UK National Screening Committee

A review of screening for autism spectrum disorders in pre-school children under the age of 5 years

External review against programme appraisal criteria for the UK National Screening Committee

Version: Public Consultation

Author: Exeter Test Group

Date: April 2022

The UK National Screening Committee secretariat is hosted by The Office for Health Improvement & Disparities.

About the UK National Screening Committee (UK NSC)

The UK NSC advises ministers and the NHS in the 4 UK countries about all aspects of population screening and supports implementation of screening programmes. Conditions are reviewed against evidence review criteria according to the UK NSC's evidence review process.

Read a complete list of UK NSC recommendations.

UK NSC, OHID, Department of Health and Social Care, 39 Victoria Street, London, SW1H 0EU
www.gov.uk/uknsc

Published: April 2022

Contents

List of tables

List of abbreviations

ADI-R = Autism Diagnostic Interview– Revised
ADOS = Autism Diagnostic Observation Schedule
ASQ-3 = Ages and Stages Questionnaire
AMSTAR = A MeaSurement Tool to Assess systematic Reviews
ART = Adapted Responsive Teaching
ASD = Autism Spectrum Disorder
CARS = Childhood Autism Rating Scale
CBCL = childhood behaviour checklist
CSBS IT = Communication and Symbolic Behaviour Scales Infant-Toddler
DISCO = the Diagnostic Instrument for Social and Communication disorders
DSM = Diagnostic and Statistical Manual of Mental Disorders
Ecl = Eclectic interventions
FUI = Follow-Up Interview
FYI = First Year Inventory
GDS = Global Developmental Screen
ICD = International Classification of Diseases
JA-OBS = Joint Attention Observation schedule
M-CHAT(-R/F) = Modified Checklist for Autism in Toddlers (Revised/ with Follow-Up)
MSEL = Mullen Scales of Early Learning
M-IL = Modified Intensive Learning programme
NICE = The National Institute for Health and Care Excellence
NVDQ = non-verbal developmental quotient
PEDS = Parents Evaluation of Developmental Status
PPV = positive predictive value
Q-CHAT = Quantitative Checklist for Autism in Toddlers
QUIPS = Quality in Prognostic Studies
QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies
PDD-NOS = Pervasive Developmental Disorder-Not Otherwise Specified
REIM = referral to early intervention and monitoring
RCT = randomised controlled trial
R-IL = Regular Intensive Learning programme
SACS-R = Social Attention Communication Surveillance- Revised
SCQ = Social Communication Questionnaire
SEND = Special Educational Needs and Disabilities
TIDOS = Three-item Direct Observation Screen
UK NSC = United Kingdom National Screening Committee
USPSTF = United States Preventive Services Taskforce

VABS = Vineland Adaptive Behaviour Scales

Plain English summary

Autism spectrum disorder (ASD) is a range of disorders of the nervous system and brain. These can affect how an individual communicates with others and how they socialise in groups. There is a lot of variation in how people are affected by ASD. About 1 in 55 school children in England have ASD. Screening has been proposed as a way of finding children with ASD early in life. The purpose of this would be to help them receive support to develop their language and social skills.

In 2011, the UK National Screening Committee (UK NSC) reviewed evidence on screening for ASD in children. Based on this, the UK NSC did not recommend screening because:

- there was not a good enough test for screening the general population
- it was not known if screening would improve long-term outcomes for children with ASD
- there was not an approach to screening which was acceptable to parents
- it was not clear why some children found to have ASD through screening at around the age of 2 years no longer have a diagnosis by the age of 4 years

The current review searched for any new evidence on whether screening for ASD would improve outcomes for children compared to those found through usual care following presentation of signs or symptoms.

The findings from the current review suggest that:

- it is still unclear how good the tests are at finding children likely to have ASD
- it is still unclear if treating children with ASD, who are found by screening, would improve their outcomes

More research is needed. This should include studies to find out:

- the proportion of children with a diagnosis of ASD at around the age of 2 years that keep their diagnosis after the age of 5 years
- what are the harms of a positive diagnosis of ASD at around the age of 2 years that is not confirmed after the age of 5 years
- if treating children with ASD, found through screening, would lead to improvements

Executive summary

Purpose of the review

The purpose of this review is to identify whether research in the last 10 years has addressed focussed gaps in the evidence as identified in the 2011 UK NSC review of screening for autism spectrum disorders in children below the age of 5 years. This is achieved by searching and synthesising evidence in young children on the diagnostic stability of ASD, the accuracy of screening tools for ASD, and the effectiveness of interventions in children with ASD who have been identified by population screening.

Background

Autism spectrum disorder (ASD) is a continuum of neurodevelopmental disorders. ASD is categorised by persistent and significant impairments in social interaction and communication and restrictive and repetitive behaviour to varying degrees, sub-classified by their severity (1, 2). In addition, a wide range of cognitive, learning, language, medical, emotional and behavioural problems (including self-injurious, challenging, and sometimes aggressive behaviours), co-occur to variable degrees. Studies suggest that >70% of individuals with ASD have other, coexisting disability, health or neurodevelopmental conditions (3).

NICE guidance lists features suggesting possible autism in preschool children. 2 years is the earliest age mentioned when characteristic symptoms or traits in behaviour are thought to differentiate affected children from typically developing children (4). However, there is uncertainty as to the stability, over time, of diagnoses of ASD at such early ages. If a diagnosis of ASD at age 2 years is considered to be stable, there would be a good basis for screening at early ages (assuming appropriate screening tools are available), and initiating effective treatment. If ASD diagnoses at early ages are not considered to be stable, this has implications for the rationale for screening and diagnosing young children.
Worldwide estimates for the prevalence of ASD are variable, ranging from <0.1% in Bangladesh in children 0-5 years old to 9.3% in Japan in 6-9 year olds (5). A study published in 2021, which included over 7 million school pupils aged 5-19 years in England, estimated the prevalence of ASD (defined using Special Educational Needs and Disabilities (SEND) registry data) as 1.76% (95%CI 1.75%, 1.77%)(6). Prevalence was higher in males than in females: 2.81% (2.79%, 2.83%) compared to 0.65% (0.64%, 0.66%).

The rationale for screening is that by screening and identifying ASD early in life, preferably before critical age-limited language development windows have closed, young children can receive ASD targeted interventions to foster their improved communication which will enable them to thrive and provide an advantage later in life. This needs to happen before developmental plasticity, the ability to acquire new skills, is lost.

Because identification and intervention provide the rationale for screening, the effectiveness of early interventions is crucial for early screening in reducing symptoms of ASD and in improving young children's life chances. However, the effectiveness of early intervention has hitherto remained unclear, due in part to the poor evidence base (as identified in the 2011 UK NSC review).

Focus of the review

The aim of this evidence summary is to find out whether the available evidence has addressed focussed gaps in the evidence identified in the 2011 UK NSC review through the following questions:

1. What is the diagnostic stability of ASD, in children diagnosed aged under 5 years?
2. What is the accuracy of screening questionnaires in children under the age of 5 to identify ASD at various ages?
3. Has the benefit of early intervention in children aged 5 years and younger, detected through screening been demonstrated?

Relevant studies were identified by searching MEDLINE, Embase, CINAHL, APA PsycInfo, Cochrane Database of Systematic Reviews, CENTRAL database and ClinicalTrials.gov. Only studies published in OECD (excluding South Korea and Mexico) and EEA countries since 1st January 2010 were included.

Recommendation under review

Based on the 2011 UK NSC review the decision to not screen for ASD was reaffirmed. The review concluded that a diagnosis of ASD in young children may not be stable, that population screening tools for ASD in young children did not report adequate sensitivity or positive predictive values, and may not be acceptable to parents, and that the effectiveness of interventions in screened populations was unclear.

Findings and gaps in the evidence of this review

What is the diagnostic stability of ASD, in children diagnosed aged under 5 years? (Question 1)

Estimates of the diagnostic stability of ASD ranged from 71.9% to 100% in the 5 studies identified that involved children detected through population screening. However, all studies raised concerns regarding their risk of bias, including a lack of blinding of follow-up diagnostic assessments in 4 of these studies. There is a lack of evidence on the stability of ASD diagnoses beyond 4 or 5 years old. This is because follow-up in the studies did not extend beyond two years from diagnosis. Further studies are needed that ensure that diagnostic evaluation at follow-up is blind to that made initially, and that follow-up is longer than 2 years, allowing assessment of the diagnosis when children are at primary school age.

What is the accuracy of screening questionnaires in children under the age of 5 to identify ASD at various ages? (Question 2)

Most of the included studies evaluated versions of the M-CHAT, which had been translated into non-English language. Estimates of sensitivity for M-CHAT(R/F) ranged from 0.67 to 1, with many studies reporting estimates of around 0.8 depending on age group or cut-off used. Little evidence was found on whether age or other characteristics impact on screening accuracy. Screening uptake was variable across studies. Only one study reported experiences by 10 nurses involved in the screening programme, which were generally positive.
The included studies suggest that the tools might have a more general purpose than just identifying ASD. Thus, if the target condition of ASD is expanded to include children who would potentially benefit from intervention, then higher positive predictive values would be observed for these tools. More studies are needed that attempt to follow-up a proportion of children who screen negative, so that reliable estimates of sensitivity and specificity can be obtained. Such studies should also ensure that diagnostic evaluation is conducted blind to the screening results. Ideally, these studies would evaluate and compare more than one tool, preferably comparing tools that involve observation of children, with tools that involve parent-completed questionnaires, for example. Evidence on factors affecting uptake or completion of ASD screening, and how better uptake/completion might be achieved would also be warranted.

Has the benefit of early intervention in children aged 5 years and younger, detected through screening been demonstrated? (Question 3)

Only 4 studies were found that evaluated interventions in young children identified through screening for ASD: 3 RCTs and one cohort study. The largest, which still only included 89 children, found that treatment effect (reduced ASD severity) was maintained at 2 years follow up, however the study sample was contaminated with referred patients in one of the research sites, making the results of this study less relevant. The other studies found no evidence of improved outcomes. As the maximum follow-up among the studies identified was just 2 years, there is limited evidence on the long-term outcomes of early intervention in these young children identified through screening. Larger studies with longer follow-up would be needed. However, due to the prevalence of ASD within the general population in the UK, and issues of attrition, such studies will need to effectively reduce the likelihood of children/families dropping out from the study at various time-points.

Recommendations on screening

Overall, the evidence reviewed here does not indicate that screening for ASD should be recommended for children aged ≤5 years.
Although there is some uncertainty as to the performance of screening tools to identify children with ASD (Question 2), the main limiting factors are uncertainty as to the stability of diagnoses of ASD when made at such young ages (Question 1), and the current lack of evidence on the effectiveness of interventions for children identified through ASD screening (Question 3).

Limitations

The available evidence relevant to all 3 questions is limited. For question 1, studies are limited by a lack of blinding of initial diagnostic assessments at the follow-up diagnostic assessment, and by a lack of follow-up.

Particular aspects of study design limited many of the studies included in this review. For instance, a lack of blinding limits the interpretation of most of the studies that evaluated diagnostic stability and many of the screening accuracy studies. While short follow-up periods limit the extent to which diagnoses can be said to be stable beyond 2 years after diagnosis, and interventions effective after 2 years. A particular limitation of many of the screening accuracy studies is to what extent, and how, children who were negative on the

screening tool were followed-up, so that reliable estimates of sensitivity and specificity estimates could be obtained.

The review is limited by the inclusion of English-language only studies, that were published after 2009. Only a proportion of articles identified from the database searches were double-screened at title and abstract, or full-text stage. Moreover, the level of agreement between reviewers on inclusion of relevant studies was generally low, due to aspects of study design not being reporting clearly. It is therefore possible that some relevant studies have not been included in the review. However, given all of the current uncertainties and limitations in the evidence across the 3 research questions, it is unlikely that omission of further studies would lead to a different recommendation at this point.

Evidence uncertainties

Further work is warranted to help address all 3 questions. To examine the stability of diagnoses of ASD made following screening (Question 1), further studies are needed that ensure that diagnostic evaluation at follow-up is blind to that made initially, and that follow-up is longer than 2 years.

To assess the performance of screening tools (Question 2), more studies are needed that attempt to follow-up a proportion of children who screen negative, so that reliable estimates of sensitivity and specificity can be obtained. Such studies should also ensure that diagnostic evaluation is conducted blind to the screening results. Ideally, these studies would evaluate and compare more than one tool, preferably comparing tools that involve observation of children, with tools that involve parent-completed questionnaires, for example. Evidence on factors affecting uptake or completion of ASD screening, and how better uptake/completion might be achieved would also be warranted.

To better evaluate the effectiveness of interventions in children with ASD identified through screening (Question 3), larger studies with longer follow-up would be needed. However, due to the relatively low prevalence of ASD, and issues of attrition, such studies will need to effectively reduce the likelihood of children/families dropping out from the study at various time-points.

Introduction and approach

Background

Autism spectrum disorder (ASD) is a continuum of neurodevelopmental disorders(1). ASD is categorised by persistent and significant impairments in social interaction and communication, and restrictive and repetitive behaviour to varying degrees, sub-classified by their severity(1). In addition, a wide range of cognitive, learning, language, medical, emotional and behavioural problems (including self-injurious, challenging, and sometimes aggressive behaviours), co-occur to variable degrees. For example, cognitive ability might range from profound intellectual disability to average or above average intelligence. Studies suggest that >70% of individuals with ASD have other, coexisting health, disability or neurodevelopmental conditions (3).

NICE guidance lists features suggesting possible autism in preschool children. 2 years is the earliest age mentioned when characteristic symptoms or traits in behaviour are thought to differentiate affected children from typically developing children (4). Risk factors that have been identified are genetic and, environmental (such as infection in pregnancy) but many are not specific to autism, and are common to a wide range of neurodevelopmental disorders (7-9). Genetic and environmental triggers are interrelated and are thought to alter brain development from very early on in life, resulting in the reorganization of neurological pathways that underlie cognition and behaviour, and affecting sensitivity to environmental and social inputs as children mature (10, 11). ASD is therefore highly heterogenous: individuals with autism have varied developmental trajectories across multiple behavioural dimensions, including core autistic traits and behaviours, cognition, sensitivities to sensory stimuli, social abilities and functional skills. ASD is thought to improve across the lifespan in most cases, but even for those individuals able to lead fulfilling independent lives at adulthood, many struggle with mental health, educational, and interactional difficulties (12).

Worldwide estimates for the prevalence of ASD are variable, ranging from <0.1% in Bangladesh in children 0-5 years old to 9.3% in Japan in 6-9 year olds (5). These differences in prevalence estimate are mostly accounted for by varying techniques to identify cases, sampling biases and the cultural frame of reference which may invoke stigmatisation of identified children, leading to under-identification, as well as a local understanding and categorisation of ASD in response to local service provision and the national educational needs context (12).

A study published in 2021, which included over 7 million school pupils aged 5-19 years in England, estimated the prevalence of ASD (defined using Special Educational Needs and

Disabilities (SEND) registry data) as 1.76% (95%CI 1.75%, 1.77%)(6). Prevalence was higher in males than in females: 2.81% (2.79%, 2.83%) compared to 0.65% (0.64%, 0.66%). Variation in prevalence was observed across ethnicity/race, with Black pupils having the highest prevalence, 2.11% (2.06%, 2.16%). Large variation was seen across geographical regions, ranging from 0.63% (0.46%, 81%) in the Cotswolds to 3.38% (3.15%, 3.61%) in Solihull. Possible reasons for such geographical variation include inconsistencies across the country in terms of the diagnostic process, variability in educational support or thresholds for accessing SEND support (6).

A global systematic review summarised time trends from over 25 studies around the world, over a 60 year time frame (13): meta-analyses of prevalence estimates from 11 European countries and the US showed significant evidence of increasing prevalence. Included studies used varied methods of case definition, including clinically diagnosed cases, and parent report of diagnosis, as well as research instruments (such as the Autism Diagnostic Interview-Revised, ADI-R). Such increases may illustrate expanding boundaries of diagnostic classification, shifts in policy and awareness, and consequent increased demand for diagnosis and service provision, but they may also relate to an underlying increase in the proportion of people with autistic difficulties. In the UK, although, Taylor et al.(14) found the incidence of recorded autism at age 8 years remained stable across a 6 year period, in an analysis that expanded on and repeated these findings utilising a more comprehensive version of the General Practice Research Database, year on year increases in application of diagnosis from 1998 to 2018 was observed in all age groups including younger children aged 2-5 (15).

Screening to identify ASD has been discussed for many years and a number of screening tools are available. These tools usually consist of questionnaires or checklists for carers or professionals to complete based on their observations of the child in question. The most commonly used questionnaire is the M-CHAT (with revisions), which has 23 items intended to be completed by the child's carer based on the child's usual behaviour (16). Revisions include a follow-up interview with the carer by a professional if responses to the questionnaire indicate the possibility of ASD. The questionnaire items are scored with yes/no responses and cover areas including joint attention skills, motor and sensory abnormalities, and the child's early language and communication skills. Other tools include checklists used by professionals in their observations of a child. For instance, the Joint Attention Observation Schedule (JA-OBS) which focusses on whether the child engages in joint attention (focussing their attention on something with another person) (17). Due to the subjective nature of these questionnaires and checklists, users should be cautious of potential misinterpretations and cultural differences when applying such tools in languages and countries that are different to where the tools were developed.

The rationale for screening is that by screening and identifying ASD early in life, preferably before critical age-limited language development windows have closed, young children can receive ASD targeted interventions to foster their improved communication which will enable them to thrive and provide an advantage later in life. This needs to happen before developmental plasticity, the ability to acquire new skills, is lost.

Because identification and intervention provide the rationale for screening, the effectiveness of early interventions is crucial for early screening in reducing symptoms of ASD and in improving young children's life chances. However, the effectiveness of early intervention has hitherto remained unclear, due in part to the poor evidence base (as identified in the 2011 UK NSC review). Interventions for ASD are almost all behavioural and as such are costly, time-consuming and almost always parent mediated with expectation of delivery often placed on mothers (who are predominantly the primary carers) (18), who may lose out on a range of career and other opportunities as a consequence. Because of the intense effort involved in parent mediated behavioural interventions it is crucial to have strong evidence justifying their use.

Furthermore it is possible a child who is slow to develop as a toddler may catch up when older, and some evidence suggests that some children who meet ASD criteria at very young ages may improve to sub-clinical levels later on (19, 20). If diagnoses can be considered stable, then making a diagnosis at earlier ages (through screening) has obvious advantages of earlier availability of any effective treatments and support to families. However, if ASD diagnoses in very young children cannot be considered stable, families may unnecessarily experience the time-consuming and complex diagnostic process, with a child receiving a diagnosis of ASD, and uncertainty as to when such a diagnosis may be removed, if ever. Thus, screening might be inappropriate if carried out too early as it might result in an overtly unstable diagnosis. Therefore, it is important to establish the stability of diagnosis into later childhood when made at very young ages.

It is important not to embark upon an endeavour that will result in many false positives and negatives, as the consequences and emotional impact of being told a child has autism are likely to be severe, ranging from shock, denial, fear, anxiety, guilt, anger, sadness, to distress (21, 22). Therefore, high sensitivity, specificity and positive predictive values are required to justify ASD screening at young ages.

Current policy context and previous reviews

In 2011, the UK NSC reviewed the evidence on screening for ASD in children ≤ 5 years of age. The review had 3 key questions:

1. Can any approach to screening for ASD demonstrate acceptable performance, in terms of both sensitivity and positive predictive value, in a general population based study?
2. Why do so many parents of children who fail initial screening tests for ASD drop out of the screening process before it has completed, and can the process be refined so that the drop-out rate is reduced?
3. Does early intervention lead to significant improvements later in childhood, or greater independence and improved vocational and social functioning in adulthood?

In response to these questions, the report concluded that

- there was no evidence on acceptable screening approaches in children ≤ 5 years old in the general population
- between a third and a half of parents dropped out of the screening process before completion
- evidence on effectiveness of early intervention was variable, with uncertainty as to whether short-term improvements continued over time.

The 2011 UK NSC review also highlighted that studies assessing the natural history of ASD in young children, by comparing initial diagnoses with diagnoses made after a period of time, suggested that diagnosis of ASD may not be stable. In other words, for children who were initially diagnosed with ASD, a proportion of these were not found to meet diagnostic criteria for ASD at a follow-up evaluation.

Based on the review, the UK NSC recommended not to screen for ASD.

The American Academy of Paediatrics proposes that children should be screened for autism at 18 and 24 month check-ups (23). However, the United States Preventive Services Taskforce (USPSTF) produced a recommendation in 2016 that screening should not be carried out in children aged 18-30 months where there is no diagnosis of developmental delay, or no concerns of ASD have previously been identified (2). This recommendation was based on an evidence review of the accuracy of screening tools, and harms and benefits of screening for ASD in children ≤ 3 years of age (24), including studies published between 2000 and August 2014.

Although the USPSTF reported “adequate evidence” for the ability of existing screening tools to identify ASD in this age group (M-CHAT(-R/F) especially), the recommendation to not screen was based on a lack of evidence for the benefits and harms of screening for ASD, in particular, a lack of evidence on the effectiveness of early interventions (2). While 46 RCTs evaluating interventions were identified in the review – with variations in study design, interventions assessed, population characteristics, and being generally small, of fair quality and limited follow-up – none of these RCTs included children identified through screening (24). The children included in these studies generally had “significant

impairments in cognition, language and behaviour", so likely to have more severe symptoms than those identified through screening, and "were older than the age group for which the screening tools were developed" (2).
Thus, although there was some evidence from some RCTs that interventions could be effective in children with ASD, the applicability of these findings to a population of young children identified through screening is uncertain.

Most interventions are behavioural in nature, and pharmacological treatments are not recommended by the National Institute for Care and Clinical Excellence (NICE) for ASD. Parent mediated behavioural interventions, are promoted, with previous reviews of behavioural interventions for the youngest children suggesting that some subgroups of ASD children display more prominent gains across studies (25), sub-group characteristics associated with greater gains are, however, not well understood. Therefore, it is not possible to state with any certainty which interventions will be useful for which children with ASD.

In England and Northern Ireland children are assessed at age 2-2.5 years through the "Healthy Child Programme" (using the Ages and Stages Questionnaire-3) and "Healthy Child, Healthy Future" respectively. The "Scottish Child Health Programme" assesses development at 27-30 months (again using Ages and Stages Questionnaire) and the "Healthy Child Wales Programme" at 27 months. These assessments are not part of a formal UK NSC recommended screening programme and although they are not designed to identify children with ASD specifically, it is possible that the behaviours and signs of undiagnosed ASD might be identified through these programmes.

Objectives

The following 3 questions are covered in this review. How they relate to the UK NSC screening criteria is shown in Table 1 below.

Question 1: What is the diagnostic stability of ASD in children diagnosed aged under 5 years?
Sub-question: Are children who screen positive, for example at age 2, still considered to have ASD after 5 years, 10 years, etc?
Question 2: What is the accuracy of screening tools in children under the age of 5 to identify ASD?
Sub-questions: Does the age at which the screening test is performed affect accuracy?
Do other characteristics affect the accuracy?

Are there incidental findings?

Question 3: Has the benefit of early intervention in children aged 5 years and younger, detected through screening been demonstrated?

Table 1. Key questions for the evidence summary, and relationship to UK NSC screening criteria

	Criterion	Key questions	Studies Included
	THE CONDITION		
1	The condition should be an important health problem as judged by its frequency and/or severity. The epidemiology, incidence, prevalence and natural history of the condition should be understood, including development from latent to declared disease and/or there should be robust evidence about the association between the risk or disease marker and serious or treatable disease.	Question 1: What is the diagnostic stability of ASD in children diagnosed aged under 5 years?	5 studies (Allison 2021(26), Pierce 2019(27), Barbaro 2017(28), Spjut Jansson 2016(29), Guthrie 2013(30))
	THE TEST		
4	There should be a simple, safe, precise and validated screening test.	Question 2: What is the accuracy of screening tools in children under the age of 5 to identify ASD?	21 articles reporting on 20 studies (Allison 2021(26), Jonsdottir 2021(31), Wieckowski 2021(32), Jonsdottir 2020(33), Kerub 2020(34), Magan-Maganto 2020(35), Oner 2020(36), Mozolic-Staunton 2020(37), Dai 2020(38), Achenie 2019(39), Suren 2019(40), Topcu 2018(41), Catino 2017(42), Baduel 2017(43), Kondolot 2016(44), Wiggins 2014(45), Robins 2014(16), Ben-Sasson 2013(46), Chlebowski 2013(47), Nygren 2012(17), Canal-Bedia 2011(48))
5	The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.		
	THE INTERVENTION		
9	There should be an effective intervention for patients identified through screening, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care. Evidence relating to wider benefits of screening, for example those relating to family members, should be taken into account where available. However, where there is no prospect of benefit for the individual screened then the screening	Question 3: Has the benefit of early intervention in children aged 5 years and younger, detected through screening been demonstrated?	4 studies (Baranek 2015(49), Watson 2017(50), Spjut Jansson 2016(29), Whitehouse 2021(51))

	Criterion	Key questions	Studies Included
	programme shouldn't be further considered.		

Methods

The current review was conducted by Exeter Test Group, in keeping with the UK NSC evidence review process. The protocol was registered on PROSPERO (CRD42021231868). Database searches were conducted on 16-19th November 2021 to identify studies relevant to the questions detailed in Table 1. Update searches were conducted on 1st July 2021. All searches were limited to the beginning of 2010 to the search dates.

Eligibility for inclusion in the review

All publications identified by the searches were screened and grouped by Review Question (using the inclusion and exclusion criteria above) by a single researcher. This was done first at title and abstract level and then, if selected in the first round, at full text. To minimise the possibility of human error, the first 10% of the titles/abstracts in the first round and 10% of the full texts in the second round were screened independently by a second researcher. Results were compared, and any discrepancies were discussed before the first researcher continued screening the rest of the papers. At the end of each round, the second researcher screened another random 10% of papers to check the overall level of agreement.

Databases/sources searched

A single search strategy covering all 3 questions was developed using a combination of free-text and medical subject headings. The search was carried out on MEDLINE (via OvidSp), EMBASE (via OvidSp), CINAHL (via EBSCOhost) and PsycINFO (via OvidSp) on 16th November 2020; Cochrane Database of Systematic Reviews and CENTRAL on 17th November 2020. Clinical trials.gov was searched on 19th November 2020. All database searches were updated on 1st July 2021. At the time of searching the WHO ICTRP was inaccessible due extremely heavy usage during the COVID pandemic. Reference lists of included studies were checked for other relevant publications. The search strategy is presented in Appendix 1.

The following review process was followed:

1. After removing duplicates across databases, the records identified were imported into EndNote X8.2 (Thomson Reuters) and combined. Each abstract was reviewed against the combined inclusion/exclusion criteria by a single reviewer (either RH, BG, JW or JP). Where the applicability of the inclusion criteria was unclear, the article was included at this stage in order to ensure that all potentially relevant studies were captured. A second independent reviewer (JP) validated 20% of the total screening decisions (but only those of RH, BG or JW). Any disagreements were resolved by discussion until a consensus was reached.
2. Full-text articles required for the full-text review stage were acquired.
3. Each full-text article was reviewed against the combined inclusion/exclusion criteria by one reviewer (either BG, JW or JP), who determined whether the article was relevant to one or more of the review questions. A second independent reviewer (either JP or BG) validated 20% of the total screening decisions. Any disagreements were resolved by discussion until a consensus was reached.

Eligibility criteria for each question are presented in **Error! Reference source not found.** below.

Table 2. Inclusion and exclusion criteria for the key questions

Key question	Inclusion criteria							Exclusion criteria
	Population	**Target condition**	**Intervention**	**Reference Standard**	**Comparator**	**Outcome**	**Study type**	
1	Children aged ≤ 5 years diagnosed with ASD (screen or clinically detected)	ASD	NA	Any validated measure		Continued diagnoses of ASD at a specified time after initial diagnosis (prioritising those studies using same measure as at initial diagnosis)	Longitudinal cohort studies, systematic reviews and/or meta-analyses of these	Non-English language, published before 2010
2	Children aged ≤ 5 not diagnosed with ASD and for whom no concerns of ASD have been raised by parents, other caregivers, or clinicians	ASD	Any specific screening tool to identify ASD, performed by health visitors, GPs, parents, other non-specialist HCPs. Also any general (multiphasic) tools to identify range of conditions including ASD as a target condition	Multidisciplinary team assessment (as defined in the original study) and clinical judgement: NICE guidelines, SIGN guidelines. Clear reference standard as defined in the study and its standing	Any other screening tool	Sensitivity Specificity PPV NPV Incidental findings	Any test accuracy study (and systematic reviews and meta-analyses of these), with concurrent validation (reference test performed at the same time as the index test)	Non-English language, published before 2010. Case-control study design, where cases are children who already have a diagnosis of ASD
3	Children aged ≤ 5 years identified with ASD through				Any intervention No treatment (control group, placebo)	Improvements in ASD core deficits/ symptom severity, including but not limited to: adaptive behaviour,	RCTs and systematic reviews of RCTs prioritised. Other study	

	screening, having had no previous concerns raised by parents, other caregivers, or clinicians, for ASD. If no or few studies in this population are found, include studies of these interventions in children under the age of 5 identified through routine practice	Where identified through screening: any intervention given after diagnostic care or routine practice (i.e. not through screening) Where identified through routine clinical practice/ diagnostically detected: any "late" intervention (started after the age of 5)	expressive language skills, receptive language skills, IQ, challenging/problem behaviour, visual spatial skills, cognitive skills, academic skills, social skills, initiative behaviours	types, including cohort studies, considered if satisfactory (for example, sufficiently powered) RCTs are not available.

Appraisal for quality/risk of bias tool

The following tools were used to assess the quality and risk of bias of each study included in the review:

- for Q1, studies reporting on diagnostic stability were assessed using a modification of the Quality in Prognostic Studies (QUIPS) tool (52).
- for Q2, the screening accuracy studies were assessed using slightly modified versions of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (53) or the QUADAS-C tool for comparative accuracy studies, for systematic reviews we used AMSTAR (54).
- for Q3, RCTs were assessed using the Cochrane Collaboration's "Risk of Bias" Tool (55), and cohort studies were assessed using the Risk of Bias in Non-randomised Studies - Interventions (ROBINS-I) checklist (56).

Data extraction

For each research question, a bespoke data extraction sheet was developed and piloted. One reviewer extracted all data which was checked by a second reviewer. Any discrepancies were resolved through discussion, and inclusion of a third reviewer if necessary.

Synthesis

A narrative synthesis of results reported for included studies is presented for each question. Where summary estimates have not been reported in studies, but raw data is available to calculate them, this has been done. This includes calculation of 95% confidence intervals, where exact binomial confidence intervals have been calculated using Stata v16 (57).

