**Systematic review and meta-analysis: Screening tools for ADHD in children and adolescents**

Melissa Mulraney\* PhD1,2,3, Gonzalo Arrondo\* PhD4,5, Hande Musullulu M.Psych4,5, Iciar Iturmendi-Sabater B.Psych4,6, Samuele Cortese MD5,7,8,9,10,, Samuel Westwood PhD11,12, Federica Donno PhD13,14, Tobias Banaschewski MD15, Emily Simonoff MD11,16, Alessandro Zuddas MD13,14, Manfred Döpfner PhD17,18, Stephen Hinshaw PhD19,20, David Coghill MD2,3

\*Denotes dual first authorship

1. Institute for Social Neuroscience, ISN Innovations, Ivanhoe, Australia
2. Murdoch Children’s Research Institute, Melbourne, Australia
3. Department of Paediatrics, University of Melbourne, Australia
4. Mind-Brain Group, Institute for Culture and Society (ICS), University of Navarra, Pamplona, Spain
5. Centre for Innovation in Mental Health, School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, UK.
6. Yale Child Study Center, University College London, UK
7. Clinical and Experimental Sciences (CNS and Psychiatry), Faculty of Medicine, University of Southampton, Southampton, UK.
8. Solent NHS Trust, Southampton, UK.
9. Hassenfeld Children's Hospital at NYU Langone, New York University Child Study Center, New York City, New York, USA.
10. Division of Psychiatry and Applied Psychology, School of Medicine, University of Nottingham, Nottingham, UK
11. Department of Child & Adolescent Psychiatry, King’s College London
12. Department of Psychology, University of Wolverhampton
13. Dept. Biomedical Sciences, Sect. Neuroscience & Clinical Pharmacology, University of Cagliari, Italy
14. Child & Adolescent Neuropsychiatry Unit, “A. Cao” Pediatric Hospital, “G. Brotzu” Hospital Trust, Cagliari, Italy
15. Department of Child & Adolescent Psychiatry and Psychotherapy, Medical Faculty Mannheim, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
16. Institute of Psychiatry, Psychology and Neuroscience, London, UK
17. School of Child and Adolescent Cognitive Behaviour Therapy at the University Cologne (AKiP), Germany
18. Department of Psychiatry, Psychosomatic and Psychotherapy of Childhood and Adolescence at the University Cologne; Germany
19. Department of Psychology, University of California, Berkeley
20. Department of Psychiatry and Behavioral Sciences, University of California, San Francisco

**Corresponding author:** Dr Melissa Mulraney, ISN Innovations, Institute for Social Neuroscience, 443 Upper Heidelberg Rd, Ivanhoe, VIC 3079, Australia. Email: mmulraney@isn.edu.au

**Word Count:** 4222

**Conflict of Interest Statement:**

Dr. Gonzalo Arrondo is funded by the Spanish Ministry of Science, Innovation and Universities to facilitate the mobility of researchers to foreign higher education and research centers (Ref. CAS19/00249)"

Professor Tobias Banaschewski reports personal fees for being on the advisory board from Lundbeck, Medice, Neurim Pharmaceuticals, Oberberg GmbH, Takeda/Shire, and Infectopharm; speaker’s fees from Medice, Takeda/Shire, and Lilly; and royalties from Hogrefe, Kohlhammer, CIP Medien, and Oxford University Press, outside the submitted work.

Professor Alessandro Zuddas reports personal fees for being on advisory boards from Angelini, Edu Pharma, and ShireTakeda; research grants from Angelini, Janssen, Lundbeck, Otsuka, and Servier; and royalties from Giunti OS and Oxford University Press, outside the submitted work.

Manfred Döpfnerreceived consulting income and research support from Lilly, Medice, Shire, Takeda, and Vifor and research support from the German Research Foundation, German Ministry of Education and Research, German Ministry of Health, and Innovation Fund. He 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 also received royalties from treatment manuals, books and psychological tests published by Beltz, Elsevier, Enke, Guilford, Hogrefe, Huber, Kohlhammer, Schattauer, Springer, Wiley.

Professor Samuele Cortese declares honoraria and reimbursement for travel and accommodation expenses for lectures from the following non-profit associations: Association for Child and Adolescent Central Health (ACAMH), Canadian ADHD Alliance Resource (CADDRA), British Association of Pharmacology (BAP), and from Healthcare Convention for educational activity on ADHD.

David Coghill has received research support and/or honoraria from Shire/Takeda, Medice, Servier and royalties from Oxford University Press and Cambridge University Press.

All other authors state they have no disclosures or potential conflicts of interest to declare.

