Supplementary Online Content

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*. Published online February 20, 2023. 10.1001/jamapediatrics.2022.5975

eMethods. Diagnostic Accuracy of the M-CHAT(-R/F)

eTable 1. Database Search Terms

eTable 2. QUADAS-2 Description and Adapted Signaling Questions

eTable 3. Quality Assessment of Studies Included in the Systematic Review

eTable 4. Additional Study Characteristics and Psychometric Properties for M-CHAT(-R/F)

eTable 5. Sensitivity and Specificity for Younger and Older Samples

eTable 6. Sensitivity and Specificity for Single and Repeated Screening

eFigure 1. Study Selection Flow Chart Following PRISMA Guidelines

eFigure 2. Overall SROC of M-CHAT(-R/F) (n = 49 Studies)

eFigure 3. SROC Plot of M-CHAT(-R/F) by M-CHAT Version (M-CHAT n = 31, M-CHAT-R n = 18)

eFigure 4. SROC Plot of M-CHAT(-R/F) by Likelihood Level of Sample (Low Likelihood n = 27, High Likelihood n = 22)

eFigure 5. SROC Plot of M-CHAT(-R/F) by Case Confirmation Strategy (Concurrent n = 40, Prospective n = 9)

eFigure 6. SROC Plot of M-CHAT(-R/F) by Sample Size (Small n = 20, Medium n = 17, Large n = 12)

eFigure 7. SROC Plot of M-CHAT(-R/F) With Follow-up vs Initial Only (Initial n = 22, Follow-up n = 27)

eFigure 8. SROC Plot of M-CHAT(-R/F) by Language (English/Primarily English n = 26, Other Language n = 23)

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Diagnostic Accuracy of the M-CHAT(-R/F)

The overall diagnostic accuracy of the M-CHAT(-R/F) was assessed using the hierarchical summary receiver operating characteristic (HSROC) model.[18,19] This model was chosen for its ability to account for the inherent relationship between sensitivity and specificity, as well as high heterogeneity of the sample (as is often the case with diagnostic test accuracy studies) through use of a Bayesian model to determine random effects. HSROC models were run on included studies (n=49) using the MetaDAS SAS macro[20], which outputs HSROC parameters, the diagnostic odds ratio (DOR; an overall estimate of diagnostic test accuracy that can be used to compare across tests and models), and pooled sensitivity and specificity. The HSROC model was run with and without covariates (added individually) to assess if these study characteristics affected the accuracy, threshold, or shape of the HSROC curve. Covariates included the ASD likelihood level of sample (low-likelihood, high-likelihood), case confirmation strategy classification (concurrent, prospective), sample size (<500, 500-5000, >5000), M-CHAT(-R) version (M-CHAT, M-CHAT-R), use of structured Follow-Up (Follow-Up, initial screen only), and language (English/ primarily English, other language). When study characteristics indicated more than one of these categories, they were re-classified based on predominant data. Four studies that reported mixed-likelihood level were reclassified as low-likelihood, as a large majority of participants were low-risk.[25,32,58,66] In addition, two studies reported mix of initial screen and use of Follow-Up interview. One study[8] was classified as initial screen only, as only 1 out of the 12 practices completed the Follow-Up in the standardized interview form. The other study[9] was classified as using structured Follow-Up, since Follow-Up was built into the electronic health record, and intended to be used when indicated based on initial score; 41% of expected sample received the Follow-Up interview, consistent with other studies that attempted to administer the interview but were not always successful. Three studies classified as "mixed" or "unknown" in any category were excluded from the analyses[62,68]. The HSROC parameters output by each model were input into RevMan 5 software to create the HSROC summary curves.

eTable 1. Database Search Terms

Database	Search terms
PubMed	((M-CHAT*[Title/Abstract]) OR (MCHAT*[Title/Abstract]) OR ("Modified Checklist for Autism in Toddlers"[Title/Abstract]) OR ("Modified-Checklist for Autism"[Title/Abstract]) OR ((screen*[Title/Abstract] AND autis*[Title/Abstract] AND toddler*)[Title/Abstract]))) AND ((Autis*[Title/Abstract]) OR (Asperger*[Title/Abstract]) OR (ASD[Title/Abstract]) OR (PDD[Title/Abstract]) OR ("Pervasive Developmental Disorder*"[Title/Abstract])) AND ("2001/01/01"[Date - Publication] : "3000"[Date - Publication])
Web of Science	AB=((M-CHAT*) OR (MCHAT*) OR ("Modified Checklist for Autism in Toddlers") OR ("Modified-Checklist for Autism") OR ((screen* AND autis* AND toddler*))) AND AB=((Autis*) OR (Asperger*) OR (ASD) OR (PDD) OR ("Pervasive Developmental Disorder*"))) AND LANGUAGE: (English); Timespan: 2001-2022
SCOPUS	TITLE-ABS-KEY((M-CHAT*) OR (MCHAT*) OR ("Modified Checklist for Autism in Toddlers") OR ("Modified-Checklist for Autism") OR (screen* AND autis* AND toddler*)) AND TITLE-ABS-KEY((Autis*) OR (Asperger*) OR (ASD) OR (PDD) OR ("Pervasive Developmental Disorder*")) AND (PUBYEAR > 2001)

eTable 2. QUADAS-2 Description and Adapted Signaling Questions

Domain 1: Patient Selection	QUADAS-2 description and signaling questions adapted for this study
Risk of Bias	**Could the selection of patients have introduced bias?** 1. Were participants randomly selected? 2. Were all exclusion appropriate and defined a priority?
Applicability	**Are there concerns that the included patients and setting do not match the review question?** *Any such studies would have been excluded for the purpose of this review, and therefore we have no applicability concerns in this area
Domain 2: Index Test	
Risk of Bias	**Could the conduct or interpretation of the index test have introduced bias?** 1. Was interpretation of M-CHAT(-R/F) done prior to knowing child's diagnosis? 2. Were standardized M-CHAT(-R/F) methods used? 3. Was an appropriate threshold used to indicate risk?
Applicability	**Are there concerns that the index test, its conduct, or its interpretation differ from the review question?** 1. Was correct criteria used for scoring? 2. Was correct threshold used to calculate sensitivity and specificity?
Domain 3: Reference Standard	
Risk of Bias	**Could the reference standard, its conduct, or its interpretation have introduced bias?** 1. Was provider who gave diagnosis blind to M-CHAT(-R/F) score? 2. Was appropriate ASD measure used to diagnose ASD (i.e., ADOS, ADI-R, or CARS), or was provider qualified to give ASD diagnosis? * If community diagnosis given and no additional information was used, marked as unclear bias.
Applicability	**Are there concerns that the target condition as defined by the reference standard does not match the question?** 1. Were all participants identified as having ASD diagnosed with ASD as based on DSM or ICD criteria? 2. Was appropriate criteria for diagnosis utilized?
Domain 4: Flow and Timing	
Risk of Bias	**Could the patient flow have introduced bias?** 1. Did all participants receive the same reference standard? 2. Was time between screen and diagnosis within 1 year? 3. Were all recruited participants screened with M-CHAT(-R/F) and evaluated for ASD?

