Movila, MS candidate, Isabella M. Piqué, BS, Patricia Rus-Ortiz, BS, and Paúl A. Yánez-Suárez, MS candidate.

Disclosure: Dr. Mulraney has received consulting income and research funds from the International Consortium for Health Outcomes Measurement, The Royal Children's Hospital Foundation, and ISN Innovations. Dr. Arrondo has received funding from the Spanish Ministry of Science, Innovation and Universities to facilitate the mobility of researchers to foreign higher education and research centers (Ref. CAS19/00249). Prof. Cortese has served on the advisory board of the Association for Child and Adolescent Mental Health (ACAMH). He has received honoraria from ACAMH and the British Association for Psychopharmacology. He has served as deputy editor of *Evidence-Based Mental Health*, associate editor of *Child and Adolescent Mental Health*, and on the editorial boards of the *Journal of Child Psychology and Psychiatry*, the *Journal of Child and Adolescent Psychopharmacology*, and *CNS Drugs*. Prof. Banaschewski has served in an advisory or consultancy role for ADHS digital, Infectopharm, Lundbeck, Medice, Neurim Pharmaceuticals, Oberberg GmbH, Roche, and Takeda/Shire. He has received conference support or speaker's fee by Medice and Takeda/Shire. He has received royalties from Hogrefe, Kohlhammer, CIP Medien, and Oxford University Press; the present work is unrelated to these relationships. Prof. Simonoff has received grant or research support from the National Institute of Health Research, the Psychiatry Research Trust, the Guy's and St. Thomas' Charitable Foundation, the Economic and Social Research Council, the Medical Research Council, the National Institute of Health Research Biomedical Research Centre at

South London and Maudsley Foundation Trust, and the European Commission. She has served on the advisory boards of the European ADHD Guidelines Group, Eunethydis, the Autistica Mental Health Steering Group, the National Autism Project Board, the Medical Research Council Neuroscience and Mental Health Board, the Central Institute for Mental Health, Manheim, Germany, and the Oak Foundation. She is author of the assessment tools *Assessment of Consuming Behaviour* (copyright, Santosh and Simonoff, manuscript in preparation) and *Observation Schedule for Children with Autism* (in preparation). She has served on the editorial board of the *British Journal of Psychiatry*. She has received honoraria from the Royal College of Physicians as Senior Clinical Advisor for the National Institute of Health and Care Excellence. Prof. Zuddas has reported personal fees for being on advisory boards from Angelini, Servier, and Shire/Takeda; research grants from Angelini, Janssen, Lundbeck, Otsuka, and Servier; and royalties from Giunti OS and Oxford University Press, outside the submitted work. Prof. Döpfner has received consulting income and research support from Lilly, Medice, Shire, Takeda, and Vifor and research support from the German Research Foundation, the German Ministry of Education and Research, the German Ministry of Health, and Innovation Fund. He has received income as head, supervisor, and lecturer of the School of Child and Adolescent Cognitive Behaviour Therapy at the University Hospital Cologne and as consultant for Child Behaviour Therapy at the National Association of Statutory Health Insurance Physicians (Kassenärztliche Bundesvereinigung). He has received royalties from treatment manuals, books, and psychological tests published by Beltz, Elsevier, Enke, Guilford, Hogrefe, Huber, Kohlhammer, Schattauer, Springer, and Wiley. Prof. Hinshaw has received grant funding from the National Institute of Mental Health (R01MH45064) and royalties from St. Martin's Press and Oxford University Press. Prof. Coghill has received research support and/or honoraria from Shire/Takeda, Medice, Novartis, and Servier and royalties from Oxford University Press and Cambridge University Press. Drs. Westwood and Donno, Mrs. Musullulu, and Ms. Iturmendi-Sabater have reported no biomedical financial interests or potential conflicts of interest.

Correspondence to Melissa Mulraney, PhD, ISN Innovations, Institute for Social Neuroscience, 443 Upper Heidelberg Rd, Ivanhoe, VIC 3079, Australia; e-mail: mmulraney@isn.edu.au

Abstract

Objective: This systematic review and meta-analysis aimed to (1) determine the accuracies of a broad range of screening tools for ADHD in children and adolescents and (2) compare the diagnostic accuracy of tools between population-based and clinical/high-risk samples, and across reporters.

Method: MEDLINE, PsycINFO, EMBASE and PubMed were searched up until February 20th, 2020 with no language restrictions. Studies reporting diagnostic accuracy of a screening tool against a diagnosis of ADHD in children <18 years were eligible for inclusion. Meta-analyses were undertaken to provide pooled estimates of the area under the curve (AUC), and sensitivity and specificity of groups of measures.

Results: Seventy-five studies published between 1985 and 2021 reporting on 41 screening tools that were grouped into four categories (ASEBA, DSM-IV symptom scales, SDQ, and Other Scales) were retained. The pooled AUC for studies using a combined ADHD symptoms score was 0.82 (95% CI 0.78-0.86), although this varied considerably across reporters (0.67-0.92) and populations (0.60-0.95). None of the measures met minimal standards for acceptable sensitivity (0.8) and specificity (0.8).

Conclusion: Most tools have excellent overall diagnostic accuracy as indicated by the AUC. However, a single measure, completed by a single reporter is unlikely to have sufficient sensitivity and specificity for clinical use or population screening.

Key words: attention-deficit/hyperactivity disorder, screening, psychometrics, rating scales

Introduction

ADHD is a common neurodevelopmental disorder with a global prevalence of approximately 5% in children and adolescents[1]. However, the degree to which ADHD is recognised varies considerably from country to country, and indeed between regions within countries[2,3]. The differences in administrative prevalence between and within countries are unlikely to reflect true geographical variability[1] and major concerns have been raised regarding both under-recognition and misdiagnosis[4,5]. Efficient screening has the potential to maximise the identification of possible cases which can then be referred for further assessment at reasonable costs to the healthcare system.

The most commonly used screening tools for ADHD include behaviour rating scales completed by parents and/or teachers (e.g., Conners' Rating Scale; Strengths and Difficulties Questionnaire (SDQ)). These sorts of measures are potentially useful as screening tools in that they are quick and easy to use, can easily be administered to large populations, and do not require clinical interpretation. With screening tools, however, there is always a tension between the identification of the highest proportion of true cases and an increase in the number of false positives. On the one hand, it is of high importance not to miss those who are at risk of ADHD, so they can undergo a more thorough evaluation. On the other hand, screening that results in high false positive rates will increase the burden on health services and may also increase the risk of overdiagnosis. However, it is also the case that many of those false positives will likely have another problem requiring assessment and treatment. In many research studies the presence or absence of ADHD is determined as meeting symptom threshold on a measure of ADHD symptoms[6] rather than by a gold standard clinical interview. This includes the use of both ADHD specific measures and, in the case of many epidemiological studies, general mental health screening measures that include an ADHD

subscale (e.g., the SDQ[7]). It is thus important to understand the accuracy of these measures in order to interpret the findings of such research.

This systematic review and meta-analysis, focuses on 'accuracy' as a multidimensional construct that includes a balance between sensitivity (the proportion of those who have ADHD who are correctly identified, also known as true positive rate) and specificity (the proportion of individuals without ADHD that are correctly identified, also known as true negative rate), which may vary due to contextual factors such as setting or informant. Moreover, our research links in a broader sense to the validity of these measures, that is, to what extent different types of screening tools assess what they are purportedly evaluating[8].