Question level synthesis

Criterion 1 — The condition should be an important health problem as judged by its frequency and/or severity. The epidemiology, incidence, prevalence and natural history of the condition should be understood, including development from latent to declared disease and/or there should be robust evidence about the association between the risk or disease marker and serious or treatable disease

Question 1 — What is the diagnostic stability of ASD in children diagnosed aged under 5 years?

Sub-question: Are children who screen positive, for example at age 2, still considered to have ASD after 5 years, 10 years, etc?

To understand the natural history of ASD in children, studies have looked at what proportion of children maintain their initial diagnosis of ASD after a length of follow-up. These studies aim to provide information on the reliability of ASD diagnoses in young children over time. Considering whether a screening programme for ASD is appropriate, studies involving children identified through screening are most relevant as these children are likely to have less severe symptoms than children who have been clinically referred. This is because children with more severe symptoms stand out more obviously from their peers as neurodevelopmentally delayed, and are therefore more often noticed by parents, pre-school providers, nursery carers and others in the child's community, entailing ongoing referral to the clinic. Therefore, diagnoses in children with less severe symptoms may be more difficult to make, potentially affecting the reliability of such diagnoses.

In the 2011 UK NSC evidence review, studies reporting the proportion of children maintaining a diagnosis of ASD over time (diagnostic stability) was reviewed. Nine studies were identified, 4 of which were in a screening population (where children had not previously been identified as high-risk of, or with concerns regarding, ASD). The follow-up diagnoses in all studies were made by individuals who were not blinded to the initial diagnosis, casting doubt on the validity of later identification. Across the studies, the proportion of children maintaining a diagnosis at follow-up (approximately 2 years later) ranged from 63-70% for autism and 33-67% for pervasive developmental disorders – not otherwise specified (PDD-NOS). For ASD, which included both autism and PDD-NOS, 75-100% of children retained their diagnosis at follow-up. The 2011 UK NSC review concluded that diagnosis of ASD may not be stable.

Eligibility for inclusion in the review

Studies involving children aged ≤ 5 years diagnosed with ASD detected either via screening or clinically, and who were followed-up with a further diagnostic assessment were included. Studies were included regardless of the approaches taken to make a diagnosis at baseline and at follow-up. Non-English language studies and those published before 2010 were excluded.
After full text review the main reasons for exclusion of studies were that studies did not evaluate diagnostic stability were neither primary studies nor systematic reviews.

Description of the evidence

0contains a full PRISMA flow diagram (Figure 1), along with a table of the included publications and details of which questions these publications were identified as being relevant to (Table 15).

Of the total 5,498 titles and abstracts from the database searches, the full-text of 27 titles were reviewed for eligibility for this question. On closer inspection, 19 of the full-texts were not eligible. This included 2 systematic reviews: one which only included studies prior to the 2010 cut-off date (58), and another that included studies in children at high-risk of, or already diagnosed with, ASD (59). Both are excluded from further discussion. The 8 remaining articles were all primary studies. From the updated searches conducted in July 2021, 2 articles were screened at full-text, with one eligible for Q1, Allison (26). Thus, 9 studies were potentially relevant to this question.

Of these 9 primary studies, 5 included children identified through screening. The remaining 4 studies recruited participants who were referred due to developmental queries(60) or who already had an ASD diagnosis (61-63). Thus, these studies are not reported here but details can be found in the appendix.

The 5 primary studies which included children identified through screening were published between 2013 – 2021. The studies included 1580 children, in total, with a baseline mean age range of 19 to 36 months. One study was based in England (26), 2 studies were based in USA (27, 30), with 1 in Australia (28), and 1 in Sweden (29). The time interval between diagnosis and final follow up assessment was approximately 24 months, except for Spjut Jansson (29) which was 60 months. All studies included less than 100 children in follow-up assessments, with the exception of Pierce (27) who followed up over 1200 children.

In 4 of the 5 studies where children were identified through screening, all children meeting their inclusion criteria, regardless of whether their initial diagnosis was ASD, were followed up. Spjut Jansson (29) is the exception to this. Thus, information on whether children not initially diagnosed as having ASD but who may then go on to receive such a diagnosis at follow-up is available. A study-level summary of data extracted from each included publication is presented in 'Summary and appraisal of individual studies **Error! Reference source not found.**' (Table 17). Where the reviewers have performed calculations on the data presented in the publications, this has been clearly indicated in the tables. See Table 3 below for a summary of study characteristics, risk of bias and results, followed by brief descriptions of each study.

Table 3 Characteristics, risk of bias and results for studies addressing diagnostic stability in a population of children identified by screening

Study	Country and population	Screening tool	N with T1 & T2 data [N with T1 only; N offered screen]	Diagnostic process at T1 and T2	Age at T1 Length of follow-up (months)	Overall Risk of Bias	Results (95% CI)
Allison 2021**(26)**	England Registered on CHSD in Luton, Bedfordshire and Cambridgeshire	Q-CHAT	81 [121; 13070]	T1: Experienced, psychologist(s) performed the ADOS, ADI-R, MSEL, VABS. ICD-10 criteria. T2: as above	~24 [median] NR (≥48 months old)	P: Low A: High DA T1: Low DA T2: Low C: High Blind*: No	100% (66.4, 100)** retained possible autism diagnosis.
Pierce 2019**(27)**	America 75% children identified as "at-risk" from a screened population, 25% referred population	CSBS IT checklist	1269 [2241; NR]	T1: Experienced, registered psychologists performed the ADOS-2, MSEL, VABS. T2: as above	19 [mean] 20.2 [mean]	P: Low A: High DA T1: Low DA T2: Low C: High Blind*: No	84% (80, 87) retained ASD dx Change to ASD: 47% ASD features, 24% DD, 16% LD, 4% TD
Barbaro 2017**(28)**	Australia Children identified as "at-risk" from a screened population	Failing 3 of 5 behavioural items from the SACS	77 [99; >20,000]	T1: Developmental history, previous check-ups, ADOS-G Module 1, MSEL, ADI– R, FYI, CSBS IT, EDI, CHAT-23,	24 [time of scheduled check-up] 24	P: Unclear A: High DA T1: High DA T2: High C: Low Blind*: Yes	71.9% (53.2, 86.2)** retained ASD dx. 40% (22.7%,

				expert clinical judgment T2: as above but excluding ADI-R			59.4%)** retained autism dx. Change to ASD: 56.6% autism, 0% DD, 0% LD.
Spjut Jansson 2016 **(29)**	Sweden Children identified as "at-risk" from routine ASD population screening	NR	71 [100; NR]	T1: Multidisciplinary assessment, including cognitive/intellectual tests, ADOS-G and DISCO (for 72% of the children). Experienced professionals. T2: As above plus ADI	Approx. 36 [mean] Approx. 60	P: High A: High DA T1: High DA T2: Low C: High Blind*: No	93% (84.3, 97.7)* retained ASD dx.
Guthrie 2013 **(30)**	Australia Two-step screened population	First step: CSCB IT or parental concern. Second step: CSBS red flags for ASD using SORF.	82 [unclear; 5419]	T1: ADOS-T, video-recordings, home observations, parent reports, MSEL, VABS, consistent with DSM-IV criteria by experienced clinician T2: as above	19 [mean] 16 [mean]	P: Low A: Unclear DA T1: Unclear DA T2: Unclear C: Low Blind*: No (all details from T1 available at T2)	100% (93.6, 100)* retained ASD dx. Change to ASD: 21% deferred dx. 0% ASD ruled-out

A, attrition; ADI-R, Autism Diagnostic Interview– Revised; C, confounding; CHAT-23, Checklist for Autism in Toddlers-23; CHSD Child Health Surveillance Database; CSBS, Communication and Symbolic Behaviour Scales ; CSBS IT, Communication and Symbolic Behaviour Scales Infant-Toddler Checklist; DA, diagnostic assessment; DD, developmental delay; dx, diagnosis; EDI, Early Development Interview; FYI, First Year Inventory; LD, language delay; MSEL, Mullen Scales of Early Learning; NR, not reported; P, participants; Q-CHAT, Quantitative Checklist for Autism in Toddlers; SACS, Social Attention and Communication Study; T1, time 1; T2, time 2; TD, typically developing; SORF, Systematic Observation of Red Flags; VABS, Vineland Adaptive Behavior Scales

*Blind, Were individuals conducting the diagnostic assessment at T2 blinded to the details and/or findings of the diagnostic assessment at T1?; **95% confidence intervals calculated by review authors

Allison 2021 (26) posted Q-CHAT questionnaires to carers of 13,070 children aged 18-30 months old registered on the Child Health Surveillance Database from Luton, Bedfordshire and Cambridgeshire. Responses were obtained from 3,770 carers. To avoid missing children with autism where questions on the returned Q-CHAT had not been completed, each child had 2 total scores: observed score (where missing questions scored 0) and an imputed score (where missing questions scored a maximum of 4). The probability of being

invited for a diagnostic assessment depended on the total observed and imputed scores. As determined by the stratified sampling approach taken by Allison (2021), children with higher scores were more likely to be invited. Diagnosis was made based on ICD-10 criteria. Eighty-one children had a diagnostic assessment at T1 (approximately, on average, 24 months old), and at T2 (when children were ≥ 48 months old). All 9 children with possible autism at T1 who were re-assessed at T2, retained their diagnosis (100% stability, 95%CI: 66.4%, 100%). Two children diagnosed as atypical at T1, and 4 children diagnosed as typical at T1, all received diagnoses of possible autism at T2. This study was deemed to be of high risk of bias as diagnosis at T2 was not blinded to diagnosis at T1.

The American study by Pierce (27) is by far the largest study identified. Of 1,269 children followed-up, 75% were identified through community screening (during routine check-ups at ages 12, 18 and 24 months old), with the remaining 25% of children from referrals. Children were, on average 19 months old at their T1 assessment, with follow-up (T2) occurring on average, 20 months later. At both assessments, diagnosis was made by experienced, registered psychologists, who assigned children to: ASD, exhibiting ASD features, developmental delay, language delay, other issue (not specified), typically developing sibling of a child with ASD, or typically developing. The psychologists making diagnoses at T2 were not blind to the diagnosis given at T1.

Of 400 children diagnosed with ASD at T1, 84% (336/400, 95%CI: 80%, 87%) retained this diagnosis at T2. No statistically significant differences in diagnostic stability were found between boys and girls. However, Pierce showed a trend for stability increasing as age at initial diagnosis of ASD increased: in children aged <14 months at initial diagnosis, stability was 50% (95%CI: 32%, 69%), with estimates of diagnostic stability >84% in children >24 months old. Of the 64 children who lost their ASD diagnosis at T2, 55% received a diagnosis of ASD features, 19% had a diagnosis of development or language delay, 16% received some other diagnosis and 11% were deemed to be typically developing at T2. At T2, 47% of children initially diagnosed as having ASD features, 24% of children initially diagnosed with development delay and 16% of children initially diagnosed with language delay, subsequently received a diagnosis of ASD. Due to the individuals undertaking diagnosis at T2 not being blind to diagnosis at T1,and that children were identified through screening and from clinical referral, this study is deemed to be at high risk of bias.

Barbaro 2017 (28) is an Australian study including 99 children identified as being “at-risk” of ASD, determined using the SACS (Social Attention and Communication Study) screening tool. Over 20,000 children were screened in the community at their 24-month routine check-up. A best-estimate diagnosis was given at T1 based on clinical judgement from 2 of the study authors, informed by a number of sources. Diagnosis at T2 was approximately 2 years later, when children were roughly 48 months old. A similar approach to diagnosis as

at T1 was taken. Importantly, clinicians undertaking the T2 diagnostic assessment were blind to the best-estimate diagnosis made at T1. At both time-points children were given a diagnosis of autism, ASD (which included Asperger's disorder or PDD-NOS), or developmental or language delay. None of the 99 children identified as at-risk for ASD at T1 were found to be typically developing. Seventy-seven children were followed-up at T2, with 71.9% (23/32, 95%CI: 53.2%, 86.2%) retaining a diagnosis of ASD, and 40% (12/30, 95%CI 22.7%, 59.4%) retaining a diagnosis of autism. Fifty-seven per cent (17/30) of children given a diagnosis of autism at T1 received a diagnosis of ASD at T2. Twenty-five per cent of children given a diagnosis of ASD at T1 (8/32) received a diagnosis of development or language delay at T2.
Post-hoc analyses identified that children who maintained their ASD diagnosis had higher ADOS social affect scores (indicating higher severity) at T1 than those who did not maintain their ASD diagnosis. Barbaro found that developing language was the most salient predictor of losing a diagnosis of ASD. Barbaro 2017 was one of the few studies to report that those involved in diagnostic assessment at T2 were blind to the diagnosis given at T1, but was found to be of high risk of bias due to attrition (22/99 were lost to follow-up), lacked detail on T1 and T2 assessment (clear definitions were not provided) and any exclusions (leading to a judgement of unclear risk of bias).

The main aim of the study by Spjut Jansson 2016 (29) was to evaluate the effectiveness of interventions in children identified as at risk of ASD through routine population screening for ASD (see section on Question 3 below). However, the authors also reported on diagnostic stability after a 2 year follow-up, but did not report the screening tool(s) used to identify children at risk of ASD. One hundred children who were screened, subsequently met diagnostic criteria for ASD at T1 (approximately 36 months old). Two years later (at approximately 60 months old), 71 children were re-assessed using some of the same tools as at T1. All children had received some form of intervention during the follow-up period. Five children no longer met ASD criteria at T2: a diagnostic stability estimate of 93% (95%CI: 84.3%, 97.7%). Although professionals conducting the diagnostic evaluation at T2 were blind to the interventions children were receiving, it is not clear whether they were also blind to diagnosis at T1. However, given the study design, this is unlikely. Other areas of study design are unclear.

Guthrie 2013 (30) included 82 children identified through community screening of 5,419 children in the FIRST WORDS project in Australia. Children who screening positive were invited for diagnostic assessment, and given a diagnosis at an average of 19 months old (T1), made by an experienced clinician. Due to lack of resources, the authors could not ensure diagnoses made at T2 (a mean of 16 months after T1) were blinded to those at T1. Therefore, all diagnostic information from T1 was available to the clinician making the diagnosis at T2. Children were categorised as either having a diagnosis of ASD, not having a diagnosis of ASD (either typically developing or having developmental delay), or they had their diagnosis deferred. Diagnoses were deferred where ASD

could not be confirmed or ruled-out due to inconsistent observations of the child being made and/or a lack, or low severity, of symptoms in the ADOS domains. At T1, 68% (56/82) of children received a diagnosis of ASD, and all retained that diagnosis at T2, providing 100% diagnostic stability (95%CI 93.6%, 100%). There were 14 children who had their diagnosis deferred at T1. At T2, 3 of these 14 children received a diagnosis of ASD, 10 had had ASD ruled out, and 1 child still had no diagnosis. None of the 12 children who had been ruled out for ASD at T1, went on to receive an ASD diagnosis at T2. The risk of bias assessment was generally unclear, however the fact that diagnostic information from T1 was available in determining diagnosis at T2, and that difficult diagnoses were allowed to be deferred, is likely to have over-estimated the stability of the ASD diagnosis.

Discussion of findings

From the 5 studies providing results on the stability of a diagnosis of ASD over time in a screened population, estimates ranged from 71.9% to 100%. In particular, the only study based in the UK (26), reported 100% stability. However, all 5 studies raised concerns regarding risk of bias. These included the lack of blinding of assessments at follow-up, participant attrition, clearly described methods of diagnosis, relatively small number of children evaluated at each time-point. One study included a mixed population of screen-detected and clinically referred cases of ASD, and for a number of studies it was unclear whether children received treatment during follow-up. Furthermore, one of the studies reporting 100% stability (30), allowed for diagnosis to be deferred, suggesting that more difficult diagnoses are not reflected in this estimate of 100% stability. In fact, at T2, 71% of those with a deferred diagnosis at T1 had been ruled out as having a diagnosis of ASD. The findings from these studies are also limited by the length of follow-up, which was at most 24 months (30).

Given that many of the concerns for risk of bias in the included studies are all likely to over-estimate diagnostic stability, we might expect the proportion of children who maintain a diagnosis of ASD to be <100%, but how much below this is unclear. Moreover, there is little evidence from this review that ASD diagnoses are maintained beyond the age of 4 or 5 years old, since the children in these studies were approximately 2 years old when they received their initial diagnosis and had a maximum follow-up of 2 years.

In the 2011 UK NSC review, none of the 4 studies involving children identified through ASD screening were blinded, and there was variability in estimates of diagnostic stability. On the basis of the studies identified in this updated 2021 review, there is still little good quality evidence to suggest that diagnoses of ASD in children ≤ 5 years old are stable.

Although a UK-based study was identified in this review (26), the lack of blinding of diagnoses, in particular, leads to concerns of risk of bias with their findings. Thus, to gain a better understanding of what proportion of children diagnosed with ASD maintain that diagnosis over time, future studies would ideally recruit children identified through screening, conduct diagnostic assessments that were blind to previous diagnoses (and blind to results from observation and diagnostic instruments), and have longer follow-up than 24 months, with clear information on any interventions received by children during that follow-up period.

Summary of Findings Relevant to Criterion 1: Not met

Five studies directly relevant to the review question were found, one based in the UK. Estimates of diagnostic stability ranged from 72% to 100%, however there are important limitations with these studies. The main limitations are a lack of blinding of T1 diagnosis (and/or measurements) when follow-up diagnostic evaluations are made, and the short follow-up periods, maximum of 2 years. Further studies in a screened population using blinding at follow-up would be required to provide more useful information on the stability of ASD diagnoses made in young children.

Criterion 4 — There should be a simple, safe, precise and validated screening test.

Criterion 5 — The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed

Question 2: What is the accuracy of screening tools in children under the age of 5 to identify ASD?
Sub-questions: Does the age at which the screening test is performed affect accuracy?
Do other characteristics affect the accuracy?
Are there incidental findings?

Although incidence has been increasing over time(15), the prevalence of ASD in young children is low. Therefore, having a screening tool with high sensitivity will be most useful so that ASD cases are identified. However, (15)given the extensive resources required in a diagnostic assessment for ASD, minimising the number of false positive cases, by having a tool with good specificity, is also important.

In the 2011 UK NSC evidence review, 10 studies assessing the accuracy of approaches to screening for ASD in children ≤5 years old were identified. These approaches included the use of trained health professionals to conduct routine surveillance, or the use of specific screening tools. This review identified the use of trained professionals, and the M-CHAT/F tool as approaches to screening that had the most potential. Sensitivity estimates for the use of health professionals of 64% were observed for assessment at age 24 months. Meaning that of all children who had a diagnosis of ASD, 64% were identified by trained health professionals. At 42 months old, the sensitivity of using trained health professionals was estimated to be 95% (i.e. 95% of children who went on to have a diagnosis of ASD had been identified by the health professionals). A PPV of 81% was reported in a second study, meaning that of all children identified as positive by the health professionals, 81% subsequently received a diagnosis of ASD. Of the screening tools which had been evaluated, the M-CHAT/F was identified as the most promising. For children > 24 months old, PPVs around 60% were reported, thus indicating that 60% of children scoring positive on the M-CHAT/F were found to have a diagnosis of ASD. It was noted that no estimates of sensitivity for the M-CHAT/F had yet been reported.

Eligibility for inclusion in the review

Only studies (or systematic reviews) involving children aged ≤ 5 years who had not been diagnosed with ASD, nor had any concerns of ASD raised by parents/carers or health

professionals were included. Furthermore, only studies where diagnostic assessment (the reference standard) was undertaken as soon as possible after the screening tool had been administered were included. Studies were not excluded on the basis of the screening tool used, with any tool used to screen for ASD being included. Non-English language studies and those published before 2010 were excluded. For systematic reviews to be included they needed to have reported full details of their search strategy, have quality appraised included studies, and included studies published after 2010.

After full paper review the main reasons for exclusion of studies included children already having a diagnosis of ASD, or that diagnoses were determined at some future time-point from medical records. For systematic reviews, the main reason studies were excluded was that they did not report quality appraisal of included studies, for example (64-66).

Description of the evidence

0contains a full PRISMA flow diagram (Figure 1), along with a table of the included publications and details of which questions these publications were identified as being relevant to (Table 15).

Of the total 5,498 titles and abstracts from the database searches conducted in November 2020, the full-text of 118 titles were reviewed for eligibility for this question. On closer inspection, 102 articles were not found to be eligible and 16 primary studies were found to be relevant to Question 2..

Six systematic reviews (5 retrieved by the search, 1, which informed the 2016 USPSTF recommendation (24),identified from the commissioning brief) all had inclusion criteria much broader than the inclusion criteria for this 2021 UK NSC update review, included studies published before 2010, and either studies where children already had a diagnosis of ASD, or where diagnosis of ASD was made at a later date, not as soon as possible after screening (for example in McPheeters (24)). Since the conclusions from these systematic reviews may not be applicable to the current review, they were ultimately excluded from the 2021 update. However, the lists of included studies in the systematic reviews were cross-checked with ours to help identify additional relevant studies published after 2010. These 6 systematic reviews are not discussed further. A summary of the systematic reviews can be found in Table 19, with quality appraisal results in Table 24.

A further 2 primary studies (16, 44) were identified from reviewing the list of included studies in these SRs. In the July 2021 update searches, of 591 titles and abstracts found, 4 were reviewed at full-text and 3 were considered eligible. Since 2 included articles report on

the same primary study by Jonsdottir (29, 31), 20 primary studies (reported in 21 articles) were relevant for this question.

Across the 20 primary studies, the performance of 11 screening tools has been evaluated. A summary of these tools is given in Table 4.

Table 4. Summary of screening tools evaluated in the included studies

Screening tool in our review	Target condition(s)	Main areas covered	Intended age	Format	Time required	Source
M-CHAT (-R/F)*	ASD	early joint-attention/theory of mind, early language and communication, motor abnormalities, sensory abnormalities and social interchange	16-30 months	2-stage: 1st parent/carer completed questionnaire 2nd parent/carer interview with health professional	5-20 minutes	Robins 2014(16), Magan-Maganto 2017(67) Thabtah 2019(68)
Quantitative Checklist for Autism in Toddlers (Q-CHAT)	ASD	Items from CHAT and additional items	18-24 months	Parent/carer-completed questionnaire (25 items)	15-20 minutes	Allison 2021(26) Thabtah 2019(68)
Global Developmental screen (GDS)	Global development	gross and fine motor skills, language and communication, and contains an emotional-social domain	3-60 months	Parent/carer interview with health professional, observation	NR	Kerub 2020(34)
Social Attention Communication Surveillance-Revised (SACS-R)	ASD	social attention and communication	12-60 months	Observation	NR	Mozolic-Staunton 2020(37)
Parents Evaluation of Developmental Status (PEDS)	Global development, ASD pathway developed	behaviour, motor skills, expressive/receptive language development, social-emotional development, and concerns around school for those children attending school	0-8 years	Parent/carer interview with health professional	5-10 minutes	Mozolic-Staunton 2020(37) Thabtah 2019(68)

Social Communication Questionnaire (SCQ)	ASD	social interaction and communication, and repetitive and stereotyped behaviours	48 months	Parent/carer questionnaire	10-20 minutes	Thabtah 2019(68)
Three-item Direct Observation Screen (TIDOS)	ASD	joint attention, eye contact and responsiveness to name	Unclear	Observation	"no additional time" to routine check	Topcu 2018(41)
Ages and Stages Questionnaire (ASQ-3)	Global development	communication, gross motor, fine motor, problem-solving and personal-social development	0-60 months	Parent/carer questionnaire	NR	Catino 2017(42)
First Year Inventory (FYI)	ASD	social–communication and sensory–regulatory domains	12 months	Parent/carer questionnaire	20-35 minutes	Ben-Sasson 2013(46) Thabtah 2019(68)
Infant Toddler Checklist (ITC)	Social and communication delays	Language predictors	6-24 months	Parent/carer questionnaire or interview format	5-10 minutes	Wieckowski 2021(32) Thabtah 2019(68)
Joint Attention Observation schedule (JA-OBS)	ASD	Joint attention	20-48 months	Observation	5-10 minutes	Nygren 2012(17) Magan-Maganto 2017(67)

*M-CHAT/F, original M-CHAT with follow-up interview; M-CHAT-R/F, revised M-CHAT with follow-up interview. The M-CHAT/F was first published in 2001, and a revised version, M-CHAT-R/F published in 2014(16).

The majority of studies reported an evaluation of versions of M-CHAT, including the revised version (M-CHAT-R) and/or with the follow-up interview (M-CHAT(-R)/F). Eleven studies evaluate non-English translations of M-CHAT(-R/F). Four studies compare M-CHAT/F with another screening tool: the Three-item Direct Observation Screen (TIDOS) (41), the Global Developmental screen (GDS) (34), Joint Attention Observation schedule (JA-OBS) (17), and Parents Evaluation of Developmental Status (PEDS) (45). The study by Wieckowski (32) evaluated the use of M-CHAT-R/F, FYI and/or ITC at different ages, but their results do not allow for evaluation of the individual tools. In the 5 studies not including M-CHAT(-R/F), 6 different tools are evaluated: Q-CHAT (26), Social Attention Communication Surveillance-Revised (SACS-R) (37), PEDS (37), Social Communication Questionnaire (SCQ) (40), Ages and Stages Questionnaire (ASQ-3) (42), First Year Inventory (FYI) (46). SACS-R, TIDOS, JA-OBS and GDS (partly or fully) involve observation of the child, with M-CHAT(R/F), PEDS, SCQ, ASQ-3, FYI based on parent/carer reports regarding the child.

Detailed characteristics for all studies are given in the appendix (Table 20), alongside risk of bias assessments (Table 25). Table 5 and Table 6 below summarise the characteristics,

results and risk of bias for studies evaluating versions of M-CHAT (Table 5) and other tools (Table 6).

Table 5 Summary of studies evaluating the screening accuracy of Q-CHAT and versions of M-CHAT

Study, Country	Intended age (months) at screening [Mean age at screening, SD]	Screening tool [language] Cut-off for referral	Uptake (%, N) [ASD prevalence]	Reference standard [Diagnostic criteria. Tools/Measures. Personnel]	Follow-up of screen negatives	Risk of bias and applicability concerns*	Results (95% CI)
Allison 2021(26) **UK**	18-30 months	Q-CHAT [English] Decreasing probability of referral with 100% referred who ≥37 (observed) & ≥ 44 (imputed), to 1% referred with scores ≤ 37	Screening: 28.8%, 3770 Uptake: 54.3%, 121 [0.98%(0.45 %, 2.16%)]	ICD-10. Consensus diagnosis as possible autism or autism spectrum (if they met the ICD-10 criteria). ADOS-G, ADI-R, MSEL, VABS. Experienced research psychologist(s) and trained research assistant.	Children re-screened ≥ 48 months using CAST. Those >15, and any where referrals for number of reasons, including autism, invited for diagnostic evaluation.	PS: Low/Low IT: Low/Low FS: Low/Low FT: High	≥39: PPV 0.17 (0.08, 0.31) Sens 1.0 (0.72, 1.0) Spec 0.95 (0.92, 0.97) NPV 1.0 (0.93, 1.0)
Jonsdottir 2020, 2021(31, 33) **Iceland**	30 [31.7, 1.72]	M-CHAT-R/F [Icelandic] >2 => FUI ≥2 => refer	Screening: 72.1%, 1586 Diagnostic evaluation: 96.2%, 25 [1.22% (0.84, 1.75)]	ICD-10. Physical and neurological examination, ADOS-2, Parent interview. Paediatrician, psychologist, social worker.	Yes. Checked databases for any ASD diagnoses up to 2 years after screening	PS: Low/Low IT: Low/High RS: High/Low FT: High	Sens: 0.62 (0.44, 0.80) Spec: 0.99 (0.99, 1.00) PPV: 0.72 (0.51, 0.88) NPV: 0.99 (0.99, 1.00)
Magan-Maganto 2020(35) **Spain**	18 and 24 [approx. 24, range 14-36]	M-CHAT-R/F [Spanish] >7 => refer 3-7 => FUI. ≥2 after FUI => refer	Screening: 56.6%, 6515 (Complete FUI 78.3%) Diagnostic evaluation: 61.3%, 19	DSM-V. Clinical history, Merril-Palmer Revised Scales, Leiter, Vineland Scales, ADOS-G module 1 and ADOS-2 module T and 1. Trained and experienced professionals	Yes. Reviewed any ASD diagnoses in children who screened negative.	PS: Low/Low IT: Low/High RS: Unclear/Low FT: High	All ages Sens 0.79 (0.54,0.93) Spec 0.99 (0.99,0.99) 14-22 months Sens 0.82 (0.48–0.97) Spec 0.99 (0.99,0.99) 23-36 months Sens 0.75 (0.36–0.96) Spec 0.99 (0.99,0.99)

			[0.29%]				
Oner 2020(36) **Turkey**	16-36 [26.75, 5.76]	M-CHAT-R/F [Turkish] >2 =>FUI. >2 FUI => refer	Screening: 74.5%, 6712 (but denominator included those out of age range) (complete FUI 84.3%) Diagnostic evaluation: 68.8%,152 [0.8%]	DSM-V. "All available information" ADOS-2, Denver Developmental Screening-II. Study author, research certified for ADOS-2 use.	No	PS: High/Unclear IT: Low/High RS: Unclear/Unclear FT: High	M-CHAT-R Sens 1.00 (0.94, 1.0)** Spec 0.91 (0.90, 0.92)** PPV 0.09 (0.07, 0.11)** NPV 1 (0.999, 1.0)** M-CHAT-R/F (calculated by 2021 review authors) Sens 1.00 (0.97, 1.00)** PPV 0.26 (0.20,0.32)**
Kerub 2020(34) **Israel**	18-36 [21.3, 3.45]	Global Developmental Screening (GDS), M-CHAT/F [Hebrew] GDS ≥ 1 => follow-up or refer. M-CHAT/F >7 => refer 3-7 => FUI. ≥2 after FUI => refer	Screening: NR, 1591 (complete FUI NR) Diagnostic evaluation: 82.3%, 70 [0.63%]	DSM-V. Child psychiatrist/neurologist.	Yes. Reviewed medical records of those screened negative (10 months later) to identify any false negatives	PS : Unclear/Unclear IT: Unclear/High RS: Unclear/Low FT: High For comparative accuracy PS: Unclear IT: Unclear RS: Unclear FT: Unclear	M-CHAT/F Sens 0.7 (0.35, 0.93) Spec 0.98 (0.97, 0.99) PPV 0.20 (0.08, 0.37) GDS Sens 0.5 (0.19, 0.81) Spec 0.998 (0.992, 0.999) M-CHAT/F plus GDS Sens 0.7 (0.35, 0.93) Spec 0.968 (0.96, 0.97)
Dai 2020(38) **US**	24 [unclear]	M-CHAT/F or M-CHAT-R/F [English, results for Spanish version not included] NR	Screening: NR, 19685 (complete FUI at 18 months 77.5%, 24 months 70.2%) Diagnostic evaluation:	DSM-IV. Demographic information, Mullen Scales of Early Learning, Vineland Adaptive Behavior Scales, ADOS-2 Toddler Module, ADOS Module 1 and 2, CARS(2). Clinical psychologist or a developmental-behavioral pediatrician	No. Re-screened those who were screen-negative at 18 months. No screen-negatives had a diagnostic	PS: Low/Low IT: Low/Low RS: High/Low FT:Low Authors co-owners of MCHAT LLC	NR Calculated by 2021 review authors Single screen at 18 months PPV 0.52 (0.47, 0.57) Negatives rescreened at 24 months PPV 0.50 (0.27, 0.73)