eTable 3. Quality Assessment of Studies Included in the Systematic Review

Reference	Risk of Bias				Concerns of Applicability		
	Participant Selection	Index Test	Reference Standard	Flow and Timing	Participant Selection	Index Test	Reference Standard
Baduel et al,[22] 2017	Low	Low	Low	High	Low	Low	Low
Beacham et al,[42] 2018	Low	Low	Unclear	Low	Low	Low	Low
Canal-Bedia et al,[58] 2011	Low	Low	Low	High	Low	Low	Low
Carbone et al,[8] 2020	Low	Low	Unclear	High	Low	Low	Low
Chang et al,[23] 2021	Low	Low	Low	High	Low	Low	Low
Charman et al,[43] 2016	Low	Low	Low	High	Low	Low	Low
Chlebowski et al,[24] 2013	Low	Low	Low	High	Low	Low	Low
Choueiri et al,[44] 2021	Low	High	Low	Low	Low	Low	Low
Christopher et al,[45] 2020	Low	Low	Low	Low	Low	Low	Low
Coelho-Medeiros et al,[25] 2019	High	Low	High	High	Low	Low	Low
Dereu et al,[26] 2012	Low	Low	Low	High	Low	Low	Low
DiGuiseppi et al,[27] 2010	Low	Low	Low	High	Low	Low	Low
Dudova et al,[28] 2014	Low	High	Low	High	Low	Low	Low
Eaves et al,[46] 2006	Low	High	Low	Low	Low	Low	Low
Guo et al,[29] 2019	Low	Low	Low	High	Low	Low	Low
Guthrie et al,[9] 2019	Low	Low	Unclear	High	Low	Low	Low
Harris et al,[30] 2021	Low	Low	Low	High	Low	Low	Low
Hoang et al,[31] 2019	Low	Low	Low	High	Low	Low	Low
Inada et al,[64] 2011	Low	Low	Low	High	Low	Low	Low
Jonsdottir et al,[60] 2021	Low	Low	High	High	Low	Unclear	Low
Kamio et al,[65] 2014	Low	High	Low	High	Low	High	Low
Kanne et al,[47] 2018	Low	Low	High	Low	Low	Low	Low
Kara et al,[32] 2014	Low	Low	Low	High	Low	Low	Low
Keehn et al,[48] 2021	Low	Low	High	Low	Low	Low	Low
Kerub et al,[61] 2020	Low	High	Low	High	Low	Low	Low
Kim et al,[67] 2016	Low	Low	Low	High	Low	Low	Low
Kleinman et al,[66] 2008	Low	Low	Low	High	Low	Low	Low

Koh et al,[49] 2014	Low	Low	Low	High	Low	Low	Low
Magan-Maganto et al,[33] 2020	Low	Low	Low	High	Low	Low	Low
Matson et al,[50] 2013	Low	Low	Low	Low	Low	Low	Low
Oner et al,[59] 2020	Low	Low	Low	High	Low	Low	Low
Robins et al,[34] 2014	Low	Low	Low	High	Low	Low	Low
Salim et al,[51] 2020	Low	Low	Unclear	Unclear	Low	Low	Low
Salisbury et al,[52] 2018	Low	Low	Low	Low	Low	Low	Low
Samadi et al,[35] 2015	Low	Low	Low	High	Low	Low	Low
Schjolberg et al,[63] 2022	Low	Low	Low	High	Low	Low	Low
Smith et al,[53] 2013	Low	Low	Low	High	Low	Low	Low
Snow et al,[54] 2008	Low	Low	Low	Low	Low	Low	Low
Srisinghasongkram et al,[21] 2016 (Sample 1)	Low	Low	Low	Low	Low	Low	Low
Srisinghasongkram et al,[21] 2016 (Sample 2)	Low	Low	Low	High	Low	Low	Low
Sturner et al,[36] 2016	Low	Low	Low	High	Low	Low	Low
Sturner et al,[39] 2022	Low	Low	Low	High	Low	Low	Low
Taylor et al,[55] 2014	Low	Low	Low	Low	Low	Low	Low
Toh et al,[62] 2018	Low	Low	Low	High	Low	Low	Low
Tsai et al,[68] 2019	Low	Low	Low	Low	Low	Low	Low
Vui et al,[40] 2022	Low	Low	Low	High	Low	Low	Low
Weitlauf et al,[37] 2015	Low	Low	Low	High	Low	Low	Low
Wieckowski et al,[38] 2021	Low	Low	Low	High	Low	Low	Low
Windiani et al,[56] 2016	Low	Low	Unclear	Low	Low	Low	Low
Wong et al,[57] 2018	Low	Low	Low	Low	Low	Low	Low
Zhang et al,[41] 2022	Low	Low	Low	High	Low	Low	Low

Note. Low: Low concern; High: High concern; Unclear: concern is unclear.

eTable 4. Additional Study Characteristics and Psychometric Properties for M-CHAT(-R/F)

Reference	Screen Age [a]	Eval. Age [a]	Sample description	Study Location	M-CHAT Version [b]	Single/ Repeat [c]	FN Strategy Description [d]	Spec. Original [e]	Spec. New [f]
Baduel et al,[22] 2017	24.2 (0.6); 22.2-26.0	25.1 (1.9); 24-34	Primary care and daycare population	France	M-CHAT/F	Single	Use of second screener	0.99	0.993
Beacham et al,[42] 2018	27.8 (6.6); 16-45	27.8 (6.6); 16-45	High likelihood for ASD	US	M-CHAT-R	Single	All evaluated	0.533	0.533
Canal-Bedia et al,[58] 2011	18–36	18-48	Primary care population and EI centers / psychiatric units	Spain	M-CHAT/F	Single	All HL children evaluated	0.98	0.980[g]
Carbone et al,[8] 2020	16-30	46.8 (17.7) [h]	Primary care population	US	M-CHAT/F	Repeat	Medical record review	0.978	0.978
Chang et al,[23] 2021	17-37	17-37	Primary care population	US	M-CHAT-R/F	Single	Physician or caregiver concern	0.988	0.988
Charman et al,[43] 2016	35.2 (8.3); 18–56	51.6 (8.8); 32–73	High likelihood for developmental concerns	London	M-CHAT	Single	All evaluated	0.5	0.500
Chlebowski et al,[24] 2013	20.4 (3.1); 16-30	25.8 (4. 5)	Primary care population	US	M-CHAT/F	Repeat	Use of second screener after concern	0.995	0.995
Choueiri et al,[44] 2021	18-36	18-36	Children in early Intervention	US	M-CHAT-R/F	Single	All evaluated	1.00	1.00
Christopher et al,[45] 2020	18-48	31.9 (8.3); 18-48	High likelihood for ASD	US	M-CHAT-R/F	Single	All evaluated	0.33	0.333
Coelho-Medeiros et al,[25] 2019	22.5 (4.2); 16-30	22.5 (4.2); 16-30	High likelihood for ASD and randomly selected controls	Chile	M-CHAT-R/F	Single	Subsample of negative screens evaluated	0.833	0.960
Dereu et al,[26] 2012	21.2 (2.6); 16.7 –31.0	28.8 (7.0); 13.5–51.4	High likelihood for ASD or language delay	Flanders, Belgium	M-CHAT	Single	Use of a second screener	0.88	0.881
DiGuiseppi et al,[27] 2010	52.7(14.8); 20-86	52.7(14.8); 20-86	Diagnosis of Down syndrome	US	M-CHAT	Single	Subsample of negative screens evaluated	0.468	0.593
Dudova et al,[28] 2014	~24	N/A	Preterm birth with low birth weight	Prague	M-CHAT	Single	Use of a second screener	0.926	0.926