A plethora of ADHD screening tools are available and are routinely used in clinical and research contexts. Whilst there have been systematic reviews and meta-analyses focusing on some of these (e.g., comparing the accuracy of the Child Behaviour Checklist to the Conners' Rating Scale Revised[9]) to the best of our knowledge there has not been a comprehensive systematic review and meta-analysis of the accuracy of a broad set of ADHD screening measures. Most screening tools can be completed by different informants, it is not, however, clear if the accuracy of screening varies across informants. Screening measures are often used in different contexts and the accuracy is likely to differ across contexts. For example, the accuracy may be higher in community samples where there is a lower baseline rate of psychopathology that could be confused with ADHD symptoms (e.g., inattentiveness due to anxiety). However, it is not clear which tools are more accurate in which settings. We address these questions through a comprehensive systematic review that examines all ADHD screening tools, including whether the accuracy of screening varies according to reporter and population, which will provide valuable information for researchers, clinicians, and health services regarding the most efficient and accurate approach to screening for ADHD across these different contexts.

The aim of this systematic review and meta-analysis was therefore to determine the accuracy of a broad range of screening tools for ADHD in children and adolescents. A secondary aim was to compare the diagnostic accuracy of groups of tools between population-based and high-risk samples (e.g., referred samples) and across reporters (i.e., parent, teacher, or self-reports).

Method

Eligibility criteria

This systematic review and meta-analysis was conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement[10]. The protocol for the study was registered in PROSPERO before the commencement of the screening process (CRD42020168091). Studies were included if they 1) were peer-reviewed, 2) included participants aged 3-18 years, 3) employed a study design that compared the ADHD diagnostic accuracy of a screening instrument to a reference standard (i.e., "clinical diagnosis with evidence of parent interview, child observation, and independent evidence of pervasiveness"; "research diagnosis with parent interview", "clinical diagnosis based on codes (ICD/DSM) in medical records/registries", "clinical diagnosis methods not specified"), 4) provided estimates – that is, true positives, true negatives, false positives, false negatives that enable the calculation of the primary outcomes for the review (sensitivity, specificity, and/or area under the curve (AUC)) - or enough information to allow for the calculation of these estimates.

Studies carried out in the general population or in psychiatric samples were both accepted, as screening approaches are typically used in the two settings. Longitudinal studies, cohort studies, and case-control studies were included as they were all adequate to infer the clinical accuracy of screening tools.

Studies were excluded on the following grounds: 1) they were a qualitative report, a review, a case report, a letter, a thesis, or conference presentation slides, 2) the mean age of participants was above 18 years, 3) there was no clinical diagnosis of ADHD (e.g., diagnosis based on rating scales only), 4) there was no assessment of ADHD in the control group, 5) they were conducted in a selected clinical population where recruitment was dependent on the presence of an additional diagnosis/disorder (e.g., children with epilepsy), 6) they evaluated an instrument that requires clinical interpretation (e.g., neuropsychological tests), or 7) they failed to provide sufficient methodological or statistical information to enable inclusion in the synthesis of findings.

Information source and search strategy

Studies were identified after searching the following psychological and medical electronic databases on February 20th, 2020: MEDLINE, PsycINFO, EMBASE and PubMed restricted to peer reviewed publications. The full electronic search strategy for each database is provided in Table S1, available online. There were no date or language restrictions.

Study selection

Two authors independently screened the titles and abstracts to eliminate those studies not relevant to this review. Full texts were retrieved for all articles deemed relevant at the title and abstract screening stage, to determine eligibility for inclusion. All full text articles were independently reviewed for eligibility by at least two authors and their reference lists were evaluated in the search of additional relevant articles. Any discrepancies were discussed, and a consensus was reached.

Data extraction

The data were independently extracted by two authors for all studies using a standardised, pilot-tested extraction sheet. Any discrepancies between authors were discussed and resolved by consensus, in cases where consensus could not be reached a third author was consulted. Data extracted included basic descriptive study information (e.g., year of publication, sample size, sample type/setting, sample age, sample gender, conflicts of interest declared by authors), screening instrument examined (e.g., number of items, cut-offs), the reference standard employed (e.g., clinical diagnosis, research diagnosis using an standardised interview), and statistical and methodological considerations, including the AUC and data needed to calculate indices of diagnostic accuracy (i.e., true positives, true negatives, false positives, false negatives).

Risk of bias in individual studies

The methodological quality of all included studies was assessed across the four domains of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2)[11] tool, rated independently by two authors with any conflicts discussed until consensus was reached. The patient selection domain assesses the potential for bias in selecting patients; the index test domain assesses whether the conduct or interpretation of the screening tool could have introduced bias; the reference standard domain assesses whether the conduct or interpretation of the reference standard could have introduced bias; and the flow and timing domain assesses whether the participant flow (e.g., all participants assessed in the same manner) or timing (e.g., the gold standard assessment was completed subsequent to the screening measure) could have introduced bias. Each included study was rated within each domain as having a low, high or unclear risk of bias. The questions used in the current systematic review and meta-analysis are shown in Table S2, available online.

Data analysis

In meta-analyses, heterogeneity refers to variation in study outcomes. Screening instruments are typically adapted for different ages, informants, languages, or evolving diagnostic criteria, to the point that it might not be easy to draw the line on what constitutes a specific scale. Conversely, when conducting a meta-analysis, the number of included effect sizes has to be balanced against their heterogeneity. Taking all this into account, screening tools found in the literature were grouped into three main categories for meta-analysis. These included two omnibus measures of child mental health from which we extracted the ADHD specific subscales, namely, the Achenbach System of Empirically Based Assessment (ASEBA) and the SDQ, and an ADHD specific group of measures: DSM-IV symptom scales. Tools not fitting within any of these general groups were considered separately. DSM-IV symptom scales included all measures that had an ADHD subscale that mapped directly onto the 18 DSM-IV ADHD symptoms and had a 4-point Likert response scale (e.g., ADHD Rating Scale IV, Swanson, Nolan and Pelham Questionnaire (SNAP-IV)). The main pooling of effect sizes was done at the level of these three groups (see Table S3, available online for a description of the specific tools found in the systematic search and their grouping). Although we originally planned to meta-analyse the Conners measures, given the broad range of measures with varying length of ADHD symptom specific subscales, and the poor reporting of these tools whereby in many cases it was not possible to determine which version was used, we were unable to meta-analyse these measures.

The accuracy indices included in the qualitative and quantitative analysis included (a) area under the curve (AUC) and (b) sensitivity and specificity. The AUC is derived from a receiver operating characteristic (ROC) analysis, whereby the true positives are plotted against the false positives for each cut-off point. The AUC can range from 0 (perfectly inaccurate) to 1 (perfectly accurate), with an AUC of 0.5 indicating the tool performs no

better than chance. An AUC of 0.7-0.8 is considered acceptable, 0.8-0.9 excellent, and above 0.9 outstanding[12]. Sensitivity and specificity relate to the ability of a specific cut-score on a tool to distinguish between cases and non-cases. Typically, a sensitivity between 0.7 and 0.8 (i.e., 80% of the true cases in the population are identified in the case of 0.8) is considered acceptable in psychiatric settings, with specificity rates as close to 0.8 or higher (80% of non-cases correctly identified) used to select the optimal cut-off score[13]. Alongside the pooled accuracy indices for each meta-analysis, we report heterogeneity in the form of I^2, the percentage of variation across studies that is due to heterogeneity rather than chance[14]. As a general rule of thumb, the Cochrane guidelines indicate that I^2 can be interpreted as: 0-40% might not be important, 30-60% may represent moderate heterogeneity, 50-90% may represent substantial heterogeneity, and 75-100% considerable heterogeneity[15].