**Acknowledgements**

We want to thank the following research assistants from the University of Navarra who collaborated during the screening and risk of bias assessment: Patricia Diaz-Sanchez, Irati García-Arbizu, Teodora Niculcea, Isabella Piqué, Patricia Rus and Paul Yánez.

**KEYWORDS:** attention-deficit/hyperactivity disorder; screening; psychometrics; rating scales

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 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 measures.

**Results:** Ninety studies reporting on 51 screening tools were retained. The pooled AUC for studies using a combined ADHD symptoms score was 0.83 (95% CI 0.80-0.86), although this varied considerably across reporters (0.69-0.86) and populations (0.75-0.93). 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 AUC. However, a single measure, completed by a single reporter is unlikely to have sufficient sensitivity and specificity for clinical use or population screening.

ADHD is a common neurodevelopmental disorder with a global prevalence of approximately 5% in children and adolescents1. However, the degree to which ADHD is recognised varies considerably from country to country, and indeed between regions within countries2,3. The differences in administrative prevalence between and within countries are unlikely to reflect true geographical variability1 and major concerns have been raised regarding both under-recognition and misdiagnosis4,5. 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). 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. Additionally, continuous performance tasks (CPTs) are computerized paradigms that assess sustained attention and inhibitory control over an extended period of time (typically, between 10 and 20 minutes). Since their administration, correction, and interpretation are highly standardized and automatized, they are frequently used as a complementary screening assessment within many clinical settings, and can potentially be used to evaluate multiple individuals concurrently.

With screening tools, however, there is always a trade-off between the identification of the highest number 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 would increase the burden on health services and the risk of overdiagnosis. In this systematic review and meta-analysis, we focus on the ‘accuracy’ of a screening tool 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 screening tools assess what they are purportedly evaluating6.

A plethora of ADHD screening tools are available. 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 Revised7) 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. Such a review would provide valuable information for researchers, clinicians, and health services regarding the most efficient and accurate approach to screening for ADHD in which contexts.

The aim of this systematic review and meta-analysis was 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 tools between population-based and high-risk samples (e.g., referred samples) and across reporters (i.e., parent, teacher, or self-reports).

**Methods**

**Eligibility criteria**

This systematic review and meta-analysis was conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement8. The protocol for the study was registered in PROSPERO before the commencement of the screening process ([CRD42020168091](https://www.crd.york.ac.uk/prospero/display_record.php?ID=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 an appropriate 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. 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 of 0.8 (i.e., 80% of the true cases in the population are identified) is considered acceptable in psychiatric settings, with specificity rates as close to 0.8 (80% of non-cases correctly identified) used to select the optimal cut-off score9. 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 outstanding10.

Screening tools must be brief, be able to be administered to large numbers of people, and scored without the need for clinical interpretation. Thus, rating scales, wearable technology, and brief computerised assessments that automatically calculate scores were all potential measures for inclusion. 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 or that would take more than 20 minutes to complete (an arbitrary definition of “short” related to the typical length of CPTs, with rating scales usually being much shorter), 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. 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 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) 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.

**Data analysis**

The accuracy indices included in the qualitative and quantitative analysis included (a) area under the curve (AUC) and (b) sensitivity and specificity.

***Area Under the Curve***

The AUC was meta-analysed using the ‘meta’ package in R employing the ‘metagen’12 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 distribution13.

In instances where multiple samples and multiple measures were reported in a single study, we followed the decision rules outlined in Table S3. That is, 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 measure. Finally, to address our second aim, we conducted subgroups analyses to explore whether the observed variance was due to reporter (parent, teacher, self-report, objective measure) or sampling (clinical, community, and case-control sample) effects.

***Sensitivity and specificity***

Diagnostic accuracy coefficients were calculated via the construction of a 2 x 2 contingency table for each screening 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 Supplementary Table S4. Sensitivity and specificity were meta-analysed using the ‘MADA’14 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 estimates15. Heterogeneity was explored through visual inspection of forest plots and the conduct of chi-squared tests for equality of sensitivities and specificities.

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). As such, we conducted separate meta-analyses for the differing thresholds. Regarding the Conners’, CPT, and DSM-IV based 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 S3. 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.

***Sensitivity Analysis***

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. Sensitivity analysis involved undertaking the meta-analyses twice: first including all studies and second only including studies where the sample did not include children <6 years.

