Spring 5-4-2020

Meta-Analysis of Diagnostic Accuracy of M-CHAT by Categorical Rank of Clinical Diagnosis

John Olmstead

Meta-Analysis of Diagnostic Accuracy of M-CHAT by Categorical Rank of Clinical Diagnosis

John Olmstead

School of Public Health – Georgia State University

Dr. Brian Barger

March 19, 2020

Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder occurring in children from a young age and has complex causation. Individuals with ASD present with atypical social-communication and cognition with restricted and repetitive behaviors (American Psychiatric Association [AAP], 2000; American Psychiatric Association [APA], 2013; Carcani-Rathwell et al., 2006; World Health Organization [WHO], 1993). U.S. based estimates show the prevalence of ASD published by the CDC each year has increased from 1 in 150 children to 1 in 59 from 2000 to 2014 (Baio et al., 2018; Center for Disease Control and Prevention [CDC], 2019). ASD is early forming and is currently considered diagnostically stable at 24 months of age (Chawarska et al., 2007; Kleinman et al., 2008; Lord et al., 2006), though the average age of diagnosis ranges from 46 to 67 months and varies between socio-demographics (Baio et al., 2018, Mandell et al., 2006). While variation in delays are not clear, a major factor in diagnostic delay is the complexity of early identification process which often requires multiple clinical appointments prior to receiving an ASD diagnosis, and much frustration amongst caretakers (Goin-Kochel et al., 2006). Ultimately, many children with ASD who are identifiable for an early diagnosis are not receiving a diagnosis and the benefits that come with early identification.

Early identification of children with ASD is important because it enables children to receive resources which can help them achieve better developmental outcomes (Green et al., 2017; National Research Council [NRC], 2001; Virues-Ortega, 2010). Research shows that intervention has larger benefits when performed on younger children with ASD compared to older children (Rogers, 1996). For example, studies suggest that interventions for children utilized earlier in life have positive effects on IQ development and adaptive behavior underscoring the need to identify earlier rather than later (Eldevik et al., 2009; Harris, 2017;

Rogers, 1996). Research also shows that children who receive treatment before 48 months are more likely to be placed into regular education class than those enrolled after 48 months of age (Kasari et al., 2012). Studies that enroll high risk children for ASD to investigate outcomes have begun to outline the benefit of the intervention groups even to the point of "reducing prodromal ASD symptoms in the second and third years of life" (Green et al., 2017, p. 1337). The evidence for the positive impact of earlier received interventions highlights the need for effective early identification of children with ASD and potential for ASD.

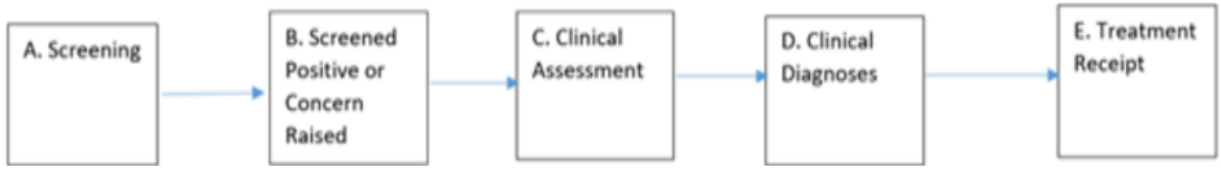
Early Identification Process Model

Early identification involves complex multi-faceted processes informed by insights from the public health, education, early intervention and psychometric/clinical literatures (Bricker et al., 2013; Sheldrick 2011, 2015, 2016, 2019). Screening/monitoring and identification of diagnosed cases are more commonly discussed in public health and epidemiology, whereas screener development, assessment tool development, specific intervention implementation, and diagnostic training are more commonly discussed in clinical science and psychometric research. The process of identification roughly follows the model seen in Figure 1 (Barger et al., 2018). The early identification process is typically initiated by monitoring (e.g. brief informal queries about development) and screening (i.e. brief formal screeners querying about development) wherein children with ASD and related conditions are observed during well-child visits (CDC: "Recommendations and Guidelines", 2019; Individuals with Disabilities Education Improvement Act [IDEA], 2004). Screening and monitoring processes (Step A) lead to the second phase where individuals who have been determined "at risk" by screening and/or monitoring (Step B) may be referred for a clinical assessment (Step C; Filipek et al., 1999). In order to render an ASD diagnosis, specialists are then required to administer and score diagnostic assessments (Step C;

Klin & Volkmar, 1995; Klin et al., 2005). The fourth phase is the diagnosis phase (Step D), wherein licensed clinicians match presenting symptoms to defined diagnostic information, preferably as part of a multi-disciplinary team (Klin & Volkmar, 1995; Klin et al., 2005; Volkmar et al., 2014a, 2014b). The final phase is treatment receipt.