Area Under the Curve

The AUC was meta-analysed using the 'meta' package in R employing the 'metagen'[16] command to conduct a generic inverse variance meta-analysis where the AUC and standard error (SE) of AUC were available. If the SE was not reported in the study paper, we estimated it by using the standard normal distribution[17].

In instances where multiple samples and multiple measures were reported in a single study, we followed the decision rules outlined in Table S4, available online. Briefly, we selected validated ADHD-specific subscales and cut-off values when available, if multiple independent samples were included (e.g., male and female participants reported separately) we included all samples but if the same sample was used for multiple measures one measure was selected pseudorandomly, if multiple reference standards were reported we selected the one that most closely approximated a 'gold standard', and where results were reported separately for an 'ADHD total score', an 'inattentive score', or a 'hyperactive/impulsivity

score' we extracted all three data points. We calculated the pooled AUC and 95% CI for all measures that reported an 'ADHD total score', an 'inattentive score', or a 'hyperactive/impulsivity' score. Due to the high level of heterogeneity in these initial analyses (see Tables 1 and 2), we also calculated the pooled AUC (95% CI) for each group of measures. Finally, to address our second aim, we conducted subgroups analyses to explore whether the observed variance was due to reporter (parent, teacher, or self-report) or sampling (clinical/high-risk, community, and case-control sample) effects. Case-control samples were those where the cases were selected from a clinically referred population and the controls were selected from a community population (e.g., recruited through schools).

Sensitivity and specificity

Diagnostic accuracy coefficients were calculated via the construction of a 2 x 2 contingency table separately for each sample for each tool. These contingency tables compare the results of the screening tool to the reference measure and provide the true positives, true negatives, false positives, and false negatives. Based on these contingency tables, we calculated other measures related to diagnostic accuracy: sensitivity, specificity, false positive rate, false negative rate, positive predictive value, negative predictive value, and overall diagnostic accuracy. For descriptive purposes, the 2×2 contingency tables and diagnostic accuracy estimates for each study are presented in Table S5, available online. Sensitivity and specificity were meta-analysed using the 'MADA'[18] package in R. A bivariate model was used to obtain pooled sensitivity and specificity along with 95% CIs; this approach uses random effects to jointly analyse pairs of sensitivity and specificity estimates whilst accounting for any correlation between these two estimates[19]. Heterogeneity was explored through visual inspection of forest plots and the I^2 statistic.

Note that many screening tools have varying cut-offs depending on the purpose of screening (e.g., the SDQ has a clinical and a borderline cut-off). Therefore, we conducted separate

meta-analyses for the differing thresholds. Regarding the DSM-IV symptom scales, insufficient independent 2 x 2 contingency tables at a consistent threshold were available, so we cannot report the pooled sensitivity and specificity. Instead, we present a summary ROC curve with the sensitivity and false positive rate (1-specificity) for each study plotted. In instances where multiple samples, multiple measures, and/or multiple thresholds were reported in a single study, we followed the decision rules outlined in Table S4, available online. Given the high level of heterogeneity observed, we conducted meta-regressions in which reporter (parent, teacher, self) and population (clinical, community, and case-control samples) were added as covariates to the bivariate model, as well as several sensitivity analyses.

Sensitivity Analysis

Sensitivity analyses were undertaken to explore whether the findings were robust to the quality of the methodological approaches taken and the effect of these choices on heterogeneity in our results. Sensitivity analysis involved undertaking the meta-analyses twice: first including all studies and second only including studies where the samples met a high standard of methodological rigour. For the purpose of this review, the sensitivity analyses were limited to studies that used a gold standard diagnosis as the reference standard, and then sequentially limited to studies that had been categorised as having a low risk of bias on the QUADAS-2 methodological criteria of patient selection, index test, reference test, and flow and timing domains. Finally, given the difficulty in diagnosing ADHD in preschool children, sensitivity analyses were conducted to determine whether results were robust to exclusion of samples that included children under 6 years of age. A detailed description of study methods and results is available in Supplement 1, available online.

Results

Search results

The PRISMA flow chart (Figure 1) describes the systematic review process. As shown in Figure 1, 7,028 references were identified through the search, 35 were identified from other sources, and 75 full texts were eligible for inclusion in this review. A list of articles excluded during full text screening with reasons for exclusion is provided in the supplementary material (Table S6, available online). The included studies provided data to enable the calculation of diagnostic accuracy coefficients for three groups of screening tools: ASEBA (includes the CBCL), DSM-IV based ratings scales (e.g., the ADHD Rating Scale IV, SNAP-IV), and the Strengths and Difficulties Questionnaires (SDQ). There were insufficient data to include the remaining identified tools in the meta-analysis. For these, we conducted a qualitative synthesis only (see Table S5, available online for a summary of accuracy statistics for all the studies included in our review). A summary description of the screening instruments is included in Table S3, available online. Briefly, the ASEBA is an omnibus measures of youth mental health comprising 112-120 items, the SDQ is brief (25-item) omnibus measure, and the DSM-IV symptom scale measures are those that include 18 items that map directly onto the DSM-IV ADHD symptoms either as the entire measure (e.g., ADHD Rating Scale IV) or as part of an omnibus measure (e.g., Early Childhood Inventory-4). All measure groupings have parent, teacher, and self-report versions available.

Characteristics of included studies

The characteristics of included study samples are described in Table S7, available online, the 75 included studies (denoted with n) included 99 samples (denoted k). Included articles were published between 1985 and 2019 with the majority of studies published from 2010 onwards (n = 43, 57.3%), followed by articles published from 2000-2009 (n = 28, 37.3%). A plurality of study samples were recruited from the US (k = 34, 34.3%), followed by the UK (k = 12, 12.1%), Canada (k = 7, 7.1%), Spain (k = 6, 6.1%) Germany (k = 5, 5.1%), and China and Switzerland (k = 4 each, 6.1%), with remaining samples coming from a range of countries.

There was a higher proportion of male participants than female participants in two thirds of the samples. Approximately half the samples focused on children (5-12 years), whilst 23% included only adolescents and 20% had a broad age range spanning childhood and adolescence (e.g., 4-18 years). A small number (8% of all samples) of community-based studies included only preschool children. The majority of studies recruited high risk or clinical samples (54, 54.5%), followed by community-based samples (k = 31, 31.3%%), a minority used a clinical/high-risk case compared to community control (k = 14, 14.1%) design. Approximately one third of samples (k = 24, 32.0%) employed a 'gold standard' reference combining information from interviews and multiple informants to diagnose ADHD. A plurality of studies used parent interview only (k = 30, 40.0%), with a mixture of structured (e.g., DISC) and semi-structured (e.g., KSADS) interviews. One fifth (k = 15) reported a reference standard of 'clinical diagnosis' with no description of how this was conducted, and a minority used medical records (k = 4), and interview with teacher (k = 1) and child (k = 1) as the reference standard. Sixty-seven studies (89.3%) used parent-reported screening tools, 32 studies (42.7%) used teacher report, and 10 studies (13.3%) used self-reported screening tools. Whilst there was an even representation of parent-reported measures across populations (65% of community, 64% of clinical, and 68% of case-control samples), teacher report was more common in community populations (32% compared to 26% in clinical and 21% in case-control), and self-report was less common in community populations (3% compared to 10% in both clinical and case-control). Similarly, the use of a gold standard diagnosis differed across populations with 57% of clinical samples employing a gold standard compared to 28% in case-control, and 17% in community samples.