			70.0%,390 (18 months) 62.5%, 20 (24 months) [1.03%]		evaluation, unless they subsequently screened positive.		
Achenie 2019(39) [based on data from Robins 2014(16)] **US**	16-30 [NR]	Machine learning applied to M-CHAT-R [English] >2 =>FUI. >2 FUI => refer.	14995 (Uptake not reported, see Robins 2014) [0.77%]	DSM-IV-TR. "ADOS, CARS-2, Toddler Autism Symptom Interview, Mullen Scales of Early Learning, Vineland Adaptive Behavior Scales–II, Behavioral Assessment System for Children–2, and developmental history" Psychologist/developmental pediatrician	Random sample of screen-negatives had diagnostic evaluation [see Robins]	PS: Unclear/Unclear IT: Low/Low RS: Unclear/Low FT: High	Comparable to M-CHAT-R/F. More results available.
Topcu 2018(41) **Turkey**	16–38 [NR]	TIDOS, M-CHAT/F [Turkish] M-CHAT/F: ≥2 of 7 critical items or ≥3 of 23 items were positive, so => refer TIDOS: refer if one of the three parameters scored ≥ 1	Screening: 40.0%, 511 Diagnostic evaluation: 91.3%, 21 [0.98%]	DSM-V NR. Child psychiatrist.	Yes. Random sample of 25 children who screened negative on M-CHAT-R/F and TIDOS. Diagnostic evaluation within 2 weeks of screen for screen-positive children and 3–9 months for screen-negative children.	PS: Low/Unclear IT: Low/ High RS: High/Low FT: High For comparative accuracy PS: Low IT: Unclear RS: High FT: Unclear	M-CHAT/F Sens 0.60 (0.15, 0.95)** Spec 0.97 (0.95, 0.99)** PPV 0.18 (0.04, 0.46)** NPV 0.995 (0.98, 1.0)** TIDOS Sens 0.80 (0.28, 0.99)** Spec 0.998 (0.989, 0.999)** PPV 0.80 (0.28, 0.99)** NPV 0.998 (0.989, 0.999)** M-CHAT/F plus TIDOS Sens 1.00 (0.48, 1.00)** Spec 0.90 (0.88, 0.93)** PPV 0.10 (0.03, 0.21)** NPV 1.00 (0.99, 1.0)**

Baduel 2017(43) **France**	24 [24]	M-CHAT/F [French] any 3 M-CHAT items or 2 of the 6 critical items => FUI. If still indicates ASD after FUI => refer [refs Robins 2001]	Screening: NR, 1227 (complete FUI 78.7%) Diagnostic evaluation: 100%, 20 [1.47%]	NR. 2-stage process 1st: ADOS-G, Psycho Educational Profile Revised, Vineland Adaptive Behavior Scales, but trained in use of ADOS. If reached ADOS-G threshold, referred to independent team to confirm diagnosis.	Those screen-negative at 24 months followed-up at 30 and 36 months. If then screen positive, they were referred for diagnostic assessment. As were any children who screened negative, but physicians had concerns.	PS: Unclear/Unclear IT: Low/High RS: High/Low FT: High	Sens 0.67 (0.41, 0.86) Spec 0.99 (0.98, 0.99) PPV 0.6 (0.36, 0.81)** NPV 0.99 (0.99, 0.99)**
Kondolot 2016(44) **Turkey**	18-30 [23, 3]	M-CHAT [Turkish] Refer if any 2 of 6 critical items or any 3 of 23 items were positive	Screening: Approx. 50.5%, 2021 Diagnostic evaluation: 100%, 17 [0.1%]	DSM-IV-TR. CARS Child psychiatrist.	Yes. Random sample (n=48) screened negative evaluated (6-12 months after screening)	PS: Low/Low IT: Low/High RS: High/Low FT: Low	PPV: 0.12 (0.01, 0.36) Sens: 1.00 (0.16, 1.00) Spec: 0.76 (0.64, 0.86)
Wiggins 2014(45) **US**	18 and 24 [21.1, range 15.2 – 27.0]	M-CHAT/F PEDS [English] M-CHAT/F: any 3 of 23 items were failed or any 2 of 6 critical	Screening: NR, 3980 Diagnostic evaluation: NR, 44 [0.75%]	NR. ADI-R, ADOS, CARS, MSEL, Vineland-II, developmental and medical history questionnaire. Experienced clinicians (blind to M-CHAT/F and PEDS score)	No. (Only screen negative children for whom clinicians had raised concern	PS: Unclear/Unclear IT: Low/Low RS: Unclear/Low FT: Low	M-CHAT/F PPV 0.61 (0.45, 0.76) PEDS Path A PPV 0.55 (0.39, 0.71) PEDS Path B PPV 0.75 (0.35, 0.97) PEDS ASD

		items were failed PEDS Path A ≥2 predictive concerns PEDS Path B 1 predictive concern noted PEDS ASD ≥3 concerns. Only children who failed M-CHAT/F were referred.			were followed)		PPV 0.59 (0.39, 0.76) [Very few of the PEDS positive were followed-up]
Robins 2014(16) **US**	18 and 24 [20.94, 3.30]	M-CHAT-R/F [English] ≥3 items on M-CHAT-R, and either ≥3 on M-CHAT-R/F, or ≥2 on M-CHAT-R/F.	Screening: NR, 16071 (complete FUI 81.9%) Diagnostic evaluation: 63.5%, 221 [0.77%]	DSM-IV-TR. "all available information and … clinical judgment". "Licensed psychologist/developmental pediatrician supervising a graduate student and research assistants."	Random sample who screened negative completed Screening Tool for Austim in Two-Year Olds (STAT) tool. If then positive offered clinical evaluation	PS: Unclear/Unclear IT: Low/Low RS: Unclear/Low FT: High	M-CHAT-R/F ≥3 Sens 0.68 (0.58, 0.75) Spec 0.99 (0.99, 0.99) PPV 0.51 (0.43, 0.59)** NPV 0.997 (0.996, 0.998)** M-CHAT-R/F ≥2 Sens 0.85 (0.79, 0.92) Spec 0.99 (0.99, 0.99) PPV 0.47 (0.41, 0.54)** NPV 0.999 (0.998, 0.999)**
Chlebowski 2013(47) **US**	18 and 24 [20.4, 3.1]	M-CHAT/F [English and Spanish] screening positive on	Screening: NR, 18989 (complete FUI 74.6%) Diagnostic evaluation: 61.5%, 171	DSM-IV. ADOS, ADI-R, Mullen Scales of Early Learning, Vineland Adaptive Behavior Scales, CARS.	Only those who screen positive on other tools or "red-flagged" by paediatrician	PS: Unclear/Low IT: Low/Low RS: Unclear/Low FT: High	PPV 0.54 (0.46, 0.61)

		2 of 6 critical items or on 3 of 23 items overall on both the M-CHAT and M-CHAT/F.	[0.5%]	Diagnosis made by clinical judgement "licensed clinical psychologist or developmental pediatrician and a psychology doctoral student."			
Nygren 2012(17) **Sweden**	30 [NR]	M-CHAT/F JA-OBS [Swedish] M-CHAT: ''failure'' on any 3 of the 23 items or on any 2 of the 6 critical items failed => FUI. If still "failed" => refer JA-OBS: failed ≥2 items	Screening: 80%, 3999 Diagnostic evaluation: 84.3%, 54 [1.2%]	DSM-IV and ICD-10 Vineland Adaptive Behavior Scales, Autism Diagnostic Observation Schedule, Diagnostic Interview for Social and COmmunication Disorders, Language assessments, 1-h observation of the child at preschool. "experienced neuropsychiatrists, neuropediatricians (4 in total) and neuropsychologists (2 in total) with expertise in autism."	Only those where a concern raised	PS: Low/Low IT: Low/High RS: Unclear/Low FT: High For comparative accuracy PS: Low IT: Unclear RS: Unclear FT: Unclear	M-CHAT/F alone Sens 0.77 (0.61, 0.88) PPV 0.92 (0.78, 0.98) JA-OBS alone Sens 0.96 (0.72, 0.95) PPV 0.92 (0.80, 0.98) M-CHAT/F plus JA-OBS Sens 0.96 (0.85, 0.99) PPV 0.90 (0.77, 0.96)
Canal-Bedia 2011(48) **Spain**	18 and 24 [range 18-36]	M-CHAT/F [Spanish] 3 out of 23 or 2 out of the 6 critical items => FUI. If still "failed" => dx eval	Screening: NR, 2055 Diagnostic evaluation: 9.2%, 31 [0.29%]	DSM-IV. ADOS-G, Vineland Adaptive Behavior Scales, Merril-Palmer Revised Scales of Development	No	PS: Unclear/Unclear IT: Low/High RS: Unclear/Low FT: Unclear	PPV 0.19 (0.05, 0.33)

ADOS-G, Autism Diagnosis Observation Schedule-Generic; CARS, Childhood autism rating scale; FUI, follow-up interview for M-CHAT(-R); PS, participant selection; IT, index test (screening tool); RS, reference standard (diagnostic evaluation); FT, flow and timing; PPV, positive predictive value; Sens, sensitivity; Spec, specificity; NR, not reported.
*Note that studies who only report PPV and who do not follow-up any children who have a negative screen are deemed to be of low risk of bias. However, if such a study reports sensitivity and specificity, then it is deemed to be of high risk of bias.
** Estimates calculated by review authors

Table 6 Summary of studies evaluating the screening accuracy of tools other than M-CHAT(-R/F)

Study, Country	Intended age (months) at screening [mean, SD]	Screening tool [language] Cut-off for referral	Uptake: %, N [ASD prevalence]	Reference standard [Dx criteria Tools/Measures Personnel]	FU screen negatives	Risk of bias and applicability concerns	Results (95% CI)
Mozolic-Staunton 2020(37) **Australia**	12, 18, 24, 36-60 [range 12 - 48]	SACS-R, PEDS [English] SACS: 3 key items of concern = high risk PEDS: PATH ASD = 3 or more concerns, Path A = 2 concerns, Path B = 1 concern	Screening: NR, 13417 Diagnostic evaluation: 83.3%, 205 [1.49%]	Bayley Scales of Infant Development (BSID), Autism Diagnostic Observation Schedule, 2nd Edition (ADOS 2, Autism Diagnostic Interview-Revised (ADI-R), clinical judgement. Paediatric health professionals	No. Negative on SACS-R and PEDS not FU	PS: Unclear/Unclear IT: Low/Low RS: High/Low FT: High For comparative accuracy PS: Unclear IT: Unclear RS: Unclear FT: Unclear	SACS-R PPV 0.83 (0.78, 0.88)* Sens 0.82 (0.76, 0.87)* Spec 0.99 (0.99, 1.00)* NPV 0.99 (0.99, 1.00)* PEDS PPV 0.88 (0.71, 0.98)* Sens 0.07 (0.03, 0.14)* Spec 0.99 (0.99, 1.00)* NPV 0.99 (0.99, 1.00)*
Suren 2019(40) **Norway**	36 [36]	SCQ [Norwegian] ≥15 for the 39 scored items. ≥11 for the 39 scored items. ≥12 for the 33 non-verbal items.	Screening: 58%, 58520 Diagnostic evaluation: NR [0.7%]	DSM-IV-TR. ADOS, ADI-R. NR.	Random sample of age-matched controls. False negative children (those with ASD who were not screen positive) were determined by checking medical records at later time-point.	PS: Low/Low IT: Low/High RS: Unclear/High FT: High	SCQ total ≥15 Sens 0.20 (0.16,0.24) Spec0.99 (0.99,0.99) PPV 0.09 (0.07, 0.11) NPV 0.99 (0.99, 1) SCQ total ≥11 Sens 0.42 (0.37,0.47) Spec 0.89 (0.89, 0.90) PPV 0.03 (0.02, 0.03) NPV 1 (1,1) SCQ total ≥12 Sens 0.25 (0.20,0.29) Spec 0.99 (0.99, 0.99) PPV 0.16 (0.13, 0.19) NPV 1 (0.99,1) Results also given by whether child had phrased speech or no.
Catino 2017(42)	42 and 48	ASQ-3 [Italian]	Screening: 88.7%, 514	"neuropsychiatric evaluation	No	PS: Low/Low	For ASD PPV 0.08 (0.01, 0.25)

Italy	[younger group 42.65, 1.82; Older group 48.08, 2.62]	scored in the clinical range in one, or more than one domain	Diagnostic evaluation: 57.5%, 40 [0.39%]	comprehensive neuropsychiatric evaluation (cognitive, neuropsychological, and psychopathological)"		IT: Unclear/High RS: Low/Low FT: Unclear	
Ben-Sasson 2013(46) **Israel**	12 [12.56]	FYI [Hebrew] 94th percentile cut-off for the social domain only, or also the 88th percentile cut-off for the sensory domain.	Screening: NR, 613 Diagnostic evaluation: NR [0.8%]	None. AOSI, MSEL. "clinician with expertise in early child Development"	Yes. 60 screen-negatives followed-up.	PS: High/Low IT: Low/High RS: Unclear/Low FT: High	Sens 0.60 (0.15, 0.95) Spec 0.753 (0.64, 0.84)
Wieckowski 2021(32) **US**	Initial screen: 12, 15, 18 Re-screens: 18, 24, 36	FYI (12 months), ITC (12 & 15 months), M-CHAT-R/F (≥ 15 months) [English and Spanish] Positive on either tool (if multiple tools used).Cut-offs NR.	Screening: 12 months NR, 1504 15 months NR, 1228 18 months NR, 3053 Diagnostic evaluation from initial screen at: 12 months 36.0%, 91 15 months 29.0%, 78 18 months 20.2%, 131 [2.35%]	ICD-10. ADOS-2, TASI or ADI-R, medical, developmental, family history. Individuals supervised by supervised by a licensed psychologist, certified school psychologist, or developmental paediatrician.	Only those for whom a concern had been raised.	PS: Low/Low IT: Unclear/High RS: Unclear/Low FT: High	**Single screen at** 12 months PPV 0.22 (0.14, 0.32)* Sens 0.64 (0.48, 0.81) Spec 0.95 (0.93, 0.96) NPV 0.991 15 months PPV 0.17 (0.09, 0.27)* Sens 0.72 (0.52, 0.93) Spec 0.94 (0.92, 0.95) NPV 0.995 18 months PPV 0.42 (0.34, 0.51)* Sens 0.74 (0.64, 0.84) Spec 0.97 (0.97, 0.98) NPV 0.993 >1 screen from 12 months PPV 0.25 (0.17, 0.35)* Sens 0.81 (0.67, 0.95) Spec 0.94 (0.93, 0.95) NPV 0.995 15 months PPV 0.19 (0.11, 0.29)* Sens 0.83 (0.66, 1.00) Spec 0.94 (0.92, 0.98) NPV 0.993 18 months

	PPV 0.44 (0.35, 0.52)* Sens 0.82 (0.74, 0.91) Spec 0.97 (0.97, 0.98) NPV 0.995

ASQ-3, Ages and Stages Questionnaire, version 3; SACS-R, Social Attention Communication Surveillance-Revised; SCQ, Social Communication Questionnaire; FYI, first Year Inventory; ITC, Infant Toddler Checklist.

*95% confidence intervals calculated by review authors

All 20 studies (reported in 21 articles) screened for ASD in a community-based population, many during existing routine surveillance appointments. One study was based in England, 5 studies were based in the USA, 3 in Turkey, 2 each in Israel and Spain, and one study in each of the following: Chile, Iceland, Japan, Australia, Italy, France and Sweden. Half of the articles were published since 2019 (2021,n=1; 2020, n=6; 2019,n=3).

Eighteen studies screen children between the ages of 12 and 36 months. Mozolic-Staunton (37) report screening children at the ages of 12, 18, 24, and 36-60 months. While Catino 2017 (42) screen children at 42 and 48 months old. Achenie (39), Dai (38) and Mozolic-Staunton (37) report re-analyses of previous cohort studies, while the remaining studies are all prospective cohort studies. The study by Achenie (39) is a retrospective analysis of prospectively collected data from Robins 2014 (16).

Included studies were generally found to be of low risk of bias on QUADAS-2 for the patient selection, index test and reference standard domains. As a number of studies were carried out with the aim of validating the M-CHAT(R/F) screening tool in populations where English is not the first language; translations of the M-CHAT(R/F) were done. Thus, there are some concerns for these studies about applicability of the index test to the UK setting. The domain where many studies were deemed to be at a high risk of bias was for timing and patient flow. In particular, children who were deemed to be negative on the screening tool were either not followed-up at all (32, 36-38, 42, 48), followed-up only if they were positive on another screening tool(s) and/or a health professional raised a concern for possible ASD diagnosis (16, 17, 26, 39, 43, 47), followed-up at a later date (6-12 months later (44), 10 months later (34), 24 months later (31, 33), unknown time-point (35, 40)), or a sample of children were followed-up (41, 46).

In many studies only those children who screened positive received a diagnostic evaluation. This is no doubt due to diagnostic evaluation being complex, and thus time and resource intensive. Therefore, in these studies the calculation of sensitivity and specificity is not possible as the total number of children in the sample with and without a diagnosis of ASD cannot be obtained. In such studies, often the PPV is the only summary estimate that can be calculated and reported. In six studies(32, 36-38, 42, 48), all, or a large proportion of children, who were negative according to the screening tool and did not receive a diagnostic assessment, were assumed to be true negatives. Clearly, this assumption leads to an overestimation of the number of true negatives and will inflate estimates of sensitivity, specificity and NPV. In other studies e.g. Topcu(36, 41), a random sample of children who screened negative are followed-up. However, it may be unclear on what basis the sample size has been chosen, or whether the analysis has been appropriately weighted(69).

In the following section, studies are grouped by the screening tool evaluated, starting with the Q-CHAT, which is the only screening tool found that was evaluated in the UK. Note that Wieckowski (32) use different screening tools at different ages, and so apart from where the initial screen is at 18 months and so only the M-CHAT-R/F is used, results specific to each screening tool are not available from the study. This study is summarised towards the end of this section.

Q-CHAT

Only one UK study(26) sought to evaluate the screening performance of Q-CHAT in children aged 18 and 30 months old and identify an optimal threshold for the Q-CHAT. Parents/carers of children who were registered on the Child Health Surveillance Database from Luton, Bedfordshire and Cambridgeshire were sent the Q-CHAT to complete and return by post. To avoid missing any children who may have autism, but may have had missing data, maximum scores were imputed for any questions with missing data for all children. Thus, all children had 2 possible scores from the Q-CHAT: their observed score (where missing items were not counted), and their imputed score (where missing items were scored as 4). The probability that a child was invited for diagnostic evaluation depended on both scores, with children having a total observed score ≥37 and total imputed score ≥ 44, all being invited for diagnostic assessment. As a stratified sampling approach was taken in this study, children with lower total observed and imputed scores had a lower probability of being invited for a diagnostic evaluation, with only 1% of children with total scores ≤ 37 being invited. Diagnosis was made by consensus of all the diagnostic information and judgement of those involved, based on ICD-10. Individuals conducting the diagnostic assessment were reported to be blind to Q-CHAT scores. Children were either diagnosed with possible autism/ASD, atypically (developmental concerns not linked to autism) or typically developing. Allison 2021 reported a PPV of 0.17 (95%CI: 0.8, 0.31) and a sensitivity of 1.0 (0.72, 1.0) for Q-CHAT using a threshold of ≥ 39. The study was deemed to be of uncertain or low risk of bias for a number of domains. Since not all children who were screened were invited for diagnostic evaluation, analyses were weighted to reflect the stratified sampling design.

M-CHAT(-R/F)

Four studies assessed the revised version with follow-up interview, M-CHAT-R/F (16, 33, 35, 36), one study (38) assessed M-CHAT/F or M-CHAT-R/F, 7 studies assessed M-CHAT/F (17, 34, 41, 43, 45, 47, 48), and one study assessed the M-CHAT without the follow-up interview (44). The study by Achenie (39) only assessed the M-CHAT-R data from

Robins (16), they did not use information from the follow-up interview (FUI). Of the 8 studies that did not assess the revised version, 5 were conducted before M-CHAT-R/F was published (17, 44, 45, 47, 48) and one specifically assessed M-CHAT as there was no Hebrew translation of M-CHAT-R (34).

M-CHAT-R(/F)

Although 5 studies reported assessing the accuracy of M-CHAT-R/F, estimates of sensitivity and specificity for the M-CHAT-R/F are only available from 4 studies: the Icelandic study by Jonsdottir (31, 33) the Spanish study by Magan-Maganto (35), the US study by Robins (16) and the Turkish study by Oner (36).

In Jonsdottir (31, 33) 2,201 children were screened at routine 30-month old check-ups in Reykjavik using an Icelandic translation of the M-CHAT-R/F. Children screened as positive were referred for diagnostic assessment, which was based on the ICD-10 classification system, and other information. Individuals conducting the diagnostic evaluations were not blind to the child's screening results. A number of diagnostic databases were examined to identify any children who screened negative but went on to receive a diagnosis of ASD within 2 years of screening. Jonsdottir (31, 33) reported a study prevalence of ASD to be 1.22%, PPV of 0.72 (0.51, 0.88) and sensitivity of 0.62 (0.44, 0.80).

Due to the Icelandic translation of the M-CHAT-R/F, there are concerns with applicability to the UK setting. Since screening results were known at the time of diagnosis it is possible that this knowledge may have impacted on the diagnostic evaluation. Furthermore, it is unclear whether and how diagnostic evaluations may have differed for children who screened positive and those who screened negative: those who screened negative had their diagnostic evaluations at an older age (by on average 10 months) than those who screened positive. It is unclear what other differences there may have been between the diagnostic evaluations for screen positive and screen negative children.

Magan-Maganto (35) screened 6,515 children during "Well Baby Check-Up Program" screening at ages 18 and 24 months. All those who screened positive on M-CHAT-R/F were referred for diagnostic evaluation based on DSM-V. For children who screened negative, any subsequent ASD diagnoses were identified from referral centres, with all other children who screened negative assumed to be true negatives. Magan-Maganto report results for the total sample, and separately for children who were screened aged 14-22 months old, and children screened aged 24-36 months old. Specificity is reported as 0.99 (95% CI: 0.99, 0.99) regardless of whether the total sample or subgroups are evaluated. Sensitivity is 0.79 (95% CI: 0.54, 0.93) for the total cohort, 0.82 (95% CI: 0.48–0.97) for the younger subgroup and 0.75 (95% CI: 0.36–0.96) for the older subgroup. Due to a Spanish translation of the M-CHAT-R/F being used, there is concern that the screening

tool is not applicable to the UK. As ASD diagnoses for screen negative children were made via the referral centres at an unknown time-point, the risk of bias regarding the flow and timing of participants is unclear.

Robins (16) screened 16,071 children during 18 and 24 month Well-Child Care routine visits. A number of cut-offs for M-CHAT-R/F were evaluated. To identify likely false negatives, a stratified random sample of children who were negative on the M-CHAT-R/F were screened with another tool (Screening Tool for Autism in Two-Year Olds (STAT)). Screen negative children who were not selected, were assumed to be true negatives. Robins reported sensitivity of 0.85 (95% CI: 0.79, 0.92) for the cut-off of M-CHAT-R/F ≥2, specificity of 0.99 (95%CI: 0.99, 0.99) regardless of the cut-offs used. There are concerns of high risk of bias, especially with estimates of specificity, since not all children received the reference standard, with most screen negative children assumed to be true negatives. The approach to finding any children falsely deemed negative on the M-CHAT-R/F relied on a second screening tool, for which the accuracy is unknown. It is unclear whether the screening result was known to those carrying out the diagnostic assessment.

Oner 2020 (36) assessed a Turkish translation of M-CHAT-R/F to screen 6,715 children aged 16 and 36 months old for ASD. Children who were negative according to the questionnaire stage of M-CHAT-R were assumed to be true negatives. However, children who were positive according to the questionnaire stage of M-CHAT-R, but after receiving the follow-up interview (M-CHAT-R/F) and subsequently found to screen negative, were followed-up. Diagnoses were based on DSM-V criteria and other information. Oner reported estimates of ASD prevalence, sensitivity and PPV for M-CHAT-R as 0.8%, 1, and 0.09, respectively. Since children who screened negative on the M-CHAT-R were not followed-up, these estimates are likely to over-estimate performance. The sensitivity and PPV of M-CHAT-R/F in the subgroup of individuals who screened positive with M-CHAT/R , are 1.00 (95% CI: 0.97, 1) (since all 57 children who were positive for M-CHAT-R/F had a diagnosis of ASD confirmed), and 0.26 (95%CI: 0.20, 0.32), respectively. As with other studies, there are applicability concerns due to a translated version being used.

Dai 2020 (38) screened 19,685 children at their 18 month routine check-up in the USA with the M-CHAT/F or M-CHAT-R/F. The authors also reported on the value of rescreening at 24 months those children who screened negative at 18 months. There was no follow-up of children who screened negative. Where evaluations were available, the PPV at 18 months can be calculated as 0.52 (95% CI: 0.47, 0.57), with a prevalence of ASD of 1.03%. In 32 children who were positive after being rescreened at 24 months after a negative screen at 18 months, the PPV was 0.5 (95% CI: 0.27, 0.73).

M-CHAT(/F)

Nine studies reported looking at the accuracy of M-CHAT(/F) (17, 34, 39, 41, 43-45, 47, 48). Five reported estimates of sensitivity; 4 also reported specificity estimates (17, 34, 39, 41, 43). The 2021 UK NSC review authors calculated sensitivity and specificity estimates for Kondolot (44), based on the 2x2 table reported in that study.

The study by Kerub (34) screened 1,591 children during routine developmental assessments in Israel using the M-CHAT/F and GDS. Children had a mean age of 21.3 months. Children who screened positive on M-CHAT/F or GDS were referred for diagnostic assessment (based on DSM-V). The medical records for all children who had a negative screen were assessed 10 months later. A study ASD prevalence of 0.63% is reported, alongside estimates of PPV, sensitivity and specificity for M-CHAT/F of 0.20 (95% CI: 0.08, 0.37), 0.70 (95% CI: 0.35, 0.93) and 0.98 (95% CI: 0.97, 0.99), respectively. It was unclear whether diagnostic assessment had been conducted without knowledge of the screening status of children, given that only screen positive children were referred for diagnosis. Similarly, there was some uncertainty as to whether the same diagnostic approach was taken for those children who screened positive, and those who screened negative.

Topcu (41) evaluated TIDOS and M-CHAT/F, in a sample of 511 children aged 16-38 months during routine well-child clinics in Turkey. All children who screened positive on either tool were referred for diagnostic assessment (using DSM-V). A random sample of 25 children who screened negative on both tools were referred for diagnostic assessment 3-9 months later. Topcu assume that children who screened negative, but were not followed-up for diagnostic assessment, are all true negatives. The reported sensitivity and specificity estimates were 0.60 (95% CI: 0.15, 0.95) and 0.97 (95% CI: 0.95, 0.99), respectively, with a study prevalence for ASD of 0.98%.

Baduel (43) screened 1,227 children aged 24 months with the M-CHAT/F during well-child visits or at daycare centres in France. To identify as many false negatives as possible children who screened negative at 24 months but then screened positive on M-CHAT/F at 30 or 36 months old, or were identified by health professionals as being of concern, were referred for diagnostic assessment. All other children who screened negative were assumed to be true negatives. Baduel reported sensitivity and specificity of 0.67 (95% CI: 0.41, 0.86) and 0.99 (95% CI: 0.98, 0.99), respectively. Given only screen positive children, or those with concerns, were referred for diagnosis, it is possible that the diagnostic assessment had been conducted with knowledge of the screening status of children.

Nygren (17) screened 3,999 children at 30 months old with the M-CHAT/F and JA-OBS during routine developmental monitoring in Sweden. Children who screened positive on either tool were referred for diagnostic assessment based on DSM-IV/ICD-10. Any children identified by health professionals as being of concern, regardless of their screening tool

score, were also referred for diagnostic assessment. Children who screened negative with M-CHAT/F and did not have any concerns raised were not offered a diagnostic evaluation. Given this study design, individuals conducting the diagnostic evaluation likely knew that children were at high risk of ASD, this information may have impacted on the diagnostic assessment. Based on diagnostic evaluation of those children with a positive M-CHAT/F or JA-OBS screen, and those who had a negative screen, but had concerns raised, Nygren estimate sensitivity of 0.77 (95% CI: 0.61, 0.88), a PPV of 0.92 (95% CI: 0.78, 0.98) with ASD prevalence of 1.2%. Due to the study design, the estimated sensitivity should be interpreted with caution.

In all 4 studies above the same cut-off for determining a positive screening result for M-CHAT/F is reported: ≥2 critical items or ≥3 total items, even though an additional item is added to the Turkish version in Topcu. Across these studies estimates of sensitivity for M-CHAT/F range from 0.6 (41) to 0.77 (0.61, 0.88) (17), with specificity ≥0.97 (34, 41, 43). In addition to the concerns already raised for each study, non-English translations of M-CHAT(/F) were used in all 4 which questions applicability to the UK setting.

Kondolot (44) used the M-CHAT to screen 2,021 children with a mean age of 23 months in Turkey. The FUI was not implemented. All children who screened positive with M-CHAT and a random sample of children who screened negative were referred for diagnostic assessment, based on DSM-IV-TR. Kondolot reported a PPV of 0.12 (95% CI: 0.01, 0.36). Based on the 2x2 table reported in their paper, estimates of sensitivity and specificity of 1.00 (95% CI: 0.16, 1.00) and 0.76 (95% CI: 0.64, 0.86), respectively, were calculated by the 2021 UK NSC review authors. As the main aim of the follow-up interview is to reduce the number of false positives, it may not be surprising that there were a large number of false positives in this study. The study was deemed to be of high risk of bias in the reference standard domain, as the psychiatrists undertaking the diagnostic evaluation were not blind to the M-CHAT results, potentially leading to overestimates of the accuracy of M-CHAT.