**Results**

**Search results**

The PRISMA flow chart (Figure 1) describes the systematic review process. As shown in Figure 1, 7,063 references were identified through the search, and 90 full texts were eligible for inclusion in this review. A list of excluded articles with reasons for exclusion is provided in the supplementary material (Table S5). The included studies provided data to enable the calculation of diagnostic accuracy coefficients for five screening tools: the Achenbach System of Empirically Based Assessment (ASEBA), Conners Scales, CPT, DSM-IV based ratings scales (e.g., the ADHD Rating Scale IV), and the SDQ. There were insufficient data to include the remaining identified tools in the meta-analysis; for these, we conducted a qualitative synthesis only (see Supplementary Table S4 for a summary of accuracy statistics). A summary description of the screening instruments is included in Table S6.

**Characteristics of included studies**

The characteristics of included study samples are described in Table S7. The majority of studies were published from 2010 onwards (k = 51, 56.7%), followed by articles published from 2000-2009 (k = 33, 36.7%). Most study samples were recruited from the US (k = 43, 43.9%), followed by Canada (k = 7, 7.1%), China (k = 6, 6.1%), Germany (k = 5, 5.1%), Sweden (k = 4, 4.1%), and Spain (k = 4, 4.1%). The majority of studies recruited high risk or clinical samples (70, 58.3%), followed by community-based samples (k = 35, 29.2%%), a minority used a high-risk/clinical case compared to community control (k = 15, 12.5%) design. Approximately one quarter of samples (k = 32, 26.7%) employed a ‘gold standard’ reference combining information from interviews and multiple informants to diagnose ADHD. The majority of studies used parent interview only (k = 61, 50.8%), with a mixture of structured (e.g., DISC) and semi-structured (e.g., KSADS) interviews. One sixth (k = 20) reported a reference standard of ‘clinical diagnosis’ with no description of how this was conducted, and a minority used medical records (k = 5), and interview with teacher (k = 1) and child (k = 1) as the reference standard.

**Risk of bias**

A summary of the risk of bias across studies and within each article is shown in Figure S1 and Table S8 respectively. Overall, the majority of studies (74%) had a low risk of bias, however, some domains were problematic. Around 25% of the studies had a high risk of bias due to non-representative participant selection and 20% had an unclear risk. 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. Risk of bias in 20% of the articles was related to the use of the screening test, and typically derived from the reporting of data driven thresholds (as opposed to a-priori ones).

**Qualitative synthesis**

The area under the curve was reported for 78 samples with estimates ranging from 0.49-0.99. The AUC for each sample is reported in Table S4. Data were available to conduct meta-analyses for ASEBA, Conners, CPT, DSM-IV symptom scales, and SDQ measures. The results of these analyses are presented below. For eight scales, insufficient independent samples were available to conduct a meta-analysis (Table 1 and Table S3). Overall, these other scales had high accuracy with AUC ranging from 0.76-1.00, with 69.2% reporting an AUC > 0.85.

Sufficient information to calculate the 2 x 2 contingency tables was reported for 85 samples. The contingency tables, including the diagnostic accuracy coefficients for each cut-off meta-analyzed, are reported for each study sample in Table S4. For twelve scales, insufficient independent samples were available to be meta-analyzed (Table 1 and Table S3). The sensitivity and specificity for these measures varied considerably across samples from unacceptably low (SE = 0.31, SP = 0.15)16, to excellent (SE = 1.00, SP = 1.00)17.

**Meta-analysis: Area Under the Curve**

Table 1 displays a summary of the meta-analyses conducted to explore the overall diagnostic accuracy of the screening measures (see Figures S2-S14 for forest plots). All of the pooled estimates 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 (I2 = 17.3%-98.5%) for all measures except the DSM-IV based inattention subscale (I2 = 17.3%) and hyperactivity/impulsivity subscale (I2 = 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 S9). This may be an indication that shared method variance is influencing the findings, given that parents are typically the main informants for diagnostic interviews. Furthermore, studies that used case-control designs tended to have the highest AUC across all measures, followed by community samples, and then studies with high-risk samples (Table S10).

**Meta-analysis: Sensitivity and Specificity**

Table 2 displays a summary of the meta-analyses conducted to determine pooled sensitivity and specificity of each measure at the most commonly used thresholds. None of the measures achieved both acceptable sensitivity and specificity. The ASEBA Attention Problems subscale at a cut-off of T > 60 and SDQ HI subscale borderline cut-off had acceptable sensitivity (>0.8), but unacceptably low specificity (0.48 and 0.64, respectively). Whilst, the ASEBA attention problems scale at a cut-off of T> 70 (SP = 0.85), and the CBCL DSM-Oriented ADHD subscale at both a 5 (SP = 0.81) or 6 (SP = 0.91) cut-off had acceptable specificity, these scales had unacceptably low sensitivity (0.38, 0.75, and 0.52, respectively). With the exception of the CBCL DSM-Oriented subscale at a cut-off of 5, there was evidence of substantial heterogeneity between studies for both sensitivity and specificity.