Risk of bias

A summary of the risk of bias across studies and within each article is shown in Figure S1, available online and Table S8, available online, respectively. Twenty-five per cent of the

articles had a high risk of bias in at least one domain, a number that increased to over 50% when articles in which the risk was unclear were also counted. Around 25% of the studies had a high risk of bias due to non-representative participant selection and 20% had an unclear risk (domain 1). For half of the studies, it was unclear whether the clinical diagnosis had been carried out without knowledge of the results of the screening test (domain 3). Risk of bias in almost 30% of the articles was related to the use of the screening test (domain 2), and due to the reporting of data-driven thresholds (as opposed to a-priori ones) with an additional 10% of articles having an unclear risk for this domain. Finally, around one fourth of the articles had a high or unclear risk of bias related to the flow and timing of participants (domain 4)

Qualitative synthesis

The area under the curve was reported for 66 samples with estimates ranging from 0.55-0.998. The AUC for each sample is reported in Table S5, available online. Data were available to conduct meta-analyses for ASEBA, DSM-IV symptom scales, and SDQ measures. The results of these analyses are presented below. For eight scales (Behavior Assessment System for Children (BASC), Brief Child and Family Phone Interview, Behavior Rating Inventory of Executive Function (BRIEF), Conners', Developmental Behaviour Checklist, HIDEA, INCLEN Diagnostic Tool for Attention Deficit Hyperactivity Disorder, interRAI Child and Youth Mental Health, Strengths and Weaknesses of Attention-Deficit/Hyperactivity-symptoms and Normal-behaviors (SWAN)), insufficient independent samples were available to conduct a meta-analysis (Table 1 and Table S5, available online). Overall, these other scales had high accuracy with AUC ranging from 0.73-1.00, with 73.5% reporting an AUC $\geq$ 0.80. In particular, the parent reported SWAN appears to be highly accurate, with reported AUCs of 0.89-0.95.

Sufficient information to calculate the 2 x 2 contingency tables was reported for 63 samples. The contingency tables, including the diagnostic accuracy coefficients for each cut-off meta-analyzed, are reported for each study sample in Table S6, available online. For thirteen scales (Behavior Assessment System for Children (BASC), Brief Child and Family Phone Interview, Behavior Rating Inventory of Executive Function (BRIEF), Brown ADD Scale for Adolescents, Conners', Developmental Behaviour Checklist, Devereux Scales of Mental Disorders, Dominic Interactive for Adolescents–Revised, HIDEA, INCLEN Diagnostic Tool for Attention Deficit Hyperactivity Disorder, interRAI Child and Youth Mental Health, MacArthur Health and Behavior Questionnaire, SWAN, Vanderbilt ADHD Rating Scale) and one family of measures (DSM-IV symptom scales), insufficient independent samples were available to be meta-analyzed (Table 1 and Table S5, available online). The sensitivity and specificity for these measures varied considerably across samples from unacceptably low (SE = 0.31, SP = 0.15)[20], to excellent (SE = 1.00, SP = 1.00)[21]. The inattentive subscale of the BASC appeared to be highly accurate (parent cutoff of 8 SE = 0.80, SP = 0.83; teacher cutoff of 9 SE = 0.90, SP = 0.86), however only a single study with a case-control design reported on this[22]. The DSM-IV symptom scales typically had good sensitivity at the expense of unacceptably low specificity, or vice versa. Although in some studies, at some cutoffs, a good balance of sensitivity and specificity was achieved, overall, these measures were not sufficiently accurate.

Meta-analysis: Area Under the Curve

Table 1 displays a summary of the meta-analyses conducted to explore the overall diagnostic accuracy of the three groups of screening measures as well as all measures combined (see Figures S2-S10, available online for forest plots). All of the pooled estimates and 95% confidence intervals (CIs) are greater than 0.5 indicating the measures are performing better than chance. For those scales that had an overall ADHD score and inattention and

hyperactive/impulsive subscales, the overall score generally had higher pooled estimates of AUC. There was a high degree of heterogeneity (I^2 = 17.3%-98.5%) for all measures except the DSM-IV based inattention subscale (I^2 = 17.3%) and hyperactivity/impulsivity subscale (I^2 = 29.5%). Subgroup analyses indicated significant differences in pooled AUC estimates across reporters, with a general pattern that parents tended to be the most accurate and teachers the least accurate (Table 2). However, in most cases the AUC for youth self-report was excellent (> 0.80) whilst teacher reports typically fell below the acceptable range (< 0.70). Furthermore, studies that used case-control designs tended to have the highest AUC across all measures (AUCs 0.84-0.95), followed by community samples (56% of pooled estimates > 0.80, with none below 0.70), and then studies with clinical/high-risk samples (33% of pooled estimates > 0.80, with 11% below 0.70; see Table 2). Despite teacher reports being more common in community samples compared to clinical samples, the accuracy of community samples was higher.

Meta-analysis: Sensitivity and Specificity

Table 3 displays a summary of the meta-analyses conducted to determine pooled sensitivity and specificity of each group of measures with sufficient data to be meta-analysed (ASEBA DSM-Oriented subscale, ASEBA Attention Problems subscale, and SDQ) at the most commonly used thresholds. None of the measures achieved both acceptable sensitivity and specificity. It is worth noting that the ASEBA DSM-Oriented ADHD subscale at a cut-off of 5 had acceptable specificity (0.81) and a reasonable sensitivity (0.75), and the ASEBA Attention Problems subscale at a cut-off of T $\geq$ 65 approached an acceptable balance of sensitivity (0.73) and specificity (0.77). However, the confidence intervals for these estimates were very wide, and in the case of ASEBA Attention Problems subscale there was a large degree of heterogeneity, so these findings should be interpreted with caution. Other scales tended to have good sensitivity at the expense of specificity (e.g., SDQ HI subscale

borderline cut-off SE = 0.8, SP =0.48) or good specificity but unacceptably low sensitivity (e.g., ASEBA DSM-Oriented ADHD subscale at 6 SP = 0.91, SE = 0.52). With the exception of the ASEBA DSM-Oriented subscale at a cut-off of 5, there was evidence of substantial heterogeneity between studies for both sensitivity and specificity.

Sufficient data was available to conduct three meta-regressions examining potential differences across reporters, and one meta-regression to examine potential differences between clinical/high-risk and community-based samples. A meta-regression comparing reporters for the ASEBA Attention Problems subscale at thresholds $T \geq 60$ and $T \geq 70$ indicated that there was no difference between parent and teacher report in terms of sensitivity or specificity (Table 4). Similarly, there was no significant difference in sensitivity or specificity between parent and self-reports on the SDQ clinical cut-off. A meta-regression comparing studies with clinical/high-risk compared to community-based populations on the SDQ clinical cut-off found no difference in sensitivity or specificity.

The summary ROC curves for the DSM-IV based measures are displayed in Figures S11-S13, available online, with descriptive data included in Table S9, available online. The SROC curves indicate a large amount of heterogeneity in estimates. Although some samples achieve acceptable sensitivity and specificity, the majority do not.