Study			Population	Location	Screener	Screen	Method		
Eaves et al,[46] 2006	37.2; 17–48	40.3 (6.9); 22–53	High likelihood for ASD	British Columbia	M-CHAT	Single	All evaluated	0.27	0.267
Guo et al,[29] 2019	22.7 (4.1); 16-30	23.2 (4.4)	Primary care population	China	M-CHAT-R/F	Single	Use of second screener after parent or provider concern	0.865	0.986
Guthrie et al,[9] 2019	16 - 26	41.3(13.6); 17.7-87.7	Primary care population	US	M-CHAT/F	Repeat	Medical record review	0.937	0.937[g]
Harris et al,[30] 2021	42.6 (2.2); 24-48	N/A	Children in Head Start	US	M-CHAT-R/F	Single	Use of two other screeners	0.99	0.99
Hoang et al,[31] 2019	18-30	18-30	Population-based sample	Vietnam	M-CHAT	Single	Subsample of negative screens evaluated	-	0.993
Inada et al,[64] 2011 [i]	18.6 (0.5); 17-23	37.1 (1.1); 35-44	Primary care population	Japan	M-CHAT	Single	Diagnosis confirmed at age 3 through interviews	0.961	0.961
Jonsdottir et al,[60] 2021	31.7(1.7)	N/A [j]	Primary care population	Iceland	M-CHAT-R/F	Single	Medical record review	0.996	0.996
Kamio et al,[65] 2014	18.7 (0.6); 17–26	49.4(11.5); 33-73	Primary care population	Japan	M-CHAT/F	Single	Follow-up primary care of all children	0.986	0.986
Kanne et al,[47] 2018	18-48	32.5 (8.3); 24-42	High likelihood for ASD	US	M-CHAT-R/F	Single	All evaluated	0.385	0.385
Kara et al,[32] 2014	18 - 36	24-42	High likelihood for ASD and primary care population	Turkey; Istanbul	M-CHAT/F	Single	All HL and subsample of LL children evaluated	-	0.973
Keehn et al,[48] 2021	30.4(6.5); 18-48	30.4(6.5); 18-48	Referred by physician for ASD concern	US	M-CHAT-R/F	Single	All evaluated	0.378	0.378
Kerub et al,[61] 2020	22.5(3.8); 18-36	N/A [k]	Primary care population	Israel	M-CHAT/F	Single	Medical record review	0.982	0.973
Kim et al,[67] 2016	24.8(2.5); 4.8-43.1 [l]	120.4(8.8); 110-151	Preterm birth	US	M-CHAT	Single	Use of second screener at 10- year follow-up	0.84	0.840
Kleinman et al,[66] 2008 [m]	16-30	52.2 (8.0)[h]	High likelihood for developmental concerns and primary care	US	M-CHAT/F	Repeat	Re-screening and surveillance at a second timepoint	N/A	0.962
Koh et al,[49] 2014	34.0 (7.9); 16.8-48.1	42.1(10.0); 17.8 – 69.2	High likelihood for developmental concerns	Singapore	M-CHAT	Single	All evaluated	0.667	0.667

Study	Age range 1	Age range 2	Population	Country	Tool	Type	Confirmation method	Value 1	Value 2
Magan-Maganto et al,[33] 2020	14-36	23-36	Primary care population	Spain	M-CHAT-R/F	Single	Surveillance from EI centers	N/A	0.996
Matson et al,[50] 2013	16–30	16–30	Enrolled in early intervention	US	M-CHAT	Single	All evaluated	0.502	0.502
Oner et al,[59] 2020	26.8 (5.8); 16-36	27.4 (5.9); 16-41	Primary care population	Turkey; Istanbul	M-CHAT-R/F	Single	Subsample of negative F/U screens evaluated	0.67	0.985
Robins et al,[34] 2014	20.9 (3.3); 16-30	26.2 (5.5)	Primary care population	US	M-CHAT-R/F	Repeat	Provider surveillance and a subsample of negative screens evaluated	0.993	0.993
Salim et al,[51] 2020	18 - 48	18-48	High likelihood for developmental concerns	Bali, Indonesia	M-CHAT	Single	All evaluated	0.786	0.786
Salisbury et al,[52] 2018	16 -48	16-48	High likelihood for developmental concerns	US	M-CHAT	Single	All evaluated	0.564	0.564
Samadi et al,[35] 2015	24-60	24-60	Primary care, preschool, and kindergarten centers	Iran	M-CHAT	Single	Use of a second screener	0.817	0.981
Schjølberg et al,[63] 2022	19.02 (1.2)	~42	Population-based study	Norway	M-CHAT	Single	Medical record review	0.925	0.925
Smith et al,[53] 2013	18–48	18–48	High likelihood for developmental concerns	US	M-CHAT	Single	All evaluated	0.617	0.617
Snow et al,[54] 2008	43.1(14.2); 18-48	43.1(14.2); 18-48	High likelihood for ASD	US	M-CHAT	Single	All evaluated	0.385	0.385
Srisinghasongkram et al,[21] 2016 (Sample 1)	31.2 (6.7); 18–48	18-48	High likelihood for language delay	Thailand	M-CHAT/F	Single	All evaluated	0.984	0.984
Srisinghasongkram et al,[21] 2016 (Sample 2)	24.6 (8.4); 18–48	18-48	Primary care population	Thailand	M-CHAT/F	Single	Telephone F/U, EHR review, or evaluation	0.999	0.999
Sturner et al,[36] 2016	18-24	22.9 (6.1); 14.7-40.8	Primary care population	US	M-CHAT/F	Single	Negative F/U screens evaluated	0.712	0.996
Sturner et al,[39] 2022	18.0 (0.53); 16-20	20.5 (1.9)	Primary care population	US	M-CHAT-R/F	Single	Subsample of negative screens evaluated	0.658	0.658