Achenie 2019 (39) reported screening data from 14,995 children screened using M-CHAT-R to evaluate whether machine learning can improve screening accuracy. The data are a sub-set from the study by Robins 2014 (16) (see below). The Machine Learning (ML) model used inputs from the M-CHAT-R (excluding any FUI answers) and ran multiple models to identify the best performing model. As in Robins (16), children who screened positive on MCHAT-R/F and a random sample of children who screened negative, were referred for diagnostic examination (based on DSM-IV-TR). All other children who screened negative were assumed to be true negatives. Achenie (39) reported results for the total group (sensitivity ranging from 0.54 to 0.74, PPV ranging from 0.79 to 0.91), and also within subgroups based on ethnicity, gender and level of maternal education (with the most

optimistic sensitivity estimate of 0.83 and PPV of 1 obtained within the female subgroup). Note that as only a small proportion of children with a negative screen result were referred for diagnostic evaluation, within subgroups there are instances where no false negative cases are observed, leading to a PPV of 1. The performance of the ML model is improved in the subgroups compared to the total group of children. The authors reported that the model was comparable on performance to that where M-CHAT-R/F is used, suggesting it could be a more efficient approach as it did not use the FUI. Due to the limited follow-up of children who screened negative, and the assumption that the majority of children screening negative were true negatives, there are concerns of high risk of bias with the results reported.

Wiggins 2014 (45) screened 3,890 children (mean age 21.1 months) during well-child visits in the US using M-CHAT/F and PEDS. Children who were positive according to the M-CHAT/F or for whom professionals had concerns were invited for follow-up diagnostic assessment. Thus, if a child was positive on the PEDS, they would only receive a diagnostic assessment if they were also positive on M-CHAT/F or where concerns had been raised. Diagnosis was made by experienced clinicians who were blind to M-CHAT/F and PEDS results, however due to the study design clinicians were aware that the child had either a positive screening result, or had raised concerns regarding ASD with a health professional. The PPV for M-CHAT/F is estimated as 0.61 (95% CI: 0.45, 0.76) by the 2021 UK NSC review authors based on the 44 children who were positive for the M-CHAT/F.

Chlebowski 2013 (47) screened 18,989 children at 18 and 24 month routine visits with M-CHAT/F in the USA. For children who screened negative on M-CHAT/F, those identified by paediatricians as having possible autism or who screened positive on the Yale screener or STAT were also offered a diagnostic evaluation. Chlebowski reported a PPV for identifying ASD of 0.54 (95% CI: 0.46, 0.61) for the M-CHAT/F.

Canal-Bedia 2011 (48) screened 2,055 children in Spain with M-CHAT/F at their mandatory vaccination appointment at 18 months old, or check-up at 24 months old. Only children positive on the M-CHAT/F were referred for diagnostic examination (based on DSM-IV). The study ASD prevalence was 0.29%, with a PPV of 0.19 (95%CI: 0.05, 0.33). Although Canal-Bedia reported estimates of sensitivity and specificity, they assumed that none of the screen-negative children have a diagnosis of ASD. As these children are not followed-up, a diagnosis of ASD cannot be ruled out and so we only focus on the PPV in this report.

Comparative accuracy of M-CHAT/F vs TIDOS, JA-OBS, GDS, PEDS

Three studies compared M-CHAT/F with other screening tools, see Table 5. Topcu (41) reported better screening performance with TIDOS than with M-CHAT/F: sensitivity of 0.60

(95%CI: 0.15, 0.95) for M-CHAT/F compared with 0.80 (95%CI: 0.28, 0.99) for TIDOS, and specificity of 0.97 (95%CI: 0.95, 0.99) for M-CHAT/F compared with 0.998 (95%CI: 0.989, 0.999) for TIDOS. When the screening tools were used in conjunction, so that children were deemed to be screen positive if positive for M-CHAT/F and/or TIDOS, sensitivity was increased to 1.00 (95%CI: 0.48, 1.00), but at the cost of specificity which reduced to 0.90 (95%CI: 0.88, 0.93).
In Nygren (17), JA-OBS was also found to have better screening performance than M-CHAT: sensitivity of 0.77 (95%CI: 0.61, 0.88) for M-CHAT/F compared with 0.86 (95%CI: 0.72, 0.95) for JA-OBS, and PPV of 0.92 (95%CI: 0.77, 0.98) for M-CHAT/F compared with 0.92 (95%CI: 0.80, 0.98) for JA-OBS. When the screening tools were used in conjunction, so children were deemed to be screen positive if positive for M-CHAT/F and/or JA-OBS, sensitivity for M-CHAT/F plus JA-OBS was similar to that for JA-OBS alone, 0.96 (95%CI: 0.85, 0.99), but PPV reduced to 0.90 (95%CI: 0.77, 0.96) (compared to 0.92 for JA-OBS alone).
Kerub (34) reported that M-CHAT/F is estimated to have higher sensitivity than GDS, 0.70 (95%CI: 0.35, 0.93) compared to 0.50 (95%CI: 0.19, 0.81), with slightly lower specificity: 0.98 (95%CI: 0.97, 0.99) compared to 0.998 (95%CI: 0.992, 0.999). Combining the results from M-CHAT/F and GDS does not improve the sensitivity or specificity from using M-CHAT/F alone (M-CHAT/F plus GDS: sensitivity 0.70 (95%CI: 0.35, 0.93), specificity 0.968 (95%CI: 0.96, 0.97)). However, the risk of bias for the evaluation of comparative accuracy was generally unclear.

Although Wiggins 2014 (45) reported results for M-CHAT/F and PEDS, because only children who were positive on M-CHAT/F were followed up, it is difficult to make comparisons between the 2 tools and results should be interpreted with care. According to the PPVs calculated by the 2021 UK NSC review authors, the PEDS Path B strategy was the best performing (PPV 0.75 (95%CI: 0.35, 0.97)), however this estimate is based on just 8 children (including 6 true positives).

Social Attention Communication Surveillance-Revised (SACS-R)

Mozolic-Staunton (37) is a retrospective analysis of 2 prospective cohort studies evaluating SACS-R and PEDS as screening tools for ASD (See Table 6). They included 13,417 children between the ages of 12 and 48 months, screened at community health check-ups. Only children who screened positive for SACS-R or PEDS were evaluated by paediatric health professionals. Diagnosis was made based on clinical judgement with ADOS and/or ADI-R. The study prevalence of ASD was 1.49%. As children who screened negative were not followed-up, the estimates of sensitivity and specificity reported by Mozolic-Staunton of 0.82 (0.76, 0.87) and 0.99 (0.99, 1.00), respectively, should be interpreted with caution. The

PPV of 0.83 (0.77, 0.88) for SACS-R can be reliably estimated from this study. The performance of SACS-R in conjunction with PEDS is not reported.

Parents Evaluation of Developmental Status (PEDS)

Mozolic-Staunton reported a PPV of 0.89 (0.47, 0.99) for PEDS ASD, with estimates of sensitivity and specificity of 0.07 (0.03, 0.14) and 0.99 (0.99, 1.00), respectively. Wiggins 2014 (45) reported results for PEDS Path A, PEDS Path B and PEDS ASD, however only those children who screened positive on M-CHAT/F were followed up. Thus, few children who were positive on the PEDS had a diagnostic assessment. We have calculated PPVs for these 3 PEDS criteria, but caution that these results are based on only a small proportion of the total number of children who were positive according to PEDS. The PPV for PEDS Path A is 0.55 (95%CI: 0.39, 0.71), 0.75 for PEDS Path B (95%CI: 0.35, 0.97), and 0.59 for PEDS ASD (95%CI: 0.39, 0.76).

Social Communication Questionnaire (SCQ)

Suren 2019 (40) involves the largest screening population of the included studies. 58,520 children aged 36 months were screened using the Social Communication Questionnaire (SCQ), see Table 6. Children referred for diagnostic assessment during the study were assessed based on DSM-IV-TR. The medical records of children who were not referred were reviewed at a later date and based on ICD-10. Sensitivity and specificity ranged from 0.42 (0.37, 0.47) and 0.89 (0.89, 0.90) with a cut-off of ≥11, to 0.25 (0.20, 0.29) and 0.99 (0.99, 0.99) with a cut-off of ≥12, respectively. Increasing the cut-off to ≥15 did not improve specificity, but reduced sensitivity compared to the cut-off of ≥12. Given that the SCQ is in Norwegian there are applicability concerns regarding the UK setting.

Ages and Stages Questionnaire 3 (ASQ-3)

Catino (42) screened 514 kindergarten children aged 42 and 48 months using the ASQ-3, see Table 6. Children who screened positive were invited for diagnostic assessment. Of the 24 children invited for diagnostic assessment, 2 received a diagnosis of ASD, giving a PPV of ASQ-3 for ASD of 0.08 (95%CI: 0.01, 0.25). Only 3 children screening positive with ASQ-3 were deemed to be typically developing. The remaining children who screened positive were found to have language disorder, developmental coordination disorder, and/or intellectual disability. Although only those children deemed to be positive on the ASQ-3 were referred for diagnostic assessment, the study was found to generally be of low risk of bias.

First Year Inventory (FYI)

Ben-Sasson 2013 (46) screened 613 12-month old children in Israel with the FYI. Children who screened positive (on the social domain alone, or social and sensory domains) were offered a further assessment at home one month later. This follow-up assessment was repeated at age 30 months. Sixty children matched on gender and socioeconomic status (presumably to those who did screen positive) who were screen negative on the FYI were also followed-up. The authors reported that the follow-up assessments formed the basis for clinical diagnosis. As the results reported by Ben-Sasson appear to include diagnoses of developmental delay as well as ASD, we have calculated sensitivity of the social-sensory cut-off for FYI based on the 5 reported ASD cases to be 0.60 (95%CI: 0.15, 0.95), and specificity as 0.75 (95%CI: 0.64, 0.84). However, it should be noted that a diagnosis in this study could have been made based on the follow-up data at 13 and/or 30 months old. Therefore, it may not accurately reflect the ability of FYI to identify ASD if signs or symptoms were not observable at 18 months prior to the follow-up assessment. The social cut-off did not identify any ASD cases.

First Year Inventory (FYI), ITC and M-CHAT-R/F

Wieckowski (32) screened 5,784 children using the FYI-L, ITC and M-CHAT-R/F (depending on the age at which they were screened). Children were initially screened at either 12, 15 or 18 months old during routine Well-Child visits. Subsequent screening was offered at 15, 18, 24 and 36 months old, depending on timing of initial screen. Children with a positive screen (on either tool if multiple tools used), or for whom concerns of ASD were raised, were refereed for diagnostic evaluation. Children who had a negative screening result but for whom concerns of ASD had been raised and subsequently received a diagnosis of ASD were all assumed to be false negatives. All children who had a negative screen and had no concerns of ASD raised were assumed to be true negatives. Due to a lack of follow-up of most children who had a negative screening result, interpretation of estimates of sensitivity, specificity and NPV are limited, even though they are reported by Wieckowski (32). Overall, PPVs and uptake were reported to be greater for children who were initially screened at 18 months old: PPV of 0.42 for a single screen at 18 months old, and PPV of 0.44 for an initial screen at 18 months old with later re-screens. Regardless of the age of initial screen, repeat screening was associated with higher estimates of PPV than single screens. Younger age at ASD diagnosis was statistically significantly associated with initial screening at age 12 months. No difference in age at diagnosis was observed between groups starting screening at 15 and 18 months old.
Although not stated, it is likely that diagnostic evaluation was not conducted blind to screening results. All screen negative children were assumed true negatives, and there are concerns regarding applicability as a proportion of the carers received a Spanish translation

of the screening tools. Due to different, and sometimes multiple, tools used for screening, it is difficult to say whether the better screening performance seen from 18 months old is due to increasing age of the child or the different screening tools used (M-CHAT-R/F rather than FYI or ITC).

Does the age at which the screening test is performed affect accuracy?

Three studies explored the accuracy of the screening tools by age at screening: Magan-Maganto 2020 (35), Catino 2017 (42) and Wieckowski (32). For the M-CHAT-R/F, Magan-Maganto (35) reported a higher sensitivity estimate for younger children than for older children: 0.82 (0.48, 0.97) in the 14-22 month age group compared to 0.75 (0.36, 0.96) in the 23-36 month age group. However, the 95% confidence intervals are wide for both subgroups and overlap, suggesting no statistically significant difference in specificity between the age groups. Specificity is reported to be 0.99 (0.99, 0.99) for both age groups. Catino (42) reported that 9 children in the 42 month age group were found to have a positive screening result, with one of these subsequently receiving a diagnosis of ASD, resulting in a PPV of 0.11. In the older 48 month age group, 1 of the 15 screen-positive children received a diagnosis of ASD, giving a PPV of 0.07.
Wieckowski reported accuracy results by the age of the child at initial screen: 12, 15 or 18 months old (32). They found that PPVs suggested better performance for children who had an initial screen at 18 months. For a single screen at 12, 15 and 18 months, PPVs were 0.22 (0.14, 0.32), 0.17 (0.09, 0.27) and 0.42 (0.34, 0.51), respectively, with no overlap between the 95% confidence intervals for 18 months and for the earlier screens. However, since different screening tools were used at different ages, this finding may reflect differences in the tools used at each age (i.e. only the M-CHAT-R/F was used at 18 months), rather than age at screening per se.

Do other characteristics affect the accuracy?

Suren (40) presented results for 3 different positivity thresholds depending on whether children subsequently diagnosed with ASD had phrase speech or not. Suren reported higher sensitivity estimates for SCQ for children with ASD who had no phrase speech. For the recommended threshold of ≥15, sensitivity was 0.46 (0.35, 0.57) for ASD children without phrase speech and 0.13 (0.9, 0.17) for ASD children with phrase speech. For both groups, specificity was estimated to be 0.99 (0.99, 0.99).
Achenie (39) reported the performance of machine learning models within different demographic subgroups described as white, black, male, female, low and high levels of maternal education. Justification for these particular subgroups was not provided, although

it is noted data from other ethnicities/races are sparse and so were not considered. The accuracy of results from the machine learning models were better within the different subgroups compared to the total group, suggesting that tailoring the tool for ethnicity, gender and level of maternal education could be justified.

Acceptability of screening

Screening uptake could be extracted or estimated from 8 of the included studies. These estimates ranged from 40% to 88.7%. Where a 2 stage screening tool had been used (i.e. M-CHAT(-R)/F), 6 studies reported the proportion with complete screening information, this ranged from 70% to 84%. Uptake of the diagnostic evaluation in those who screened positive was available from 14 studies. Uptake ranged from 57.5% to 100%, with the study by Canal-Bedia (48), reporting just 9.2% uptake of the diagnostic evaluation. The 2 studies reporting uptake of 100% evaluated ≤ 20 children.

None of the included studies reported details on reasons for families not taking part in the screening, or the subsequent diagnostic evaluation when it was offered. Jonsdottir 2020 (33) surveyed an unreported number of nurses involved in the study about their experiences of screening. The 10 nurses who responded, reported positively about the willingness of parents to answer the screening questions (mean score 4.9, SD 0.32, on Likert scale 1-5, with 5 being strongly agree). They reported positively that parents were able to answer the questions without assistance (mean 4.7, SD 0.48), and that the ASD screening was “easily integrated into the scheduled visit” (mean 4.5, SD 0.53). Jonsdottir (33) also reported that the nurses were very positive about a universal ASD screening programme: mean 4.8, SD 0.42 for the question “Screening all young children for ASD should be formally adopted”.

Are there any incidental findings?

Thirteen studies reported on incidental findings from the screening tools. From these studies, the proportion of false positive screens that were subsequently determined to have a non-ASD diagnosis or concern for developmental delay ranged from 7% to 100%, see Table 7. The non-ASD diagnoses are defined across the studies as global developmental disorders, language disorders, developmental coordination disorders, unspecified neurodevelopmental disorders and intellectual disability. The reported delays include language and psychomotor delays, and more general developmental concerns. For M-CHAT(R/F), all studies, except Kondolot (44) reported that ≥50% of false positives either have a non-ASD diagnosis or developmental delay concern.

Table 7 Quantity and type of incidental findings in those studies reporting such results

Study	Screening tool	Total number of false positive screening results	Percentage of false positives with atypical development	Diagnoses and/or concerns in false positives with atypical development
Allison 2021(26)	Q-CHAT	110 (from study design, not reported Q-CHAT threshold)	15%	Language delay, developmental delay, other atypical
Jonsdottir 2021, 2020(31, 33)	M-CHAT-R/F	7	86%	Non-ASD DSM diagnoses
Wieckowski 2021(32)	FYI, ITC and/or M-CHAT-R/F	71 (12 months only) 65 (15 months only) 76 (18 months only)	48% 51% 81%	Developmental disability (not defined)
Magan-Maganto 2020(35)	M-CHAT-R/F	10 (younger group) 14 (older group)	90% 71%	Disorders of language or global development. Diagnoses of unspecified neurodevelopmental or systemic disease. Delays of language and psychomotor skills
Oner 2020(36)	M-CHAT-R/F	95	41%	Developmental delay
Robins 2014(16)	M-CHAT-R/F	116	90%	Delays. Developmental concerns, with no diagnosis.
Kerub 2020(34)	M-CHAT/F GDS	28 53	68% 36%	Delays.
Baduel 2017(43)	M-CHAT/F	8	100%	Delays.
Kondolot 2017(44)	M-CHAT	15	7%	Developmental delay.
Chlebowski 2014(47)	M-CHAT(/F)	79	95%	Non-ASD diagnoses. Developmental concerns, with no diagnosis.
Wiggins 2014(45)	M-CHAT/F PEDS Path A	17 18	88% 89%	Global developmental disorder.
Nygren 2012(17)	M-CHAT/F plus JA-OBS	6	50%	Language disorder.
Catino 2014(42)	ASQ-3	8 (younger group) 12 (older group)	88% 68%	Disorders of language or developmental coordination. Intellectual disability.
Ben-Sasson 2013(46)	FYI	22	68%	Delays of developmental and language.

Discussion of findings

A study-level summary of data extracted from each included publication is presented in 'Summary and appraisal of individual studies **Error! Reference source not found.**'. Where the reviewers have performed calculations on the data presented in the publications, this has been clearly indicated in the tables.

Although a number of systematic reviews on the accuracy of screening tools for ASD in children have been published, their inclusion criteria were much broader than in this review. In particular, many studies included in those reviews were excluded here because they assessed accuracy in children already diagnosed with ASD, determined diagnosis well after screening had been conducted, or included children >5 years of age.

Of the 20 primary studies included, 15 evaluated a version of the M-CHAT(-R/F). Nine of these were carried out where English is not the first language; thus translations of the M-CHAT(R/F) were used. PEDS was evaluated in 2 studies; GDS, SACS-R, SCQ, TIDOS, ASQ-3, FYI and JA-OBS were all evaluated in one study each.

In the 2011 UK NSC review, the M-CHAT/F was identified as the most promising tool for a potential screening programme, but no estimates of sensitivity had been reported. In this 2021 review, a number of studies reported estimates of sensitivity for M-CHAT(R/F) ranging from 0.67 to 1. With many studies reporting sensitivity estimates of around 0.8 depending on age group or cut-off used.

Where comparative accuracy between M-CHAT/F and other screening tools had been conducted, the tools that incorporated observation of the child (TIDOS and JA-OBS) tended to perform better than the M-CHAT/F which relies on parent/carer reported questionnaires. Moreover, TIDOS and JA-OBS were the only non-M-CHAT(-R/F) tools with estimates of sensitivity above 0.5: 0.8 for TIDOS and 0.86 for JA-OBS. This finding has some consistency with the 2011 UK NSC review where surveillance of children by trained professionals was reported to have high sensitivity estimates (0.94 for children aged 3.5 years) (70). However, the resource implications for using observational screening tools compared to parent/carer-completed screening questionnaires should be kept in mind.

The ASQ-3 performed particularly poorly in identifying children with ASD, in terms of the reported PPV (the only accuracy metric calculated). It identified many incidental findings, but given the tool is for screening global development, rather than for ASD in particular, this is not surprising.

Little evidence was found on whether age or other characteristics impact on screening accuracy. Catino (42) and Magan-Maganto (35) investigated screening performance by age group, but neither reported any evidence to suggest there was a difference in accuracy between the age groups assessed. Suren (40) reported higher sensitivity of the SCQ to identify children with ASD who had not developed phrase speech compared to those who had. However, they also report that many of these children had already started the referral process, thus suggesting that screening may not help identify children with ASD without

phrase speech. Suren argued that screening could still be helpful in these children to speed up the diagnostic process, but also to identify children with other developmental concerns.

In line with findings from the 2011 UK NSC review, screening uptake is variable across studies. Completion of 2 stage screening tools was not 100% in the studies reporting this information. Thus, it would be important to consider ways to improve completion of 2 stage screening tools (as well as improve overall uptake), so that any advantages with the use of 2 stage screening tools are not outweighed by the disadvantages of non-completion. For instance, the M-CHAT(-R)/F includes a follow-up interview for those children found to be at high risk on the questionnaire, which has been shown to reduce the number of false positives. Although Jonsdottir (33) reported positive experiences of 10 nurses involved in the screening programme, none of the studies reported on the experiences of the parents. Further work in this area would be warranted if a screening programme for ASD in young children were to be considered.

Across studies and tools, the proportion of typically developing children who screen positive is <50% for all studies reporting incidental findings. This suggests that the tools might have a more general purpose than just identifying ASD. An exception to this is Kondolot (44) who used the M-CHAT without follow-up, where 93% of false positives were subsequently found to be children who are typically developing. Thus, emphasising the benefits of the two stage screening approach.

The included studies were generally well-conducted. A particular area of concern and variation was in whether and how diagnostic evaluation of children who screened negative took place. Conducting diagnostic assessments for all children, regardless of screening result, will be resource and time intensive for any study. A reasonable approach to this challenge is to assess a random sample of children who screened negative, which was done in 2 studies (41, 46). Alternatively, assessing medical records for diagnoses of ASD in screen negative children might be useful, but there are more threats with such an approach. Firstly, the timing of when medical records are reviewed is important. If such a review is done too long after the screening, any diagnoses found may not reflect children being false negatives as symptoms may not have been present at the time screening was undertaken. However, defining an appropriate time-period for reducing such risks is challenging. To best evaluate the screening tools, having as short an interval between application of the screening tool and diagnostic evaluation minimises any potential risks for the misclassification of children. Secondly, the diagnostic assessment for screen positives in the study may be different to the diagnostic assessment screen negatives go through outside of the study setting.

A further area of concern is whether the diagnostic evaluation was performed without knowledge of the screening results. If only children screening positive are referred and the professionals conducting the diagnostic assessment are aware of the study design, they will be aware of the screening results.

To improve understanding of the accuracy of screening tools to identify ASD in children ≤ 5 years in the UK, an ideal study would use English language tools and be based in the UK. Given the fact that so many studies evaluating M-CHAT(-R/F) used non-English translations, the applicability of their results to the UK is unclear. Identifying the screening tools to be evaluated in such a study would not be straightforward given many tools, in addition to M-CHAT(-R/F), have been developed and evaluated to some extent. However, since there is some evidence to suggest that observational tools may perform well in identifying children with ASD, comparison of an observational tool with a less resource-intensive parent/carer questionnaire could be useful. A subsample of screen negative children should be offered the same diagnostic assessment that screen positive children receive. The diagnostic assessment should be conducted blind to the screening results. In addition to measures of accuracy, time and resource use for each screening tool would also be helpful to inform consideration of the feasibility of incorporating the different tools into a screening programme. Parent/carer questionnaires may require relatively low levels of resource, while observations by health professionals would require more resources, not just in undertaking these assessments but also in any training required.

Summary of Findings Relevant to Criterion 4 and 5: Criteria not met

Versions of the M-CHAT have most commonly been evaluated, with estimates of sensitivity of approximately 0.8. However, such studies are generally at high risk of bias, mainly due to approaches to follow-up of screen negative children, and a lack of blinding of screening results in diagnostic evaluations for ASD. There is some evidence to suggest that tools based on observation of the child by professionals, could lead to improved estimates of sensitivity (over parent-completed questionnaires such as M-CHAT). There is little evidence to indicate that certain characteristics, such as age, affect screening performance.

The evidence reviewed here suggests that screening for ASD leads to incidental findings such as other developmental/language disorders or delays.

Criterion 9 – There should be an effective intervention for patients identified through screening, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care. Evidence relating to wider benefits of screening, for example those relating to family members, should be taken into account where available. However, where there is no prospect of benefit for the individual screened then the screening programme should not be further considered.

Question 3 – Has the benefit of early intervention in children aged 5 years and younger, detected through screening been demonstrated?

In the 2011 evidence review, 14 RCTs were identified that evaluated interventions for children ≤5 years old diagnosed with ASD. Three RCTs, including 100 participants in total, evaluated Early Intensive Behavioural Interventions (EIBI) or Applied Behavioural Analysis (ABA). These interventions aim to address a number of deficits including language, social and cognitive issues. Findings from these RCTs were mixed, with 2 reporting improvements in IQ, one of which also reported positive findings with visual-spatial skills and language, the other in adaptive behaviour associated with the intervention. The third study indicated that the interventions were not effective. The largest trial, which showed the greatest improvement, was limited to a 2 year follow-up. Eleven additional RCTs evaluated other focused behavioural interventions, which were generally less intensive than EIBI/ABA. Many of these studies reported some level of effectiveness of the interventions. However, again, the number of participants was small, with one study including 152 children (the rest included ≤ 60 children), and follow-up was limited. It was also noted that studies with more participants generally reported small effect sizes. Overall, the 2011 UK NSC evidence review concluded that the evidence on effectiveness of early intervention was variable, with uncertainty as to whether short-term improvements continued over time.

Eligibility for inclusion in the review

RCTs and cohort studies involving children aged ≤ 5 years diagnosed with ASD detected via screening. Studies were not excluded because of the type of intervention evaluated, however only those evaluating ASD core deficits or symptom severity were sought. Non-English language studies and those published before 2010 were excluded.

After full paper review the main reasons for exclusion of studies were that interventions were not implemented in populations identified through screening.

Description of the evidence

0contains a full PRISMA flow diagram (Figure 1), along with a table of the included publications and details of which questions these publications were identified as being relevant to (Table 15).

Of the total 5,498 titles and abstracts from the database searches, the full-text of 114 titles were reviewed for eligibility for this question. After reviewing the full-texts, only 3 studies were eligible to address this review question: 2 RCTs (49, 50) and a prospective cohort study (29). A fourth study was identified from other sources (51).

A summary of study characteristics, risk of bias and results is given in Table 8 for the 4 included studies that focussed on children who had been identified through screening for ASD. Tables in Appendix 3 provide more details on study characteristics (Table 21) and risk of bias (Table 27 and Table 28).

Table 8 Characteristics, risk of bias and results for studies evaluating the effectiveness of early interventions

Study, Country, Design	Screening	Sample size and follow-up	Intervention and Control	Overall RoB	Results*
Baranek 2015(49) **USA** **RCT**	Community sample, 12 months old (N=12,000). Screened positive on FYI, or parental concerns (N=59/2261 responses).	24 agreed to RCT. 18 eligible. 16 randomised (11 ART, 5 REIM) ~15 months old at randomisation. FU post-intervention (~22 months old), diagnostic evaluation (~32 months old)	ART, parent administered (after training). Mean of 33.5 (range 20–39) total contacts (in-home + phone/email) across a 6- to 8-month period. REIM	R: Low I: Some MD: Low M: Low RB: Low Overall: Some	ART significantly associated with improved receptive language, socialisation, sensory hyporesponsiveness and “less directive parental interactive style” during the intervention period. Little evidence of any difference at 32 month FU. ASD dx at 32 months old: 36% ART, 40% REIM, 100% not randomised.

Watson 2017(50) **USA** **RCT**	Community sample, 12 months old (N=61,437). Screened positive on FYI (N=280/8709 responses).	109 decline and 74 ineligible. 97 eligible and agreed to RCT. 87 consented to randomisation (45 ART, 42 REIM) ~13.7 months old at randomisation. FU post-intervention (~22 months old)	ART, parent administered (after training). Mean of 24.9 (*sd* = 5.2, range 12–32) in-home sessions and 2.4 (*sd* = 3.6, range 0–15) other contacts. REIM	R: Low I: Some MD: Low M: Low RB: Low Overall: Some	No evidence that ART associated with Improved Social-Communication, Sensory‑Regulatory, Adaptive, and Autism Symptom Outcomes. ART was associated with improvements in motor skills, but the finding could just reflect regression-to the mean. Across both groups, 41% met criteria for ASD,
Spjut Jansson 2016(29) **Sweden** **Prospective naturalistic cohort**	2.5 year old children referred following routine screening for ASD in Gothenburg (tool NR). From 2009 – 2011, 134 <4 years referred with positive screening result.	129 consented to assessment. 100 met ASD dx criteria. 71 received interventions. Approx. 36 months old at ASD dx evaluation. FU after 2 years (approx. 60 months old).	Regular Intensive Learning programme Modified intensive learning programme Usual care	C: Low P: Low Class: Low I: Low MD: Low M: Low RB: Low Overall: Low	Adaptive composite scores: No evidence of increase in scores over time across total sample. No evidence of any of the interventions increased scores more than another intervention. Global functioning: Evidence that scores increased over time, but no evidence that greater increases seen with any of the interventions over another.
Whitehouse 2021(51) **Australia** **RCT**	12 months old infants referred following community wide screening in Perth and Melbourne	104 were randomised 89 included in intention-to-treat analysis at 24 months (3 years of age)	iBASIS–Video Interaction to Promote Positive Parenting (iBASIS-VIPP) Usual care	R: Some I: Some MD: Low M: Low RB: Low Overall: Some	Combined treatment effect on reducing ASD symptom severity across time points favoured the intervention

											(ABC, −5.53; 95%CI, −∞ to −0.28; P = .04).

*Cochrane Risk of Bias tool for RCTs, ROBINS-I for cohort study.

ART, Adapted Responsive Teaching; C, confounding; FU, follow-up; FYI, First Year Inventory; P, participants; R, randomisation; I, intervention(s); MD, missing data; NA, not applicable; NR, not reported; OM, outcome measurement; RB, bias in reporting outcomes; RCT, randomised controlled trial; REIM, referral to early intervention and monitoring;

In their RCTs, Baranek and Watson (49, 50) included children who were found to be at risk of ASD according to the First Year Inventory (FYI) from a community population in USA. Baranek also included children who had not screened positive on FYI, but had concerns raised by their parents. In both RCTs, children deemed to be at-risk of ASD were randomised to either receive a 24-week parent-led intervention (Adapted Responsive Teaching (ART)) or were referred to existing services in the local communities (referral to early intervention and monitoring (REIM)). ART is a home-based relationship-focussed intervention aiming to improve outcomes in children through the encouragement of parents "to use responsive strategies during daily routines with their children, … designed to target "pivotal" behaviours (for example, social play, joint attention, arousal and attention, engagement, adaptability, and coping)" (49). Both RCTs provided the intervention over 6 months, consisting of 36 planned contacts (mainly home sessions, with additional phone calls and emails) between parents and professionals experienced in child development. Although participants were identified from relatively large community populations (2,261 and, respectively, 8,709 screen results were available), studies included relatively low samples (16 and 87 respectively). After contacting the authors, it was confirmed that the 2 studies used different samples, so were independent of each other.