Sensitivity analyses

Limiting the meta-analyses to study samples that only included child aged ≥ 6 years of age did not change the results. Tables S10 and S11, available online show that the AUC and pooled sensitivity and specificity for all measures remains very similar. The one exception being DSM-IV based hyperactivity subscales whereby the AUC decreased from 0.66 (0.61; 0.72) to 0.60 (0.53; 0.68) when excluding studies with participants less than 6 years.

Limiting the meta-analyses to samples where a 'gold standard' reference standard was used for the AUC generally resulted in a slight increase of the pooled AUC estimate across measures (Table S12, available online), with the exception of the SDQ where the pooled estimate reduced from excellent (AUC = 0.82, 95% CI 0.78-0.86) to poor (AUC = 0.65, 95% CI 0.44-0.85). The results did not change for sensitivity and specificity of the ASEBA measures (Table S13, available online). There were insufficient samples to conduct the sensitivity analysis for the SDQ sensitivity and specificity.

Limiting the meta-analyses to samples with a low risk of bias in relation to patient selection resulted in a reduction in the pooled AUC estimate across measures of 0.01-0.09, though all pooled estimates and 95% confidence intervals remained greater than 0.5 (Table S14, available online). This reduction in accuracy is likely due to the exclusion of case-control designs, which generally had higher AUCs (Table 2). The results in relation to sensitivity and specificity remained similar (Table S15, available online). However, the specificity of the ASEBA Attention Problem subscale at $T \geq 65$ fell to 0.61, meaning this measure no longer approached an acceptable standard for screening. Limiting the analyses to those studies with low risk of bias in the index test domain, the reference test domain, and flow and timing domain did not change the results, with pooled estimates of AUC, sensitivity, and specificity remaining very similar. There were insufficient studies to conduct the sensitivity analysis for the meta-analysis of the SDQ AUC in the reference test domain, and for the ASEBA DSM-Oriented scale in the index test and reference test domains.

Discussion

In this systematic review and meta-analysis, we have described the accuracy of a broad range of ADHD screening tools. We have provided pooled estimates of the AUC and of the sensitivity and specificity of groupings of several commonly used measures, as well as how

these vary according to reporter and population. Overall, the results indicate that a single measure, completed by a single reporter is unlikely to be sufficiently accurate for large scale screening. Increasing cutoffs result in high specificity however, this is at the expense of sensitivity and means that a large proportion of true cases would be missed. Conversely, reducing cutoffs increases sensitivity but would be likely, by virtue of having a low specificity, to result in a significant proportion of false positives and therefore place excessive burden on health care systems. Furthermore, the low agreement and lack of reproducibility of estimates of sensitivity and specificity between studies mean that is not possible to be sure how any of the measures would perform as screeners in real world settings.

The findings in relation to AUC indicate that all the included screening tools performed better than chance, and most had excellent overall diagnostic accuracy. There was a trend whereby the AUC was generally lower for high-risk (samples drawn from the community with an oversampling of individuals with a high degree of symptomatology) and clinical samples compared to community-based samples and case-control studies. In community-based and case-control studies, the control group is from the general population and as such would typically have a low level of psychopathology. High-risk and clinical samples are more likely to be experiencing some ADHD symptoms that do not meet full criteria for an ADHD diagnosis. Moreover, there are many symptoms of other disorders that are similar to symptoms of ADHD, such that a reporter is likely to rate these as high on a subjective measure of ADHD symptoms (e.g., children with depression often have difficulties concentrating).

There were differences in accuracy of measures across reporters, which likely interacted with study design to influence the pooled estimates of accuracy. Parents were typically the most accurate reporters however, this may be an indication that shared method variance is influencing the findings, given that parents are typically the main informants for diagnostic

interviews. Teachers were generally the least accurate reporters, and in most cases the pooled AUC was below an acceptable range. However, teachers were more likely to be a reporter in community samples, and in studies focusing exclusively on pre-schoolers for whom it is difficult to diagnose ADHD. Estimates of AUC for self-reported measures tended to be excellent. This finding contradicts a body of literature that suggests youth are not accurate reporters of their own ADHD symptoms[23]. However, it is important to note that very few community samples included self-report.

Studies which included clinical samples were the most likely to employ a gold standard diagnosis as the reference, thus we can be more confident about the results from clinical samples than from community samples which had a tendency to rely more on a diagnostic interview with parents. It may be that the shared method variance has contributed to the community samples having a higher overall accuracy than clinical/high-risk samples.

Whilst the AUC findings indicate good overall accuracy of screening tools, this is less important to clinicians or for real world population-based screening aimed at identifying individuals who should be considered for further assessment than is the accuracy of the measure at a specific cut-off. For this, busy clinicians need to know that at the cut off that achieves optimal sensitivity and specificity the scale they are using is good enough to identify most of those with ADHD and not throw up too many false positives which would unnecessarily increase clinical load. However, none of the tools in this systematic review and meta-analysis met the generally agreed minimally acceptable balance of sensitivity (0.8) and specificity (0.8). This finding is particularly concerning when considering population-based screening for ADHD. The population prevalence of ADHD is approximately 5%[1]. Thus, although a screening tool with a sensitivity of 0.8 would identify four out of five true cases in a population sample, a measure with 0.8 specificity would result in a false positive rate of 19%. If such a tool were to be implemented as a population-level screen further, more

detailed assessments would need to be conducted with seven screen-positive cases for every true case identified which is likely not feasible or sustainable for health services. A recent paper indicated that specificity could be increased substantially by implementing a second stage to screening. Coghill and colleagues[24] trained teachers to administer the SNAP-IV as a semi-structured interview. Parents and teachers of all students in a school in Hunan Province, China completed the SNAP-IV questionnaire (Stage 1 screening) after which the teachers of all screen-positive children were interviewed by the SNAP-IV trained teacher (Stage 2 screening). Stage 1 sensitivity and specificity were 0.83 and 0.80, respectively. The addition of a second stage of screening resulted in a sensitivity of 0.83 and a specificity of 0.97. Still, these impressive findings need replication. A similar approach has been successfully adopted in autism spectrum disorders, whereby the Modified Checklist for Autism in Toddlers, Revised with Follow-Up has a first stage of screening designed to maximise sensitivity, and a second stage to maximise sensitivity amongst those who screen positive at the first stage[25]. Another approach that may increase accuracy, would be to include questions about impairment across settings, age of onset, and duration of symptoms to the screening scales. There is some evidence that suggests measuring impairment in addition to symptoms increases specificity[26,27], although more research is needed. With the increasingly widespread use of computerised screening, such measures could be designed so that these questions are only asked of those who indicate a threshold level of symptoms. Further, the use of multi-informant approaches to screening may improve accuracy. For example, combined parent and teacher ratings on the SDQ have been shown to increase prediction of ADHD above parent or teacher ratings alone[28].

There was a very large degree of heterogeneity between studies included in this systematic review and meta-analysis making it difficult to draw firm conclusions about how any of the included measures would perform in the real world, or across different settings and

populations. In addition, there was a high risk of bias in a large proportion of studies related to selection of participation population and specifically to the usage of case-control studies. Indeed, their elimination in a sensitivity analysis led to a reduction of the overall accuracy for the identification of individuals with ADHD. Of concern, there was insufficient information in 50% of the articles to determine whether the diagnosis was made without knowing the results of the screening test. However, when this was taken into account through a sensitivity analysis, results did not change. Similarly, many of the included studies reported only the diagnostic accuracy of the cut-off with the best balance of sensitivity and specificity in their own sample (i.e., they used a non-a priori threshold). Such procedures can introduce bias, making it difficult to make comparisons across the literature, and limiting conclusions about the performance of the measure across diverse populations. A clear need exists for more rigorous reporting standards in relation to diagnostic accuracy of screening tools. Despite the high degree of heterogeneity, results were remarkably similar across the various sensitivity analyses undertaken.

