

Meta-analysis of the Modified Checklist for Autism in Toddlers, Revised/Follow-up for Screening

Ramkumar Aishworiya, MMed,^{a,b,c} Van Kim Ma, MD,^{a,d} Susan Stewart, PhD,^e Randi Hagerman, MD,^{a,d} Heidi M. Feldman, MD, PhD^f

CONTEXT: The Modified Checklist for Autism in Toddlers, Revised with Follow-up (M-CHAT-R/F) abstract is used worldwide to screen for autism spectrum disorder (ASD).

OBJECTIVE: To calculate psychometric properties of the M-CHAT-R/F for subsequent diagnosis of ASD.

DATA SOURCES: Systematic searches of Medline, Embase, SCOPUS, and Trip Pro databases from January 2014 to November 2021.

STUDY SELECTION: Studies were included if they (1) used the M-CHAT-R/F (2) applied standard scoring protocol, (3) used a diagnostic assessment for ASD, and (4) reported at least 1 psychometric property of the M-CHAT-R/F.

DATA EXTRACTION: Two independent reviewers completed screening, full-text review, data extraction, and quality assessment, following Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines. A random-effects model was used to derive pooled estimates and assess for between-study heterogeneity.

RESULTS: Of 667 studies identified, 15 with 18 distinct samples from 10 countries (49 841 children) were used in the meta-analysis. Pooled positive predictive value (PPV), was 57.7% (95% confidence interval [CI] 48.6–66.8, $\tau^2 = 0.031$). PPV was higher among high-risk (75.6% [95% CI 66.0–85.2]) than low-risk samples (51.2% [95% CI 43.0–59.5]). Pooled negative predictive value was 72.5% (95% CI 62.5–82.4 $\tau^2 = 0.031$), sensitivity was 82.6% (95% CI 76.2–88.9) and specificity 45.7% (95% CI 25.0–66.4).

LIMITATIONS: Negative predictive value, sensitivity, and specificity were calculated based on small sample sizes because of limited or no evaluation of screen-negative children.

CONCLUSIONS: These results support use of the M-CHAT-R/F as a screening tool for ASD. Caregiver counseling regarding likelihood of an ASD diagnosis after positive screen should acknowledge the moderate PPV.



^aMedical Investigation of Neurodevelopmental Disorders (MIND) Institute, ^dDepartment of Pediatrics, School of Medicine, and ^eDivision of Biostatistics, Department of Public Health Sciences, University of California Davis, Davis, California; ^bKhoo Teck Puat-National University Children's Medical Institute, National University Health System, Singapore; ^cDepartment of Pediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; and ^fDepartment of Pediatrics, Stanford University, Palo Alto, California

Dr Aishworiya conceptualized the study, critically screened and appraised potential studies, performed data extraction, and wrote the first draft and subsequent versions of the manuscript; Dr Ma critically screened and appraised potential studies, performed data extraction, and contributed to writing the manuscript and subsequent revisions; Dr Stewart conducted data analysis and contributed to writing the manuscript; Dr Hagerman contributed to the manuscript and critically reviewed revised versions; Dr Feldman conceptualized the study, resolved conflicts during the screening and appraisal process, and (Continued)

To cite: Aishworiya R, Ma VK, Stewart S, et al. Meta-analysis of the Modified Checklist for Autism in Toddlers, Revised/Follow-up for Screening. *Pediatrics*. 2023;151(6):e2022059393

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impairments in social communication, and restricted, repetitive patterns of behavior.¹ The prevalence rate of ASD is estimated to be 1.0% globally (median value, range 0.01% to 4.4% in a recent systematic review of studies since 2012) and 2.3% in the United States.^{2,3} Behavioral intervention initiated before 24 months of age has been shown to improve cognitive and adaptive functioning more than intervention initiated later in life.⁴⁻⁸ One of the focal points of research has been on developing effective screening tools to make the diagnosis of ASD at the youngest possible age, in tandem with the drive toward earlier initiation of appropriate therapeutic and educational services. In line with these priorities, the American Academy of Pediatrics advocates for routine ASD screening in children ages 18 and 24 months⁹ given the superior effectiveness of early behavioral therapy and the accuracy and diagnostic stability of ASD diagnosis as early as 18 months.¹⁰⁻¹²

Among the available screening tools, the Modified Checklist for Autism in Toddlers, Revised with Follow-Up (M-CHAT-R/F) has been one of the most studied and implemented screening tools.^{13,14} It has been translated into 58 different languages and dialects. It has been recommended as a screening tool by several national guidelines, especially for children at high-risk of autism, based on its high sensitivity in the original sample and the ease of administration.^{9,15,16} The M-CHAT-R/F is a 2-stage parent-report screening tool for ASD¹⁷ designed for use in children ages 16 to 30 months. In the norming sample, sensitivity was reported to be 0.83 to 0.85, specificity 0.95 to 0.99 and positive predictive value (PPV) was 0.465.¹⁷ The M-CHAT-R/F reported a lower false positive rate compared with its predecessor, the Modified Checklist for

Autism in Toddlers with Follow-Up (M-CHAT/F)¹⁸ and has replaced the older version as the preferable screening instrument.

Given the widespread use of the M-CHAT-R/F as a screening instrument for ASD around the globe, it is imperative to understand its psychometric properties beyond the original validation study and to determine if the M-CHAT-R/F is sensitive and specific in other samples, including non-English speaking samples worldwide using translated versions of the M-CHAT-R/F. The PPV and the negative predictive value (NPV) are both of importance in a screening tool for ASD given that false positive and negative screening test results could be highly detrimental for children and their families. Previous systematic reviews to date have focused on the original Modified Checklist for Autism in Toddlers (M-CHAT) questionnaire or other screening tools,¹⁹⁻²¹ not the M-CHAT-R/F. A recent meta-analysis evaluated the original M-CHAT and the M-CHAT-R/F without distinguishing between the 2 versions.²² That review has limited applicability for evaluating the M-CHAT-R/F, even though this version is the preferred screening instrument in the present era. Hence, we performed a systematic review and meta-analysis to evaluate the utility of the M-CHAT-R/F exclusively to screen for ASD in young children. We intentionally screened studies irrespective of language or country of use. We planned to report PPV and NPV in addition to sensitivity and specificity because of the importance of these psychometric values for interpreting the results of screening. In primary analyses, to maintain equipoise and increase the accuracy of sensitivity, specificity, and NPV, we required that screen-negative children received a diagnostic evaluation rather than assumed they were condition-free.