A meta-regression comparing reporters for the ASEBA Attention Problems subscale at threshold T >70 indicated that there was no difference between reporters in terms of sensitivity. However, compared to parents, both self-report (p=0.001) and teacher-reporters (p=0.041) had poorer specificity. In contrast, a meta-regression comparing high-risk to community-based study populations on the ASEBA Attention Problems subscale found no significant difference in sensitivity or specificity. A meta-regression comparing reporters on the SDQ clinical cut-off indicated no significant difference in sensitivity or specificity between parent and self-reports. Similarly, there was no significant difference between studies with high-risk compared to community-based populations in sensitivity or specificity.

The summary ROC curves for the Conners, CPT, and DSM-IV based measures are displayed in Figure 2, with descriptive data included in Table 3. The SROC are shown in Figures 2a, 2b, and 2c for measures using an ADHD total symptoms subscale, hyperactive/impulsive subscales (errors of commission for CPT), and inattentive subscales (errors of omission for CPT), respectively. 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 S11 and S12 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.

**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 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 likely, by virtue of having a to 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 between studies make it difficult to be sure how any of the measures would perform in real world screening.

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. This finding may well be due to the prevalence of ADHD-like symptoms in the differing groups and the increasing rates of other diagnoses who also yield higher results in the ADHD screening tools. 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. However, in high-risk and clinical samples there are high rates of psychopathology. These populations are more likely to be experiencing some ADHD symptoms that do not meet full criteria for an ADHD diagnosis. Further, 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).

Whilst the AUC findings indicate good overall accuracy of screening tools, this is less important to clinicians or for population-based screening than is the accuracy of the measure at a specific cut-off. Clinicians need to know which measure they should be using, related to its optimal sensitivity and specificity at a pre-specified cut-off to indicate whether an individual should receive further assessment and potentially treatment. However, none of the tools in this systematic review and meta-analysis met a 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 but this is 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 colleagues18 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 after which the teachers of all screen-positive children were interviewed by the SNAP-IV trained teacher. 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 stage19.

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. 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. 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. 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.

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 25% of the included articles used such a standard. 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.

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 world2,3 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 reported survey combined with CPT), or a two-stage screening process.

**References**

1. Polanczyk GV, Willcutt EG, Salum GA, Kieling C, Rohde LAJIjoe. ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. 2014;43(2):434-442.

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.

4. Thomas R, Mitchell GK, Batstra L. Attention-deficit/hyperactivity disorder: are we helping or harming? *BMJ.* 2013;347: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. 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.

7. 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.

8. 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.

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

10. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5(9):1315-1316.

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.

12. 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.

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

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

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

Table 1 Meta-analytic estimates of the Area Under the Curve of ADHD screening tools

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Screening tool | Number of samples | Combined N cases | Combined N controls | Pooled AUC (95% CI)  | Heterogeneity (I2) | Heterogeneity (H) |
| All measures ADHD subscale | 66 | 8499 | 79323 | 0.83 [0.80; 0.86] | 98.5% [98.4%; 98.7%] | 8.29 [7.89; 8.72] |
| All measures IA subscale | 18 | 1516 | 1702 | 0.77 [0.73; 0.81] | 82.0% [72.5%; 88.2%] | 2.35 [1.91; 2.91] |
| All measures HI subscale | 18 | 1451 | 1671 | 0.76 [0.70; 0.82] | 91.7% [88.3%; 94.0%] | 3.46 [2.93; 4.10] |
| ASEBA attention problems subscale | 24 | 2292 | 3910 | 0.77 [0.72; 0.83] | 96.6% [95.8%; 97.3%] | 5.45 [4.89; 6.09] |
| CBCL DSM oriented subscale | 8 | 596 | 2319 | 0.81 [0.76; 0.86] | 70.4% [38.6%; 85.7%] | 1.84 [1.28; 2.65] |
| SDQ | 20 | 3095 | 72406 | 0.82 [0.78; 0.86] | 95.8% [94.6%; 96.8%] | 4.89 [4.30; 5.57] |
| CPT total | 10 | 655 | 626 | 0.87 [0.83; 0.92] | 85.3% [74.7%; 91.4%] | 2.61 [1.99; 3.42] |
| Omissions | 10 | 715 | 747 | 0.75 [0.68; 0.82] | 88.2% [80.3%; 92.9%] | 2.91 [2.26; 3.75] |
| Commissions | 9 | 602 | 644 | 0.73 [0.67; 0.80] | 80.3% [63.4%; 89.4%] | 2.25 [1.65; 3.07] |
| DSM-IV oriented scales ADHD subscale | 8 | 790 | 1031 | 0.87 [0.81; 0.94] | 90.4% [83.5%; 94.4%] | 3.22 [2.46; 4.23] |
| DSM-IV oriented scales inattention subscale | 4 | 421 | 592 | 0.75 [0.71; 0.80] | 17.3% [0.0%; 87.3%] | 1.10 [1.00; 2.81] |
| DSM-IV oriented scales hyperactivity subscale | 4 | 421 | 592 | 0.66 [0.61; 0.72] | 29.5% [0.0%; 74.2%] | 1.19 [1.00; 1.97] |
| Conners Revised Short Form-ADHD | 3 | 175 | 138 | 0.89 [0.77; 1.01] | 95.0% [88.8%; 97.8%] | 4.48 [2.98; 6.73] |