A limitation of this systematic review and meta-analysis is the inclusion of multiple reference standards of ADHD diagnosis. This heterogeneity of criterion measures may well be responsible for some of the heterogeneity observed between studies. However, the scope of the review would have greatly limited if we had accepted only a 'gold standard' reference to diagnose ADHD as only 32% of the included articles used such a standard. Further, we conducted a sensitivity analysis restricting to those studies that used a gold standard reference which largely did not change the findings. Half of the included articles had a reference standard of ADHD diagnosis based on parent report only. As noted earlier, such shared method variance may well have impacted the accuracy of parent-rated screening measures. The inclusion of articles published between 1985 and 2021 means that there were changes in the diagnostic criteria used in studies included in the review, it is likely that this contributed

to the heterogeneity across studies. Our grouping of measures is likely to have introduced heterogeneity, however without this grouping we would have largely been restricted to a qualitative description of the literature rather than a meta-analysis. These groupings enabled us to meta-analyse outcomes from two tools that have few differences between versions (SDQ, ASEBA), and also DSM-IV symptom-based questionnaires that are highly similar regarding wording and structure. Although we combined several different measures under the broad grouping of DSM-IV symptom scales, there was far less heterogeneity in these analyses than in analyses of other measures (e.g., SDQ). A strength of the review is that by not limiting to English language articles we were able to include papers from a diverse range of countries. Whilst data from African nations is under-represented and the US is over-represented, we have data from many European, South American, Asian, and Australasian countries. Whilst this has likely contributed to the degree of heterogeneity of findings, it does ensure a good representation of participants from varying cultural and ethnic backgrounds. This is particularly important given that the majority of papers did not report, or included very limited information on the ethnic/racial composition of their samples, although it should be noted that almost half (46%) of the samples were from the US or UK. As is the case with the majority of ADHD research, two thirds of the samples had a higher proportion of male participants than female participants.

Taken together these findings reinforce the need to apply caution when using questionnaires when screening for ADHD. While this caution needs to be applied across both clinical and research settings the issues differ for each setting. In a research context it was interesting to note that the sensitivities and specificities were, contrary to expectation, better in unselected community samples than in clinical/high-risk samples where prevalence of ADHD is expected to be higher. In both settings the trade-off between sensitivity and specificity is considerable. This means that the cost of adjusting the cut off down to achieve high

sensitivity (maximizing the identification of true positives) is a real drop off in specificity with increased false positives. While the impact of this on results will very much depend on the research question being addressed, we believe it is essential for researchers to address this when they interpret the findings from studies that have used screening questionnaires to identify cases.

In the USA the administrative prevalence of ADHD, number of cases with a clinical diagnosis, is greater than the epidemiological prevalence. In most other countries the rates of diagnosis fall well below what would be expected. In both situations a case could be made that community based screening, either whole population or targeted on those at increased risk, could improve the accuracy of identification and ensure that, where under recognition is an issue, a greater proportion of those with ADHD are recognised while over diagnosis is avoided. Unfortunately, the data suggest that current single stage screening approaches, using either general or ADHD specific questionnaires, would not be efficient. As noted above there is a real trade-off between sensitivity and specificity. The recommended cut-off scores for ADHD screening instruments are associated with reasonable sensitivity that will correctly pick up around 80% of those with ADHD. However, at this cut-off the false positive rate is high with 20% of those who do not have ADHD being identified as possible cases. At a population level this rate of false positives is much too high with too many potential cases identified as needing further assessment. This would overload the already stretched clinical services and, in some settings, where assessment is less robust, result in over diagnosis. We had anticipated that the false positive rate would be lower in enriched clinical samples, however this was not the case. A potential solution to this would be to consider adopting a relatively low cost two-stage screening process similar to that employed by Coghill and colleagues[24]. We recommend that future research studies investigate combinations of screening measures across different populations and settings in order to identify the most

effective combination for each setting. Cost effectiveness is also an issue in many settings. While several of the measures included in this review are free to use in both research and clinical settings, other are not. As many clinicians use screening questionnaires as a part of their case identification process these studies should consider accuracy at a single case as well as population settings. These single case applications of screening measures are subject to many of the same limitations seen in community screening. While this is a valid approach it must be used with some degree of caution, and we do not recommend that clinicians over rely on these screeners when making clinical decisions. While screening measures can help identify those who may have ADHD, accurate assessment needs be based on a clinical interview.

Based on the findings of this systematic review and meta-analysis, we recommend that wherever possible screening measures are completed by multiple reporters. If only one reporter can be used for screening this should not be teacher report, though when conducting a full assessment, it is still very important to gather information from teachers and schools. Another consideration related to obtaining teacher report is the age of the child. In secondary school students often have multiple teachers meaning that each teacher has limited opportunities to observe behaviour and thus may be less accurate in their ratings. In community-based samples parent reported measures appear to be the most accurate, though it should be noted that the diagnostic reference standards relied heavily on parent-reported interviews in this setting. In clinical/high-risk samples there appears to be little difference between the accuracy of parent- and self-reported measures. Thus, in young children it is recommended that parent-reported measures are used, whilst in adolescents it appears that self-report or parent-report measures are likely to perform similarly for screening purposes.

In conclusion, in this systematic review and meta-analysis of screening tools for ADHD we have found that although most tools have excellent overall diagnostic accuracy, a single

measure, completed by a single reporter is unlikely to have sufficient sensitivity and specificity for clinical use or population screening. Further, the very high degree of heterogeneity between studies means that we cannot be confident of how the screening tools would perform in the real world. The variation in ADHD diagnostic rates around the world[2,3] points to the need to identify and implement efficient screening to both increase detection of cases and reduce misdiagnosis. Further research is required to identify the optimal approach to screening for ADHD. It is likely that this would include data from multiple sources (e.g., a parent- and self-reported reported survey), or a two-stage screening process.