Baranek reported that ART was significantly associated with improved receptive language, socialisation and sensory hyporesponsiveness in children, and "less directive parental interactive style" during the intervention period, compared to REIM. However, the larger RCT by Watson found no evidence that ART was associated Improved Social‐Communication, Sensory‐Regulatory, Adaptive, and Autism Symptom Outcomes compared to REIM.

The Swedish prospective cohort study by Spjut Jansson (29) included children aged 2 and a half years who were referred to the Child Neuropsychiatric Clinic following a positive ASD screen result from routine ASD screening. Of the 134 children referred, consent was provided for 129, and 100 of these subsequently received a diagnosis of ASD. Evaluation after 2 years did not show any significant differences between interventions on the Vineland Adaptive Behavior Scale, Second Edition (VABS-II) (71) and the Children's Global Assessment Scale (C-GAS) (72).

Whitehouse(51) included children aged 12 months who were referred mostly as a result of a positive result following community wide screening. 104 children were randomised to either the iBASIS–Video Interaction to Promote Positive Parenting (iBASIS-VIPP) intervention or usual care. Data were available for 89 children 2 years after baseline, showing a reduction in ASD symptom severity (ABC, −5.53; 95%CI, −∞ to −0.28; P = .04) and reduced odds of ASD classification (odds ratio, 0.18; 95%CI, 0-0.68; P = .02).

Quality of the included studies was acceptable, with some concerns due to the lack of blinding to the intervention type for the RCTs. Spjut Jansson et al. (29) presented the least concern of bias, however it was a non-randomised study. Whitehouse(51) included a mix of screened and referred participants. We contacted the authors for clarification, and they confirmed that the majority of the children were included following their positive screen result, with only a minority in the Perth trial site being referred beside the screened participants.

Discussion of findings

A study-level summary of data extracted from each included publication is presented in ‘Summary and appraisal of individual studies **Error! Reference source not found.**’.

Of 114 full-text articles reviewed, only 4 were found to evaluate interventions in young children identified through screening for ASD. Two RCTs took a similar approach and screened children at 12 months old, and evaluated the impact of ART (over a 6 month duration) compared to REIM. The largest RCT, which still only included 89 children, found that treatment effect (reduced ASD severity) of iBASIS-VIPP maintained at 2 years follow up, however the study sample was contaminated with referred patients in one of the research sites, making the results of this study less relevant. The other studies found no evidence of improved outcomes. The Swedish cohort study found no difference between interventions in global functioning or adaptive composite scores.

In the 2011 UK NSC review, although 14 RCTs were included, it is not clear whether children included in these RCTs had been identified through screening. A quick review of the 14 RCTs indicates that only one study was conducted in a screened population(73). In this RCT, 24 families of children with a median age of 23 months (identified through use of CHAT) were randomised to receive training on joint attention skills and action routines, and visits from speech and language therapists or usual care. After 12 months of receiving the intervention, “marginal improvements in words understood” were reported in the intervention group (70).

Bringing together the evidence identified in the 2021 update review, and that from the USPSTF (24) (who did not identify any studies in a screened population) adds little to the previous 2011 UK NSC review. Very few studies were identified, and these were small, affected by attrition to varying degrees, and provided inconclusive evidence for the effectiveness of interventions in children identified as at risk of ASD through screening. As the maximum follow-up among the studies identified was just 2 years, there is limited evidence on the long-term outcomes of early intervention in these young children identified through screening.

There are many challenges to the design of the ideal study to evaluate effectiveness of interventions in this population. Larger studies would be preferable, so that they could be adequately powered to detect intervention effects. However, since the target population is individuals identified through ASD screening, and the prevalence of ASD is 1-2%, very large populations would need to be screened, so that larger numbers of children would be available for randomisation. For instance, Watson 2017 screened >8700 children, and 87 were eligible and consented to randomisation (they reported requiring a sample size of 102 to detect a statistically significant difference).

Even considering the low prevalence of ASD, attrition is a particular challenge where the population is screened detected children as there are many points in the pathway where participants could drop-out. As highlighted in the review of screening accuracy studies above, where reported, uptake of screening ranged from 40% to 88%. While uptake of subsequent diagnostic evaluation for those who were identified as being at high risk from the screening tool(s) ranged from 57.5% to 100%. At this point in the pathway, for RCTs, it is likely that more participants will drop-out before randomisation and commencement of treatment.

Ideal future studies would include longer follow-up, so that the long-term outcomes of early intervention can be evaluated. However, given the challenges of attrition in these studies generally, following-up participants for longer time-points is also likely to increase the probability of participants dropping-out.

Studies involving screening of larger populations, or multiple populations might be required to have sufficient power to evaluate the effectiveness of interventions. While efforts to reduce the number of participants dropping out at different points would be required.

Summary of Findings Relevant to Criterion 9: Not met

Only 4 studies,3 RCTs and a cohort study, were identified that evaluated the effectiveness of early intervention in young children with ASD identified through population screening. Thus, the available evidence is limited by a small number of studies that include children identified through ASD screening. Moreover, the studies identified have insufficient power, are affected by attrition, report mixed findings and only have short-term follow-up.

Review summary

Conclusions and implications for policy

Overall, the evidence reviewed here do not indicate that screening for ASD should be recommended for children aged ≤5 years.
Although there is some uncertainty as to the performance of screening tools to identify children with ASD (Question 2), the main limiting factors are uncertainty as to the stability of diagnoses of ASD when made at such young ages (Question 1), and the current lack of evidence on the effectiveness of interventions for children identified through ASD screening (Question 3).

In particular, although estimates of stability ranged from 72% to 100%, all studies raised concerns regarding risk of bias, which importantly included a lack of blinding in follow-up assessments. Notwithstanding this, there is little evidence that ASD diagnoses are maintained beyond the age of 4 or 5 years old, since the children in the included studies were approximately 2 years old when they received their initial diagnosis and most had a maximum follow-up of 2 years. In terms of evaluating effectiveness of interventions, very few studies were identified, and these were small, affected by attrition to varying degrees, and provided inconclusive evidence for the effectiveness of interventions in children identified as at risk of ASD through screening. As the maximum follow-up among the studies identified was just 2 years, there is limited evidence on the long-term outcomes of early intervention in these young children identified through screening.

Further work is warranted to help address all 3 questions. To examine the stability of diagnoses of screen-detected ASD (Question 1), further studies are needed that ensure that diagnostic evaluation at follow-up is blind to that made initially, and that follow-up is longer than 2 years. To assess the performance of screening tools (Question 2), more studies are needed that attempt to follow-up a proportion of children who screen negative, so that reliable estimates of sensitivity and specificity can be obtained. Such studies should also ensure that diagnostic evaluation is conducted blind to the screening results. Ideally, these studies would evaluate and compare more than one tool, preferably comparing tools that involve observation of children, with tools that involve parent-completed questionnaires, for example. Evidence on factors affecting uptake or completion of ASD screening, and how better uptake or completion might be achieved would also be warranted.

To better evaluate the effectiveness of interventions in children with ASD identified through screening (Question 3), larger studies with longer follow-up would be needed. However,

due to the relatively low prevalence of ASD, and issues of attrition, such studies will need to effectively reduce the likelihood of children/families dropping out from the study at various time-points.

Limitations

The available evidence relevant to all 3 questions are limited. For question 1, studies are limited by a lack of blinding of initial diagnostic assessments at the follow-up diagnostic assessment, and by a lack of follow-up.

Particular aspects of study design limited many of the studies included in this review. For instance, a lack of blinding limits the interpretation of most of the studies that evaluated diagnostic stability and many of the screening accuracy studies. While short follow-up periods limit the extent to which diagnoses can be said to be stable beyond 2 years after diagnosis, and interventions effective after 2 years. A particular limitation of many of the screening accuracy studies is to what extent, and how, children who were negative on the screening tool were followed-up, so that reliable estimates of sensitivity and specificity estimates could be obtained.

The review is limited by the inclusion of English-language only studies, that were published since 2010. Only a proportion of articles identified from the database searches were double-screened at title and abstract, or full-text stage. Moreover, the level of agreement between reviewers on inclusion of relevant studies was generally low, due to aspects of study design not being reporting clearly. It is therefore possible that some relevant studies have not been included in the review. However, given all of the current uncertainties and limitations in the evidence across the 3 research questions, it is unlikely that omission of some further studies would lead to a different recommendation at this point.

Appendix 1 — Search strategy

Electronic databases

The search strategy included searches of the databases shown in Table 9. MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print and Embase.

Table 9. Summary of electronic database searches and dates

Database	Platform	Searched on date	Date range of search	Hits
MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print	Ovid SP	16/11/20	1946 to November 13, 2020	2980
Embase	Ovid SP	16/11/20	1974 to 2020 November 13	3015
CINAHL	EBSCOhost	16/11/20		1111
APA PsycInfo			1806 to November Week 2 2020	2673
The Cochrane Library, including: – Cochrane Database of Systematic Reviews (CDSR) – Cochrane Central Register of Controlled Trials (CENTRAL)	Wiley Online	17/11/20	CDSR: Issue 11 of 12, November 2020 CENTRAL: Issue 11 of 12, November 2020	16 622
ClinicalTrial.gov		19/11/20		107
ICTRP		N/A	Currently unavailable	N/A

Search Terms

Search terms included combinations of free text and subject headings (Medical Subject Headings [MeSH] for MEDLINE, and Emtree terms for Embase), grouped into the following categories:

- disease area: autism spectrum disorder, developmental disabilities, neurodevelopmental disorders
- population: child, preschool, young child, toddler
- intervention: test, screening, questionnaire, M-CHAT
- outcomes: diagnostic stability, sensitivity, specificity, accuracy, predictive value, attention intervention, applied behavioural intervention

Search terms for MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print are shown in Table 10; search terms for Embase, APA PsycInfo CINAHL and the Cochrane Library databases are shown in Table 11, Table 12, Table 13 and Table 14, respectively.

Table 10. Search strategy for MEDLINE, MEDLINE In-Process, MEDLINE Daily, Epub Ahead of Print

N	Terms (N of hits)
1	exp Autism Spectrum Disorder/di, rh, th [Diagnosis, Rehabilitation, Therapy] (10860)

2	autis*.ti,jn. (33767)
3	ASD.ti. (1736)
4	asperger*.ti. (1126)
5	Developmental Disabilities/di [Diagnosis] (5139)
6	Neurodevelopmental Disorders/di [Diagnosis] (439)
7	Child Development Disorders, Pervasive/di [Diagnosis] (2110)
8	((developmental or neurodevelopment*) adj (condition* or disorder*)).ti,ab. (19779)
9	or/1-4 (37836)
10	or/1-8 (58015)
11	Child, Preschool/ (925828)
12	(child* or preschool* or toddler* or infant*).ti,ab. (1695387)
13	("young child*" or toddler* or infant* or preschool*).ti,ab. (477666)
14	11 or 12 (2112390)
15	11 or 13 (1297535)
16	test*.ti,ab. (3264770)
17	M CHAT.ti,ab. (144)
18	PDDST.ti,ab. (3)
19	autism spectrum quotient.ti,ab. (479)
20	scale.ti,ab. (744294)
21	checklist.ti,ab. (36389)
22	tool*.ti,ab. (751429)
23	index.ti,ab. (797646)
24	questionnaire*.ti,ab. (525966)
25	screening.ti,ab. (536834)
26	instrument*.ti,ab. (288019)
27	(measure or measures or measuring).ti,ab. (1450192)
28	observation schedule.ti,ab. (694)
29	or/16-28 (6521202)
30	"sensitivity and specificity"/ or "predictive value of tests"/ (521015)
31	stability.ti,ab. (436796)
32	diagnostic value.ti,ab. (34170)
33	sensitivity.ti,ab. (815702)
34	specificity.ti,ab. (471234)
35	(validity or validation).ti,ab. (364353)
36	reliability.ti,ab. (168571)
37	(utility or utilisation or utilization).ti,ab. (413254)
38	predictive value.ti,ab. (93819)
39	accuracy.ti,ab. (414727)
40	acceptability.ti,ab. (37814)
41	feasibility.ti,ab. (188628)
42	false positives.ti,ab. (13433)
43	false negatives.ti,ab. (6254)
44	or/30-43 (2988350)
45	(early adj3 intervention*).ti,ab. (34422)
46	play therapy.ti,ab. (392)
47	attention intervention*.ti,ab. (58)
48	communication intervention*.ti,ab. (872)
49	language intervention*.ti,ab. (469)
50	(play adj2 intervention*).ti,ab. (585)
51	pivotal response.ti,ab. (79)
52	occupational therap*.ti,ab. (13889)
53	applied behavio?r analysis.ti,ab. (550)
54	focus?ed behavio?ral intervention*.ti,ab. (28)
55	psychosocial intervention*.ti,ab. (5464)
56	or/45-55 (56129)
57	10 and 14 and 29 and 44 (3311)

58	9 and 15 and 56 (885)
59	57 or 58 (4076)
60	limit 59 to yr="2010-Current" (2980)

Table 11.Search strategy for Embase <1974 to 2020 November 13>

N	Terms (N of hits)
1	exp autism/di [Diagnosis] (8154)
2	autis*.ti,jn. (41834)
3	ASD.ti. (3042)
4	asperger.ti. (873)
5	developmental disorder/di [Diagnosis] (3716)
6	1 or 2 or 3 or 4 (47146)
7	1 or 2 or 3 or 4 or 5 (50331)
8	preschool child/ (554471)
9	(child* or preschool* or toddler* or infant*).ti,ab. (2082239)
10	(young child* or toddler* or infant* or preschool*).ti,ab. (551188)
11	8 or 9 (2300984)
12	8 or 10 (1022028)
13	exp autism assessment/ (2387)
14	test*.ti,ab. (4435690)
15	M CHAT.ti,ab. (223)
16	pddst.ti,ab. (5)
17	autism spectrum quotient.ti,ab. (600)
18	scale.ti,ab. (1006270)
19	checklist.ti,ab. (49785)
20	tool*.ti,ab. (1018185)
21	questionnaire*.ti,ab. (771359)
22	screening.ti,ab. (757364)
23	instrument*.ti,ab. (374360)
24	(measure or measures or measuring).ti,ab. (1912691)
25	observation schedule.ti,ab. (921)
26	or/13-25 (8083317)
27	"sensitivity and specificity"/ (375166)
28	diagnostic value/ (191675)
29	predictive value/ (180074)
30	stability.ti,ab. (502558)
31	diagnostic value.ti,ab. (45177)
32	sensitivity.ti,ab. (1063611)
33	specificity.ti,ab. (611490)
34	(validity or validation).ti,ab. (495009)
35	reliability.ti,ab. (209048)
36	(utility or utilisation or utilization).ti,ab. (565201)
37	predictive value.ti,ab. (143699)
38	accuracy.ti,ab. (537700)
39	acceptability.ti,ab. (49072)
40	feasibility.ti,ab. (267784)
41	false positives.ti,ab. (18496)
42	false negatives.ti,ab. (8725)
43	or/27-42 (3729059)
44	(early adj3 intervention*).ti,ab. (52087)
45	play therapy.ti,ab. (506)
46	attention intervention*.ti,ab. (73)
47	communication intervention*.ti,ab. (1143)
48	language intervention*.ti,ab. (564)

49	(play adj2 intervention*).ti,ab. (818)
50	pivotal response.ti,ab. (97)
51	occupational therap*.ti,ab. (20451)
52	applied behavio?r analysis.ti,ab. (504)
53	focus?ed behavio?ral intervention*.ti,ab. (40)
54	psychosocial intervention*.ti,ab. (7911)
55	or/44-54 (83080)
56	7 and 11 and 26 and 43 (3034)
57	6 and 12 and 55 (967)
58	56 or 57 (3870)
59	limit 58 to yr="2010-Current" (3015)

Table 12. Search strategy for APA PsycInfo <1806 to November Week 2 2020>

N	Terms (N of hits)
1	autism spectrum disorders/ (44700)
2	autis*.ti,jn. (36049)
3	ASD.ti. (1586)
4	asperger*.ti. (1955)
5	developmental disabilities/ (12483)
6	((developmental or neurodevelopment*) adj (condition* or disorder* or disabilit*)).ti,ab. (21312)
7	1 or 2 or 3 or 4 (46609)
8	1 or 2 or 3 or 4 or 5 or 6 (65981)
9	exp preschool students/ (11882)
10	(child* or preschool* or toddler* or infant*).ti,ab. (748517)
11	("young child*" or toddler* or infant* or preschool*).ti,ab. (153233)
12	9 or 10 (748763)
13	9 or 11 (154849)
14	M CHAT.ti,ab. (140)
15	PDDST.ti,ab. (6)
16	autism spectrum quotient.ti,ab. (432)
17	scale.ti,ab. (318583)
18	checklist.ti,ab. (26568)
19	tool*.ti,ab. (155799)
20	index.ti,ab. (102522)
21	questionnaire*.ti,ab. (280262)
22	screening.ti,ab. (65318)
23	instrument*.ti,ab. (139778)
24	(measure or measures or measuring).ti,ab. (553395)
25	observation schedule.ti,ab. (749)
26	or/14-25 (1199756)
27	test reliability/ (54772)
28	test performance/ (4840)
29	test validity/ (79303)
30	stability.ti,ab. (45321)
31	diagnostic value.ti,ab. (1433)
32	sensitivity.ti,ab. (93311)
33	specificity.ti,ab. (37213)
34	(validity or validation).ti,ab. (169924)
35	reliability.ti,ab. (86436)
36	(utility or utilisation or utilization).ti,ab. (92000)
37	predictive value.ti,ab. (7850)
38	accuracy.ti,ab. (74361)
39	acceptability.ti,ab. (13869)

40	feasibility.ti,ab. (22897)
41	false positives.ti,ab. (1628)
42	false negatives.ti,ab. (670)
43	or/27-42 (533772)
44	(early adj3 intervention*).ti,ab. (19381)
45	play therapy.ti,ab. (3056)
46	attention intervention*.ti,ab. (68)
47	communication intervention*.ti,ab. (784)
48	language intervention*.ti,ab. (910)
49	(play adj2 intervention*).ti,ab. (762)
50	pivotal response.ti,ab. (183)
51	occupational therap*.ti,ab. (10345)
52	applied behavio?r analysis.ti,ab. (1869)
53	focus?ed behavio?ral intervention*.ti,ab. (14)
54	psychosocial intervention*.ti,ab. (5497)
55	or/44-54 (41727)
56	8 and 12 and 26 and 43 (2920)
57	7 and 13 and 55 (1020)
58	56 or 57 (3817)
59	limit 58 to yr="2010 -Current" (2673)

Table 13. Search strategy for CINAHL

#	Query	Results
S57	S55 OR S56	1,111
S56	S8 AND S13 AND S54	369
S55	S9 AND S14 AND S28 AND S43	1,086
S54	S44 OR S45 OR S46 OR S47 OR S48 OR S49 OR S50 OR S51 OR S52 OR S53	39,673
S53	TI "psychosocial intervention*" OR AB "psychosocial intervention*"	3,296
S52	TI ("focused behavior" or "focused behaviour" or "focussed behavior" or "focussed behaviour") OR AB ("focused behavior" or "focused behaviour" or "focussed behavior" or "focussed behaviour")	25
S51	TI ("applied behavior analysis" or "applied behaviour analysis") OR AB ("applied behavior analysis" or "applied behaviour analysis")	298
S50	TI "occupational therap*" OR AB "occupational therap*"	24,349
S49	TI "pivotal response" OR AB "pivotal response"	69
S48	TI play N2 intervention OR AB play N2 intervention	519
S47	TI "early intervention*" OR AB "early intervention*"	10,404
S46	TI "attention intervention*" OR AB "attention intervention*"	41
S45	TI "communication intervention*" OR AB "communication intervention*"	727
S44	TI "play therapy" OR AB "play therapy"	422

S43	S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42	474,710
S42	TI "false negatives" OR AB "false negatives"	910
S41	TI "false positives" OR AB "false positives"	1,885
S40	TI "feasibility" OR AB "feasibility"	47,818
S39	TI "acceptability" OR AB "acceptability"	15,529
S38	TI "accuracy" OR AB "accuracy"	75,004
S37	TI "predictive value" OR AB "predictive value"	22,794
S36	TI ("utility" or "utilisation" or "utilization") OR AB ("utility" or "utilisation" or "utilization")	101,858
S35	TI "reliability" OR AB "reliability"	61,558
S34	TI ("validity" or "validation") OR AB ("validity" or "validation")	108,295
S33	TI "specificity" OR AB "specificity"	57,182
S32	TI "sensitivity" OR AB "sensitivity"	113,584
S31	TI "diagnostic value" OR AB "diagnostic value"	4,923
S30	TI "stability" OR AB "stability"	37,139
S29	(MM "Sensitivity and Specificity")	1,894
S28	S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27	1,047,705
S27	TI "observation schedule" OR AB "observation schedule"	326
S26	TI ("measure" or "measures" or "measuring") OR AB ("measure" or "measures" or "measuring")	421,034
S25	TI "instrument" OR AB "instrument"	44,204
S24	TI "screening" OR AB "screening"	134,809
S23	TI "questionnaire" OR AB "questionnaire"	171,214
S22	TI "index" OR AB "index"	203,292
S21	TI "tool*" OR AB "tool*"	178,549
S20	TI "checklist" OR AB "checklist"	17,142
S19	TI "scale" OR AB "scale"	204,357
S18	TI "autism spectrum quotient" OR AB "autism spectrum quotient"	178
S17	TI "PDDST" OR AB "PDDST"	2
S16	TI "M Chat" OR AB "M Chat"	83
S15	(MM "Diagnosis, Developmental")	254
S14	S10 OR S11	572,626
S13	S10 OR S12	139,222
S12	TI ("young child*" or preschool* or toddler* or infant*) OR AB ("young child*" or preschool* or toddler* or infant*)	139,085
S11	TI (child* or preschool* or toddler* or infant*) OR AB (child* or preschool* or toddler* or infant*)	572,562
S10	(MM "Child, Preschool")	385
S9	S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7	25,862
S8	S1 OR S2 OR S3 OR S4	23,675
S7	TI "developmental disorder*" OR AB "developmental disorder*"	2,207

S6	(MM "Child Development Disorders, Pervasive/DI")	323
S5	(MM "Developmental Disabilities/DI")	881
S4	TI asperger*	982
S3	TI ASD	1,418
S2	TI autis* OR SO autis*	22,192
S1	(MM "Autistic Disorder/DI")	2,496

Table 14. Search strategy for the Cochrane Library Databases (Searched via the Wiley Online platform)

ID	Search terms	Results
#1	MeSH descriptor: [Autistic Disorder] explode all trees	1001
#2	autis*:ti,ab	3484
#3	ASD:ti	321
#4	asperger*:ti	40
#5	MeSH descriptor: [Developmental Disabilities] explode all trees	635
#6	MeSH descriptor: [Neurodevelopmental Disorders] explode all trees and with qualifier(s): [diagnosis - DI]	1026
#7	MeSH descriptor: [Child Development Disorders, Pervasive] explode all trees and with qualifier(s): [diagnosis - DI]	176
#8	((developmental or neurodevelopment*) NEXT (condition* or disorder*)):ti,ab	784
#9	#1 or #2 or #3 or #4	3624
#10	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8	5330
#11	MeSH descriptor: [Child, Preschool] explode all trees	29316
#12	(child* or preschool* or toddler* or infant*):ti,ab	153535
#13	("young child*" or toddler* or infant* or preschool*):ti,ab	47657
#14	#11 or #12	159396
#15	#11 or #13	71371
#16	test*:ti,ab	332015
#17	M CHAT:ti,ab	175
#18	PDDST:ti,ab	0
#19	"autism spectrum quotient":ti,ab	34
#20	scale:ti,ab	157276
#21	checklist:ti,ab	6494
#22	tool*:ti,ab	31669
#23	index:ti,ab	132731
#24	questionnaire*:ti,ab	106632
#25	screening:ti,ab	51010
#26	instrument*:ti,ab	21709
#27	(measure or measures or measuring):ti,ab	203711
#28	"observation schedule":ti,ab	114
#29	#16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24 or #25 or #26 or #27 or #28	685023
#30	MeSH descriptor: [Sensitivity and Specificity] explode all trees	15343
#31	MeSH descriptor: [Predictive Value of Tests] explode all trees	6960
#32	stability:ti,ab	14138
#33	"diagnostic value":ti,ab	801
#34	sensitivity:ti,ab	45526

#35	specificity:ti,ab	10947
#36	(validity or validation):ti,ab	18779
#37	reliability:ti,ab	8781
#38	(utility or utilisation or utilization):ti,ab	27621
#39	"Predictive value":ti,ab	5441
#40	accuracy:ti,ab	17227
#41	acceptability:ti,ab	14637
#42	feasibility:ti,ab	34185
#43	"false positives":ti,ab	461
#44	"false negatives":ti,ab	250
#45	#30 or #31 or #32 or #33 or #34 or #35 or #36 or #37 or #38 or #39 or #40 or #41 or #42 or #43 or #44	167707
#46	early NEXT/3 intervention*:ti,ab	4818
#47	"play therapy":ti,ab	98
#48	"attention intervention*":ti,ab	30
#49	"communication intervention*":ti,ab	258
#50	"language intervention*":ti,ab	108
#51	play NEXT/2 intervention:ti,ab	126
#52	"pivotal response":ti,ab	44
#53	"occupational therapies":ti,ab	13
#54	"applied behaviour analysis":ti,ab or "applied behavior analysis":ti,ab	65
#55	("focused behaviour intervention*" or "focused behavior intervention*" or "focussed behaviour intervention*" or "focussed behavior intervention*"):ti,ab	0
#56	"psychosocial intervention*":ti,ab	1118
#57	#46 or #47 or #48 or #49 or #50 or #51 or #52 or #53 or #54 or #55 or #56	6539
#58	#10 and #14 and #29 and #45	441
#59	#9 and #15 and #57	223
#60	#58 or #59	638

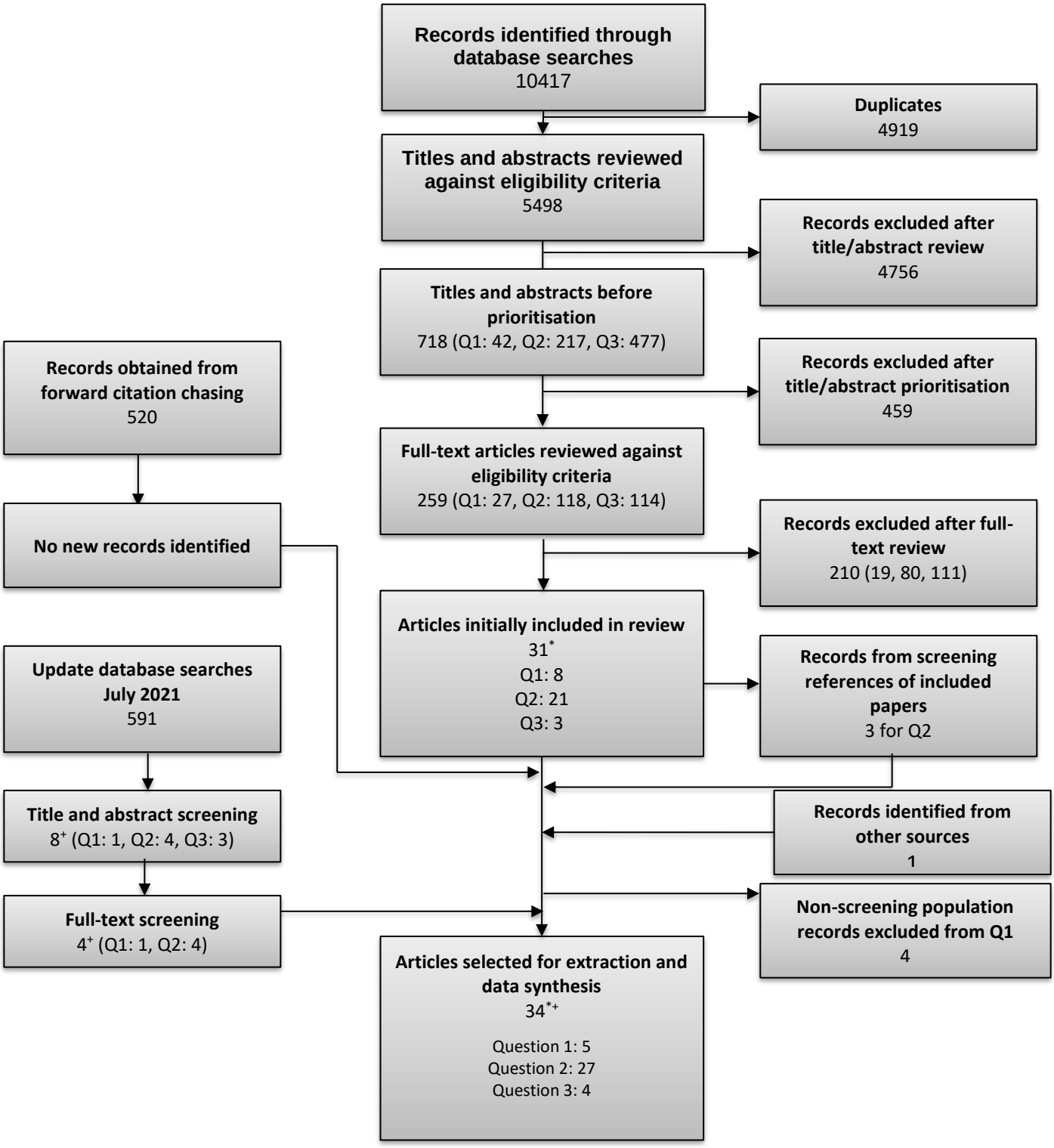
Results were imported into EndNote and de-duplicated.

Appendix 2 — Included and excluded studies

PRISMA flowchart

Figure 1 shows the volume of publications included and excluded at each stage of the review. 34 publications were ultimately judged to be relevant to one or more review questions and were considered for extraction. Publications that were included or excluded after the review of full-text articles are detailed below.