Study	Age of screen [a]	Age of evaluation [a]	Population	Country	Version [b]	Schedule [c]	FN detection strategy [d]	Spec [e]	Spec [f]
Taylor et al,[55] 2014	28.1 (4.8); <36	28.1 (4.8); <36	High likelihood for developmental delay	US	M-CHAT	Single	All evaluated	0.559	0.559
Toh et al,[62] 2018	20.8 (4.1); 15.0–36.0	N/A	Primary care population	Malaysia	M-CHAT	Single	Medical record review	0.999	0.999
Tsai et al,[68] 2019	24.3 (4.4); 16-32	36.0 (0.1); 36-37	Community and clinical settings	Taiwan	M-CHAT-R/F	Single	All evaluated	0.935	0.935
Vui et al,[40] 2022	18-30	N/A	Population-based study	Vietnam	M-CHAT	Single	Subsample of negative screens evaluated	0.977	.995
Weitlauf et al,[37] 2015	16-36	18–43	Younger siblings of children with ASD	US	M-CHAT-R/F	Single	Subsample of negative screens evaluated	0.806	0.806
Wieckowski et al,[38] 2021 [n]	18.8(0.93); 17-22	23.8(6.87); 18-60	Primary care population	US	M-CHAT-R/F	Repeat	Physician concern	0.972	0.972
Windiani et al,[56] 2016	30.6 (9.6); 18-48	30.6 (9.6); 18-48	High likelihood for developmental delay	Indonesia	M-CHAT-R/F	Single	All evaluated	0.946	0.946
Wong et al,[57] 2018	18-47	30.2 (8.1); 18-47	High likelihood for developmental delay	Taiwan	M-CHAT	Single	All evaluated	.528	.528
Zhang et al,[41] 2022	18-24	23.1(4.6)	Primary care population	China	M-CHAT-R/F	Single	Use of a second screener and follow-up of negative screens	0.995	0.995

Note: [a] Mean, Standard Deviation, and Range in months reported for entire sample that received M-CHAT(-R/F) or evaluation, when available. If not available, an estimate from the manuscript or from communication with authors is reported in months. [b] Version of M-CHAT; M-CHAT = original M-CHAT without Follow-Up; M-CHAT/F = original M-CHAT with Follow-up; M-CHAT-R = revised version of M-CHAT without Follow-Up; M-CHAT-R/F = revised version of M-CHAT with Follow-Up [c] Single or repeat timepoint screening schedule. Studies were classified as repeat even if only a subset of children completed screener more than once. [d] Description of strategy used to detect False Negative (FN) cases. [e] Original specificity reported in the manuscript. [f] Specificity was recalculated using the recalculated TN. TN was recalculated to include presumed true negatives (i.e., including children who screened negative but were not further evaluated), for consistency across studies, unless noted otherwise. Negative screens were presumed to be TN unless there was other presented evidence. [g] TN and Spec taken from directly from paper and not recalculated due to missing information. [h] Age of evaluation is for ASD sample only; age for non-ASD sample is unknown. [i] Discriminant validity sample only reported due to not enough information provided for the concurrent validity sample. [j] Evaluation occurred up to 18 months after the screening. [k] Age of evaluation was within 10 months of screen. [l] Age reported is uncorrected for prematurity. [m] Study 2 sample only presented and analyzed due to overlap of sample 1 with Chlebowski et al. [n] Information is reported for 18 month screening start age only.

eTable 5. Sensitivity and Specificity for Younger and Older Samples

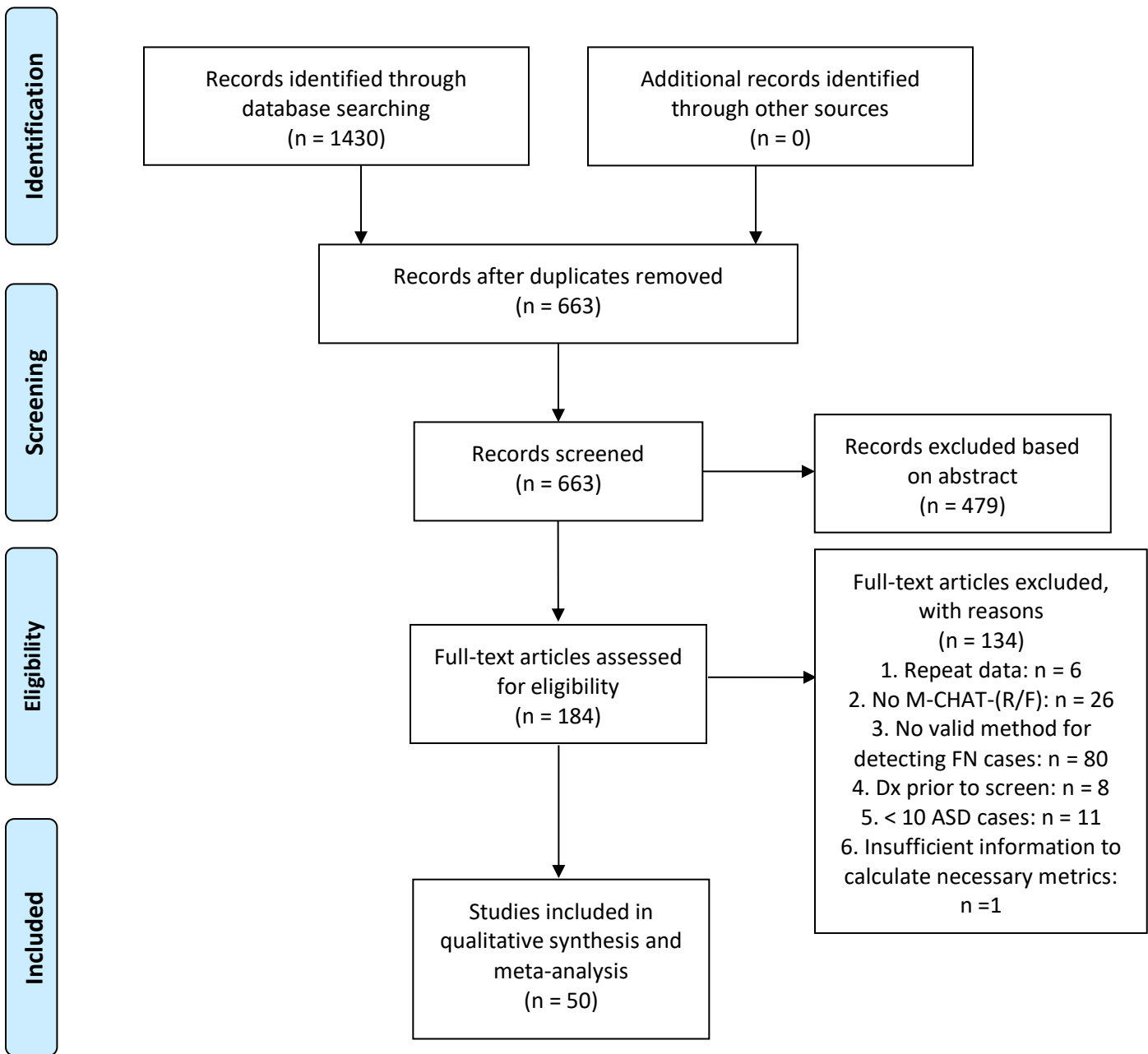| References	FN Strategy	Screen Age (m)	N	Sens.	Spec.	TP	TN	FP	FN
Beacham et al,[42] 2018	C: All Eval	16-30	99	.87	.58	65	14	10	10
		31-45	55	.82	.33	40	2	4	9
Christopher et al,[45] 2020*	C: All Eval	18-30	115	.90	.22	79	6	21	9
		31-48	173	.70	.41	90	18	26	39
Kanne et al,[47] 2018	C: All Eval	18-30	72	.90	.30	47	6	14	5
		31-48	86	.73	.47	49	9	10	18
Koh et al,[49] 2014	C: All Eval	18-30	173	.89	.59	47	71	49	6
		31-48	407	.76	.72	111	187	74	35
Salisbury et al,[52] 2018	C: All Eval	16-30	271	.78	.54	134	53	45	39
		31-48	214	.69	.59	86	53	37	38