Figure 1

Steps in the Process of Early Identification and Intervention of Autism Spectrum Disorders



Early Identification: Monitoring and Screening

Monitoring and screening are the first community based intentional efforts to identify any atypical development in children (APA, 2006; Hirai et al., 2018). Monitoring refers to the continued surveillance of children's developmental status by health care providers that may lead to referral for services or diagnosis for developmental delay (Barger et al., 2018; Bright Futures, 2006). Screening refers to a method of determining if a child suspected of having a disability, by parent or health care provider, needs early intervention services (Hirai et al., 2018; National Research Council, 2001). A U.S. study examining the prevalence of screening in early childhood reported that in 2016 fewer than one third of children 9 through 35 months in age had received any form of screening (Hirai et al., 2018). Between screening and monitoring, screening is the more studied of the two (Barger et al., 2018).

ASD screening is considered within the broader context of developmental screening wherein a number of social and developmental milestones, including those specific to ASD, are considered (NRC, 2001). Developmental screening and monitoring is recommended by several federal and professional groups (Bright Futures, 2007; Office of Disease Prevention and Health Promotion [ODPHP], 2010; Committee on Children, 2001; Johnson & Myers, 2007). ASD specific screenings are recommended to occur around 18 to 24 months of age as this timeframe coincides with the ability to recognize symptoms and the appropriate age of diagnosis (Lord et al., 2006; Chawarska et al., 2007; Bright Futures, 2007).

The tools informing the screening process are referred to as screeners and are designed to be reliable for identifying probable cases of their intended population and screening out non-cases (U.S. Department of Health and Human Services [HHS], 2014). Screeners are meant to be brief and used to indicate that an assessment is necessary and not necessarily to provide a diagnosis (CDC, 2020; Robins, 2008). Ideally, screeners can be made more accessible by making sure they accomodate eight-grade reading levels and are informative enough for assessment referral (Arnold et al., 2006). Despite widespread recommendations for screening, screeners are often underutilized by pediatricians during scheduled check-ups even though they are an important part of early intervention systems (Hirai et al., 2018; NRC, 2001; Sand, 2005).

There exist different instruments when considering ASD identification screeners; in particular, whether screeners are designed for use in the general population or in “at risk” populations (Norris & Lecavalier, 2010). Both the general and at-risk populations should be screened for developmental delays as cases are present in both levels. Level 1, population-based, samples require screeners relevant for low-risk populations where they can be implemented in areas like typical well-child visits to try to identify risk for atypical development (Barger et al.,

2018; Robins, 2008; Zwaigenbaum et al., 2015). The purpose of Level 1 screeners is to increase the number of identified children from a population sample who would otherwise continue to be unidentified. Level 2, "high-risk", populations are fundamentally different than Level 1 screening because these populations typically are already being served or monitored by health care providers due to risk for developmental delays (Robins, 2008; Zwaigenbaum et al., 2015).

Early Identification: Diagnostic Assessment and Diagnosis

After the screening process, infants and children are either determined to be low risk for a delay or carry moderate/high risk (Klin et al., 2005; Volkmar et al., 2014a). Children with moderate or high risk continue from screening clinical assessment phases where relevant information such as medical history and diagnostic assessments are gathered (Klin et al., 2005). Notably, although ASD can be diagnosed by a single trained clinician, "the clinical assessment of individuals with this disorder is most effectively conducted by an experienced interdisciplinary team" (Klin & Volkmar, 1995, p. 5; Klin et al., 2005). Once an individual continues from the assessment step to the diagnosis step all relevant information from the screening and assessment phase becomes evidentiary to support or deny a diagnosis. At this stage, diagnosis may be conducted by a single clinician or multiple clinicians who make the ultimate clinical decision for a child's outcome. The gold standard is to incorporate all available resources to acquire the most accurate diagnosis which includes multi-disciplinary clinicians as well as diagnostic scored assessments (Klin & Volkmar, 1995; Klin et al., 2005).

The tools of the assessment phase can be used to aid the clinician to determine if individuals meet the criteria for ASD (LeCouteur et al., 2007; Klin et al., 2005; Volkmar et al., 2014a). Two widely recommended instruments in aiding autism assessment and diagnosis are the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule

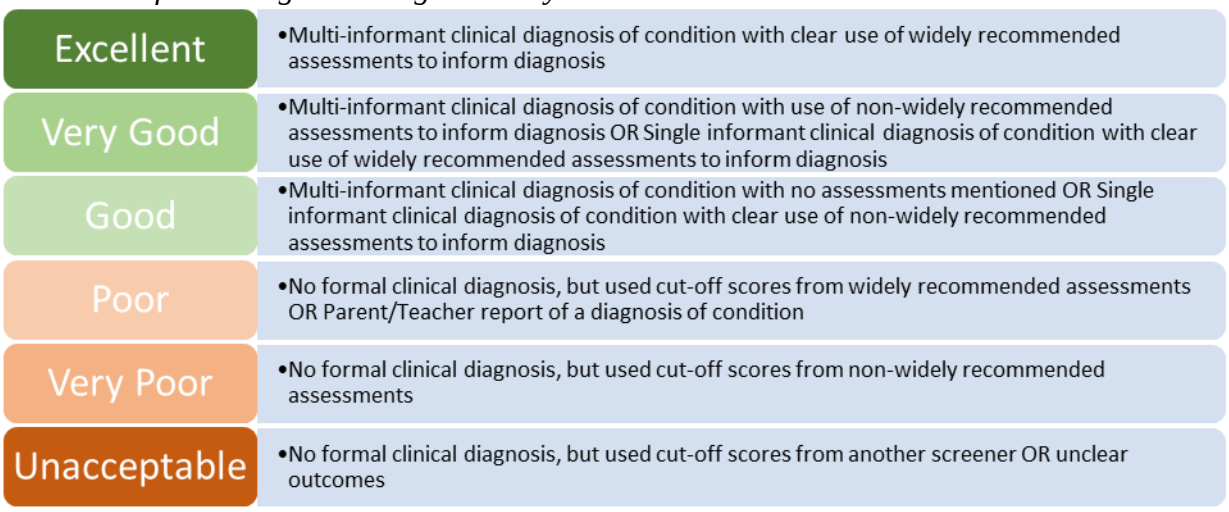
(ADOS) (LeCouteur et al., 2007). The ADOS is a semi-structured standardized assessment of four modules on which an individual is observed by clinicians which is intended to "complement information obtained from developmental tests and caregiver history" (Gotham et al., 2007). The ADI-R is a semi-structured investigator-based interview for caregivers in reference to ICD-10 and/or DSM criteria (Lord et al., 1994). These instruments have been reported to match characteristics in children with DSM-IV diagnosis of autism traits, however DSM-V criteria leads to a drop off in diagnosis compared to DSM-IV when using ADOS (Foley-Nicpon et al., 2017; LeCouteur et al., 2007; Mazefsky et al., 2013; Ventola et al., 2006). The differences in criteria between DSM-IV criteria and DSM-V criteria is estimated to lead to fewer diagnoses of children with PDD-NOS and Asperger's disorder (Smith et al., 2015).

Since an ASD diagnosis is heavily reliant on the instruments used and the clinical judgement and experience of those rendering the diagnosis (Klin et al., 2005), it is reasonable that early identification studies take diagnostic approaches into account. These diagnostic approaches include both the multi-disciplinary teams and highly diagnostic instruments like ADOS/ADI-R being used in tandem to create a high quality and robust diagnosis. Figure 2 displays a framework from Barger (2018) for ranking screening accuracy studies. Two subranges exist within this framework: clinical diagnosis exists in the range of *adequate* to *excellent* while *unacceptable* to *poor* exists to explain research studies that measure diagnostic accuracy without clear clinical input. The different tiers represent expected differences of quality of diagnosis. Starting from the bottom of the framework, *unacceptable* indicates that a study does not clearly indicate that the child received a formal diagnosis (e.g., assessment scores were used to denote a diagnosis, but not clinician); *very poor* indicates diagnosis based on non-recommended assessment, but does not indicate a clinician provided a diagnosis; and *poor*

represents no clinical diagnosis but a positive diagnostic result from recommended assessments such as the ADOS or ADI-R. Continuing up the framework we reach 3 levels that require clinical diagnosis *good*, *very good*, and *excellent.* The difference between these tiers, all of which use multi-informant clinical diagnoses, ranges from: *good* where no assessments were referenced along with the clinical diagnosis; *very good* where non-widely recommended assessments were used and along with a clinical diagnosis, and *excellent* where multi-informant clinical diagnosis method is used in conjunction with widely recommended assessments. Single clinician diagnosis can only be considered *good* if it is accompanied by an assessment that is not widely recommended or *very good* if it is informed by a widely recommended screener. The tiering aims to lay out the difference in quality and why studies should seek to achieve for *excellent* reference standard. This framework is a proposed model and the tiering is based on expert reasoning.

Figure 2

Framework for Sorting Screening Accuracy Studies



Diagnostic Accuracy and Meta-analysis of Diagnostic Accuracy for Screeners

ASD screeners and assessments seek to correctly classify individuals with or without a condition (Robins 2008). Following the intended purpose, the best measure of a screener and assessment is to determine their ability to correctly predict a particular clinical outcome. Diagnostic accuracy is the measure of a test's ability to correctly identify when a condition is present or not in an individual (Gatsonis & Paliwal, 2006). Diagnostic accuracy metrics indicate the degree to which tools correctly classify people into cases and non-cases.

A variety of diagnostic accuracy metrics can be developed using the confusion matrix seen in Figure 3. Understanding the multi-faceted aspects of accuracy can be aided by filling out a confusion matrix like the one shown in Figure 3. Individuals who are screened fall into 1 of 4 categories based on their screening outcome (positive or negative) and their true condition status. The ideal test would correctly show a positive result for all individuals with the condition and a negative result for all individuals without the condition. This describes true positives (TP) which are positive screens for those with the condition and true negatives (TN) which are negative screens for those without the condition respectively. The other two classifications are individuals who screen positive but do not truly have the condition (i.e., false positives; FP), and negative screens who truly do have the condition, false negatives (FN).