METHODS

Search Strategy

This systematic review was conducted in accordance with Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines²³ from January 2014, date of the publication of the instrument, to 19th November 2021 across Medline, Embase, SCOPUS, and Trip Pro database by a health sciences librarian (Supplemental Table 3). We reviewed papers written in Spanish, French, and Russian during the initial search, though none met criteria for full text review.

Study Inclusion and Exclusion Criteria

Inclusion criteria were: (1) an original sample of children \leq 48 months of age from a low-risk (ie, community) sample or high-risk (ie, children of siblings with ASD and premature infants) sample, (2) use of the 20-question M-CHAT-R/F questionnaire and clarifying interview and standard scoring protocol, (3) report of at least 1 psychometric property, and use of a diagnostic evaluation for diagnosis of ASD. Studies that reported only rates of the positive questionnaire screens without data from clarifying interviews were excluded. Studies with samples that overlapped other included studies in the meta-analysis were excluded (Fig 1). Two researchers (R.A. and V.K.M.) independently screened titles and abstracts of studies identified from the initial search. Discrepancies in decisions were discussed with a third researcher (H.M.F.) and consensus was reached following group discussion. Included articles then underwent full text review by the same independent process with all conflicts resolved by consensus following group discussion. We contacted corresponding authors of 8 studies via e-mail for additional information to clarify if the study met inclusion criteria; we requested raw data on screening numbers in 5 cases and details of the administration

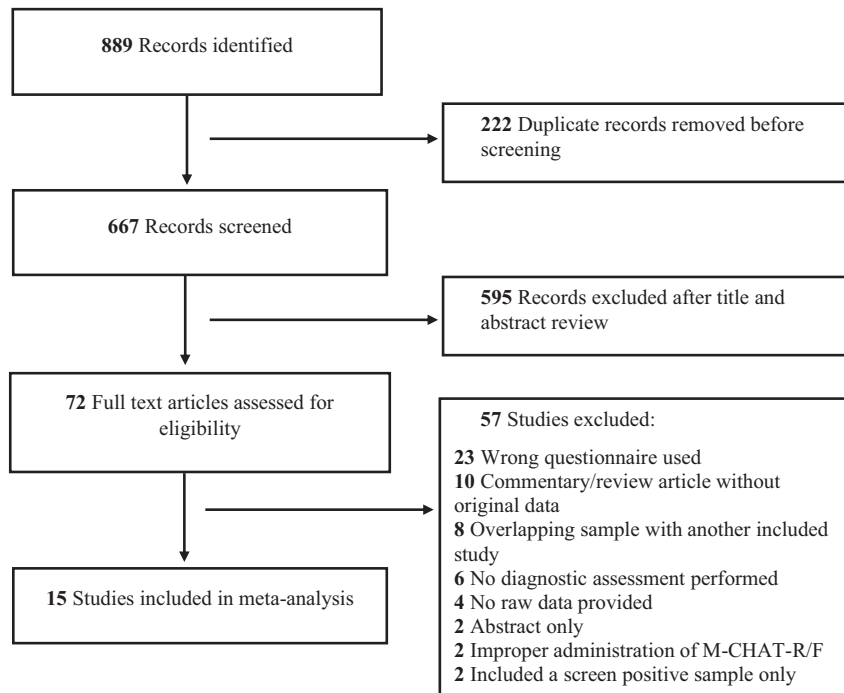


FIGURE 1
PRISMA diagram for study selection

process of the M-CHAT-R/F in 3 cases. We received responses from 5 corresponding authors.

Data Extraction

Data collected included: demographics, risk profile of children studied, country of administration, socioeconomic status, administrator and interviewer of the M-CHAT-R/F questionnaire, language used and translation process, diagnostic measure of ASD, results of screening and diagnostic assessments, and number of individuals lost to follow-up at any stage within each study.

Study Quality Assessment

The quality of studies in the final sample was evaluated based on the US Preventive Services Task Force framework (Table 1).²⁴ Ratings of “good,” “fair,” and “poor,” were assigned based on sample size, validity of diagnostic evaluation, diagnostic evaluation applied to at

least a proportion of the screen-negative participants, and administration of the screening and diagnostic evaluations by different personnel. When study attributes fell across 2 quality categories, the lower category was assigned. This rule was applied to 3 studies, which were ultimately rated “fair.” Quality ratings of each study were assigned

independently by 2 researchers (R.A. and V.K.M.) and discrepancies were resolved following discussion with a third researcher (H.M.F.).

Data Analysis

We applied a rigorous approach to the M-CHAT-R/F scoring criteria; to be counted as a screen-positive, the child must have been classified

TABLE 1 Quality Rating Definition

Quality	Characteristics
Good	Evaluates M-CHAT-R/F questionnaire with standard scoring protocol Reliable reference standard used as per DSM-5 criteria Independent assessor for reference standard and screening questionnaire Reliability of M-CHAT-R/F assessed Handles indeterminate results in a reasonable manner or has few such results Includes large number (greater than 1000) of subjects
Fair	Evaluates M-CHAT-R/F questionnaire with standard scoring protocol Uses reasonable although not best reference standard Independent assessor for reference standard and screening questionnaire Includes moderate number (500 to 1000) of subjects
Poor	Has a major flaw, such as: - uses inappropriate reference standard - improperly administers screening test - biased ascertainment of reference standard - has very small (<500) number of subjects

either as high-risk (score ≥ 8) or scored above the cut-off after the follow-up interview (post interview score ≥ 2). For each distinct study sample, PPV was estimated as the proportion of participants with a positive screen who were confirmed to have ASD on the diagnostic evaluation. In the primary analyses, for each distinct study sample in which all, or a random sample, of participants who screened negative were evaluated by the diagnostic evaluation, the NPV was estimated as the proportion of participants with a negative screen who were confirmed not to have ASD. If all evaluated participants with a given screening outcome had the same diagnostic outcome (eg, all negative screens were deemed negative by the diagnostic evaluation), a continuity correction of 0.5 was added to the numerator and 1.0 to the denominator of the proportion. The variance of each PPV and NPV estimate was computed as $p(1-p)/n$, where p is the proportion, and n is the denominator of the proportion; binomial confidence intervals were estimated as specified by Fleiss (1981).²⁵ Pooled estimates of overall PPV and for subgroups (risk-status, United States versus non-United States country and good or fair quality studies only), and NPV of selected study samples, were estimated using random effects models according to the method of DerSimonian and Laird.²⁶ Pooled estimates of sensitivity and specificity were similarly calculated.

In the primary analysis, screen-negative participants who did not have a diagnostic evaluation were assumed to be of unknown ASD status. We conducted secondary analyses for NPV and specificity, assuming that all screen-negative participants without a diagnostic evaluation did not have ASD (ie, true negatives).