Figure 1. Summary of publications included and excluded at each stage of the review



Note: *One paper included in both Q1 and Q3; + One paper included in both Q1 and Q2

Publications included after review of full-text articles

The 34 publications included after review of full-texts are summarised in Table 15 below.

Studies were prioritised for extraction and data synthesis. It was planned *a priori* that the following approach would be taken to prioritise studies for extraction:

1. Systematic reviews and meta-analyses would be considered the highest quality of evidence if any were found. Following this, for Q1 prospective studies were prioritised, while RCTs were prioritised for Q3.
2. Studies based in the UK were prioritised.

In addition, the following criteria were applied after assessing the overall volume of evidence identified in the review:

3. For Q1 and Q3, only studies conducted in populations identified through screening for ASD were considered.

Publications not selected for extraction and data synthesis are clearly detailed in Table 15 below.

Table 15. Summary of publications included after review of full-text articles, and the question(s) each publication was identified as being relevant to

Study	Q1: Diagnostic stability over time	Q2: Screening accuracy of tools	Q3: Effectiveness of interventions
Achenie 2019**(39)**	No	Yes	No
Allison 2021**(26)**	Yes	Yes	No
Baduel 2017**(43)**	No	Yes	No
Baranek 2015**(49)**	No	No	Yes
Barbaro 2017**(28)**	Yes	No	No
Ben-Sasson 2013**(46)**	No	Yes	No
Canal-Bedia 2011(48)	No	Yes	No
Catino 2017**(42)**	No	Yes	No
Chlebowski 2013**(47)**	No	Yes	No
Dai 2020**(38)**	No	Yes	No
Guthrie 2013**(30)**	Yes	No	No
Jonsdottir 2021**(31)**	No	Yes	No
Jonsdottir 2020**(33)**	No	Yes	No
Kerub 2020**(34)**	No	Yes	No
Kondolot 2016**(44)**	No	Yes	No
Levy 2020**(74)**	No	Yes	No
Magan-Maganto 2020**(35)**	No	Yes	No
McPheeters 2016**(24)**	No	Yes	No
Mozolic-Staunton 2020**(37)**	No	Yes	No
Nygren 2012**(17)**	No	Yes	No
Oner 2020**(36)**	No	Yes	No
Petrocchi 2020**(75)**	No	Yes	No

Pierce 2019**(27)**	Yes	No	No
Robins 2014**(16)**	No	Yes	No
Sanchez-Garcia 2019**(76)**	No	Yes	No
Spjut Jansson 2016**(29)**	Yes	No	Yes
Suren 2019**(40)**	No	Yes	No
Topcu 2018**(41)**	No	Yes	No
Towle 2016**(77)**	No	Yes	No
Watson 2017**(50)**	No	No	Yes
Wiggins 2014**(45)**	No	Yes	No
Wieckowski 2021**(32)**	No	Yes	No
Whitehouse 2021**(51)**	No	No	Yes
Yuen 2018**(78)**	No	Yes	No

Publications excluded after review of full-text articles

Of the 254 publications included after the prioritisation of titles and abstracts, 220 were ultimately judged not to be relevant to this review. These publications, along with reasons for exclusion, are listed in Table 16.

Table 16. Publications excluded after review of full-text articles

Reference	Reason for exclusion
Q1	
Papavasiliou 2011	Not stability of ASD diagnosis
Thurm 2011	Not SR, or primary study
Whitehouse 2011	Stability of autistic traits
Worley 2011	Unable to retrieve
Anglim 2012	Conference abstract
Fisch 2012	Commentary
Dix 2015	No follow-up
Olsson 2015	Not primary study or SR
Yaari 2016	Stability of ASD risk
Bieleninik 2017	SR of studies in children already diagnosed with, or at risk of, ASD
Jiang 2017	Stability of symptoms
Solerdelcoll Arimany 2017	Conference abstract
Boone 2018	No follow-up
Kenny 2019	Not ASD diagnosis
Levante 2019	SR protocol
Rescorla 2019	Not stability of ASD diagnosis
Solari 2019	Older children
Dai 2020	Follow-up is not for diagnosis
Pender 2020	Not stability of ASD diagnosis
Q2	
Duyme 2010	No accuracy data
Jee 2010	No ASD diagnosis
Zwaigenbaum 2010	Review
Dall'Oglio 2010	No ASD diagnosis
Grossman 2010	No ASD diagnosis
Kapci 2010	No ASD diagnosis
Koyama 2010	Referred population
Moricke 2010	No accuracy data
Norris 2010	Referred population
Oosterling 2010	Referred population
Pandolfi 2010	No accuracy data
Baduel 2011	Non-English
Sheldrick 2011	Before 2010
Miller 2011	No ASD diagnosis

Bilenberg 2011	Conference abstract
Canals 2011	No ASD diagnosis
Charman 2011	Editorial
Goodwin 2011	Not peer-reviewed
Inada 2011	ASD diagnosis at later date
Matson 2011	Referred population
Nordenbaek 2011	ASD diagnosis known at screen
Beranova 2012	Non-English
Ellis 2012	No ASD diagnosis
Soares 2012	Narrative review
Ben-Sasson 2012	No ASD diagnosis
Dereu 2012	High-risk group
Fisch 2012	Commentary
Kozlowski 2012	Referred population
Macari 2012	ASD diagnosis at later date
Roux 2012	No ASD diagnosis
Guevara 2013	No ASD diagnosis
Matson 2013	Case-control
Plumb 2013	ASD diagnosis at later date
Dawson 2013	Not SR
de Bildt 2013	Referred population
de Wolff 2013	Outcome not ASD
Deconinck 2013	Not SR
Gardner 2013	No ASD diagnoses
Scarpa 2013	No ASD diagnoses
Smith 2013	Mixed group
Turner-Brown 2013	ASD diagnosis at later date
Girio-Herrera 2015	Not specific to ASD
Choueiri 2015	Case-control
Dix 2015	Referred population
Hampton 2015	Referred population
Hirota 2016	Conference abstract
Young 2016	Referred population
Velikonja 2017	Not ASD
Abbas 2018	Referred population
Adachi 2018	No ASD assessment
Kanne 2018	Referred population
Zahorodny 2018	Case-control
WHO protocol 2019	Trial protocol
Bong 2019	High risk population
Ahlers 2019	Referred population
Finlay-Jones 2019	Not a SR
Gulsrud 2019	No ASD diagnosis
Ibanez 2019	Protocol
Jang 2019	Referred population

Janvier 2019	No accuracy data
Lee 2019	ASD diagnosis at later date
Polo-DeSantos 2019	Conference abstract
Rescorla 2019	No ASD diagnosis
Roy 2019	ASD diagnosis at later date
Kara Uzun 2019	Non-English
Bussu 2020	ASD diagnosis at later date
Baadel 2020	No ASD diagnosis
Chung 2020	Case-control
Dickinson 2020	No ASD diagnosis
Jones 2020	Case-control
Lee 2020	Case-control
Levante 2020	ASD diagnosis at later date
McCarty 2020	Description of screening programme
Parikh 2020	ASD diagnosis at later date
Rescorla 2020	No accuracy data
Sacrey 2020	ASD diagnosis at later date
Carbone 2020	ASD diagnosis at later date
Bevan 2020	SR in Spanish-speaking populations
Meera 2020	High risk group
Geng 2020	No critical appraisal
Q3	
Aldred 2012	not a screened population
Barrett 2020	not a screened population
Bauminger-Zviely 2020	not a screened population
Beaudoin 2014	not a screened population
Bejarano-Martin 2020	not a screened population
Bent 2011	not a screened population
Bond 2016	not a screened population
Boyd 2018	not a screened population
Bradshaw 2019	not a screened population
Bradshaw 2015	not a screened population
Brignell 2018	not a screened population
Buie 2013	not a screened population
Byford 2015	not a screened population
Caron 2017	not a screened population
Carruthers 2020	not a screened population
Carter 2011	not a screened population
Chapin 2018	not a screened population
Dababnah 2016	not a screened population
Dawson 2012	not a screened population
De Korte 2020	not a screened population
Dean 2019	not a screened population
Engelstad 2020	not a screened population
Felzer-Kim 2020	not a screened population

Fletcher-Watson 2016	not a screened population
French 2018	not a screened population
Fuller 2020	not a screened population
Gengoux 2015	not a screened population
Geretsegger 2016	not a screened population
Green 2010	not a screened population
Green 2017	not a screened population
Griffin 2010	not a screened population
Gulsrud 2010	not a screened population
Hampton 2020	not a screened population
Hardan 2012	not a screened population
Hardan 2015	not a screened population
Harrop 2017	not a screened population
Holzinger 2019	not a screened population
Howard 2014	not a screened population
Hunter 2020	not a screened population
Ibanez 2019	protocol, no full results reported
Kaale 2014	not a screened population
Kaale 2012	not a screened population
Kaiser 2013	not a screened population
Kasari 2015	not a screened population
Kasari 2010	not a screened population
Kasari 2010	not a screened population
Kitzerow 2020	not a screened population
Kovshoff 2011	not a screened population
Kruizinga 2015	no post-screen intervention, screening alone was considered an intervention
Landa 2011	not a screened population
Lawton 2012	not a screened population
Lawton 2012	not a screened population
Leaf 2017	not a screened population
Lee 2020	not a screened population
Lemonnier 2012	not a screened population
Lieberman 2012	not a screened population
Mankad 2015	not a screened population
Marshall 2015	not a screened population
Mazon 2019	not a screened population
McConkey 2010	not a screened population
McKenzie 2020	not a screened population
Murza 2016	not a screened population
Nahmias 2019	not a screened population
Oono 2013	not a screened population
Palmer 2019	not a screened population
Parr 2010	not a screened population
Parsons 2019	not a screened population

Parsons 2017	not a screened population
Parsons 2017	not a screened population
Paul 2013	not a screened population
Pickles 2016	not a screened population
Poslawsky 2015	not a screened population
Reed 2010	not a screened population
Reichow 2018	not a screened population
Reitzel 2013	not a screened population
Roberts 2011	not a screened population
Rodgers 2020	not a screened population
Rogers 2012	not a screened population
Rogers 2019	not a screened population
Rollins 2020	not a screened population
Saleh 2020	not a screened population
Sandbank 2020	not a screened population
Schaaf 2014	not a screened population
Schreibman 2014	not a screened population
Shire 2017	not a screened population
Shire 2020	not a screened population
Smith 2015	not a screened population
Smith 2010	not a screened population
Sokhadze 2018	not a screened population
Solomon 2014	not a screened population
Stavropoulos 2013	not a screened population
Strain 2011	not a screened population
Strauss 2013	not a screened population
Su Maw 2018	not a screened population
Valeri 2020	not a screened population
Venker 2012	not a screened population
Vernon 2019	not a screened population
Virues-Ortega 2010	not a screened population
Virues-Ortega 2013	not a screened population
Vivanti 2019	not a screened population
Weitlauf 2020	not a screened population
Wetherby 2014	not a screened population
Whitehouse 2017	not a screened population
Williams 2012	not a screened population
Williams 2020	not a screened population
Wong 2013	not a screened population
Yatawara 2016	not a screened population
Yoder 2013	not a screened population
Yoder 2010	not a screened population
Yu 2020	not a screened population
Zheng 2020	not a screened population

Appendix 3 — Summary and appraisal of individual studies

Data Extraction

Table 17. Studies relevant to criterion 1 in a screened population

Author Country Study design	Population	Screening tool	Sample size at T1 and T2	Age (months) at T1 and T2. Length of follow-up (FU)	Diagnostic criteria at T1 and T2	Results: % retaining ASD diagnosis
Allison 2021(26) **England** **Prospective cohort**	Registered on CHSD in Luton, Bedfordshire and Cambridgeshire	Q-CHAT	T1: 121 T2: 81	T1: ~24 [median] T2: ≥48 FU: NR (approx. 24 months))	T1: Experienced, psychologist(s) performed the ADOS, ADI-R, MSEL, VABS. ICD-10 criteria. T2: as above	100% (66.4, 100)** retained possible autism diagnosis.
Pierce 2019(27) **America** **Prospective cohort**	75% children identified as "at-risk" from a screened population, 25% referred population	CSBS IT checklist	T1: 1269 T2: 1269	T1: 19 [mean] 12 to 36 [range] T2: 40 [mean] FU: 20 [mean], 17 [median]	T1: Highly experienced, licensed psychologists with PhD degrees performed diagnostic and psychometric tests, including the ADOS-2 (module T, 1, or 2 as appropriate), Mullen Scales of Early Learning and Vineland Adaptive Behaviour Scales T2: as above	84% (80, 87) retained ASD dx Change to ASD: 47% ASD features, 24% DD, 16% LD, 4% TD "Only 7 toddlers (1.8%) initially considered to have ASD transitioned into a final diagnosis of typical development. Diagnostic stability of ASD within the youngest age band (12-13 months) was lowest at 0.50 (95%CI, 0.32-0.69) but increased to 0.79 by 14 months and 0.83 by 16

						months (age bands of 12 vs 14 and 16 months; odds ratio, 4.25; 95%CI, 1.59-11.74). A total of 105 toddlers (23.8%) were not designated as having ASD at their first visit but were identified at a later visit."
Barbaro 2017(28) **Australia** **Prospective cohort**	Children identified as "at-risk" from a screened population	Failing 3 of 5 behavioural items from the SACS	T1: 99 T2: 77	T1: 24 [mean] T2: 48 [mean] FU: 24 [mean]	T1: Developmental history, children's MCH records, ADOS-G, the Autism Diagnostic Interview–Revised, results from various questionnaires (First Year Inventory, Communication and Symbolic Behaviour Scales–Infant Toddler Checklist, Early Development Interview, Checklist for Autism in Toddlers-23), expert clinical judgment by both authors. T2: as above but not ADI-R	71.9% (53.2, 86.2)** retained ASD dx. "diagnoses of ASD by 24 months of age are stable across time, with nearly 86% of children retaining an ASD diagnosis between 2 and 4 years of age"
Spjut Jansson 2016(29) **Sweden** **Prospective cohort**	Children identified as "at-risk" from routine ASD population screening	NR	T1: 71 T2: 100	T1: 36 [mean] T2: approx. 96 FU: approx. 60	T1: Multidisciplinary assessment, including cognitive/intellectual tests, ADOS-G and DISCO (for 72% of the children). Experienced professionals. T2: As above plus ADI	93% (84.3, 97.7)* retained ASD dx.
Guthrie 2013(30) **Australia** **Prospective cohort**	Two-step screened population	First step: CSCB IT or parental concern. Second step: CSBS red	T1: Unclear T2: 82	T1: 19 [mean], 15 to 24 [range] T2: 37 [mean], 30 to 46 [range] FU 16 [mean]	T1: ADOS-T and evaluation by 2 clinicians T2: as above	100% (93.6, 100)* retained ASD dx. Change to ASD: 21% deferred dx. 0% ASD ruled-out

Screened	flags for ASD using SORF.		Short-term stability was documented for children diagnosed at 19 months on average. Findings highlight utility of the ADOS-T in making early diagnoses.

A, attrition; ADI-R, Autism Diagnostic Interview– Revised; C, confounding; CHAT-23, Checklist for Autism in Toddlers-23; CHSD Child Health Surveillance Database; CSBS, Communication and Symbolic Behaviour Scales ; CSBS IT, Communication and Symbolic Behaviour Scales Infant-Toddler Checklist; DA, diagnostic assessment; DD, developmental delay; dx, diagnosis; EDI, Early Development Interview; FYI, First Year Inventory; LD, language delay; MSEL, Mullen Scales of Early Learning; NR, not reported; P, participants; Q-CHAT, Quantitative Checklist for Autism in Toddlers; SACS, Social Attention and Communication Study; T1, time 1; T2, time 2; TD, typically developing; SORF, Systematic Observation of Red Flags; VABS, Vineland Adaptive Behavior Scales

*Blind, Were individuals conducting the diagnostic assessment at T2 blinded to the details and/or findings of the diagnostic assessment at T1?; **95% confidence intervals calculated by review authors

Table 18. Studies relevant to criterion 1 not in a screened population

Author Country Study design	Population	Sample size at T1 and T2	Age (months) at T1, T2 Length of follow-up (FU)	Diagnostic criteria at T1 and T2	% retaining ASD diagnosis
McDonald 2020(61) **Australia** **Prospective cohort**	Referred population	T1: 145 T2: 54	T1: 55.3 [mean], 5.3 [SD] T2: 120 [mean], 28[SD] FU: 64 [mean], 29[SD]	T1: Developmental and medical history, CARS, CBCL. DSM-V was used to confirm diagnoses, and the ADOS-2 was used where there was any uncertainty T2: ADOS-2	100%
Anglim 2020(60) **Ireland** **Prospective cohort**	Referred population diagnosed by psychiatrist opinion, DSM-V, ICD-10, ADOS, and/or DISCO	T1: NR T2: 29	T1: 46 [mean] , 30 to 76 [range] T2: 58 [mean], 48 to 72 [range] FU: 24 to 60 [range]	T1: DISCO, psychiatrist opinion, DSM-5, ICD-10, ADOS T2: DISCO	"No statistically significant difference" in number of diagnosis at T1 and T2
Pellicano 2012(62) **Australia**	Referred population diagnosed according to	T1: 45 T2: 37	T1: 67 [mean], 48 to 84 [range]	T1: ADI-R and DSM-IV T2: ADOS-G and SCQ	81%

Prospective cohort	DSM-V and ADI-R		T2: 100 [mean], 11 [SD] FU: 33 [mean]		
Soke 2011(63) **USA** **Prospective cohort**	Referred population	T1: 36 T2: 28	T1: 33 [mean], 4 [SD] T2: 55 [mean] FU: 22	T1: ADI-R, ADOS-G, Mullen scale of early learning by 2 clinical psychologists T2: ADI-R	67% of children meeting ADI-R cut-off for autism at T1 also did at T2

Table 19 Systematic reviews relevant to criterions 4 and 5

Author, year	Focus	Inclusion criteria	Databases Search period Search limits	Quality appraisal	No. includes relevant to Q2* Screening tools evaluated	Author's conclusions
Levy 2020(74)	Universal screening of children for ASD in primary care	P: Clinical definition of ASD, Children 0-12yrs undergoing screening in primary care (or similar). Not preselected or diagnosed with ASD. I: Any approaches to screen for ASD C: no, or alternative, screening RS: O: accuracy, timing of referral or diagnosis D: >100 participants in countries rated as "very high human development"	Medline through OVID, PsycINFO, ERIC, CINAHL January 2000 to June 2018 English language only	Tool developed through key stakeholder appraisal of current tools and processes: 7 studies good quality; 19 fair quality; 1 poor quality	27 CHAT, M-CHAT, PDQ-1, SCQ, PDDSTII	Moderate to high PPVs for children aged 16 – 40 months old. Limited evidence on sensitivity, specificity, and NPVs in the literature.
Petrocchi 2020(75)	Level 1 and 2 screening measures to detect early signs of risk of ASD in children < 24 months of age	P: children under 24 months I: level 1 and level 2 screening measures of ASD; questionnaires, interviews and observation procedures only D: validation studies, standardization of measures, cross-cultural comparisons, longitudinal, or follow-up studies; published papers in peer- reviewed journals; papers written in English;	PsychINFO, PBSC, CINAHL, Scopus, ERIC, Google Scholar, and Pubmed (MEDLINE) 1990 and October 2019 English only, no dissertation theses or	COSMIN checklist. Across all included studies, criterion validity (most relevant to the accuracy studies) rated fair or poor, with	9 CESDD, CHAT, ESAT, FYI, JA-OBS, M-CHAT, M-CHAT-R/F, Q-CHAT, SEEK, YACHT-18	M-CHAT, FYI and Q-CHAT have promise as screening tools for ASD

			conference papers	some exceptions		
Sanchez-Garcia 2019(76)	To evaluate accuracy of different screening tools for ASD	P: general population of children aged 14 to 36 months I: screening and diagnosis of ASD and other developmental disorders (level 1 screening) O: sufficient data to construct a 2 × 2 contingency table D: not rated as low quality in quality assessment	CINHAL, ERIC, PsycINFO, PubMed, WoS January 1992 and April 2015 English language	QUADAS-2. Ratings are unclear as the table of risk of bias results and accompanying text are inconsistent.	14 MCHAT, JOBS, CHAT, PEDS, PATH, MCHAT-JV, YALE SCREENER, STAT, SACS, CESDD, ITC, MCHATR/F, YACHT-18	Pooled estimates (95% CI) of sensitivity and specificity were 0.72 (0.61, 0.81) and 0.98 (0.97, 0.99), respectively, indicating ability to adequately screen for ASD.
Yuen 2018(78)	Accuracy of M-CHAT	P: children screened for ASD; sample population was not selected on the basis of any medical condition other than developmental delays (e.g. low birthweight, Down syndrome) I: original M-CHAT, screened more than 90% of children using the English M-CHAT O: such that sensitivity, specificity, and PPV could be calculated	MEDLINE, PsycINFO, CINAHL, Embase Jan 2001 to May 2016 English language	QUADAS-2. The 1 study relevant to the 2021 update review was rated as high concern for risk of bias for reference standard domain, and applicability in index test and reference standard domains.	1 M-CHAT	Lack of evidence on the performance of M-CHAT in low-risk population.
Towle 2016(77)	Screening instruments for children < 18 months	P: children < 18 months old I: any screening tool C: NR RS: NR O: receiver operator characteristics (ROC) or performance measures of sensitivity and specificity and/or positive predictive and negative predictive values	Included PsycInfo and Medline, and others not specified. Dates NR English language	Adaptation of QUADAS-2. No reporting of quality results.	3 ITC, FYI	ITC and FYI “demonstrated some success” as screening tools for ASD. Showed better success at identifying multiple

		D: prospective, published in peer-review journal				conditions (not just ASD)
McPheeters 2016(24)	Review evidence on benefits and harms of ASD screening	P: children 12 - 36 months old in populations without suspected ASD or developmental delay, not already diagnosed with ASD I: Any conducted in primary care C: NR RS: NR O: timing of referral and diagnosis and timing of access to intervention D: primary studies with at least 2 participants, and good quality systematic reviews	MEDLINE (via PubMed), PsycINFO, ERIC, CINAHL. Reference lists of included studies and relevant reviews 2000 – Aug 2014 English language	"quality criteria based on the USPSTF methods."	5 M-CHAT-R/F, M-CHAT/F, FYI	The M-CHAT(-R/F) was the most commonly evaluated tool, with PPVs of 0.48 for the age group of interest

* Number of includes in each SR that meet criteria for the 2021 update review, i.e. published since 2010, set in OECD and EEA countries, screen children for whom no concerns of ASD have been raised by parents/carers/ professionals, diagnosis of ASD conducted as soon after screening as possible.

C, comparator; CINAHL, Cumulative Index of Nursing and Allied Health Literature; COSMIN, COnsus-based Standards for the selection of health Measurement INstrument; D, design; ERIC, Educational Resources Informational Clearinghouse; I, intervention; NR, not reported; O, outcomes; P, population; PBSC, the Psychology and Behavioral Sciences Collection; PDDSTII, Pervasive Developmental Disorders Screening Test II; PPV, positive predictive value; PsycINFO, psychology and psychiatry literature; RS, reference standard; SCQ, Social Communication Questionnaire; USPSTF, US Preventative Services Taskforce; WoS, Web of Science

Table 20. Studies relevant to criterions 4 and 5

Author Year Country Design Aim Study ASD prevalence	Sample	Screening setting and tool(s)	Reference standard	Follow-up of screen negatives	Results (95%CI)	Proportion of false positive screening results with atypical development
Allison 2021(26) **UK** **Prospective** **Aim: Evaluate the Q-CHAT** **Prevalence: 0.98%(0.45%, 2.16%)**	Source: children registered on the Child Health Surveillance Database at the primary care trusts (PCT) Excludes: NR Total sample: 3770 Age (months): intended 18-30	Setting: Postal questionnaire Personnel: carer self-complete Training: NR Screening tool: Q-CHAT [English] Threshold: Multiple, including ≥39 => referred	Setting and timing: NR. Timing assumed as soon as possible. Diagnostic criteria: ICD-10. Tools/measures: Consensus diagnosis as possible autism or autism spectrum (if they met the ICD-10 criteria). ADOS-G, ADI-R, MSEL, VABS. Personnel: Experienced research psychologist(s) and trained research assistant.	Children re-screened ≥ 48 months using CAST. Those >15, and any where referrals for number of reasons, including autism, invited for diagnostic evaluation.	≥39: PPV 0.17 (0.08, 0.31)	15% Language delay, developmental delay, other atypical
Jonsdottir 2020, 2021(31, 33) **Iceland,** **Prospective** **Aim: Evaluate early detection program for ASD within well-child care in primary healthcare centers (PHCs)** **Prevalence: 0.8% (0.45, 1.15)**	Source: 9 primary healthcare centres in the capital area of Reykjavik, Iceland (randomly selected from among the 17 centres in that area) Excludes: children with previous referrals for ASD. Total sample: 1586 Age (months): mean 31.66, SD 1.72	Setting: 30-month well-child visit. Personnel: Carer self-complete; FUI - first author over phone. Training: Professionals offered half-day course on ASD. Screening tools: M-CHAT-R/F [Icelandic] Thresholds: M-CHAT-R >2 => FUI.	Setting and timing: A tertiary institution that receives referrals. Timing not reported, assumed as soon as possible Diagnostic criteria: ICD-10. Tools/measures: Physical and neurological examination, ADOS-2, Carer interview. Personnel: Paediatrician, psychologist, social worker. .	Identified any ASD diagnoses up to a maximum of 2 years after screening.	Sens: 0.62 (0.44, 0.80) Spec: 0.99 (0.99, 1.00) PPV: 0.72 (0.51, 0.88) NPV: 0.99 (0.99, 1.00)	86% Non-ASD DSM diagnoses (not defined further)

		FUI ≥ 2 => refer				
Wieckowski 2021(32) **US** **Prospective** **Aim: Examine early and repeated screening** **Prevalence: 2.35%**	Source: Well-child visits Excludes: screening not completed, first screen outside screening age range Total sample: 5784 Age: Initial screen - 12, 15, 18; Re-screens - 18, 24, 36	Setting: 12-, 15- or 18-month well-child visits Personnel: NR Training: NR Screening tools: FYI, ITC, M-CHAT-R/F Thresholds: Positive on either tool (if multiple tools used).Cut-offs NR.	Setting and timing: University clinic or paediatric office. Timing not reported but assumed to be as soon as possible. Diagnostic criteria: ICD-10. Tools/measures: ADOS-2, TASI or ADI-R, medical, developmental, family history. Personnel: Individuals supervised by supervised by a licensed psychologist, certified school psychologist, or developmental paediatrician.	Only those for whom a concern had been raised.	**Single screen at** 12 months PPV 0.22 (0.14, 0.32) Sens 0.64 (0.48, 0.81) Spec 0.95 (0.93, 0.96) NPV 0.991 15 months PPV 0.17 (0.09, 0.27) Sens 0.72 (0.52, 0.93) Spec 0.94 (0.92, 0.95) NPV 0.995 18 months PPV 0.42 (0.34, 0.51) Sens 0.74 (0.64, 0.84) Spec 0.97 (0.97, 0.98) NPV 0.993 >1 screen from 12 months PPV 0.25 (0.17, 0.35) Sens 0.81 (0.67, 0.95) Spec 0.94 (0.93, 0.95) NPV 0.995 15 months PPV 0.19 (0.11, 0.29) Sens 0.83 (0.66, 1.00)	48% - 81% depending on age of screen. Developmental disability (not defined further)

					Spec 0.94 (0.92, 0.98) NPV 0.993 18 months PPV 0.44 (0.35, 0.52) Sens 0.82 (0.74, 0.91) Spec 0.97 (0.97, 0.98) NPV 0.995	
Kerub 2020(34) **Israel** **Prospective** **Aim: Compare GDS and M-CHAT/F** **Prevalence: 0.63%**	Source: During routine monitoring at 35 randomly selected government-funded clinics. Excludes: NR Total sample: 1591 Age: mean 21.30, SD 3.45	Setting: 18 and 36 months routine assessments at Maternal child-health centres. Personnel: Nurses at the clinics completed GDS & M-CHAT. FUI = ASD specialist nurse. Training: Nurses had 1 day workshop on ASD & specific to M-CHAT. All had experience with GDS. Screening tools: Global Developmental Screening (GDS), M-CHAT/F. [Hebrew] Thresholds: GDS ≥ 1 => follow-up or refer. M-CHAT >7 => refer. 3-7 => FUI. FUI ≥2 => refer.	Setting and timing: Soroka University Medical Center. Timing not reported, assumed as soon as possible. Diagnostic criteria: DSM-V. Tools/measures: NR Personnel: Child psychiatrist/neurologist.	Yes. Reviewed medical records of those screened negative (10 months later) to identify any false negatives	M-CHAT/F Sens 0.7 (0.35, 0.93) Spec 0.98 (0.97, 0.99) PPV 0.20 (0.08, 0.37) GDS Sens 0.5 (0.19, 0.81) Spec 0.998 (0.992, 0.999) M-CHAT/F plus GDS Sens 0.7 (0.35, 0.93) Spec 0.968 (0.96, 0.97)	M-CHAT/F: 68% GDS: 36% Delays (not defined further).
Magan-Maganto 2020(35)	Source: routine 18 and 24 months “Well Baby Check-	Setting: Routine 24 month check-ups.	Setting and timing: University of Salamanca, and psychiatric units of the	Yes. For children who	All ages Sens 0.79 (0.54,0.93)	71% - 90% depending on age