References

1. Polanczyk GV, Willcutt EG, Salum GA, Kieling C, Rohde LA. ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *Int J Epidemiol.* 2014;43(2):434-442. 10.1093/ije/dyt261
2. Australian Commission on Safety Quality in Health Care. *Australian atlas of healthcare variation.* Australian Commission on Safety and Quality in Health Care; 2015.
3. Visser SN, Danielson ML, Bitsko RH, et al. Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003–2011. *J Am Acad Child Adolesc Psychiatry.* 2014;53(1):34-46. e32. 10.1016/j.jaac.2013.09.001
4. Thomas R, Mitchell GK, Batstra L. Attention-deficit/hyperactivity disorder: are we helping or harming? *BMJ.* 2013;347:f6172. 10.1136/bmj.f6172
5. Hinshaw SP, Scheffler RM, Fulton BD, et al. International variation in treatment procedures for ADHD: social context and recent trends. *Psychiatr Serv.* 2011;62(5):459-464.
6. Ruiz-Goikoetxea M, Cortese S, Aznarez-Sanado M, et al. Risk of unintentional injuries in children and adolescents with ADHD and the impact of ADHD medications: A systematic review and meta-analysis. *Neurosci Biobehav Rev.* 2018;84:63-71. 10.1016/j.neubiorev.2017.11.007
7. Mulraney M, Giallo R, Efron D, Brown S, Nicholson JM, Sciberras E. Maternal postnatal mental health and offspring symptoms of ADHD at 8–9 years: pathways via parenting behavior. *Eur Child Adolesc Psychiatry.* 2019;28(7):923-932. 10.1007/s00787-018-1254-5

8. Flake JK, Fried EI. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science.* 2020;3(4):456-465. 10.1177/2515245920952393

9. Chang L-Y, Wang M-Y, Tsai P-S. Diagnostic accuracy of rating scales for attention-deficit/hyperactivity disorder: a meta-analysis. *Pediatrics.* 2016;137(3):e20152749. 10.1542/peds.2015-2749

10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6(7):e1000097. 10.1371/journal.pmed.1000097

11. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529-536. 10.7326/0003-4819-155-8-201110180-00009

12. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5(9):1315-1316. 10.1097/JTO.0b013e3181ec173d

13. Glascoe FP. Screening for developmental and behavioral problems. *Mental Retardation and Developmental Disabilities Research Reviews.* 2005;11(3):173-179. 10.1002/mrdd.20068

14. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta- analysis. *Stat Med.* 2002;21(11):1539-1558. 10.1002/sim.1186

15. Higgins JP, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions.* John Wiley & Sons; 2019.

16. Schwarzer G. Meta: General Package for Meta-Analysis. R package version 4.15-1. https://www.rdocumentation.org/packages/meta/versions/4.9-6/topics/metagen. Published 2020. Accessed.

17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36. 10.1148/radiology.143.1.7063747

18. Doebler P. MADA: Meta-analysis of diagnostic accuracy. R package version 0.5.10. https://cran.r-project.org/web/packages/mada/mada.pdf. Published 2020. Accessed.

19. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.* 2005;58(10):982-990. 10.1016/j.jclinepi.2005.02.022

20. Rucklidge JJ, Tannock R, Rucklidge J, Tannock R. Validity of the Brown ADD scales: an investigation in a predominantly inattentive ADHD adolescent sample with and without reading disabilities. *Journal of Attention Disorders.* 2002;5(3):155-164. 10.1177/108705470200500303

21. Gargaro BA, May T, Tonge BJ, Sheppard DM, Bradshaw JL, Rinehart NJ. Using the DBC-P Hyperactivity Index to screen for ADHD in young people with autism and ADHD: A pilot study. *Research in Autism Spectrum Disorders.* 2014;8(9):1008-1015. 10.1016/j.rasd.2014.05.004

22. Pineda DA, Aguirre DC, Garcia MA, Lopera FJ, Palacio LG, Kamphaus RW. Validation of two rating scales for attention-deficit hyperactivity disorder diagnosis in Colombian children. *Pediatr Neurol.* 2005;33(1):15-25. 10.1016/j.pediatrneurol.2005.02.001

23. Sibley MH, Pelham Jr WE, Molina BS, et al. When diagnosing ADHD in young adults emphasize informant reports, DSM items, and impairment. *J Consult Clin Psychol.* 2012;80(6):1052. 10.1037/a0029098

24. Coghill D, Du Y, Jiang W, et al. A novel school-based approach to screening for attention deficit hyperactivity disorder. *Eur Child Adolesc Psychiatry.* 2021:1-9. 10.1007/s00787-021-01721-w

25. Robins D, Fein D, Barton M. Modified checklist for autism in toddlers, Revised with Follow-Up. *Georgia: Self-published.* 2009.

26. Healey DM, Miller CJ, Castelli KL, Marks DJ, Halperin JM. The impact of impairment criteria on rates of ADHD diagnoses in preschoolers. *J Abnorm Child Psychol.* 2008;36(5):771-778. 10.1007/s10802-007-9209-1

27. Fabiano GA, Pelham J, William E, Waschbusch DA, et al. A practical measure of impairment: Psychometric properties of the impairment rating scale in samples of children with attention deficit hyperactivity disorder and two school-based samples. *J Clin Child Adolesc Psychol.* 2006;35(3):369-385. 10.1207/s15374424jccp3503_3

28. Johnson S, Hollis C, Marlow N, Simms V, Wolke D. Screening for childhood mental health disorders using the Strengths and Difficulties Questionnaire: the validity of multi- informant reports. *Dev Med Child Neurol.* 2014;56(5):453-459. 10.1111/dmcn.12360

Table 1 Meta-analytic Estimates of the Area Under the Curve of Attention-Deficit/Hyperactivity (ADHD) Disorder Screening Tools

Screening tool	Number of samples	Combined N cases	Combined N controls	Pooled AUC (95% CI)	Heterogeneity (I^2)
All measures combined					
ADHD subscale	55	7711	78654	0.82 [0.78; 0.86]	98.8 [98.6; 98.9]
Inattentive subscale	8	801	955	0.80 [0.76; 0.83]	56.1 [3.2; 80.1]
Hyperactive/impulsive subscale	9	849	1027	0.79 [0.71; 0.86]	92.9 [88.6; 95.5]
ASEBA					
Attention problems subscale	24	2292	3910	0.77 [0.72; 0.83]	96.6 [95.8; 97.3]
DSM oriented subscale	8	596	2319	0.81 [0.76; 0.86]	70.4 [38.6; 85.7]
SDQ	20	3095	72406	0.82 [0.78; 0.86]	95.8 [94.6; 96.8]
DSM-IV scales					
ADHD subscale	8	790	1031	0.87 [0.81; 0.94]	90.4 [83.5; 94.4]
Inattention subscale	4	421	592	0.75 [0.71; 0.80]	17.3 [0.0; 87.3]
Hyperactivity subscale	4	421	592	0.66 [0.61; 0.72]	29.5 [0.0; 74.2]

Note: ASEBA = Achenbach System of Empirically Based Assessment; SDQ = Strengths and Difficulties Questionnaire

Table 2 Subgroup Analyses of Area Under the Curve (AUV) by Reporter and Population

Screening tool	Number of samples	Combined N cases	Combined N controls	Pooled AUC (95% CI)	I^2	Difference between subgroups (Q)
Reporter						
All measures combined						
ADHD subscale						8.31, p=0.016
Parent-report	41	4216	2997	0.84 [0.80; 0.89]	99.0	
Self-report	9	1365	3040	0.80 [0.71; 0.88]	96.0	
Teacher-report	4	621	355	0.69 [0.60; 0.78]	83.2	
Inattentive subscale						14.33, p<0.001
Parent-report	3	396	557	0.74 [0.70; 0.78]	0.0	
Self-report	1	55	55	0.82 [0.74; 0.90]	-	
Teacher-report	4	350	343	0.84 [0.81; 0.87]	0.0	
Hyperactive/impulsive subscale						26.59, p<0.001
Parent-report	5	438	455	0.88 [0.83; 0.94]	82.7	
Teacher-report	4	411	572	0.67 [0.61; 0.73]	36.1	
ASEBA						
Attention Problems subscale						4.90, *p*=.09
Parent-report	15	1323	2997	0.78 [0.72; 0.84]	95.6	
Self-report	6	366	631	0.81 [0.67; 0.95]	96.4	
Teacher-report	3	603	282	0.66 [0.56; 0.76]	84.7	
SDQ						33.71, *p*<.001
Parent-report	15	1948	69580	0.85 [0.82; 0.88]	93.9	
Self-report	3	956	2383	0.74 [0.72; 0.76]	0.0	
Teacher-report	2	191	443	0.67 [0.41; 0.93]	90.8	
DSM-IV scales						
ADHD subscale						10.51, *p*=.01