The SAS macro METAANAL²⁷ was used to compute estimates and associated statistics: SE, Q, and I^2 (ie, τ^2), and forest plots were created. A meta-regression of PPV versus mean participant age was performed using a generalized linear mixed model specifying a binomial outcome with identity link and a random intercept for study sample. All analyses were performed with SAS Enterprise Guide version 7.15; a graph of PPV versus mean participant age was created with Excel.

RESULTS

Our search yielded 667 nonduplicate records. A total of 72 full-text articles were reviewed. A total of 15 studies,^{17,28–41} comprised of 18 distinct samples, with a total of 49 841 participants met full inclusion criteria. Two studies did not meet inclusion criteria as they did not administer the follow-up interview after an initial positive screen.

Study Characteristics

Table 2 shows the characteristics of the study sample from included studies (published between 2014 and 2021). The majority were of children 16 to 30 months of age; 2 studies^{33,37} included 14-month-old children and 3 studies included children up to 36 months^{30,33,34} (mean age of sample 23.3 months). Two studies reported results of screening of more narrow age-range subcohorts and their results were analyzed as 2 discrete samples.^{33,37} There was a good representation of international studies; 9 of 15 studies were conducted in countries other than the United States, including 4 in Asia,^{29,31,34,38} 3 in Europe,^{30,32,33} and 1 each in Africa²⁸ and South America.³⁹ The mean age of the sample from studies conducted in the United States was 22.3 months as compared with 23.9 months for that from other countries. All

international studies used the M-CHAT-R/F in languages other than English; forward and backward translation completion was specifically mentioned in 6 of the 9 studies.

Most of the studies were conducted solely on community samples (10 of 15). Two were conducted solely on a high-risk sample of siblings of children previously diagnosed with ASD.^{35,36} One study was conducted solely on children referred for concerns of ASD⁴⁰ and 2 studies included a mixture of a low-risk community sample and a high-risk sample of children with developmental concerns referred for further evaluation.^{38,39} Among the studies which included both a community and a high-risk sample, 1 did not report findings for each sample separately,³⁸ hence it was not included in the risk-based subgroup analysis. No other high-risk samples were included (eg, extreme premature infants). Thirteen of 15 studies applied a diagnostic evaluation for ASD on at least a portion of screen-negative patients, although this step was applied to the entire sample or a completely random screen-negative sample in only 5 studies. The remainder applied the diagnostic assessment on screen-negative children who had other concerns for an eventual ASD diagnosis (such as provider clinical concerns or failure of another screening test). Across all 15 included studies, only 780 screen-negative children had a diagnostic evaluation (1.67% of the total screen-negative sample). The majority of the studies (12 of 15) had a quality rating of “good” or “fair”; 3 studies had a “poor” rating. The reasons for a poor rating included lack of information on diagnostic assessments used, small sample size (<500 children), and same personnel administering both the M-CHAT-R/F and diagnostic assessment.

TABLE 2 Characteristics of Included Studies

Source and Country	Total Number	Age Range (Months)	Mean Age (SD) ^b	Males (%)	Type of Sample	Initial Interviewer	Follow up Interviewer	Language	Diagnostic Evaluation Method	Quality Indicator	Screen-Negative Evaluated
Guo et al ³¹ 2019; China	7928	16–30	22.7 (4.1)	55.6	Community	Healthcare	NS	Mandarin ^e	Clinical evaluation	Good	Some ^g
Jonsdottir et al ³² 2021; Iceland	1586	NS	31.7 (1.7)	50.5	Community	Healthcare	Research personnel	Icelandish ^e	ADOS	Good	Some ^g
Magano Maganto et al ³³ 2020; Spain	6625	14–36	14–22 mo: 18.2 (0.7); 23–36 mo: 24.5 (1.2)	51.2	Community	Healthcare	Research personnel, Healthcare	Spanish ^e	ADOS	Good	Some ^g
Oner et al ³⁴ 2020; Turkey	6712	16–36	26.8 (5.8)	51.5	Community	Healthcare	Healthcare	Turkish ^e	ADOS	Good	Some ^h
Robins et al ¹⁷ 2014; US	16071	16–30.9	21.0 (3.3)	50.12	Community	Healthcare	Research personnel	English	ADOS	Good	Some ^h
Wieckowski et al ^{37a} 2021; US	4281	14.0–21.9	NA	51.2 (15 mo); 49.3 (18 mo)	Community	Healthcare	NS	English	ADOS	Good	Some ^g
Bradbury et al ³⁵ 2020; US	187	16–30	21.2 (4.1)	58.4	High-risk ^c	Healthcare	Research personnel	English	ADOS	Fair	Some ^g
Brennan et al ³⁰ 2016; Albania	2594	16–36	24 (2.8)	50.1	Community	Healthcare	Research personnel	Albanian ^e	ADOS	Fair	None
Christopher et al ⁴⁰ 2021; US	290	18–48	32.7 (7.7)	79.6	High-risk	NS	NS	English	ADOS	Fair	All
McNally Keehn et al ⁴¹ 2021; US	605	18–48	30.4 (6.5)	NS	Community	Healthcare	Healthcare	English	Clinical evaluation	Fair	Some ⁱ
Tsai et al ³⁸ 2019; Taiwan	317	16–32	24.3 (4.4)	52.7	Community, High-risk ^c	Parent-completed via mail	NS	Mandarin ^e	Clinical evaluation	Fair	All
Weitlauf et al ³⁶ 2015; US	74	16–21	18.0 (NS)	63	High-risk ^c	NS	Research personnel	English	ADOS	Fair	Some ⁱ
Coelho-Medeiros et al ³⁹ 2019; Chile	120	16–30	22.47 (4.22)	64.2	Community high-risk ^c	Healthcare	Healthcare provider	Spanish ^f	ADOS	Poor	Some ^h
Manzouri et al ²⁹ 2019; Iran	1504	16–30	20.26 (3.74)	49.8	Community	Healthcare	Healthcare provider	Persian ^d	Clinical evaluation	Poor	None
Sangare et al ²⁸ 2019; Mali	947	NS	NS	NS	Community	Research	Research personnel	French ^f	Clinical evaluation	Poor	Some ⁱ

Healthcare provider refers to personnel at healthcare facility including but not limited to physicians and nurses. ADOS, Autism Diagnostic Observation Schedule; NS, not specified.

^a Distinct sample of 15- and 18-mo children screened with M-CHAT-R/F and rescreened at older ages; only results of initial screen included in analysis.

^b Ages and standard deviations rounded to the nearest tenth.

^c High risk samples include: siblings of children with ASD and children referred for developmental concerns for ASD.

^d Forward translated.

^e Forward and backward translated.