Figure 3

Confusion Matrix Accuracy Metrics

	True Positive (TP)	False Positive (FP)	
This row represents all individuals who are indicated to have the condition.	False Negative (FN)	True Negative (TN)	Positive Predictive Value = TP/TP+FP
This row represents all individuals who are indicated to not have the condition.	Sensitivity = TP/TP+FN	Specificity = TN/FP+TN	Negative Predictive Value = TN/FN+TN
	This column represents all individuals truly with the condition	This column represents all individuals truly without the condition	

The most common measurements of accuracy are sensitivity and specificity, respectively the ability to correctly determine those with the condition among those who truly have the condition and the ability to determine those without the condition among those who truly do not have the condition (Rothman, 2012). Other metrics for determining accuracy include the positive predictive value (PPV) and the negative predictive value (NPV). The PPV finds among the number of individuals referred, how many were accurately identified and met criteria to receive such referral for a diagnosis, while the NPV does the opposite in that it determines among individuals not referred how many correctly needed no referral. Sensitivity, specificity, PPV, and NPV can be combined to get metrics that represent the diagnostic ability of the test overall such as the positive likelihood ratio (LR+), the negative likelihood ratio (LR), and the Diagnostic Odds Ratio (DOR). The positive likelihood ratio (LR+) divides sensitivity by 1-specificity and the negative likelihood ratio (LR-) is measured using the reciprocal of LR+. The DOR is a ratio of the (LR+) and the (LR-). The usefulness of these tests are their interpretation which for LR+ is: the likelihood of a positive test result in a person with a disease is more (or less, in a poorly

designed test) likely in a person with the condition than a person without the condition. Similarly, the LR- shows the opposite result: the likelihood of a negative test in a person without the condition is more (or less) likely in a person without the condition than a person with the condition. Finally, DOR indicates the odds of a person with a condition being correctly classified when taking a test.

Systematic reviews and meta-analyses provide "a transparent and replicable method for summarizing the literature" (Pigott 2012, p. 1). A systematic review is a structured review of published literature to bring together all relevant studies (Uman, 2011; Siddaway et al., 2019). A meta-analysis assesses studies collected from a systematic review quantitatively or qualitatively and their quality and findings are reported (Armstrong, 2011; Siddaway et al., 2019; Grant & Booth, 2009). The extraction of effect sizes from a systematic review may be combined and synthesized into a meta-analysis (Walker et al., 2008). These reproducible studies and combined effect estimates provide ways to assemble common studies to drive research and policy forward (Walker et al., 2008).

The best method for screener comparison is a review of the published literature to determine the screener with the highest diagnostic accuracy (Gatsonis & Paliwal, 2006). When performing a quantitative review such as a meta-analysis, fixed effects or random effects will be used. A fixed effects model assumes that there is a true effect size that studies are estimating; these models may be used when a meta-analysis combines estimates of closely related studies (Borenstein, 2009; Pigott and Polanin, 2020). The corollary to fixed effects is random effects where an effect size is "similar but not identical across studies" (Borenstein et al., 2009, p. 69). When there is more expected variability between studies a random effects model may be better to use; for example, variability may result from differences in study settings or sampling

procedures (Borenstein et al. 2009, p. 69). Random effects are recommended for use in most meta-analyses (Pigott & Polanin, 2020). When combining the previous topics of meta-analysis, diagnostic accuracy, and modeling techniques there are important choices to be made.

While the choice to utilize random or fixed effects modeling must be made prior to running a model, the choice of how to analyze your results (univariate versus multivariate) can be made after carefully reviewing the studies. Starting with the fewest variables, the univariate approach is conceptually the simplest approach only using one variable to describe an outcome. This type of approach can be useful in areas where a controlled system is possible. For example, traditional physics and chemistry experiments studying the effect on pressure as temperature (one variable) increases. This approach's strength is that you can make direct assertions about the relationship between the dependent and independent variable. The weaknesses are that the simplicity of this approach cannot adequately address the complexity of studies that cannot control all variables except the ones of interest.

The multivariate approach is similar to the univariate approach, but instead of one variable, more than one variable is used to analyze an outcome or set of data. By increasing the number of variables considered, analyses can more accurately reflect the multiple factors impacting an outcome, thus more closely reflect a complex reality. However, this has the potential for negative consequences and predictors for a multivariate model should be chosen carefully. If confounded variables are included in a multivariate model misleading results can occur. That said, the benefits of multivariate analysis over univariate cannot be understated as multivariate approaches can more robustly describe complex relationships, a noted weakness of the univariate approach (Jackson et al., 2010).

According to Walter and Jadad (1999) meta-analysts have historically employed univariate approaches by analyzing sensitivity or specificity independently, but univariate approaches do not account for their non-independent negative correlations; multivariate approaches do account for these correlations. Two approaches have been proposed to address univariate limitations: the hierarchical receiver operating curve (HSROC) and the Reitsma (Harbord & Whiting, 2009). The HSROC is a Bayesian approach that accounts for the correlations between sensitivity and specificity (Rutter & Gatsonis, 2001). The Reitsma random effects model allows the sensitivity and specificity to be modeled simultaneously. Both approaches can be visualized with graphs displaying sensitivity by 1 – specificity (i.e., FPR). Due to their addressing the fundamental limitation of correlated variables and imperfect visualizations of traditional univariate analyses, these multivariate approaches are strongly recommended (Jackson et al., 2010; Reitsma et al., 2005).