Screening tool	Number of samples	Combined N cases	Combined N controls	Pooled AUC (95% CI)	I^2	Difference between subgroups (Q)
Parent-report	5	645	708	0.92 [0.86; 0.97]	82.6	
Self-report	2	127	250	0.79 [0.74; 0.84]	0.0	
Teacher-report	1	18	73	0.82 [0.69; 0.94]	-	
Population						
All measures combined						
ADHD subscale						78.47, p<0.001
Clinical/high-risk	30	4441	5473	0.76 [0.72, 0.80]	95.0	
Community	18	2509	72410	0.88 [0.85, 0.92]	97.0	
Case-control	7	761	771	0.95 [0.93, 0.97]	57.8	
Inattentive subscale						9.56, p=0.008
Clinical/high-risk	4	471	180	0.80 [0.75; 0.85]	46.3	
Community	2	130	521	0.75 [0.70; 0.79]	0.0	
Case-control	2	200	254	0.84 [0.81; 0.88]	0.0	
Hyperactive/impulsive subscale						5.33, p=0.070
Clinical/high-risk		559	292	0.81 [0.70; 0.91]	94.0	
Community		130	521	0.70 [0.64; 0.75]	0.0	
Case-control		160	214	0.84 [0.69; 0.98]	64.7	
ASEBA						
DSM-Oriented ADHD subscale						15.54, p<0.001
Clinical/high-risk	4	474	483	0.75 [0.71, 0.79]	18.2	
Community	4	122	1836	0.87 [0.83, 0.91]	0.0	
Attention Problems subscale						45.25, p<0.001
Clinical/high-risk	15	2005	2143	0.71 [0.66; 0.76]	94.7	
Community	5	132	1460	0.86 [0.82; 0.90]	0.0	
Case-control	4	155	307	0.93 [0.90, 0.97]	28.4	
SDQ						36.04, p<0.001
Clinical/high-risk	12	1727	4363	0.77 [0.75; 0.79]	44.0	

Screening tool	Number of samples	Combined N cases	Combined N controls	Pooled AUC (95% CI)	I^2	Difference between subgroups (Q)
Community	8	1368	68043	0.90 [0.86; 0.93]	94.3	
DSM-IV scales						
Hyperactive/impulsive subscale						3.38, p=0.066
Clinical/high-risk	2	291	71	0.60 [0.53; 0.68]	0.0	
Community	2	130	521	0.70 [0.64; 0.75]	0.0	
Inattentive subscale						0.20, p=0.657
Clinical/high-risk	2	291	71	0.77 [0.66; 0.89]	68.1	
Community	2	130	521	0.75 [0.70; 0.79]	0.0	
ADHD subscale						5.08, p=0.079
Clinical/high-risk	3	134	381	0.83 [0.79; 0.88]	7.3	
Community	3	107	285	0.87 [0.74; 1.00]	88.1	
Case-control	2	549	365	0.94 [0.86; 1.01]	82.1	

Note: ADHD = attention-deficit/hyperactivity disorder; ASEBA = Achenbach System of Empirically Based Assessment; SDQ = Strengths and Difficulties Questionnaire

Table 3 Meta-analytic Estimates of Attention-Deficit/Hyperactivity Disorder (ADHD) Screening Tools Diagnostic Accuracy

Screening tool	Number of samples	Combined N cases	Combined N controls	Pooled sensitivity (95% CI)	Heterogeneity (I^2)	Pooled Specificity (95% CI)	Heterogeneity (I^2)	AUC
ASEBA DSM-IV ADHD subscale								
Cut-off = 5	2	51	340	0.75 [0.60, 0.85]	0.0	0.81 [0.72, 0.88]	39.4	.826
Cut-off = 6	2	61	330	0.52 [0.38, 0.66]	0.0	0.91 [0.75, 0.97]	78.5	.641
ASEBA Attention Problems subscale								
T $\geq$60	4	464	274	0.89 [0.65, 0.97]	92.1	0.48 [0.40, 0.57]	23.7	.565
T $\geq$65	3	239	106	0.73 [0.45, 0.90]	81.8	0.77 [0.41, 0.94]	82.7	.808
T $\geq$70	8	889	733	0.38 [0.27, 0.50]	89.9	0.85 [0.75, 0.91]	93.3	.658
SDQ								
Borderline cut-off	3	258	636	0.80 [0.62, 0.91]	81.5	0.64 [0.45, 0.80]	91.9	.786
Clinical cut-off	6	2159	53327	0.59 [0.46, 0.70]	95.5	0.79 [0.65, 0.89]	99.3	.726

Note: ASEBA = Achenbach System of Empirically Based Assessment; CBCL = Child Behavior Checklist; SDQ = Strengths and Difficulties Questionnaire

Table 4 Subgroup Analyses of Sensitivity and Specificity by Reporter and Population

Screening tool	Number of samples	Combined N cases	Combined N controls	Pooled sensitivity (95% CI)	Pooled specificity (95% CI)	Sensitivity p-value	Specificity p-value
Reporter							
ASEBA Attention Problems subscale							
Parent T ≥60	2	206	49	0.92 (0.37, 1.00)	0.42 (0.28, 0.57)	Ref	Ref
Teacher T ≥60	2	258	225	0.84 (0.46, 0.97)	0.48 (0.34, 0.63)	0.692	0.674
Parent T ≥70	5	414	174	0.45 (0.31, 0.61)	0.78 (0.65, 0.88)	Ref	Ref
Teacher T ≥70	2	320	252	0.28 (0.20, 0.38)	0.91 (0.87, 0.94)	0.138	0.066
SDQ – clinical cut-off							
Parent	2	1154	50760	0.63 (0.38, 0.82)	0.84 (0.53, 0.96)	Ref	Ref
Self	2	1005	2567	0.54 (0.42, 0.66)	0.75 (0.71, 0.78)	0.577	0.433
Population							
SDQ – clinical cut-off							
Clinical/high-risk	2	1131	2755	0.61 (0.46, 0.74)	0.72 (0.67, 0.77)	Ref	Ref

Community	2	1028	50572	0.56 (0.24, 0.83)	0.90 (0.54, 0.98)	0.753	0.059

Note: Table 4 displays the result of subgroup analyses comparing the sensitivity and specificity between reporters and across populations. The data included in the table includes the N of cases and controls, as well as pooled sensitivity and specificity for each measure at each cutoff across the different raters and populations for which there was sufficient data to run subgroup analyses. For the reporter subgroup analyses parent-report is the reference to which the sensitivity and specificity of teacher and self-report were compared to, the p-value column indicates if the pooled sensitivity or pooled specificity differs according to reporter. For the population subgroup analysis clinical/high-risk was the reference group and the p-values indicate whether there was a significant difference in the pooled sensitivity and pooled specificity between these populations and community samples. ASEBA = Achenbach System of Empirically Based Assessment; SDQ = Strengths and Difficulties Questionnaire.

Figure 1. PRISMA Diagram

Note: Where studies were excluded based on more than one criterion, the first exclusion criterion that was met is displayed in the PRISMA flowchart.