^f Translation not specified.

^g Evaluated based on screening tool results, physician concern, or referral for diagnostic assessment.

^h Evaluated based on random selection.

ⁱ Evaluation reason not specified.

Data on family socioeconomic status or parental education was not reported uniformly across studies, and thus was not adequate for analysis within the meta-analysis. There was significant loss

to follow-up of screen-positive children who did not have subsequent evaluations in each study (range 3.2% to 71.0%, mean 28.8%, median 32.3%). None of the studies compared children who were lost to

follow-up and those who received diagnostic assessment. Further, a portion of children in each study did not complete the full 2-stage screening process, with indeterminate screening results following a failed

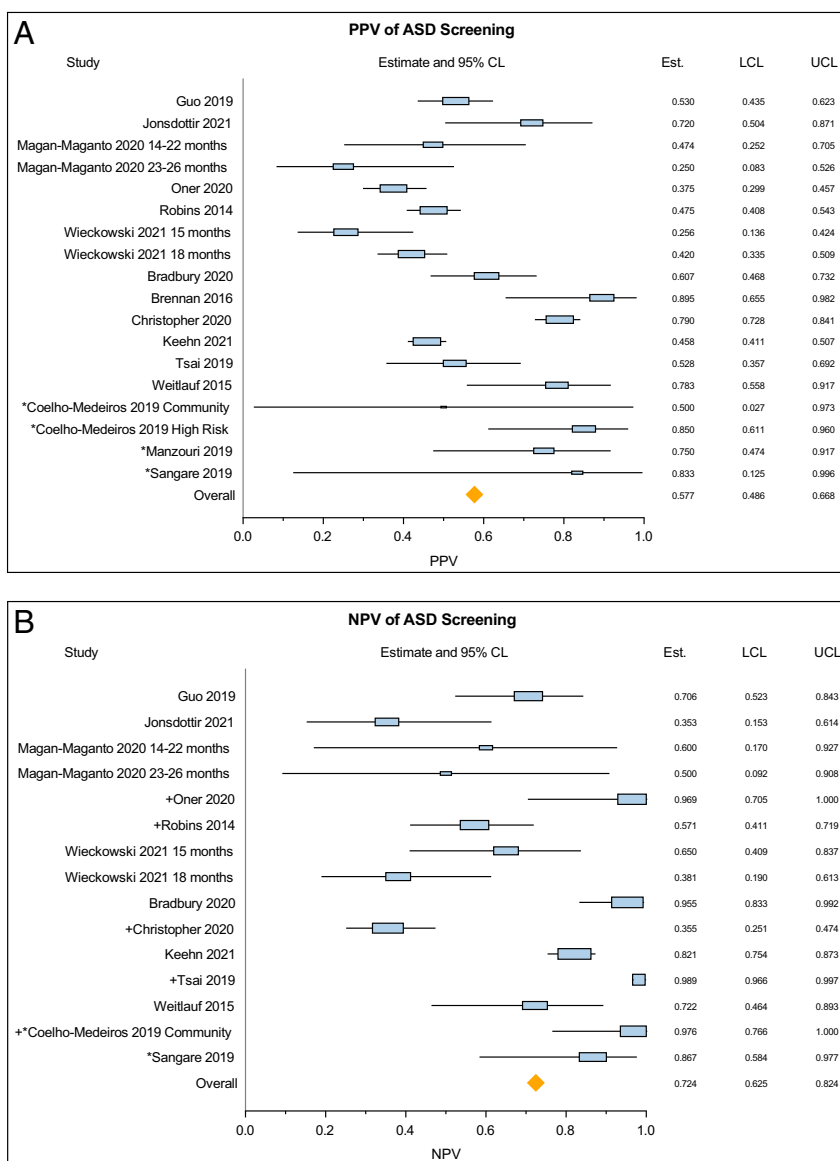


FIGURE 2

(A) Pooled estimates of positive predictive value of the M-CHAT-R/F questionnaire – all studies. (B) Pooled estimates of negative predictive value of the M-CHAT-R/F questionnaire. Studies are listed based on quality from “good” to “poor.” *, Denotes studies related as “poor” quality. +, Denotes studies that assessed a random sample of screen-negative children.

first stage (mean 25.7%, median 18.6%).

Overall Findings From All Studies

The meta-analysis of 15 studies covering 18 distinct samples showed a pooled PPV of 57.7% (95% CI 48.6–66.8) with significant between-study heterogeneity ($Q = 203.54, P < .0001, \tau^2 = 0.031$) (Fig 2A). The pooled NPV derived from the primary analysis of screen-negative

children who had a diagnostic evaluation (from 13 studies) was 72.5% (95% CI 62.5–82.4), with significant between-study heterogeneity ($Q = 280.27, P < .0001, \tau^2 = 0.031$) (Fig 2B). The pooled NPV based on the 5 studies that evaluated a random sample of screen-negative children was 78.0% (95% CI 59.5–96.5, $Q = 160.25, P < .0001$). In the secondary analysis (assuming that screen-

negative participants who did not have a diagnostic evaluation were true negatives) the pooled NPV was 99.7% (95% CI 99.6–99.9). The pooled sensitivity was 82.6% (95% CI 76.2–88.9), with significant between-study heterogeneity ($Q = 106.49, P < .0001, \tau^2 = 0.010$). Pooled specificity in the primary analysis was 45.7% (95% CI 25.0–66.4), with significant between-study heterogeneity ($Q =$

1532.70, $P < .0001$, $\tau^2 = 0.163$) (Figs 3, A and B). Pooled specificity in the secondary analysis was 97.1% (95% CI 96.4–97.9).

A meta-regression for PPV versus mean age of sample estimated an increase of 0.012 (SE 0.010) in PPV per month of mean age, but this trend was not statistically significant ($P = .23$) (Supplemental Fig 5). A funnel plot depicting number of screen-positive cases evaluated versus PPV was symmetrical suggesting low possibility of publication bias (Supplemental Fig 6). We calculated a pooled PPV for the presence of any developmental disorder (eg, global developmental delay or language delay) from 8 studies that provided these data. The PPV was higher at 89.0% (95% CI 82.9–94.6), with significant between-study heterogeneity ($Q = 60.21$, $P < .0001$). Because of inconsistent and/or absent reporting of these data across studies, including for screen-negative children, we could not derive sensitivity, specificity, or NPV for the presence of any developmental disorder.