Modified Checklist for Autism in Toddlers

As mentioned above there are multiple screeners for ASD; however, few if any are as widely researched as the Modified-Checklist for Autism in Toddlers (M-CHAT) (McPheeters et al., 2016). The M-CHAT is a population level screener that was developed at Georgia State University by Diana Robins PhD. The M-CHAT's seminal study was published in 2001 and since has been suggested for use as a population-based tool (Zwaigenbaum et al., 2015; Robins et al., 2001, 2014). The M-CHAT was developed for identifying ASD in population samples and uses questions relevant to infant and toddler development to identify red flags consistent with ASD indicators. The M-CHAT was revised from its initial 23 questions to 20 questions in the population-based version. A second, follow-up phase was added to confirm the results of the first phase (Robins et al., 2014). The original M-CHAT has been translated into 69 different

languages as of 2016, adjusted culturally, and validated home and abroad (Brennan et al., 2016; Robins et al., 2001, 2014). A core strength of this screener is that the M-CHAT is free to access online (in English and Spanish) and provides an instant feedback score for curious parents (Robins et al., 2014). The psychometrics of the M-CHAT are reported to be .911 and .955 for sensitivity and specificity for the initial screening stage without follow-up. With follow up the sensitivity decreases to .854 and specificity increases to .993 (Robins et al., 2014).

To date, there have been two systematic reviews and meta-analysis of diagnostic accuracy conducted on autism screeners (Yuen et al., 2018; Sanchez-Garcia et al., 2019). Yuen (2018) reviews the M-CHAT critically concluding there is a lack of evidence for use on children 18 to 24 months of age and high risk children screened with the M-CHAT have a pooled sensitivity of 0.83 and pooled specificity of 0.51. Sanchez-Garcia (2019) includes the M-CHAT in their conclusion that population based screeners are effective for low risk children under the age of 3. Sanchez-Garcia looked at 9 total screeners and found a pooled sensitivity of 0.72 and pooled specificity of 0.98 across multiple screeners. These studies somewhat complemented one another while Yuen called for more evidence for a specific population, Sanchez-Garcia provided broad evidence of screener usage on that population, but not any particular screener.

The methods of these studies showed strengths in similar and different ways. Yuen's study provided insights by reviewing the grey literature and extracting sample characteristics from the studies. Sanchez-Garcia provided a thorough review when it evaluated its study sample for publication bias and performed a subgroup of analysis. Both studies used a Bayesian approach to their meta-analysis which is a strong method of analysis for this body of work. Despite a number of strengths, the studies by Yuen and Sanchez-Garcia have a number of weaknesses. Common weaknesses between the studies include both their data sets suffered from

data heterogeneity, indicating substantial between-study differences. Further, Sanchez-Garcia missed a population-based study by Magan-Maganto et al. (2018) which matched their study criteria. Yuen did not include some large studies such as Stenberg et al. (2014) that includes a large low-risk sample of children screened by the M-CHAT. While no study is without weaknesses, it is important to consider how these could have impacted their results; by leaving out studies that matched their respective criteria, these analyses did not consider all available information.

While there are reviews analyzing the true diagnostic ability of the M-CHAT, to our knowledge there is no study investigating the impact of reference standard categories on accuracy metrics. This study proposes to categorically analyze studies that screen and diagnose children with autism using the Barger reference standard framework. This study seeks to answer if there is a relationship between reference standard category and reported screening accuracy. Our specific hypothesis is that there is an inverse relationship between screening accuracy and reference standard category – in that Multi-Clinician reference standard tier will report the lowest diagnostic accuracy and Other will report the highest.

Methods

The structure of this systematic review was based on the "Preferred Reporting Items for Systematic Reviews and Meta-Analysis" (PRISMA) (Moher, 2009).

Study Criteria

The papers selected for this study contained information to determine diagnostic accuracy for unique ASD screenings using the M-CHAT and with diagnosis information using a clear reference standard in both population based and high-risk samples. A requirement of these studies is that the study is in English and TP, TN, FP, FN must be made available and clear in the study, either in a flowchart or clearly described within the study. For studies whose population samples were identical, only one study was included where the population size was the largest; if studies were the same size the original study results was given preference.

Search Criteria

Three searches were performed for complete coverage of the timeline April 2001 to December 31, 2019 the initial on August 2nd, 2019, a prospective search on October 17, 2019 and a final review search on February 1, 2020. The search terms for this study were "Modified checklist for autism in toddlers" OR MCHAT OR M-CHAT. The initial and final search used Web of Science, EBSCO, and ProQuest. In EBSCO the following were searched: Academic search complete, Alt health watch, Child Development and Adolescent Studies, CINAHL Plus with full text, consumer health complete, Education source, ERIC, family and society studies worldwide, Fuente academia premier, Health source nursing academic edition, medline, medline with full text, mental measurements yearbook, professional development collection, psych articles, psychology and behavioral sciences collection, psychinfo, psychtests, and social work

abstracts; ProQuest: dissertations and thesis at GSU, ProQuest central, arts and humanities, New Zealand, biological science, consumer health database, continental Europe, east Europe, central Europe, education database, health and medical collection, India, Latin American, Iberian, middle east and African, nursing and allied health, public health database, publicly available science, social science, UK and Ireland, ProQuest dissertation and thesis, social science premium, and education. All articles starting from April 2001 were reviewed for relevance in all countries, in English. The ancestral search included a traditional bibliography search of articles ultimately deemed to fit inclusion criteria. A prospective search was also performed on all articles deemed to fit inclusion criteria via identifying relevant articles using the "cited by" feature of Google Scholar.