*Data obtained from personal communication with one of the authors.

eTable 6. Sensitivity and Specificity for Single and Repeated Screening

|		Single (18 or 24 months)	Repeat (18 and 24 months)	Combined (18 and/or 24 months)
Carbone et al,[8] 2020	Sensitivity	.28	.41	.33
	Specificity	.98	.98	.98
Guthrie et al,[9] 2019 [a]	Sensitivity	.39	.51	.50
	Specificity	.95	.95	.94
Wieckowski et al,[38] 2021 [b]	Sensitivity	.74	-	.82
	Specificity	.97	-	.97

[a] Data obtained from personal communication with one of the authors.
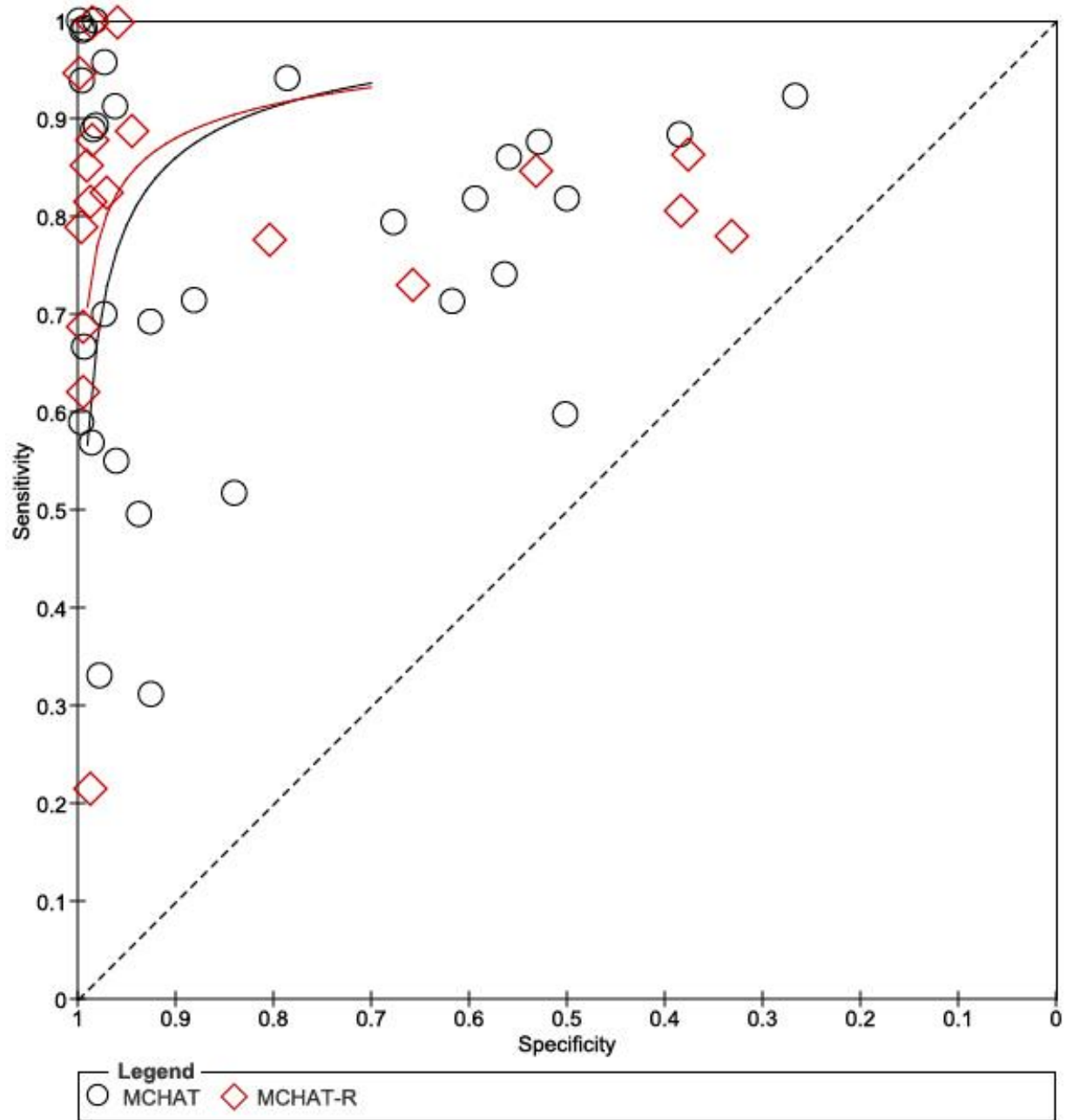[b] Data for 18 month screening and rescreening at 24 and/or 36 months is reported

Records identified through
database searching
(n = 1430)

Additional records identified
through other sources
(n = 0)

Records after duplicates removed
(n = 663)

Records screened
(n = 663)

Records excluded based
on abstract
(n = 479)

Full-text articles assessed
for eligibility
(n = 184)

Full-text articles excluded,
with reasons
(n = 134)
1. Repeat data: n = 6
2. No M-CHAT-(R/F): n = 26
3. No valid method for
detecting FN cases: n = 80
4. Dx prior to screen: n = 8
5. < 10 ASD cases: n = 11
6. Insufficient information to
calculate necessary metrics:
n =1