Additional Findings From Subgroup Analyses

Subgroup analysis was conducted based on risk status and country of origin of the study (United States versus non-United States), as planned a priori. PPV of the high-risk sample based on 4 studies was 75.6% (95% CI 66.0–85.2) with no significant between-study heterogeneity ($Q = 57.75$, $P = .051$, $\tau^2 = 0.006$) (Supplemental Fig 7). In contrast, PPV for the low-risk sample was 51.2% (95% CI 43.0–59.5, $Q = 76.28$, $P < .0001$, $\tau^2 = 0.016$). Pooled PPV of studies conducted in the United States was lower than PPV of those conducted in other countries (PPV 54.0% [95% CI 40.2–67.7] versus 60.8% [95% CI 47.3–74.3]) (Fig 4). There was significant between-study

heterogeneity in both of these subgroups.

DISCUSSION

This study is a meta-analysis of the specific performance of the M-CHAT-R/F as a screening tool for ASD from a global sample that employed the tool in various languages.

The findings of our meta-analysis, from 10 different countries and using 8 different languages, including English, showed an overall pooled sensitivity of the M-CHAT-R/F at 82.5%. Sensitivity is above the recommended 70% to 80% expected of a good screening tool.⁴² Thus, the proportion of young children with ASD who are missed by screening with the M-CHAT-R/F is small. Some children with ASD who have subtle social communication difficulties may not be detected by screening before 36 months of age.

Our results reveal a pooled estimated PPV of 57.7% for the subsequent diagnosis of ASD following a positive screen. In a low-risk community sample, for whom the tool is most appropriately used, the chance of an eventual ASD diagnosis following a positive M-CHAT-R/F screen decreases to 51.2%. This finding is similar to that reported in the original validation study of the M-CHAT-R/F of 46.5%¹⁷ and to that of the original M-CHAT.²¹ It is also in line with the findings of the US Preventive Services Task Force that looked at both the M-CHAT and M-CHAT-R/F.⁴³ The pooled estimate of the PPV of the M-CHAT-R/F in relation to the presence of any developmental disorder, such as global developmental delay, increases to 89.0%, in keeping with previous studies.^{17,33,36}

Real-world performance of a screening tool must consider PPV and NPV. The pooled PPV estimate that

we derived is in the moderate range at 57.5%. The PPV for ASD is relevant when counseling parents of children with a positive screen with respect to the likelihood of an eventual ASD diagnosis; only approximately half of those children with a positive screen will ultimately receive the diagnosis of ASD; the other half may have no or another developmental disorder. Many factors affect PPV. PPV is related to the population prevalence of the condition; lower prevalence is associated with lower PPV, regardless of sensitivity.⁴⁴ Given the prevalence rate of ASD at 1% to 2.3%, even a screening tool with 99% sensitivity would result in a PPV of 67% at best.⁴⁴ In addition, given that the M-CHAT-R/F is a parent-reported screening tool, lack of parental awareness of expected social-communication milestones may lead to false responses on the M-CHAT-R/F, compromising PPV. PPV rates would also be influenced by screen-positive children who were lost to follow-up; these children may have sought diagnostic evaluation elsewhere, lowering the reported PPV by the study. The moderate PPV emphasizes the need for further evaluations for the presence of ASD and/or any other developmental conditions rather than reliance on the results of this screening tool for diagnosis. Repeated screening of the same child at multiple timepoints, such as at 18 and 24 months as recommended by the American Academy of Pediatrics, may improve the accuracy of ASD screening,³⁷ though this recommendation bears further evaluation. A positive screening result may also indicate a developmental disorder other than ASD. The diagnostic evaluation after a positive screening test should consider a range of diagnostic possibilities in addition to ASD.

Our finding of a higher PPV for studies originating from non-United States countries compared with

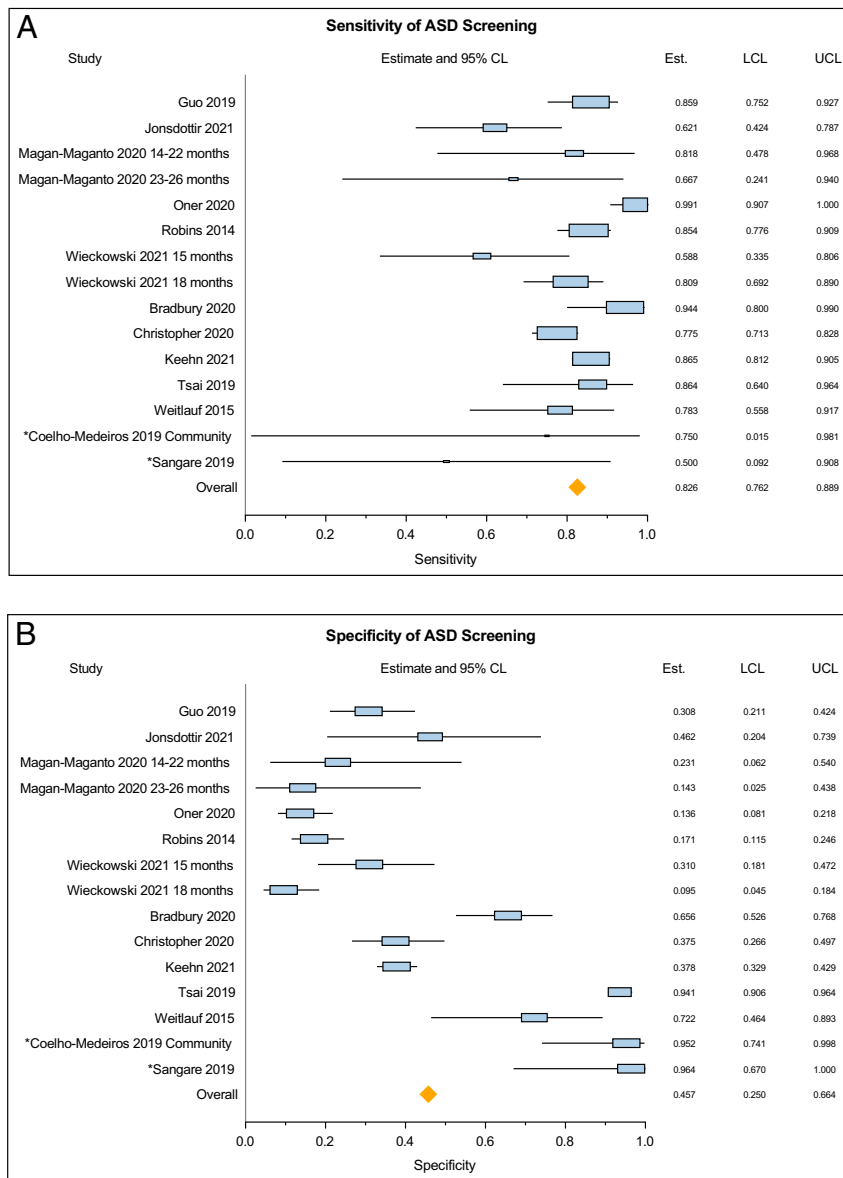


FIGURE 3

(A) Pooled estimates of sensitivity of the M-CHAT-R/F questionnaire. (B) Pooled estimates of specificity of the M-CHAT-R/F questionnaire.