The following describes the process for data abstraction of the initial, ancestral and prospective searches. After collecting the titles and abstracts in the three searches the collected studies underwent a title/abstract review for M-CHAT screening studies. Studies were screened green to be included, yellow to be reviewed a second time and red to be screened out. To check on this initial screen-out process, 26 titles and abstracts were randomly selected and coded by an independent reviewer - good agreement was found ($K = 0.77$). The process then follow that pdfs from yellow and green highlighted studies had a full text review performed for the inclusion/exclusion criteria, the studies were organized in keep remove computer files for quality control.

Inclusion Criteria

Studies were included if they fit the following criteria: they included M-CHAT data, they were in English, their TP, TN, FP, and FN data was included, and had at least 10 participants. The characteristics of these studies were independently reviewed by two separate researchers in

order to verify their qualities and assure that all studies that match inclusion criteria are included. Search terms and date information was preserved through the search engine account features using the “save search” tool. Disagreements between study qualities were resolved by a third-party researcher blind to either initial party’s decision and the third party decision served as tiebreaker.

Methodologic Quality

These studies entered an assessment known as Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) which assesses the quality of the study based on potential sources of bias in diagnostic accuracy studies. Using appendix F questions of the QUADAS-2 the domains of patient selection, interpretive bias, criterion assessment, and Domain 4: Flow and Timing are reported for low, high, or unclear amount of bias. A “low” bias score was determined for all domains where all questions receive a “yes” score and a “high” bias score was determined for all domains where one or more questions receive a “no” or “unclear” bias score. The process followed this method: All information of the QUADAS-2 appendix F was coded by myself for all studies included in the sample set, this file was preserved for quality control. The next round was conducted by Dr. Barger where a randomized sample of 10 studies were selected and coded, this file was preserved for quality control. The preserved files between myself and Dr. Barger were compared for any discrepancies between individual QUADAS-2 domain items that did not meet 100% agreement. QUADAS-2 domain items that did not meet 100% agreement resulted in a focused review of all 24 studies for the particular domain items in that were not in agreement. For QUADAS-2 sections about appropriate interval between index test and reference standard the reviewers determined that any amount of time passed 1 year for positive screens was not appropriate and would be marked with the high bias option. Any review questions asked that

were not directly answered by the study in clear terms was marked unclear. During the QUADAS-2 review only the methods and results sections of included studies were reviewed for relevant material to answer the QUADAS-2 domain items.

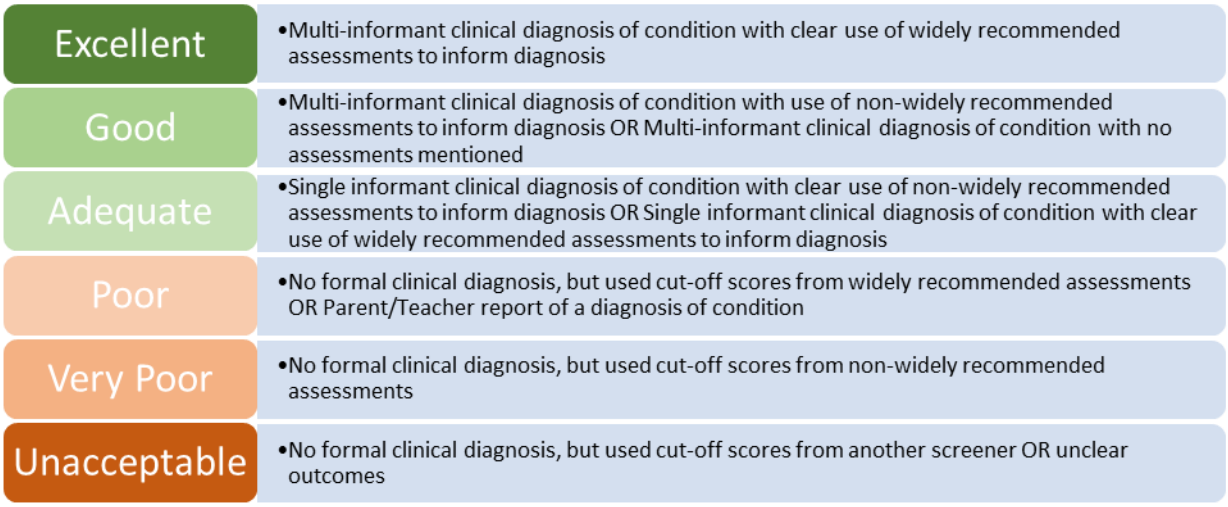
Reporting

The information used for calculating diagnostic accuracy was collected from the selected studies. The information of TP, TN, FP, and FN, the screening level, and reference standard from the original studies was directly transcribed and recorded for data analysis. As seen below in Figure 4, the scale used in Barger et al. (2018), was adapted for this study and reference standard rank was recorded where each study matched the reference standard grouping. The original framework mixed multi-clinician and single-clinician criteria in the *Excellent, Good* and *Very Good* rankings. This adapted framework distinguishes multi and single from one another placing any multi-clinician criteria above that of any single-clinician criteria.