Table 2 Meta-analytic estimates of ADHD screening tools diagnostic accuracy

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Screening tool | Number of samples | Combined N cases | Combined N controls | Pooled sensitivity (95% CI) | Heterogeneity (X2) | Pooled Specificity (95% CI) | Heterogeneity (X2) | AUC |
| **CBCL DSM-IV ADHD subscale** |
| Cut-off = 5 | 2 | 51 | 340 | 0.75 (0.60, 0.85) | 0.01 | 0.81 (0.72, 0.88) | 2.83 | .826 |
| Cut-off = 6 | 2 | 61 | 330 | 0.52 (0.38, 0.66) | 0.10 | 0.91 (0.75, 0.97) | 8.83\* | .641 |
| **ASEBA Attention Problems subscale**  |
| T >60 | 4 | 464 | 274 | 0.89 (0.65, 0.97) | 63.22\*\* | 0.48 (0.40, 0.57) | 4.62 | .565 |
| T >65 | 3 | 239 | 106 | 0.73 (0.45, 0.90) | 14.24\*\* | 0.77 (0.41, 0.94) | 11.97\*\* | .808 |
| T >70 | 8 | 889 | 733 | 0.38 (0.27, 0.50) | 70.77\*\* | 0.85 (0.75, 0.91) | 106.95\*\* | .658 |
| **SDQ** |  |
| Borderline cut-off | 3 | 258 | 636 | 0.80 (0.62, 0.91) | 17.075\*\* | 0.64 (0.45, 0.80) | 42.34\*\* | .786 |
| Clinical cut-off | 6 | 2159 | 53327 | 0.59 (0.46, 0.70) | 131.76\*\* | 0.793 (0.65, 0.89) | 2687.42\*\* | .726 |

\*p<.01; \*\*p<.001

Table 3 Summary data for SROC analyses

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Screening tool | Number of samples | Combined N cases | Combined N controls | AUC |
| **All Conners measures** |
| ADHD subscale | 8 | 624 | 530 | 0.85 |
| Inattention subscale | 3 | 477 | 271 | 0.78 |
| Hyperactivity/impulsivity subscale | 6 | 584 | 545 | 0.75 |
| **Conners Revised Short Form** |
| ADHD subscale | 5 | 514 | 382 | 0.80 |
| Hyperactivity/impulsivity subscale | 4 | 526 | 449 | 0.81 |
| **CPT** |
| ADHD total | 14 | 810 | 765 | 0.83 |
| Errors of omission | 5 | 417 | 378 | 0.75 |
| Errors of commission | 4 | 304 | 275 | 0.74 |
| **DSM-IV based scales** |
| ADHD subscale | 17 | 1885 | 2259 | 0.84 |
| Inattention subscale | 4 | 715 | 811 | 0.67 |
| Hyperactivity/impulsivity subscale | 4 | 712 | 1037 | 0.74 |

Figure 1. PRISMA diagram

Records identified through database searches
n = 8094

EMBASE n = 2050

MEDLINE n = 813

PsycINFO n = 1608

PubMed n = 3623

Records identified through searching reference lists
n = 35

Records after duplicates removed
n = 7063

Records excluded
n = 6672

Records screened
n = 7063

Full-text articles excluded
n = 301

Not a screening tool (n = 65)

Participants from a highly specific population (n = 16)

No control group (n = 29)

No ADHD group (n = 24)

No clinical diagnosis of ADHD (n = 41)

Accuracy statistics not reported (n = 119)

Could not locate PDF (n = 6)

Unable to translate (n=1)

Full-text articles assessed for eligibility
n = 391

Studies included in qualitative synthesis
n = 90

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