those within the United States is surprising. This finding could have been related to the varying prevalence of ASD within each country given the variation within global prevalence rates.^{2,45} Further, when using the questionnaire in different languages, the translation process could have unintentionally altered the phrasing of the questions or interpretation of the questions could have been influenced by

culture and local contexts, such that the nature of a positive screen is different from that when used in English. Another possibility could also be related to the older mean age of the sample from non-United States studies as compared with those from the United States, with the PPV of the M-CHAT-R/F being better among older children. However, the meta-regression found that the increase in PPV with age of

sample was not statistically significant. It is also possible that within each country, a selection bias for being screened is at play such that children who are at higher risk for ASD are inadvertently more likely to participate in research studies on screening. We also note that of 9 international studies that were included, only 4 had “good” quality ratings, 2 had “fair,” and 3 had “poor” quality ratings. The

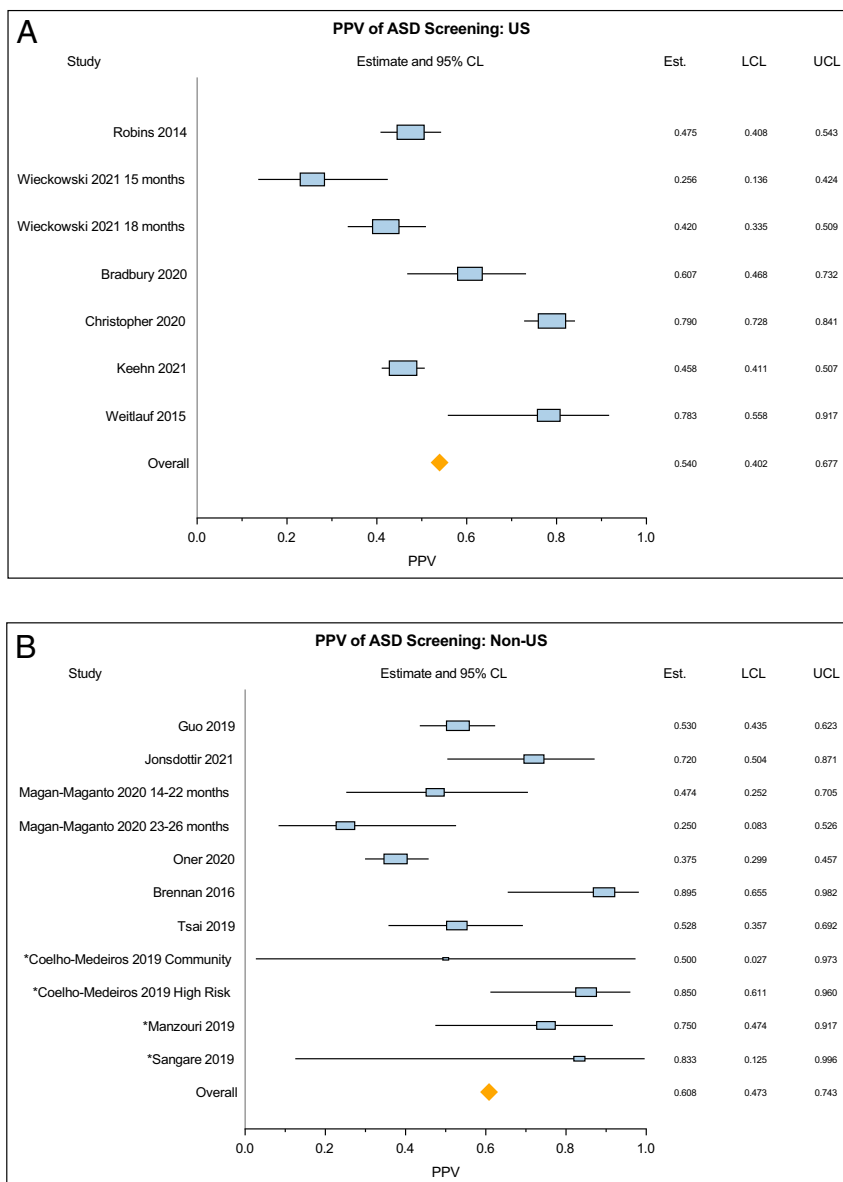


FIGURE 4

Pooled estimates of positive predictive value of the M-CHAT-R/F questionnaire – stratified by country of origin of studies (United States versus non-United States).

studies that received lower ratings had serious methodological flaws, including small sample sizes or the same personnel administering both the M-CHAT-R/F and the diagnostic evaluation. Hence, the PPV rate could potentially have been artificially inflated because of study-related biases. However, in the overall sample, elimination of studies rated as “poor” quality did not significantly alter the pooled PPV.

The pooled NPV of 72.5% in the primary analyses indicates the need for a screening tool for ASD with a smaller proportion of false negative screens. We recognize that the estimates for NPV, sensitivity, and specificity were based on the small proportion of screen-negative children who had a diagnostic evaluation. When we assumed that all nonevaluated screen-negative children did not have ASD, the NPV unsurprisingly increased to 99.7%.

However, we are concerned about the appropriateness of assuming that nonevaluated screen-negative children were condition-free. The assumption would lead to significant overestimation of the true NPV, even if the pooled estimate based on data of screen-negative children who had a diagnostic evaluation may be an underestimate based on the small sample size. This difference in analytic approach explains why 3 of the studies^{17,31,32}

included, reported a higher NPV in their manuscript as compared with what we used in our meta-analysis. Conversely, 4 of the studies^{28,36,38,40} used the same analytical method as we did in the primary analysis and these studies had reported NPV values that were similar (lower) to those that we derived. This limitation regarding calculation of NPV is inherent in almost all studies examining screening tools in developmental disorders because of the large numbers of children who have negative screening tests, the financial costs involved in testing screen-negative children, and the long time-frame within which the disorder could potentially develop.^{44,46-48} The other unanticipated limitation was that even though the majority (13 of 15) of the final included studies applied the diagnostic assessment to a sample of screen-negative patients, the majority of these patients had bias toward an eventual positive diagnosis (ie, it was a not a randomly selected sample). This bias likely contributed to the widely varying estimates of NPV across studies. Cultural factors influencing parental reporting of developmental milestones may also affect NPV rates. For example, parents may over-report the presence of developmental skills, leading to false negative screens because of cultural stigma associated with developmental delays. Lack of understanding of the screening questions because of language or translation-related difficulties may also lead to false responses. Overall, the true NPV is likely to lie somewhere in between 72.5% and 78.0% and closer to 78.0%, which was estimated based on a less-biased, random sample.