Studies included in
qualitative synthesis and
meta-analysis
(n = 50)

eFigure 1. Study Selection Flow Chart Following PRISMA Guidelines

eFigure 2. Overall SROC of M-CHAT(-R/F) (n = 49[1] Studies)
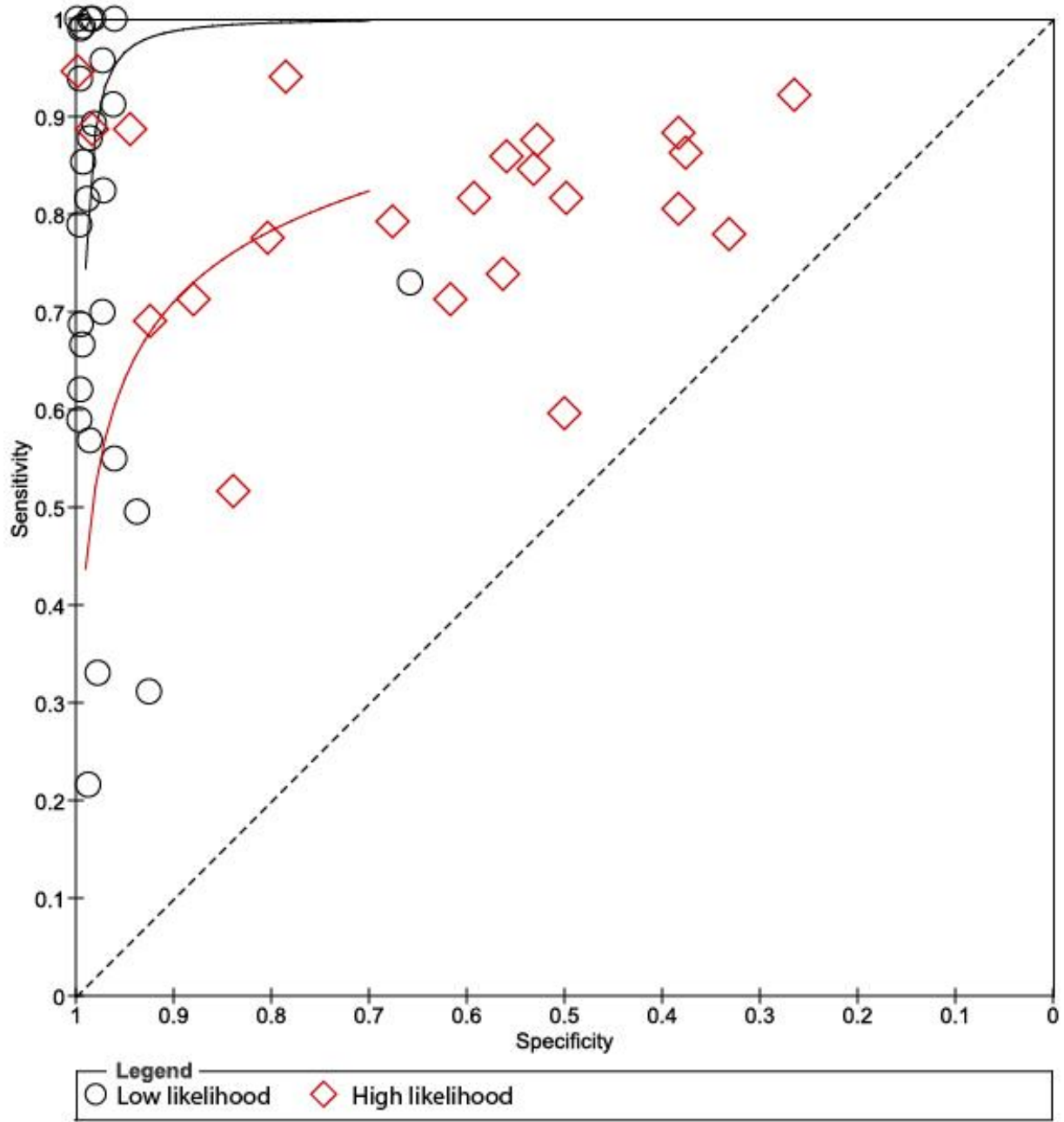Diamond width reflects sample size.
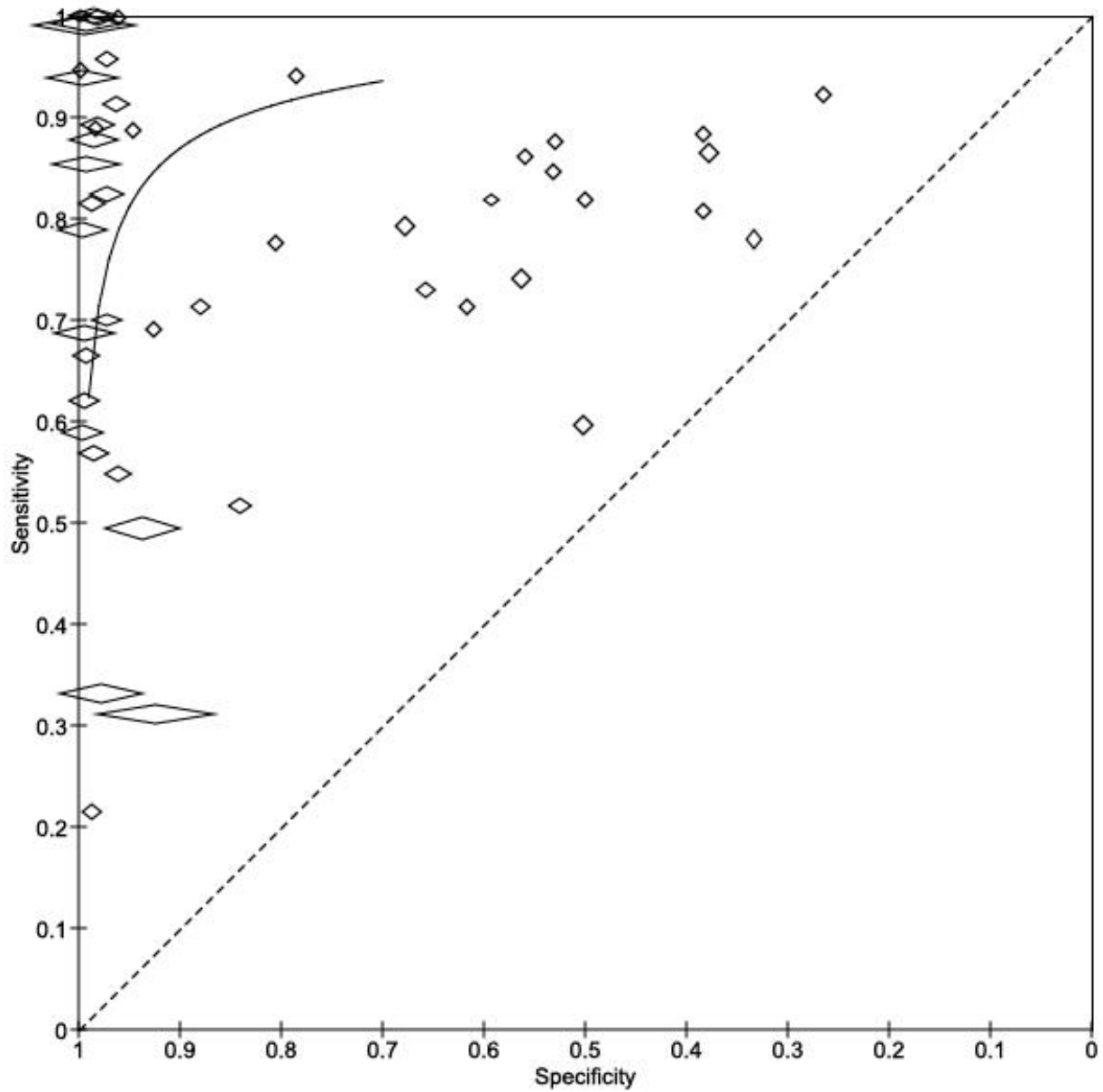
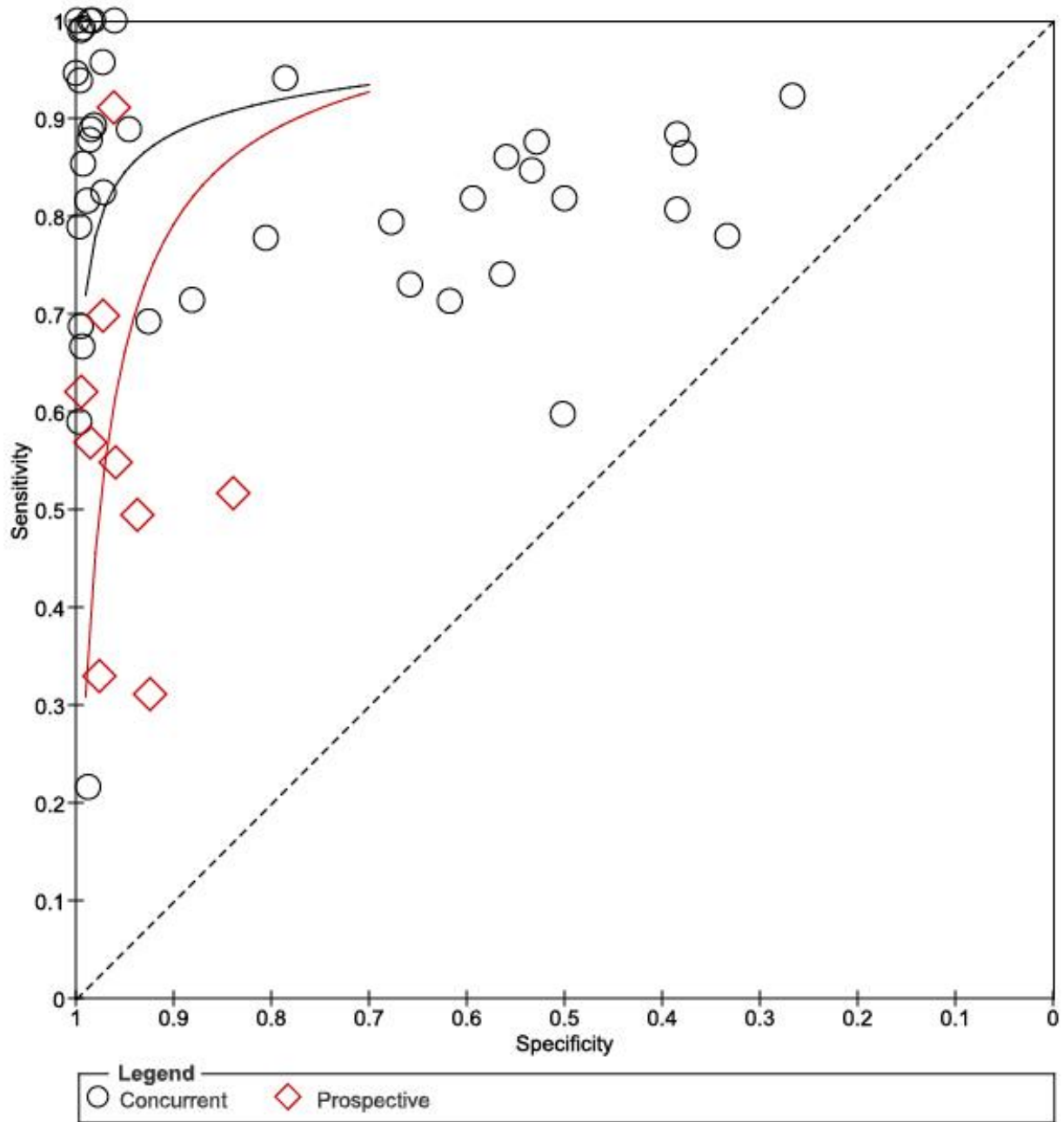[1] Studies classified as "mixed" or "unknown" in any category were excluded from the analysis (n=2)[62,68]