Spain **Prospective** **Aim: Validate Spanish version of the M-CHAT-R/F in the public health system** **Prevalence: 0.3% in 14-22 month age grp; 0.26% in 23-36 month grp;** 0.29% in 14-36 month grp	up Program" screenings. Excludes: incomplete FUI or evaluation due to problems of communication with the families Total sample: 6585 Age: 14-22 month grp - Mean 18.22, SD 0.72; 23-36 month grp - mean 24.47, SD 1.23	Personnel: FUI: pediatricians and pediatric nurses. Training: Offered on the screening programme. Screening tool: M-CHAT-R/F [Spanish] Thresholds: >7 => refer 3-7 => FUI. ≥2 after FUI => refer	NHS. Timing not reported, assumed as soon as possible. Diagnostic criteria: DSM-V. Tools/measures: Clinical history, Merril-Palmer Revised Scales, Leiter, Vineland Scales, ADOS-G module 1 and ADOS-2 module T and 1. Personnel: Trained and experienced professionals	screened negative, any subsequent ASD diagnoses were identified from referral centres, with all other children who screened negative assumed to be true negatives	Spec 0.99 (0.99,0.99) 14-22 months Sens 0.82 (0.48–0.97) Spec 0.99 (0.99,0.99) 23-36 months Sens 0.75 (0.36–0.96) Spec 0.99 (0.99,0.99)	Disorders of language or global development, diagnoses of unspecified neurodevelopmental or systemic disease, delays of language and psychomotor skills
Oner 2020(36) **Turkey** **Prospective** **Aim: Evaluate feasibility of Turkish version of the M-CHAT-R/F in an urban low risk population** **Prevalence: 0.8%; 95% CI 0.063–1.05%**	Source: Family Healthcare centres. 75 FHCs sites comprising 148 practitioners who volunteered. Excludes: NR Total sample: 6712 Age: mean 26.75, SD 5.76	Setting: Family Healthcare centres Personnel: M-CHAT-R read aloud to carers by participating practitioners. FUI conducted by psychologists. Training: Practitioners were trained for use of M-CHAT-R/F. Screening tool: M-CHAT-R/F [Turkish] Threshold: >7 => FUI. ≥2 after FUI => refer	Setting and timing: NR. Timing assumed as soon as possible. Diagnostic criteria: DSM-V. Tools/measures: "All available information" ADOS-2, Denver Developmental Screening-II. Personnel: Study author, research certified for ADOS-2 use.	No.	M-CHAT-R Sens 1.00 (0.94, 1.0) Spec 0.91 (0.90, 0.92) PPV 0.09 (0.07, 0.11) NPV 1 (0.999, 1.0) M-CHAT-R/F (calculated by 2021 review authors) Sens 1.00 (0.97, 1.00) PPV 0.26 (0.20,0.32)	41% Developmental delay
Mozolic-Staunton 2020(37) **Australia**	Source: children in the general population who were attending either a routine visit	Setting: Routine 'well child' checks at 12, 18, 24 and 42 months, and childcare centres (aged 12 - 48 months).	Setting and timing: Southern Cross University Health, Wellbeing Clinic, Gold Coast and Olga Tennison Autism research	No FU of screen negatives.	SACS-R PPV 0.83 (0.78, 0.88) PEDS	1 case of developmental delay, 1 case with sensory processing concerns

Retrospective analyses of 2 prospective cohort studies **Aim: Compare SACS-R and PEDS** **Prevalence: 1.49%**	at an MCH centre or were enrolled at a participating early childhood education and care centre. Excludes: NR Total sample: 13417 Age: range 12-48	Personnel: Educators and nurses who had training implemented the SACS-R, caregivers completed the PEDS. Training: to monitor general development Screening tools: SACS-R, PEDS [English] Thresholds: SACS: 3 items => high risk PEDS: Path ASD = ≥3, Path A = 2, Path B = 1 => high risk	Centre, Melbourne. Typically within 2 months. Diagnostic criteria: NR Tools/measures: BSID, ADOS 2, ADI-R, clinical judgement. Personnel: Paediatric health professionals		PPV 0.88 (0.71, 0.98)	
Dai 2020(38) **US** **Retrospective analysis of prospective study** **Aim: Evaluate utility of rescreening at 24 months, after a negative 18-month screening** **Prevalence: 1.03%**	Source: recruited in a previous study from community Excludes: If used Spanish M-CHAT versions Total sample: 19685 Age: 18	Setting: 18-month pediatric well-child care visit. Personnel: M-CHAT(-R) completed by care-giver, FUI NR Training: NR Screening tool: M-CHAT/F or M-CHAT-R/F [English] Threshold: NR.	Setting and timing: University clinics and paediatricians office or family home. Diagnostic criteria: DSM-IV. Tools/measures: Demographic information, MSEL, VABS, ADOS-2 Toddler Module, ADOS Module 1 and 2, CARS(2). Personnel: Clinical psychologist or a developmental-behavioral pediatrician	No. Re-screened those who were screen-negative at 18 months. No screen-negatives had a diagnostic evaluation, unless they subsequently screened positive.	NR Calculated by 2021 review authors Single screen at 18 months PPV 0.52 (0.47, 0.57) Negatives rescreened at 24 months PPV 0.50 (0.27, 0.73)	NR
Achenie 2019(39) **US**	Source: toddlers screened in metropolitan Atlanta (Georgia State University) or	Setting: 18- and 24-month well-child care visits.	Setting and timing: NR. Timing assumed as soon as possible	Partial. Random sample of screen-negatives	Comparable to M-CHAT-R/F.	NR

Retrospective analysis of prospectively collected data **Aim: Examine a potential alternative to assessment barriers by using machine learning** **Prevalence: NR – assume same as Robins 2014 (0.77%)**	Connecticut (University of Connecticut). Excludes: Participants with missing responses to M-CHAT Total sample: 14995 Age: range 16-30	Personnel: Carers self-complete. Training: NR Screening tool: M-CHAT-R [English] Thresholds: >2 =>FUI. >2 FUI => refer.	Diagnostic criteria: DSM-IV-TR. Tools/measures: ADOS, CARS-2, TASI, MSEL, VABS, BASC, and developmental history Personnel: Psychologist/developmental pediatrician	had diagnostic evaluation	More results available.	
Suren 2019(40) **Norway** **Prospective** **Aim: Evaluate performance of early population based screening for ASDs** **Prevalence: 0.70%**	Source: Norwegian Mother and Child Cohort Study Excludes: Lack of consent Total sample: 58520 Age: mean 36	Setting: Postal questionnaires. Personnel: Completed by carers. Training: NR Screening tool: SCQ [Norwegian] Threshold: A score of ≥12 on the 33 non-verbal SCQ items	Setting and timing: NR. Timing assumed as soon as possible. Diagnostic criteria: DSM-IV-TR or ICD-10. Tools/measures: ADOS, ADI-R. Personnel: NR.	Yes. Random sample of age-matched controls. False negative children (those with ASD who were not screen positive) were determined by checking medical records at later time-point.	SCQ total ≥15 Sens 0.20 (0.16,0.24) Spec0.99 (0.99,0.99) PPV 0.09 (0.07, 0.11) NPV 0.99 (0.99, 1) SCQ total ≥11 Sens 0.42 (0.37,0.47) Spec 0.89 (0.89, 0.90) PPV 0.03 (0.02, 0.03) NPV 1 (1,1) SCQ total ≥12 Sens 0.25 (0.20,0.29) Spec 0.99 (0.99, 0.99)	NR

					PPV 0.16 (0.13, 0.19) NPV 1 (0.99,1) Results also given by whether child had phrased speech or no.	
Topcu 2018(41) **Turkey** **Prospective** **Aim: Compare TIDOS and M-CHAT** **Prevalence: 0.98%**	Source: children presenting for well-child visits at the Social Pediatrics Department of Ankara University Excludes: NR Total sample: 511 Age: range 16-38	Setting: Well-child visit at the Social Pediatrics Department of Ankara University. Personnel: M-CHAT completed by carers, FUI by study author. Training: NR Screening tools: TIDOS, M-CHAT [Turkish] Thresholds: M-CHAT/F: ≥2 of 7 critical items or ≥3 of 23 items were positive, so => refer TIDOS: refer if one of the 3 parameters scored ≥ 1	Setting and timing: Child Psychiatry Clinic. Clinical evaluation was performed within 2 weeks of the initial screening positive children, within 3–9 months for screening randomly selected negative children. Diagnostic criteria: DSM-V Tools/measures: NR. Personnel: Child psychiatrist.	Yes. Random sample of 25 children who screened negative on M-CHAT-R/F and TIDOS.	M-CHAT/F Sens 0.60 (0.15, 0.95) Spec 0.97 (0.95, 0.99) PPV 0.18 (0.04, 0.46) NPV 0.995 (0.98, 1.0) TIDOS Sens 0.80 (0.28, 0.99) Spec 0.998 (0.989, 0.999) PPV 0.80 (0.28, 0.99) NPV 0.998 (0.989, 0.999) M-CHAT/F plus TIDOS Sens 1.00 (0.48, 1.00) Spec 0.90 (0.88, 0.93) PPV 0.10 (0.03, 0.21) NPV 1.00 (0.99, 1.0)	NR

Catino 2017(42) **Italy** **Prospective** **Aim: Validate Italian version of ASQ** **Prevalence: 0.39%**	Source: 15 kindergarten schools of Rome. Excludes: refusal to participate, incorrect questionnaires Total sample: 514 Age: Younger group Mean 42.65, SD 1.82; Older group Mean 48.08, SD 2.62	Setting: Kindergarten school. Personnel: Completed by carers. Training: NR Screening tool: ASQ-3 [Italian] Threshold: Score in the clinical range in one, or more than one domain	Setting and timing: NR. Timing unclear. Diagnostic criteria: NR Tools/measures: Neuropsychiatric evaluation comprehensive neuropsychiatric evaluation (cognitive, neuropsychological, and psychopathological). Personnel: Clinician of the neuropsychiatric service.	No	For ASD PPV 0.08 (0.01, 0.25)	68%-88% depending on age Disorders of language or developmental coordination. Intellectual disability.
Baduel 2017(43) **France** **Prospective** **Aim: Validate French version of M-CHAT to provide decision rules regarding a child risk status for French primary care providers** **Prevalence: 1.47%**	Source: 24 month-old children living in the Midi-Pyrénées area. Excludes: high risk: (1) prior diagnosis of ASD, (2) preterm birth, and (3) severe sensory or motor impairments. Total sample: 1227 Age: mean 24	Setting: Well-child visit or daycare centre. Personnel: Carers completed. Training: Professionals had 2-hour course on ASD Screening tool: M-CHAT [French] Thresholds: any 3 M-CHAT items or 2 of the 6 critical items	Setting and timing: laboratory or at the child's daycare centre. Timing NR, assumed as soon as possible. Diagnostic criteria: NR. Tools/measures: 2-stage process 1st: ADOS-G, PEPR, VABS. If reached ADOS-G threshold, referred to independent team to confirm diagnosis. Personnel: One of the authors, all trained in the use and scoring of the ADOS-G in young children.	Partial. Those screen-negative at 24 months followed-up at 30 and 36 months. If then screen positive, they were referred for diagnostic assessment. As were any children who screened negative, but physicians had concerns.	Sens 0.67 (0.41, 0.86) Spec 0.99 (0.98, 0.99) PPV 0.6 (0.36, 0.81) NPV 0.99 (0.99, 0.99)	100% Delays (not defined further)
Kondolot 2016(44) **Turkey** **Prospective** **Aim: Adapt the M-CHAT to healthy**	Source: Healthy toddlers aged 18–30 months from Kayseri, Turkey between June 2011 and June 2012. Excludes: diagnosed with any	Setting: Usual 18-30 month screen at family health centres. Personnel: Students by face-to-face interview. Training: yes	Setting and timing: Child Psychiatry Clinic. Within 2 months of the initial M-CHAT screening for screen positive children and within 6-12 months for randomly selected screen negative children.	Yes. Random sample (n=48) screened negative evaluated (6-12 months	PPV: 0.12 (0.01, 0.36) Sens: 1.00 (0.16, 1.00) Spec: 0.76 (0.64, 0.86)	7% Developmental delays

18–30-month-old toddlers in Turkey **Prevalence: 0.10%**	neurodevelopmental disease or ASD before, those who had a severe sensory or motor disability, or whose carers did not want to participate in the study Total sample: 2021 Age: mean 23	Screening tool: M-CHAT [Turkish] Thresholds: Any 2 of 6 critical items or any 3 of 23 items were positive	Diagnostic criteria: DSM-IV-TR. Tools/measures: CARS Personnel: Child psychiatrist.	after screening)		
Wiggins 2014(45) **US** **Prospective** **Aim: "Compare agreement between ASD diagnosis and outcome of the M-CHAT and PEDS and examine specific concerns noted for toddlers who screened negative on the M-CHAT or PEDS but were later diagnosed with ASD"** **Prevalence: 0.75%**	Source: Children attending well-child visits Excludes: Unclear Total sample: 3980 Age: mean 21.1, range 15.2–27.0	Setting: Routine 18- or 24-month well-child visits. Personnel: Completed by physician office staff. Training: NR Screening tools: M-CHAT, PEDS [English] Thresholds: M-CHAT, any 2 of 6 critical items or any 3 of 23 items were positive. PEDS, "if predictive concerns are noted"	Setting and timing: Clinic (2 evaluations at home). Timing not reported, assumed as soon as possible. Diagnostic criteria: NR. Tools/measures: ADI-R, ADOS, CARS, MSEL, Vineland-II, developmental and medical history questionnaire. Personnel: Experienced clinicians (blind to M-CHAT/F and PEDS score)	No. (Only screen negative children for whom clinicians had raised concern were followed)	M-CHAT/F PPV 0.61 (0.45, 0.76) PEDS Path A PPV 0.55 (0.39, 0.71) PEDS Path B PPV 0.75 (0.35, 0.97) PEDS ASD PPV 0.59 (0.39, 0.76) [Very few of the PEDS positive were followed-up]	88% Global development disorder
Robins 2014(16) **US** **Prospective**	Source: toddlers screened in metropolitan Atlanta (Georgia State University) or Connecticut (University of Connecticut).	Setting: 18- and 24-month well-child care visits. Personnel: Carers completed. Training: NR	Setting and timing: NR. Timing assumed as soon as possible. Diagnostic criteria: DSM-IV-TR.	Partial. Random sample who screened negative completed Screening Tool for	M-CHAT-R/F ≥3 Sens 0.68 (0.58, 0.75) Spec 0.99 (0.99, 0.99) PPV 0.51 (0.43, 0.59)	90% Delays and developmental concerns, with no diagnosis.

Aim: Validate the M-CHAT-R/F in a low-risk sample **Prevalence: 0.77%**	Excludes: incomplete data, insufficient English proficiency, previous ASD diagnosis, a medical condition that precluded evaluation, withdrawal from the study, or being outside the study's screening age. Total sample: 16071 Age: mean 20.94, SD 3.30	Screening tool: M-CHAT-R/F [English] Thresholds: positive on 3 or more items. And other threshold reported based on this study	Tools/measures: "all available information and … clinical judgment". Personnel: "Licensed psychologist/developmental pediatrician supervising a graduate student and research assistants."	Austim in Two-Year Olds (STAT) tool. If then positive offered clinical evaluation	NPV 0.997 (0.996, 0.998) M-CHAT-R/F ≥2 Sens 0.85 (0.79, 0.92) Spec 0.99 (0.99, 0.99) PPV 0.47 (0.41, 0.54) NPV 0.999 (0.998, 0.999)	
Ben-Sasson 2013(46) **Israel** **Prospective** **Aim: Evaluate combining sensory-regulatory markers with social–communication markers in 12-month ASD screening** **Prevalence: 0.80%**	Source: 4 public daycare organizations in Israel. Excludes: families who lacked Hebrew proficiency and families where the child age could not be determined accurately Total sample: 613 Age: mean 12.56	Setting: Questionnaires mailed to home. Personnel: carers to complete. Training: NR Screening tool: FYI [Hebrew] Threshold: 94th percentile cut-off for the social domain only, or also the 88th percentile cut-off for the sensory domain.	Setting and timing: NR Diagnostic criteria: NR Tools/measures: AOSI, MSEL. Personnel: "clinician with expertise in early child development"	Yes. 60 screen-negatives followed-up.	Sens 0.60 (0.15, 0.95) Spec 0.753 (0.64, 0.84)	68% Delays of developmental and language.
Chlebowski 2013(47) **US** **Prospective**	Source: children who participated in the large-scale M-CHAT screening studies conducted at the University of	Setting: 18- and 24-month wellchild visits at pediatric offices Personnel: NR	Setting and timing: NR. Timing assumed as soon as possible after screening. Diagnostic criteria: DSM-IV.	Partial. (Only those who screen positive on other tools or "red-flagged"	PPV 0.54 (0.46, 0.61)	95% Non-ASD diagnoses and developmental concerns, with no diagnosis.

	Connecticut and Georgia State University Excludes: screened by an early intervention provider, screened as part of an autism sibling study, or if they were self-referred by their caregivers with autism-related concerns. Received an ASD diagnosis before being screened with the M-CHAT, had a severe sensory or motor disability (eg, blindness or deafness) that prevented them from completing study assessments, or if the child's caregivers were not fluent in English or Spanish. Total sample: 18989 Age: mean 20.41, SD 3.1	Training: NR Screening tool: M-CHAT/F [English or Spanish] Thresholds: screening positive on 2 of 6 critical items or on 3 of 23 items overall on both the M-CHAT and M-CHAT/F.	Tools/measures: ADOS, ADI-R, MSEL, VABS, CARS. Personnel: Diagnosis made by clinical judgement "licensed clinical psychologist or developmental pediatrician and a psychology doctoral student."	by paediatrician)		
Nygren 2012(17) **Sweden** **Prospective** **Aim: Evaluate psychometric properties of M-**	Source: children from Gothenburg (Sweden) coming for their 2.5-year-old check-up, and all other children (regardless of age) raising any suspicion of ASD	Setting: Routine 30 month check-ups. Personnel: Nurses completed. Training: yes	Setting and timing: Neuropsychiatric specialist clinic. Timing NR, assumed as soon as possible. Diagnostic criteria: DSM-IV and ICD-10	Partial. (only those where a concern raised)	M-CHAT/F alone Sens 0.77 (0.61, 0.88) PPV 0.92 (0.78, 0.98) JA-OBS alone	50% Language disorder

CHAT and JA-OBS **Prevalence: 1.20%**	Excludes: NR Total sample: 3999 Age: intended to be 30	Screening tools: M-CHAT, JA-OBS [Swedish] Thresholds: either a (i) definitive failure on the M-CHAT or (ii) failure on 2 or more of the items of the JA-OBS or, (iii) both	Tools/measures: VABS, ADOS, DISCO, Language assessments, 1-h observation of the child at preschool. Personnel: "experienced neuropsychiatrists, neuropediatricians (4 in total) and neuropsychologists (2 in total) with expertise in autism."		Sens 0.96 (0.72, 0.95) PPV 0.92 (0.80, 0.98) M-CHAT/F plus JA-OBS Sens 0.96 (0.85, 0.99) PPV 0.90 (0.77, 0.96)	
Canal-Bedia 2011(48) **Spain** **Prospective** **Aim: Adapt and validate the Spanish version of the M-CHAT** **Prevalence: 0.29%**	Source: children 18-36 month old in Salamanca and Zamora provinces, Spain. Excludes: NR. Total sample: 2055 Age: range 18-36	Setting: Mandatory vaccination program at 18 months, and/or the general well-baby check-up at 24 months. Personnel: primary care paediatricians and nurses Training: Yes Screening tool: M-CHAT [Spanish] Threshold: 3 out of 23 or 2 out of the 6 critical items, confirmed by follow-up interviews	Setting and timing: Salamanca University ASD unit. Timing assumed as soon as possible. Diagnostic criteria: DSM-IV. Tools/measures: ADOS-G, VABS, MPRSD . Personnel: Pediatrician or nurse.	No	PPV 0.19 (0.05, 0.33)	NR

ADI-R, Autism Diagnostic Interview-Revised; ADOS-2, Autism Diagnostic Observation Schedule, 2nd Edition; ADOS-G, Autism Diagnostic Observation Schedule-General; AOSI, Autism Observation Scale in Infants; ASQ, Ages and Stages Questionnaire; BASC, Behavioral Assessment System for Children; BSID, Bayley Scales of Infant Development; CARS, Childhood Autism Rating Scale; CAST, Childhood Autism Spectrum Test; DISCO, Diagnostic Interview for Social and Communication Disorders; DSM, Diagnostic and Statistical Manual of Mental Disorders; FUI, Follow-up interview; FYI, First Year Inventory; GDS, Global Developmental Screen; ICD-10, International Classification of Diseases-10; JA-OBS, Joint Attention Observation schedule; M-CHAT-R/F, Modified Checklist for Autism in Toddlers (Revised/ with Follow-Up); MPRSD, Merril-Palmer Revised Scales of Development; MSEL, Mullen Scales of Early Learning; NPV, negative predictive value; NR, not report; PEDS, Parents Evaluation of Developmental Status; PEPR, Psycho Educational Profile Revised; PPV, positive predictive value; Q-CHAT, Quantitative Checklist for Autism in Toddlers; SACS-R, Social Attention

Communication Surveillance- Revised; SCQ, Social Communication Questionnaire; SD, standard deviation; Sens, sensitivity; Spec, specificity; TASI, Toddler Autism Symptom Interview; TIDOS, Three-item Direct Observation Screen; VABS, Vineland Adaptive Behavior Scales

Table 21. Studies relevant to criterion 9

Author year Design	Population	Sample size	Intervention(s)	Control	Outcomes - children	Results
Baranek 2015(49) **RCT**	Source: 5 counties in central North Carolina, USA Screening tool: FYI version 2.0	n=16: ART 11, REIM 5	Adapted Responsive Teaching (ART) Administered by parents (trained by 3 interventionists)	referral to early intervention and monitoring (REIM)	Children: MSEL, VABS-II, CSBS, SPA, SEQ evaluated at baseline (T1), 8 months (T2) and ~14 months later (T3) Family members: MBRS	ART significantly associated with improved receptive language, socialisation, sensory hyporesponsiveness and “less directive parental interactive style” during the intervention period. Little evidence of any difference at 32 month FU. ASD diagnosis at 32 months old: 36% ART, 40% REIM, 100% not randomised.
Watson 2017(50) **RCT**	Source: 6 central counties of North Carolina, USA	n=87 : ART 45, REIM42	Adapted Responsive Teaching (ART) Administered by parents (trained by 6 interventionists)	Referral to early intervention and monitoring (REIM)	Children: CSBS, SPA, MSEL, VABS-II, ADOS Family members: PRCS, MBRS	No evidence that ART associated with Improved Social-Communication, Sensory‑Regulatory,

	Screening tool: First Year Inventory					Adaptive, and Autism Symptom Outcomes. ART was associated with improvements in motor skills, but the finding could just reflect regression-to the mean. Across both groups, 41% met criteria for ASD,
Spjut Jansson 2016(29) **Prospective naturalistic cohort**	Source: 2.5 year old children referred following screening in Gothenburg, Sweden Screening tool: NR	n=71	Regular Intensive Learning programme, modified intensive learning programme or usual care Administered by habilitation centre professionals	all participants received one of the interventions	Children: VABS-II, C-GAS, 2 years after initial assessment Family members: None	Adaptive composite scores: No evidence of increase in scores over time across total sample. No evidence of any of the interventions increased scores more than another intervention. Global functioning: Evidence that scores increased over time, but no evidence that greater increases seen with any of the interventions over another.
Whitehouse 2021(51) **RCT**	12-month old infants (mostly) referred following screening in Perth and	N=89	iBASIS–Video Interaction to Promote Positive Parenting (iBASIS-VIPP)	Usual care	Children: -primary outcome - ASD symptom severity over time measured by: AOSI at baseline and treatment end; ADOS-2 at 12-month and 24-month post baseline -secondary outcome - clinical ASD diagnoses according to DSM-5,	

	Melbourne, Australia	MSEL, VABS-II Parents: MACI, MCDI

C-GAS, Children's Global Assessment Scale; CSBS, Communication and Symbolic Behaviour Scales; FU, follow-up; FYI, First Year Inventory; MACI, Manchester Assessment of Caregiver-Child Interaction; MBRS, Maternal Behavior Rating Scale; MCDI, The MacArthur-Bates Communicative Development Inventory; MSEL, Mullen Scales of Early Learning; NR, not reported; PRCS, Parent Responsiveness Coding System ; SEQ, sensory experiences questionnaire; SPA, Sensory Processing Assessment; VABS, Vineland Adaptive Behavior Scales

Appraisal for quality and risk of bias

Quality assessments of included studies are reported below.

Table 22. Quality assessment of studies relevant to criterion 1 (after QUIPS)

Reference	Participants			Attrition						Diagnosis Assessment at T1		
	Sample	**Avoided inappropriate exclusions**	**Overall RoB**	**Adequate response rate**	**Details of drop-outs**	**Reasons for loss-to-follow-up**	**Described lost participants**	**No important differences**	**Overall RoB**	**Clear definition/ description**	**Method and setting same**	**Overall RoB**
Screened population												
Allison 2021(26)	Yes	Yes	Low	No	Yes	Yes	Yes	Unclear	No	Yes	Yes	Low
Pierce 2019(27)	Yes	Yes	Low	Yes	No	Yes	Unclear	Unclear	High	Yes	Yes	Low
Barbaro 2017(28)	Yes	Unclear	Unclear	No	Unclear	No	Unclear	Unclear	High	No	Yes	High
Spjut Jansson 2016(29)	Yes	No	High	No	No	Yes	No	Unclear	High	No	No	High
Guthrie 2013(30)	Yes	Yes	Low	Yes	NA	NA	NA	NA	Low	Yes	Unclear	Unclear
Non-screened population												
McDonald 2020(61)	No	Unclear	High	No	NR	Yes	Yes	Unclear	High	No	Yes	High
Pellicano 2012(62)	No	Unclear	High	No	?	Yes	Unclear	Yes	High	Yes	Yes	Low
Soke 2011(63)	No	No	High	No	No	No	Unclear	Unclear	High	Yes	Yes	Low
Anglim 2020(60)	No	No	High	No	Unclear	No	No	Unclear	High	No	No	High

RoB, risk of bias

Table 23. Quality assessment of studies relevant to criterion 1 (after QUIPS) continued

Reference	Diagnostic assessment at T2			Confounding				
	Clear definition/description	**Method and setting same**	**Overall RoB**	**Measured important confounders**	**Clear definitions**	**Overall RoB**	**Blinding**	**Pre-specified design**

Screened population									
Allison 2021(26)	Yes	No	High	No	No	High	Yes	Yes	
Pierce 2019(27)	Yes	Yes	Low	No	No	High	No	Unclear	
Barbaro 2017(28)	No	Yes	High	Yes	Yes	Low	Yes	Yes	
Spjut Jansson 2016(29)	Yes	Yes	Low	No	No	High	No	Unclear	
Guthrie 2013(30)	Yes	Unclear	Unclear	Yes	Yes	Low	No	Yes	
Non-screened population									
McDonald 2020(61)	Yes	No	High	No	No	High	Yes	Yes	
Pellicano 2012(62)	Yes	No	High	Yes	No	High	Unclear	Yes	
Soke 2011(63)	Yes	Yes	Low	No	No	High	No	Yes	
Anglim 2020(60)	No	No	High	No	No	High	Unclear	Yes	

RoB, risk of bias

Table 24. Quality assessment of systematic reviews relevant to criterions 4 and 5

Study	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Levy 2020(74)	N	N	Y	PY	Y	Y	N	PY	Y	N	NA	NA	Y	Y	NA	Y
Petrocchi 2020(75)	N	PY	Y	N	Y	Unclear	N	N	Y	N	NA	NA	Y	N	NA	Y
Sanchez-Garcia 2019(76)	N	N	N	N	N	N	PY	N	Y	N	PY	N	Y	N	Y	N
Yuen 2018(78)	N	Unclear	N	PY	Unclear	Unclear	N	PY	Y	N	Y	N	Y	Y	N	Y
Towle 2016(77)	N	N	Y	N	Unclear	Unclear	N	Y	Y	N	NA	NA	Y	N	NA	Y
McPheeters 2016(24)	N	N	Y	N	PY	N	PY	Y	Y	N	NA	NA	Y	Y	NA	Y

N, no; NA, not applicable; PY, partly yes; Y, yes:

1. Did the research questions and inclusion criteria for the review include the components of PICO?
2. Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?
3. Did the review authors explain their selection of the study designs for inclusion in the review?
4. Did the review authors use a comprehensive literature search strategy?

5. Did the review authors perform study selection in duplicate?
6. Did the review authors perform data extraction in duplicate?
7. Did the review authors provide a list of excluded studies and justify the exclusions?
8. Did the review authors describe the included studies in adequate detail?
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?
10. Did the review authors report on the sources of funding for the studies included in the review?
11. If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?
12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?
13. Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?
14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?
15. If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?
16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?