Future studies of screening measures for ASD or other developmental disorders should evaluate a larger

and random sample of children who screen-negative to ensure that a nonbiased subsample has the same reference test applied as screen-positive children. This improvement would be important to allow reliable interpretation of the tool's performance, especially in terms of NPV, sensitivity, and specificity. It would also be ideal to have more detailed reporting with respect to the presence of other developmental disorders apart from ASD, if any, following diagnostic evaluations of both screen-positive and screen-negative children. This calculation would allow greater inference of the utility of the M-CHAT-R/F for the presence of general developmental disorders apart from ASD. The substantial number of children who were screen-positive but were lost to follow-up and those who did not complete the full 2 stage screening also add caution to the overall estimates derived in our meta-analysis. The reason for these high attrition rates within each study context were not fully discussed in the studies. However, possible reasons include that the families of children who screened positive were able to access services for ASD without a diagnostic evaluation or alternatively, that caregivers were reluctant to continue follow-up evaluations following an initial positive screen for ASD for cultural or other reasons. This selective sample loss and the individual study's handling of this sample could have distorted estimates of PPV across studies. For example, a previous study that examined a previous version of the M-CHAT in a real-world setting assumed that all screen-positive children who were lost to follow-up did not have ASD, affecting its PPV estimate.⁴⁹

Strengths and Limitations

Strengths of this meta-analysis include the intentional focus solely on the M-CHAT-R/F, the preferred screening tool for ASD screening in many healthcare settings. By calculating pooled sensitivity, specificity, and NPV based on screen-negative children who received a diagnostic evaluation, the estimates provided here are likely accurate, and not an over-estimation of the true values. The diverse nature of the samples allowed us to capture the performance of the M-CHAT-R/F in countries and languages that the questionnaire was not originally developed or normed. We were also able to obtain relevant data from studies that employed the M-CHAT-R/F for purposes other than the detection of ASD but included data pertinent to ASD screening.

Limitations include the lack of a manual search of conference papers and proceedings, although we tried to circumvent this limitation by reviewing the references of all included studies to minimize potentially missed ones. Although we did not intentionally exclude any papers, it is possible that those without an abstract or keywords in English were not identified in the systematic search. Lastly, we were not able to analyze results based on family socioeconomic status and demographic variables; as a result, findings may not be generalizable across all populations of children within the studied age ranges.

CONCLUSIONS

This meta-analysis supports the use of the M-CHAT-R/F as a screening questionnaire for ASD, given favorable sensitivity and acceptable PPV in the context of a low prevalence condition. A positive screen with the M-CHAT-R/F is predictive of the diagnosis of ASD in approximately 50% of children and

a predictive diagnosis of any developmental disorder in about 90%. The PPV should be considered in the clinical application of the M-CHAT-R/F, particularly when counseling caregivers with respect to the presence of ASD. Future research should evaluate a substantial, random sample of children who screen-negative to increase the accuracy of estimates of NPV, sensitivity, and specificity. Studies examining the performance of the M-CHAT-R/F with outcome

data from longitudinal follow-up of a cohort of screen-positive and screen-negative children will improve our assessment of the M-CHAT-R/F as a screening tool for ASD.

ACKNOWLEDGMENTS

We thank Bruce T Abbott, MLS, the health sciences librarian who conducted the systematic search; and Robin Hansen, MD for critically reviewing the manuscript before submission.

ABBREVIATIONS

ASD: autism spectrum disorder
M-CHAT-R/F: Modified Checklist for Autism in Toddlers, Revised with Follow-Up
NPV: negative predictive value
PPV: positive predictive value

critically reviewed and revised versions of the manuscript; and all authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DOI: <https://doi.org/10.1542/peds.2022-059393>

Accepted for publication Feb 27, 2023

Address correspondence to Ramkumar Aishworiya, MMed, Consultant, Khoo Teck Puat – National University Children's Medical Institute, National University Health System, 5 Lower Kent Ridge Rd, Singapore 119074. E-mail: paearam@nus.edu.sg

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275).

Copyright © 2023 by the American Academy of Pediatrics

FUNDING: The project described was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR001860. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Partial salary support was received for VKM from T77MC25733 and P50HD103526, and for HMF from T77MC09796 from the Maternal Child Health Bureau Grant, Health Resources and Services Administration.

CONFLICT OF INTEREST DISCLOSURES: The authors have indicated they have no conflicts of interest relevant to this article to disclose.

REFERENCES

1. Association D-AP. *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, TX: American Psychiatric Publishing; 2013
2. Zeidan J, Fombonne E, Scorah J, et al. Global prevalence of autism: a systematic review update. *Autism Res*. 2022;15(5):778–790
3. Maenner MJ, Shaw KA, Bakian AV, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2018. *MMWR Surveill Summ*. 2021;70(11):1–16
4. Ben Itzhak E, Zachor DA. Who benefits from early intervention in autism spectrum disorders? *Res Autism Spectr Disord*. 2011;5(1):345–350
5. Virues-Ortega J, Rodríguez V, Yu CT. Prediction of treatment outcomes and longitudinal analysis in children with autism undergoing intensive behavioral intervention. *Int J Clin Health Psychol*. 2013;13(2):91–100
6. Granpeesheh D, Dixon DR, Tarbox J, Kaplan AM, Wilke AE. The effects of age and treatment intensity on behavioral intervention outcomes for children with autism spectrum disorders. *Res Autism Spectr Disord*. 2009;3(4):1014–1022
7. Cicchetti D. *Developmental Psychopathology, Developmental Neuroscience*. New York, NY: John Wiley and Sons; 2016
8. Dawson G. Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Dev Psychopathol*. 2008;20(3):775–803
9. Hyman SL, Levy SE, Myers SM. Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics*. 2020;145(1):e20193447
10. Zwaigenbaum L, Bryson SE, Brian J, et al. Stability of diagnostic assessment for autism spectrum disorder between 18 and 36 months in a high-risk cohort. *Autism Res*. 2016;9(7):790–800
11. Pierce K, Gazestani VH, Bacon E, et al. Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. *JAMA Pediatr*. 2019;173(6):578–587