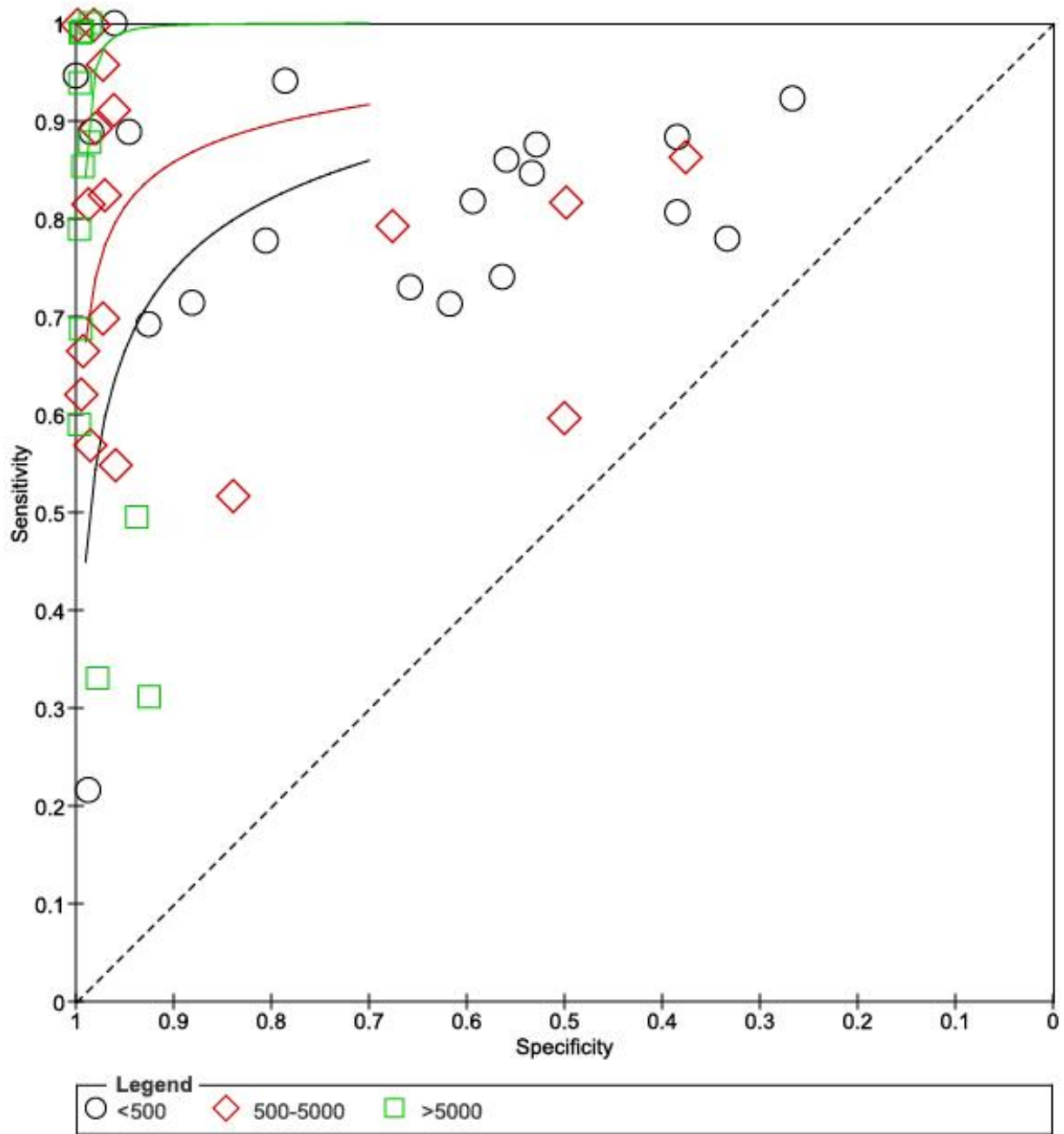**eFigure 3.** SROC Plot of M-CHAT(-R/F) by M-CHAT Version (M-CHAT n = 31, M-CHAT-R n = 18)