Table 25. Quality assessment of screening accuracy studies relevant to criterions 4 and 5

	Consecutive or random sample?	Avoided a case-control design?	Avoid inappropriate exclusions?	Risk of bias for selection of patients?	Concerns of applicability of included patients?	Index test results blind to reference standard?	Pre-specified threshold?	Risk of bias for index test?	Concerns of applicability of index test?	Reference standard correctly classify target condition?	Reference standard results blind to index test?	Risk of bias for reference standard?	Concerns of applicability of reference standard?	Appropriate interval between index test(s) and reference standard?	All patients receive a reference standard?	Did patients receive the same reference standard?	Were all patients included in the analysis?	Risk of bias for patient flow?	Was the study method pre-specified?	Is the study funded/conducted by screening tool developers?
Allison 2021(26)	Y	Y	Y	Low	Low	Y	Y	Low	Low	Y	Y	Low	Low	U	N	Y	N	High	Y	N
Jonsdottir 2021, 2020(31, 33)	Y	Y	Y	Low	Low	Y	Y	Low	High	Y	N	High	Low	N	N	Y	N	High	U	N
Wieckowski 2021(32)	Y	Y	Y	Low	Low	U	U	U	High	Y	U	U	Low	U	N	Y	Y	High	Y	N
Kerub 2020(34)	Y	Y	U	U	U	Y	U	U	High	Y	U	U	Low	U	N	U	N	High	U	N
Magan-Maganto 2020(35)	Y	Y	Y	Low	Low	Y	Y	Low	High	Y	U	U	Low	U	N	U	N	High	U	N

Oner 2020(36)	N	Y	U	High	U	Y	Y	Low	High	U	U	U	U	U	N	Y	N	High	U	N
Mozolic-Staunton 2020(37)	U	Y	U	U	U	Y	Y	Low	Low	Y	N	High	Low	Y	N	U	N	High	U	N
Dai 2020(38)	Y	Y	Y	Low	Low	Y	Y	Low	Low	Y	N	High	Low	Y	N	Y	Y	Low*	U	Y
Achenie 2019(39)	Y	Y	U	U	U	Y	Y	Low	Low	Y	U	Unclear	Low	U	N	Y	N	High	U	N
Suren 2019(40)	Y	Y	Y	Low	Low	Y	Y	Low	High	N	U	U	High	U	N	N	Y	High	U	N
Topcu 2018(41)	Y	Y	Y	Low	U	Y	Y	Low	High	Y	N	High	Low	Y	N	Y	N	High	U	N
Catino 2017(42)	Y	Y	Y	Low	Low	Y	N	U	High	Y	U	Low	Low	U	N	Y	U	U*	U	N
Baduel 2017(43)	Y	Y	U	U	U	Y	Y	Low	High	Y	N	High	Low	U	N	Y	N	High	U	N
Kondolot 2016(44)	Y	Y	Y	Low	Low	Y	Y	Low	High	Y	N	High	Low	Y	N	Y	Y	Low	U	N
Wiggins 2014(45)	Y	Y	U	U	U	Y	Y	Low	Low	Y	U	U	Low	U	N	Y	Y	Low*	U	N
Robins 2014(16)	Y	Y	U	U	U	Y	Y	Low	Low	Y	U	U	Low	U	N	Y	N	High	U	N
Ben-Sasson 2013(46)	Y	Y	N	High	Low	Y	Y	Low	High	U	N	U	Low	U	N	Y	N	High	U	N
Chlebowski 2013(47)	Y	Y	N	U	Low	Y	Y	Low	Low	Y	U	U	Low	U	N	Y	N	High	U	N
Nygren 2012(17)	Y	Y	Y	Low	Low	Y	Y	Low	High	Y	U	U	Low	U	N	Y	N	High	U	N
Canal-Bedia 2011(48)	Y	Y	U	U	U	Y	Y	Low	High	Y	U	U	Low	U	N	Y	U	U	U	N

N, no; U, unclear; Y, yes; *Although not all participants received the reference standard these studies were deemed to be at low risk of bias as only PPVs were reported

Table 26. Quality assessment of the comparative screening accuracy aspect of studies relevant to criterion 4 and 5

	Low risk of bias for selection of participants for all index tests?	Receive all index tests or to be randomly allocated to index	If randomized, was the allocation sequence random?	If randomized, was the allocation sequence concealed?	Risk of bias for selection of patients?	Low risk of bias for index tests?	If patients received multiple index tests, were test results interpreted without knowledge of	If patients received multiple index tests, is undergoing one index test unlikely to affect the performance of the other index	Were differences in the conduct or interpretation between the index tests unlikely to advantage	Could the conduct or interpretation of the index tests have introduced bias in the	Was risk of bias for this domain judged ' low' for all index tests?	Did the reference standard avoid incorporating any of the index	Could the reference standard, its conduct, or its interpretation have introduced bias in the comparison?	Was risk of bias for this domain judged ' low' for all index tests?	Was there an appropriate interval between the index tests?	Was the same reference standard used for all index tests?	Are the proportions and reasons for missing data similar across index tests?	Could the patient flow have introduced bias in the comparison?
Kerub 2020(34)	N	Y	NA	NA	U	N	N	U	U	U	N	U	U	N	U	Y	NA	U
Topcu 2018(41)	Y	Y	NA	NA	Low	Y	U	U	U	U	N	Y	High	N	Y	Y	Y	U
Nygren 2012(17)	Y	Y	NA	NA	Low	Y	N	U	U	U	Y	Y	U	N	U	Y	NA	U

N, no; U, unclear; Y, yes

Table 27. Quality assessment of randomised controlled trials relevant to criterion 9 (Cochrane RoB)

	Baranek 2015**(49)**	Watson 2017**(50)**	Whitehouse 2021**(51)**
1.1 Was the allocation sequence random?	Y (Randomization was conducted using a random number generator in Excel by an investigator blind to the assessment results)	Y (a randomization sequenc was generated using a randomization method for small samples that mixes simple randomization with permuted block randomization)	Y (was performed by minimization stratified by site, sex, number of relevant behaviors, and age range at recruitment, with randomization determined by a biased coin with a probability of 0.7)
1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?	Y (families were notified of their assignment following randomization)	Y (team members who interacted with participants were not privy to randomization method details)	Y
1.3 Did baseline differences between intervention groups suggest a problem with the randomization process?	N	N	Y (beside screened population the sample from one site also included referred infants)
D1. Risk-of-bias judgement	Low concern	Low concern	Some concern

2.1. Were participants aware of their assigned intervention during the trial?	Y (families were notified of their assignment following randomization)	Y	Y (families could not be blidend to group allocation)
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	Y	Y (Of necessity, intervention team staff then learned if a family was allocated to the ART group)	Y
2.3. If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?	NI	NI	NI
2.4 If Y/PY to 2.3: Were these deviations likely to have affected the outcome?	NA	NA	NA
2.5. If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?	NA	NA	NA
2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?	Y	Y	Y
2.7 If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomized?	NA	NA	NA
D2. Risk-of-bias judgement	Some concerns	Some concerns	Some concerns
3.1 Were data for this outcome available for all, or nearly all, participants randomized?	Y	Y	Y
3.2 If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?	NA	NA	NA
3.3 If N/PN to 3.2: Could missingness in the outcome depend on its true value?	NA	NA	NA
3.4 If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?	NA	NA	NA
D3. Risk-of-bias judgement	Low concern	Low concern	Low concern
4.1 Was the method of measuring the outcome inappropriate?	PN	PN	PN
4.2 Could measurement or ascertainment of the outcome have	PN	PN	PN

differed between intervention groups?			
4.3 If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?	N (The assessment team was blinded to group assignment; parents were instructed to not share information regarding EI services or group assignment)	N (assessment team staff remained blind to allocation throughout the project)	N (research staff conducting the assessments were independent of the clinical teams administering the iBASIS-VIPP intervention)
4.4 If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?	NA	NA	NA
4.5 If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?	NA	NA	NA
D4. Risk-of-bias judgement	Low concern	Low concern	Low concern
5.1 Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis?	PY	PY	PY
5.2. Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain?	PN	PN	PN
5.3 Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible analyses of the data?	PN	PN	PN
D5. Risk-of-bias judgement	Low concern	Low concern	Low concern

PN, partial no, PY, partial yes; N, no; NA, not applicable; U, unclear; Y, yes

Table 28. Quality assessment of non-randomised controlled trials relevant to criterion 9 (ROBINS-I)

Reference	Spjut Jansson 2016**(29)**
1.1 Is there potential for confounding of the effect of intervention in this study?	N
1.2. Was the analysis based on splitting participants' follow up time according to intervention received?	NA
1.3. Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome?	NA
1.4. Did the authors use an appropriate analysis method that controlled for all the important confounding domains?	Y (adaptive composite score and C-GAS before vs after treatment were used as dependent variables in 2 separate mixed analysis of variance (ANOVA) and intellectual level as independent variables in the intervention groups)
1.5. [If Y/PY to 1.4]: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?	Y
1.6. Did the authors control for any post-intervention variables that could have been affected by the intervention?	N
1.7. Did the authors use an appropriate analysis method that controlled for all the important confounding domains and for time-varying confounding?	Y
Bias due to confounding judgement	Low
2.1. Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention?	N (all eligible children were referred)
2.2. [If Y/PY to 2.1]: Were the post-intervention variables that influenced selection likely to be associated with intervention?	NA
2.3 [If Y/PY to 2.2]: Were the post-intervention variables that influenced selection likely to be	NA

influenced by the outcome or a cause of the outcome?	
2.4. Do start of follow-up and start of intervention coincide for most participants?	Y
2.5. [If Y/PY to 2.2 and 2.3, or N/PN to 2.4]: Were adjustment techniques used that are likely to correct for the presence of selection biases?	NA
Bias in selection of participants into the study judgement	Low
3.1 Were intervention groups clearly defined?	Y
3.2 Was the information used to define intervention groups recorded at the start of the intervention?	Y
3.3 Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?	PN
Bias in classification of interventions judgement	Low
4.1. Were there deviations from the intended intervention beyond what would be expected in usual practice?	PN
4.2. [If Y/PY to 4.1]: Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome?	NA
4.3. Were important co-interventions balanced across intervention groups?	PY
4.4. Was the intervention implemented successfully for most participants?	Y
4.5. Did study participants adhere to the assigned intervention regimen?	PY
4.6. [If N/PN to 4.3, 4.4 or 4.5]: Was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?	NA

Bias due to deviations from intended interventions judgement	Low
5.1 Were outcome data available for all, or nearly all, participants?	Y
5.2 Were participants excluded due to missing data on intervention status?	N
5.3 Were participants excluded due to missing data on other variables needed for the analysis?	N
5.4 [If PN/N to 5.1, or Y/PY to 5.2 or 5.3]: Are the proportion of participants and reasons for missing data similar across interventions?	NA
5.5 [If PN/N to 5.1, or Y/PY to 5.2 or 5.3]: Is there evidence that results were robust to the presence of missing data?	NA
Bias due to missing data judgement	Low
6.1 Could the outcome measure have been influenced by knowledge of the intervention received?	PN
6.2 Were outcome assessors aware of the intervention received by study participants?	N (All the professionals were blinded to the type of intervention received by the children.)
6.3 Were the methods of outcome assessment comparable across intervention groups?	Y
6.4 Were any systematic errors in measurement of the outcome related to intervention received?	PN
Bias in measurement of outcomes judgement	Low
7.1.Is the reported effect estimate likely to be selected, on the basis of the results, from multiple outcome measurements within the outcome domain?	N
7.2. Is the reported effect estimate likely to be selected, on the basis of the results, from multiple analyses of the intervention-outcome relationship?	N

7.3. Is the reported effect estimate likely to be selected, on the basis of the results, from different subgroups?	N
Bias in selection of the reported result judgement	Low

PN, partial no; PY, partial yes; N, no; NA, not applicable; Y, yes

Appendix 4 – UK NSC reporting checklist for evidence summaries

All items on the UK NSC Reporting Checklist for Evidence Summaries have been addressed in this report. A summary of the checklist, along with the page or pages where each item can be found in this report, is presented in Table 29.

Table 29. UK NSC reporting checklist for evidence summaries

	Section	Item	Page no.
1.	TITLE AND SUMMARIES		
1.1	Title sheet	Identify the review as a UK NSC evidence summary.	Title page
1.2	Plain English summary	Plain English description of the executive summary.	6
1.3	Executive summary	Structured overview of the whole report. To include: the purpose/aim of the review; background; previous recommendations; findings and gaps in the evidence; recommendations on the screening that can or cannot be made on the basis of the review.	7-11
2.	INTRODUCTION AND APPROACH		
2.1	Background and objectives	Background – Current policy context and rationale for the current review – for example, reference to details of previous reviews, basis for current recommendation, recommendations made, gaps identified, drivers for new reviews Objectives – What are the questions the current evidence summary intends to answer? – statement of the key questions for the current evidence summary, criteria they address, and number of studies included per question, description of the overall results of the literature search. Method – briefly outline the rapid review methods used.	12-17
2.2	Eligibility for inclusion in the review	State all criteria for inclusion and exclusion of studies to the review clearly (PICO, dates, language, study type, publication type, publication status etc.) To be decided *a priori*.	17-20
2.3	Appraisal for quality/risk of bias tool	Details of tool/checklist used to assess quality, e.g. QUADAS 2, CASP, SIGN, AMSTAR.	20
3.	SEARCH STRATEGY AND STUDY SELECTION (FOR EACH KEY QUESTION)		
3.1	Databases/ sources searched	Give details of all databases searched (including platform/interface and coverage dates) and date of final search.	18

3.2	Search strategy and results	Present the full search strategy for at least one database (usually a version of Medline), including limits and search filters if used. Provide details of the total number of (results from each database searched), number of duplicates removed, and the final number of unique records to consider for inclusion.	Appendix 1, Appendix 2
3.3	Study selection	State the process for selecting studies – inclusion and exclusion criteria, number of studies screened by title/abstract and full text, number of reviewers, any cross checking carried out.	Q1: 23 Q2: 30-31 Q3: 61-62
4.	STUDY LEVEL REPORTING OF RESULTS (FOR EACH KEY QUESTION)		
4.1	Study level reporting, results and risk of bias assessment	For each study, produce a table that includes the full citation and a summary of the data relevant to the question (for example, study size, PICO, follow-up period, outcomes reported, statistical analyses etc.). Provide a simple summary of key measures, effect estimates and confidence intervals for each study where available. For each study, present the results of any assessment of quality/risk of bias.	Study level reporting: Q1: 24 Q2: 35-44 Q3: 62-63 Quality assessment: Q1: 24 Q2: 35-44 Q3: 65
5.	QUESTION LEVEL SYNTHESIS		
5.1	Description of the evidence	For each question, give numbers of studies screened, assessed for eligibility, and included in the review, with summary reasons for exclusion.	Q1: 23 Q2: 30-31 Q3: 61-62
5.2	Combining and presenting the findings	Provide a balanced discussion of the body of evidence which avoids over reliance on one study or set of studies. Consideration of 4 components should inform the reviewer's judgement on whether the criterion is 'met', 'not met' or 'uncertain': quantity; quality; applicability and consistency.	Q1: 25-28 Q2: 45-57 Q3: 64-65
5.3	Summary of findings	Provide a description of the evidence reviewed and included for each question, with reference to their eligibility for inclusion. Summarise the main findings including the quality/risk of bias issues for each question. Have the criteria addressed been 'met', 'not met' or 'uncertain'?	Q1: 28-29 Q2: 57-60 Q3: 65-67
6.	REVIEW SUMMARY		
6.1	Conclusions and implications for policy	Do findings indicate whether screening should be recommended? Is further work warranted? Are there gaps in the evidence highlighted by the review?	68-69

6.2	Limitations	Discuss limitations of the available evidence and of the review methodology if relevant.	69

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 2013.
2. Siu AL, Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, Ebell M, et al. Screening for Autism Spectrum Disorder in Young Children: US Preventive Services Task Force Recommendation Statement. Jama. 2016;315(7):691-6.
3. Lai MC, Lombardo MV, Baron-Cohen S. Autism. Lancet (London, England). 2014;383(9920):896-910.
4. National Institute for Health and Care Excellence. Autism spectrum disorder in under 19s. Autism spectrum disorder in under 19s: recognition, referral and diagnosis. NICE; 2011.
5. Chiarotti F, Venerosi A. Epidemiology of Autism Spectrum Disorders: A Review of Worldwide Prevalence Estimates Since 2014. Brain Sci. 2020;10(5).
6. Roman-Urrestarazu A, van Kessel R, Allison C, Matthews FE, Brayne C, Baron-Cohen S. Association of Race/Ethnicity and Social Disadvantage With Autism Prevalence in 7 Million School Children in England. JAMA Pediatr. 2021;175(6):e210054.
7. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics. 2014;133(1):e54-63.
8. Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, Duncan L, et al. Analysis of shared heritability in common disorders of the brain. Science (New York, NY). 2018;360(6395).
9. Jokiranta-Olkoniemi E, Cheslack-Postava K, Sucksdorff D, Suominen A, Gyllenberg D, Chudal R, et al. Risk of Psychiatric and Neurodevelopmental Disorders Among Siblings of Probands With Autism Spectrum Disorders. JAMA psychiatry. 2016;73(6):622-9.
10. Jones EJ, Gliga T, Bedford R, Charman T, Johnson MH. Developmental pathways to autism: a review of prospective studies of infants at risk. Neuroscience and biobehavioral reviews. 2014;39(100):1-33.
11. Elsabbagh M, Johnson MH. Autism and the Social Brain: The First-Year Puzzle. Biological psychiatry. 2016;80(2):94-9.
12. Elsabbagh M. Linking risk factors and outcomes in autism spectrum disorder: is there evidence for resilience? BMJ (Clinical research ed). 2020;368:l6880.
13. Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcín C, et al. Global prevalence of autism and other pervasive developmental disorders. Autism research : official journal of the International Society for Autism Research. 2012;5(3):160-79.
14. Taylor B, Jick H, Maclaughlin D. Prevalence and incidence rates of autism in the UK: time trend from 2004-2010 in children aged 8 years. BMJ open. 2013;3(10):e003219.
15. Russell G, Stapley S, Newlove-Delgado T, Salmon A, White R, Warren F, et al. Time trends in autism diagnosis over 20 years: a UK population-based cohort study. Journal of Child Psychology and Psychiatry. 2021 (early online).
16. Robins DL, Casagrande K, Barton M, Chen CM, Dumont-Mathieu T, Fein D. Validation of the modified checklist for Autism in toddlers, revised with follow-up (M-CHAT-R/F). Pediatrics. 2014;133(1):37-45.

17. Nygren G--S, Eva-//-Gillstedt, Fredrik-//-Ekeroth, Gunnar-//-Arvidsson, Thomas-//-Gillberg, Christopher. A new screening programme for autism in a general population of Swedish toddlers. Research in developmental disabilities. 2012;33(4):1200-10.
18. Catalano D, Holloway L, Mpofu E. Mental health interventions for parent carers of children with autistic spectrum disorder: Practice guidelines from a critical interpretive synthesis (CIS) systematic review. International journal of environmental research and public health. 2018;15(2).
19. Turner LM, Stone WL. Variability in outcome for children with an ASD diagnosis at age 2. J Child Psychol Psychiatry. 2007;48(8):793-802.
20. Russell G, Golding J, Norwich B, Emond A, Ford T, Steer C. Social and behavioural outcomes in children diagnosed with autism spectrum disorders: a longitudinal cohort study. J Child Psychol Psychiatry. 2012;53(7):735-44.
21. Poslawsky IE, Naber FB, Van Daalen E, Van Engeland H. Parental reaction to early diagnosis of their children's autism spectrum disorder: an exploratory study. Child psychiatry and human development. 2014;45(3):294-305.
22. Russell G, Norwich B. Dilemmas, diagnosis and de-stigmatization: parental perspectives on the diagnosis of autism spectrum disorders. Clinical child psychology and psychiatry. 2012;17(2):229-45.
23. Hyman SL, Levy SE, Myers SM. Identification, Evaluation, and Management of Children With Autism Spectrum Disorder. Pediatrics. 2020;145(1).
24. McPheeters ML, Weitlauf A, Vehorn A, Taylor C, Sathe NA, Krishnaswami S, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Screening for Autism Spectrum Disorder in Young Children: A Systematic Evidence Review for the US Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016.
25. Warren Z, McPheeters ML, Sathe N, Foss-Feig JH, Glasser A, Veenstra-Vanderweele J. A systematic review of early intensive intervention for autism spectrum disorders. Pediatrics. 2011;127(5):e1303-11.
26. Allison C, Soufer R, Baron-Cohen S, Matthews FE, Ruta L, Pasco G, et al. Quantitative checklist for autism in toddlers (Q-CHAT). A population screening study with follow-up: The case for multiple time-point screening for autism. BMJ Paediatrics Open. 2021;5(1):e000700.
27. Pierce K, Gazestani VH, Bacon E, Barnes CC, Cha D, Nalabolu S, et al. Evaluation of the Diagnostic Stability of the Early Autism Spectrum Disorder Phenotype in the General Population Starting at 12 Months. JAMA pediatrics. 2019;173(6):578-87.
28. Barbaro J, Dissanayake C. Diagnostic stability of autism spectrum disorder in toddlers prospectively identified in a community-based setting: Behavioural characteristics and predictors of change over time. Autism : the international journal of research and practice. 2017;21(7):830-40.
29. Spjut Jansson B, Miniscalco C, Westerlund J, Kantzer A-K, Fernell E, Gillberg C. Children who screen positive for autism at 2.5 years and receive early intervention: a prospective naturalistic 2-year outcome study. Neuropsychiatric disease and treatment. 2016;12:2255-63.
30. Guthrie W, Swineford LB, Nottke C, Wetherby AM. Early diagnosis of autism spectrum disorder: stability and change in clinical diagnosis and symptom presentation. Journal of child psychology and psychiatry, and allied disciplines. 2013;54(5):582-90.
31. Jonsdottir SL, Saemundsen E, Jonsson BG, Rafnsson V. Validation of the modified checklist for autism in toddlers, revised with follow-up in a population sample of 30-month-old children in iceland: A prospective approach. Journal of Autism and Developmental Disorders. 2021:No-Specified.

32. Wieckowski AT, Hamner T, Nanovic S, Porto KS, Coulter KL, Eldeeb SY, et al. Early and Repeated Screening Detects Autism Spectrum Disorder. The Journal of pediatrics. 2021;234:227-35.
33. Jonsdottir SL--S, E.-//-Gudmundsdottir, S.-//-Haraldsdottir, G. S.-//-Palsdottir, A. H.-//-Rafnsson, V. Implementing an early detection program for autism in primary healthcare: Screening, education of healthcare professionals, referrals for diagnostic evaluation, and early intervention. Research in Autism Spectrum Disorders. 2020;77:101616.
34. Kerub O--H, Eric J.-//-Meiri, Gal-//-Davidovitch, Nadav-//-Menashe, Idan. A Comparison Between Two Screening Approaches for ASD Among Toddlers in Israel. Journal of autism and developmental disorders. 2020;50(5):1553-60.
35. Magan-Maganto M--C-B, Ricardo-//-Hernandez-Fabian, Aranzazu-//-Bejarano-Martin, Alvaro-//-Fernandez-Alvarez, Clara J.-//-Martinez-Velarte, Maria-//-Martin-Cilleros, Maria V.-//-Flores-Robaina, Noelia-//-Roeyers, Herbert-//-Posada de la Paz, Manuel. Spanish Cultural Validation of the Modified Checklist for Autism in Toddlers, Revised. Journal of autism and developmental disorders. 2020;50(7):2412-23.
36. Oner O--M, Kerim M. Modified Checklist for Autism in Toddlers Revised (MCHAT-R/F) in an Urban Metropolitan Sample of Young Children in Turkey. Journal of autism and developmental disorders. 2020;50(9):3312-9.
37. Mozolic-Staunton B--D, M.-//-Yoxall, J.-//-Barbaro, J. Early detection for better outcomes: Universal developmental surveillance for autism across health and early childhood education settings. Research in Autism Spectrum Disorders. 2020;71:101496.
38. Dai YG--M, Lauren E.-//-Ramsey, Riane K.-//-Robins, Diana L.-//-Fein, Deborah A.-//-Dumont-Mathieu, Thyde. Incremental Utility of 24-Month Autism Spectrum Disorder Screening After Negative 18-Month Screening. Journal of autism and developmental disorders. 2020;50(6):2030-40.
39. Achenie LEK--S, Angela-//-Factor, Reina S.-//-Wang, Tao-//-Robins, Diana L.-//-McCrickard, D. Scott. A Machine Learning Strategy for Autism Screening in Toddlers. Journal of developmental and behavioral pediatrics : JDBP. 2019;40(5):369-76.
40. Suren P--S-H, Alexandra-//-Bresnahan, Michaeline-//-Hirtz, Deborah-//-Hornig, Mady-//-Lord, Catherine-//-Reichborn-Kjennerud, Ted-//-Schjolberg, Synnve-//-Oyen, Anne-Siri-//-Magnus, Per-//-Susser, Ezra-//-Lipkin, W. Ian-//-Stoltenberg, Camilla. Sensitivity and specificity of early screening for autism. BJPsych open. 2019;5(3):e41.
41. Topcu S--U, B.-//-Oner, O.-//-Simsek Orhon, F.-//-Baskan, S. Comparison of tidos with m-chat for screening autism spectrum disorder. Psychiatry and Clinical Psychopharmacology. 2018;28(4):416-22.
42. Catino E--DT, Michela-//-Giovannone, Federica-//-Manti, Filippo-//-Nunziata, Letizia-//-Piccari, Francesca-//-Sirchia, Virginia-//-Vannucci, Lucia-//-Sogos, Carla. Screening for Developmental Disorders in 3- and 4-Year-Old Italian Children: A Preliminary Study. Frontiers in pediatrics. 2017;5:181.
43. Baduel S--G, Quentin-//-Afzali, Mohammad H.-//-Foudon, Nadege-//-Kruck, Jeanne-//-Roge, Bernadette. The French Version of the Modified-Checklist for Autism in Toddlers (M-CHAT): A Validation Study on a French Sample of 24 Month-Old Children. Journal of autism and developmental disorders. 2017;47(2):297-304.
44. Kondolot M, Ozmert EN, Oztop DB, Mazicioglu MM, Gumus H, Elmali F. The modified checklist for autism in Turkish toddlers: A different cultural adaptation sample. Research in Autism Spectrum Disorders. 2016;21:121-7.

45. Wiggins LD--P, Vivian-//-Robins, Diana L. Comparison of a broad-based screen versus disorder-specific screen in detecting young children with an autism spectrum disorder. Autism : the international journal of research and practice. 2014;18(2):76-84.
46. Ben-Sasson A--C, Alice S. The application of the first year inventory for ASD screening in Israel. Journal of autism and developmental disorders. 2012;42(9):1906-16.
47. Chlebowski C--R, Diana L.-//-Barton, Marianne L.-//-Fein, Deborah. Large-scale use of the modified checklist for autism in low-risk toddlers. Pediatrics. 2013;131(4):e1121-7.
48. Canal-Bedia R--G-P, Patricia-//-Martin-Cilleros, Maria Victoria-//-Santos-Borbujo, Jose-//-Guisuraga-Fernandez, Zoila-//-Herraez-Garcia, Lorena-//-Herraez-Garcia, Maria del Mar-//-Boada-Munoz, Leticia-//-Fuentes-Biggi, Joaquin-//-Posada-de la Paz, Manuel. Modified checklist for autism in toddlers: cross-cultural adaptation and validation in Spain. Journal of autism and developmental disorders. 2011;41(10):1342-51.
49. Baranek GT, Watson LR, Turner-Brown L, Field SH, Crais ER, Wakeford L, et al. Preliminary efficacy of adapted responsive teaching for infants at risk of autism spectrum disorder in a community sample. Autism research and treatment. 2015;2015:386951.
50. Watson LR, Crais ER, Baranek GT, Turner-Brown L, Sideris J, Wakeford L, et al. Parent-Mediated Intervention for One-Year-Olds Screened as At-Risk for Autism Spectrum Disorder: A Randomized Controlled Trial. J Autism Dev Disord. 2017;47(11):3520-40.
51. Whitehouse AJO, Varcin KJ, Pillar S, Billingham W, Alvares GA, Barbaro J, et al. Effect of Preemptive Intervention on Developmental Outcomes Among Infants Showing Early Signs of Autism: A Randomized Clinical Trial of Outcomes to Diagnosis. JAMA Pediatr. 2021:e213298.
52. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;158(4):280-6.
53. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.
54. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ (Clinical research ed). 2017;358:j4008.
55. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ (Clinical research ed). 2011;343:d5928.
56. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ (Clinical research ed). 2016;355:i4919.
57. StataCorp. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC; 2019.
58. Woolfenden S, Sarkozy V, Ridley G, Williams K. A systematic review of the diagnostic stability of Autism Spectrum Disorder. Research in Autism Spectrum Disorders. 2012;6(1):345-54.
59. Bieleninik Ł, Posserud MB, Geretsegger M, Thompson G, Elefant C, Gold C. Tracing the temporal stability of autism spectrum diagnosis and severity as measured by the Autism Diagnostic Observation Schedule: A systematic review and meta-analysis. PloS one. 2017;12(9):e0183160.
60. Anglim M, Conway EV, Barry M, Kashif M, Ackermann P, Moran A, et al. An initial examination of the psychometric properties of the Diagnostic Instrument for Social and Communication Disorders (DISCO-11) in a clinical sample of children with a diagnosis of Autism spectrum disorder. Irish journal of psychological medicine. 2020:1-10.

61. McDonald J, Milne S, Masi A, Zieba J, Eapen V. Where are they now? An autism follow-up study. Journal of paediatrics and child health. 2020.
62. Pellicano E. Do autistic symptoms persist across time? Evidence of substantial change in symptomatology over a 3-year period in cognitively able children with autism. American journal on intellectual and developmental disabilities. 2012;117(2):156-66.
63. Soke GN, Philofsky A, Diguiseppi C, Lezotte D, Rogers S, Hepburn S. Longitudinal changes in scores on the Autism Diagnostic Interview-Revised (ADI-R) in pre-school children with autism: Implications for diagnostic classification and symptom stability. Autism. 2011;15(5):545-62.
64. Ferre Rey G--SR, Josefina-//-Llorca Linares, Miguel-//-Vicens, Paloma-//-Camps, Misericordia-//-Torrente, Margarita-//-Morales Vives, Fabia. A systematic review of instruments for early detection of autism spectrum disorders. International Journal of Psychology & Psychological Therapy. 2019;19(1):29-38.
65. Chesnut SR--W, Tianlan-//-Barnard-Brak, Lucy-//-Richman, David M. A meta-analysis of the social communication questionnaire: Screening for autism spectrum disorder. Autism : the international journal of research and practice. 2017;21(8):920-8.
66. Sunita --B, Justin L. C. Early identification of autism: a comparison of the Checklist for Autism in Toddlers and the Modified Checklist for Autism in Toddlers. Journal of paediatrics and child health. 2013;49(6):438-44.
67. Magán-Maganto M, Bejarano-Martín Á, Fernández-Alvarez C, Narzisi A, García-Primo P, Kawa R, et al. Early Detection and Intervention of ASD: A European Overview. Brain Sci. 2017;7(12).
68. Thabtah F, Peebles D. Early Autism Screening: A Comprehensive Review. International journal of environmental research and public health. 2019;16(18).
69. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. American journal of epidemiology. 1994;140(8):759-69.
70. Allaby M, Sharma M. Screening for Autism Spectrum Disorders on Children below the age of 5 years: A draft report for the UK National Screening Committee. 2011.
71. Sparrow SS, Cichetti DV, Balla DA. Vineland Adaptive Behaviour Scales. 2nd ed: Circle Pines (MN): American Guidance Service; 2005.
72. Schorre BE, Vandvik IH. Global assessment of psychosocial functioning in child and adolescent psychiatry. A review of three unidimensional scales (CGAS, GAF, GAPD). European child & adolescent psychiatry. 2004;13(5):273-86.
73. Drew A, Baird G, Baron-Cohen S, Cox A, Slonims V, Wheelwright S, et al. A pilot randomised control trial of a parent training intervention for pre-school children with autism. Preliminary findings and methodological challenges. European child & adolescent psychiatry. 2002;11(6):266-72.
74. Levy SE--W, Audrey-//-Coury, Daniel-//-Duby, John-//-Farmer, Justin-//-Schor, Edward-//-Van Cleave, Jeanne-//-Warren, Zachary. Screening Tools for Autism Spectrum Disorder in Primary Care: A Systematic Evidence Review. Pediatrics. 2020;145(Suppl 1):S47-S59.
75. Petrocchi S--L, Annalisa-//-Lecciso, Flavia. Systematic Review of Level 1 and Level 2 Screening Tools for Autism Spectrum Disorders in Toddlers. Brain sciences. 2020;10(3).
76. Sanchez-Garcia AB--G-V, Purificacion-//-Nieto-Librero, Ana B.-//-Martin-Rodero, Helena-//-Robins, Diana L. Toddler Screening for Autism Spectrum Disorder: A Meta-Analysis of Diagnostic Accuracy. Journal of autism and developmental disorders. 2019;49(5):1837-52.
77. Towle PO--P, Patricia A. Autism Spectrum Disorder Screening Instruments for Very Young Children: A Systematic Review. Autism research and treatment. 2016;2016:4624829.

78. Yuen T--P, Melanie-//-Carter, Melissa T.-//-Szatmari, Peter-//-Ungar, Wendy J. Assessing the accuracy of the Modified Checklist for Autism in Toddlers: a systematic review and meta-analysis. Developmental medicine and child neurology. 2018;60(11):1093-100.