12. Zwaigenbaum L, Bauman ML, Choueiri R, et al. Early intervention for children with autism spectrum disorder under 3 years of age: recommendations for practice and research. *Pediatrics*. 2015;136 Suppl 1(Suppl 1):S60–S81
13. Marlow M, Servili C, Tomlinson M. A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: recommendations for use in low- and middle-income countries. *Autism Res*. 2019; 12(2):176–199
14. Campbell K, Carpenter KL, Espinosa S, Hashemi J, Qiu Q, Tepper M, et al. Use of a digital modified checklist for autism in toddlers—revised with follow-up to improve quality of screening for autism. *J Pediatr*. 2017;183:133–139. e131
15. Zwaigenbaum L, Brian JA, Ip A. Early detection for autism spectrum disorder in young children. *Paediatr Child Health*. 2019;24(7):424–443
16. Fuentes J, Hervás A, Howlin P; ESCAP ASD Working Party. ESCAP practice guidance for autism: a summary of evidence-based recommendations for diagnosis and treatment. *Eur Child Adolesc Psychiatry*. 2021;30(6):961–984
17. Robins DL, Casagrande K, Barton M, Chen C-MA, 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
18. Chlebowski C, Robins DL, Barton ML, Fein D. Large-scale use of the modified checklist for autism in low-risk toddlers. *Pediatrics*. 2013;131(4): e1121–e1127
19. Petrocchi S, Levante A, Lecciso F. Systematic review of level 1 and level 2 screening tools for autism spectrum disorders in toddlers. *Brain Sci*. 2020; 10(3):180
20. Levy SE, Wolfe A, Coury D, et al. Screening tools for autism spectrum disorder in primary care: a systematic evidence review. *Pediatrics*. 2020;145(Suppl 1): S47–S59
21. Yuen T, Penner M, Carter MT, Szatmari P, Ungar WJ. Assessing the accuracy of the modified checklist for autism in toddlers: a systematic review and meta-analysis. *Dev Med Child Neurol*. 2018;60(11):1093–1100
22. Wieckowski AT, Williams LN, Rando J, Lyall K, Robins DL. Sensitivity and specificity of the modified checklist for autism in toddlers (Original and Revised): a systematic review and meta-analysis. *JAMA Pediatr*. 2023; e225975
23. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372(71):n71
24. US Preventative Services Task Force. *Procedure Manual*. Rockville, MD: US Preventative Services Task Force; 2008
25. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. New York, NY: John Wiley and Sons; 2013
26. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177–188
27. Hertzmark E, Spiegelman D. *The SAS METAANAL Macro*. Boston, MA: Channing Laboratory; 2012:24
28. Sangare M, Toure HB, Toure A, et al. Validation of two parent-reported autism spectrum disorders screening tools M-CHAT-R and SCQ in Bamako, Mali. *eNeurologicalSci*. 2019;15:100188
29. Manzouri L, Yousefian S, Keshtkari A, Hashemi N. Advanced parental age and risk of positive autism spectrum disorders screening. *Int J Prev Med*. 2019;10:135
30. Brennan L, Fein D, Como A, Rathwell IC, Chen CM. Use of the Modified Checklist for Autism, Revised with Follow Up-Albanian to screen for ASD in Albania. *J Autism Dev Disord*. 2016;46(11):3392–3407
31. Guo C, Luo M, Wang X, et al. Reliability and validity of the Chinese version of Modified Checklist for Autism in Toddlers, Revised, with Follow-Up (M-CHAT-R/F). *J Autism Dev Disord*. 2019;49(1):185–196
32. 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. *J Autism Dev Disord*. 2022;52(4):1507–1522
33. Magán-Maganto M, Canal-Bedia R, Hernández-Fabián A, et al. Spanish cultural validation of the Modified Checklist for Autism in Toddlers, Revised. *J Autism Dev Disord*. 2020;50(9):3312–3319
34. Oner O, Munir KM. Modified Checklist for Autism in Toddlers Revised (MCHAT-R/F) in an urban metropolitan sample of young children in Turkey. *J Autism Dev Disord*. 2020;50(9):3312–3319
35. Bradbury K, Robins DL, Barton M, et al. Screening for autism spectrum disorder in high-risk younger siblings. *J Dev Behav Pediatr*. 2020;41(8):596–604
36. Weitlauf AS, Vehorn AC, Stone WL, Fein D, Warren ZE. Using the M-CHAT-R/F to identify developmental concerns in a high-risk 18-month-old sibling sample. *J Dev Behav Pediatr*. 2015;36(7):497–502
37. Wieckowski AT, Hamner T, Nanovic S, et al. Early and repeated screening detects autism spectrum disorder. *J Pediatr*. 2021;234:227–235
38. Tsai JM, Lu L, Jeng SF, et al. Validation of the modified checklist for autism in toddlers, revised with follow-up in Taiwanese toddlers. *Res Dev Disabil*. 2019;85:205–216
39. Coelho-Medeiros ME, Bronstein J, Aedo K, et al. M-CHAT-R/F Validation as a screening tool for early detection in children with autism spectrum disorder. *Rev Chil Pediatr*. 2019;90(5):492–499
40. Christopher K, Bishop S, Carpenter LA, Warren Z, Kanne S. The implications of parent-reported emotional and behavioral problems on the Modified Checklist for Autism in Toddlers. *J Autism Dev Disord*. 2021;51(3): 884–891
41. McNally Keehn R, Tang Q, Swigonski N, Ciccarella M. Associations among referral concerns, screening results, and diagnostic outcomes of young children assessed in a statewide Early Autism Evaluation Network. *J Pediatr*. 2021;233:74–81.e8
42. Council on Children With Disabilities; Section on Developmental Behavioral Pediatrics; Bright Futures Steering Committee; Medical Home Initiatives for Children With Special Needs Project Advisory Committee. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*. 2006;118(1):405–420

43. McPheeters ML, Weitlauf A, Vehorn A, 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; 2016
44. McCarty P, Frye RE. *Early Detection and Diagnosis of Autism Spectrum Disorder: Why is it so Difficult? Seminars in Pediatric Neurology.* New York, NY: Elsevier; 2020:100831
45. Fombonne E. *The Rising Prevalence of Autism.* New York, NY: Wiley Online Library; 2018:717–720
46. Glascoe FP, Foster EM, Wolraich ML. An economic analysis of developmental detection methods. *Pediatrics.* 1997; 99(6):830–837
47. Randall M, Egberts KJ, Samtani A, Scholten RJ, Hooft L, Livingstone N, et al. Diagnostic tests for autism spectrum disorder (ASD) in preschool children. *Cochrane Database Syst Rev.* 2018;7(7):CD009044
48. Robins DL. How do we determine the utility of screening tools? *Autism.* 2020; 24(2):271–273
49. Guthrie W, Wallis K, Bennett A, et al. Accuracy of autism screening in a large pediatric network. *Pediatrics.* 2019;144(4):e20183963