eFigure 4. SROC Plot of M-CHAT(-R/F) by Likelihood Level of Sample (Low Likelihood n = 27, High Likelihood n = 22)
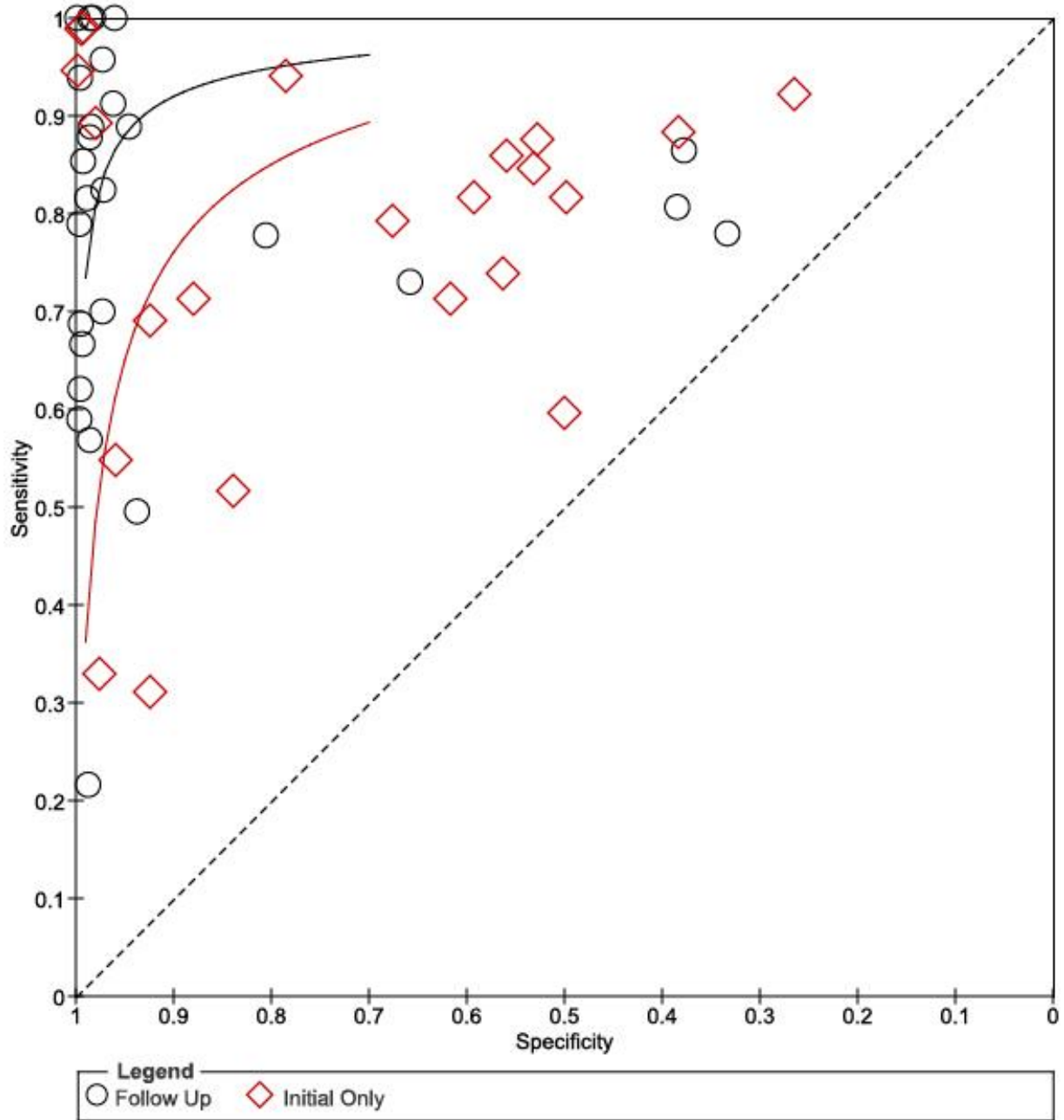
eFigure 5. SROC Plot of M-CHAT(-R/F) by Case Confirmation Strategy (Concurrent n = 40, Prospective n = 9)

eFigure 6. SROC Plot of M-CHAT(-R/F) by Sample Size (Small n = 20, Medium n = 17, Large n = 12)

eFigure 7. SROC Plot of M-CHAT(-R/F) With Follow-up vs Initial Only (Initial n = 22, Follow-up n = 27)

eFigure 8. SROC Plot of M-CHAT(-R/F) by Language (English/Primarily English n = 26, Other Language n = 23)