Starting from the bottom of the framework, *unacceptable, very poor,* and *poor* have not been changed from the original. Continuing up the framework have been changed *good, very good,* and *excellent.* The difference between these tiers, are that multi-clinician diagnosis is now only in *excellent* and *good* while single-clinician now only exists in a new category *Adequate.* Specifically the categories are: *good* where single clinician diagnosis was used in conjunction with or without recommended assessments; *very good* where multi-clinician diagnosis was performed with non-widely recommended assessments or no assessments; and *excellent* where multi-informant clinical diagnosis method is used in conjunction with widely recommended assessments.

Figure 4

Redesigned Framework for Categorical Assessment of Reference Standards



Study Variables

M-CHAT. The primary outcome variables included diagnostic accuracy outcomes that could be derived from confusion matrices. The M-CHAT has an initial screening phase and a follow up phase as well as two methods for determining positive criteria. The data for the initial phase and follow-up stages were combined in studies where it was necessary, creating only 1 data set per M-CHAT screening sample. To do this, data was assessed from flow charts to accurately report the TP, TN, FP, and FN of the final results and not any one stage specifically. The data reported by the study was transcribed for data analysis. The complete original 23 question M-CHAT can be found in Appendix C.

Other variables abstracted included: Author, study year, M-CHAT version (original or revised), study country, initial screen or follow up, reported screen cut-off, tested positive, tested negative, loss to follow up, TP, TN, FP, FN, number with condition, number without condition, total N, study reported sensitivity, study reported specificity, study reported PPV, study reported

NPV, study reported LR+, diagnostic system, diagnostic assessment tool used, diagnosis determined by, publication style, study level, and QUADAS-2 criteria. Diagnostic system categories included DSM-IV, DSM-V, ICD-9, ICD-10. Diagnostic assessment tools included ADOS, ADI-R, Social Communication Questionnaire (SCQ), Vineland Adaptive Behavior Scale (VABS), Childhood Autism Rating Scale (CARS), Bayley Scales of Infant Development (BSID), and Parents' Evaluation of Developmental Status (PEDS). Determination of diagnosis was noted as multiple clinicians, a single clinician, ADOS or ADI-R cut-off, other screener cut-off, or unclearly reported. Study levels were reported as population based or high-risk. Data on diagnostic measure and diagnostic determination were combined to develop proxy variables measuring Barger et al.'s (2018) reference rating metric. Two independent raters coded 9 randomly selected studies and had perfect reliability on Study Year (K = 1.0), good agreement on Diagnostic System (K = .75), Moderate to Perfect agreement on Diagnostic Assessment tools (*Krange* = .41-1.0), Moderate agreement for multiple/single clinician (K = .50), and Substantial agreement for population/high-risk categories (K = .94).

Analysis

Using the Meta-Analysis of Diagnostic Accuracy (MADA) package in R, the diagnostic accuracy metrics of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Diagnostic Odds Ratio (DOR) were assessed using the data collected from the original studies. Bivariate Reitsma models were selected for this analysis which is a random effects model and was specified in the MADA package by specifying the Reitsma function. For reference standard analysis two groups were made, a combination of the multi-clinician groups of *excellent* and *good*, referred to as Multi-Clinician, and a combination of *adequate*, *poor*, and *unacceptable*, referred to as Other. Results were recorded and reported using Tables 2-7 and

Figures 4-15 below. The relatively small universe of available studies resulted in the decision to use $\alpha = 0.10$ to indicate meaningful statistical results.

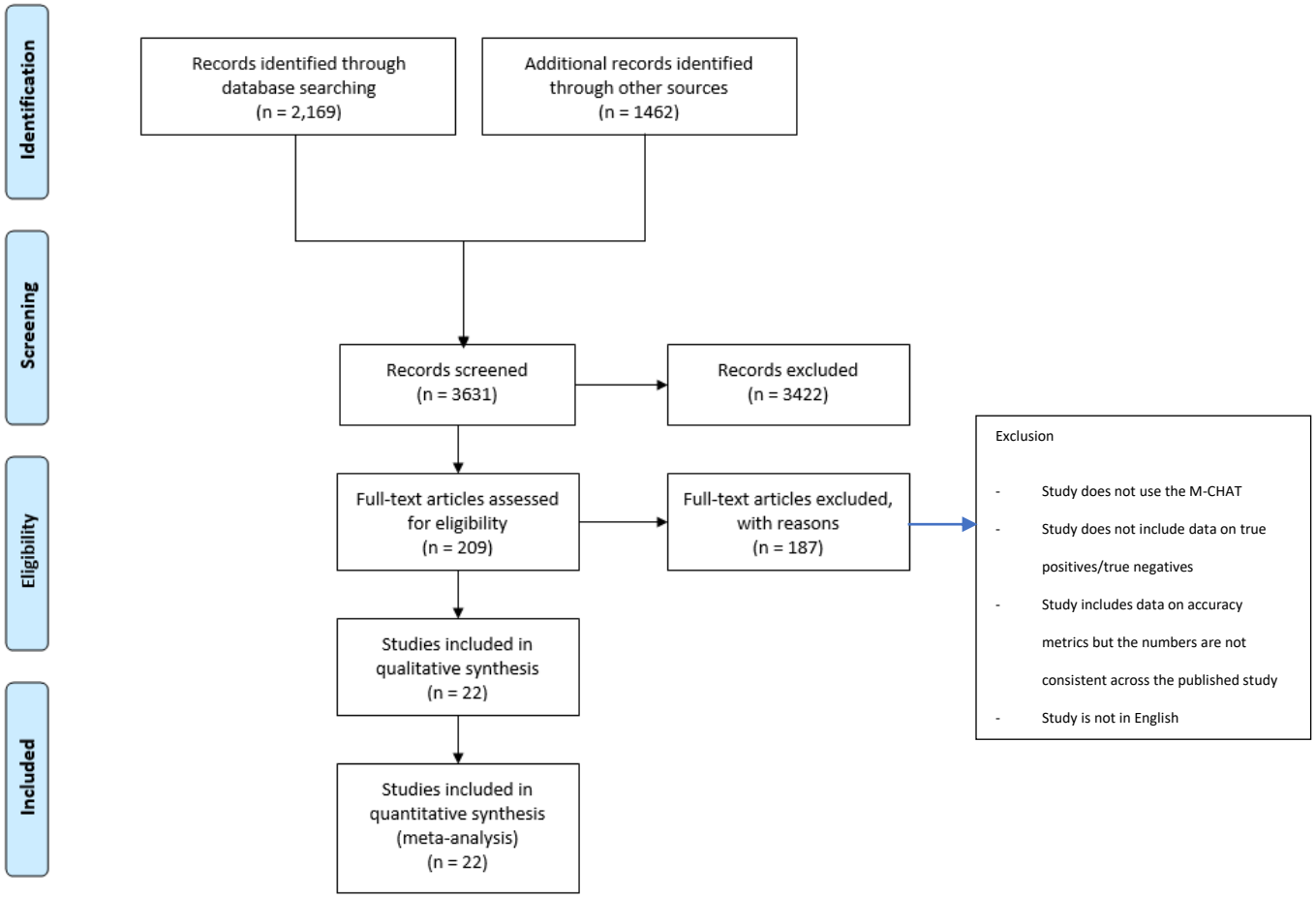
Results

Qualitative Review

As described in Figure 5 the database search yielded 2,169 studies and the ancestral search yielded 1,462 studies. 3,422 studies were either duplicates or were excluded through the title and abstract search and 209 studies were accessed for their full text qualities. 22 studies were finally included in the analysis. Two publications reported data on multiple independent samples within each manuscript (each had two independent samples). Thus, we report analyses on 24 data-sets across 22 total studies. Of these 24 sample sets 15/24 (62%) are population based and 9/24 (38%) are clinical/high risk samples.

Figure 5

PRISMA Flowchart of Study Selection



The characteristics of the included studies are presented in Table 1. All identified studies were published between 2008 and 2019, and were conducted across 14 countries. The wide range of diagnostic measures used are also available for analysis. The ADI-R and ADOS are the most commonly reported 13/22 (59%) studies that use them because they are widely recommended instruments. There are 24 sample sets of which 15/24 (62%) are population based and 9/24 (38%) are clinical/high risk samples. There are 4 *Excellent*, 3 *Good*, 11 *Adequate*, 2 *Poor*, and 2 *Unacceptable* studies ranked by reference standard criteria.

Table 1

Summary of Included Study Characteristics

Study	Year	Country	Study Type	Clinical Description	Diagnostic Measure	Diagnosis	Rank Based on Figure 4
Baduel et al. (2017)	2017	France	Population Based/General Population		ADOS, VABS, PEP-R	Single Clinician	*Adequate*
Canal-Bedia et al. (2011)	2011	Spain	Population Based/General Population		ADOS, VABS	Multiple Clinicians	*Excellent*
Chlebowski et al. (2013)	2013	USA	Population Based/General Population		ADOS, ADI-R, VABS, CARS	Single Clinician	*Adequate*
Coelho-Medeiros et al. (2019)	2019	Chile	High-Risk	Flagged by pediatricians	ADOS	Unsure - Unclearly Reported	*Poor*
Cuesta-Gomez et al. (2016)	2016	Argentina	Population Based/General Population		None	Unsure - Unclearly Reported	*Unacceptable*
Guthrie et al. (2019)	2019	USA	Population Based/General Population		Unclear	Single Clinician	*Adequate*
Hoang et al. (2019)	2019	Vietnam	Population Based/General Population		None	Multiple Clinicians	*Good*
Kamio et al. (2014)	2014	Japan	Population Based/General Population		ADOS, ADI-R, CARS	Multiple Clinicians	*Excellent*

Table 1 (Continued)

Study	Year	Country	Study Type	Clinical Description	Diagnostic Measure	Diagnosis	Rank Based on Figure 4
Kerub et. al. (2018)	2018	Israel	Population Based/General Population		None	Unsure - Unclearly Reported	*Unacceptable*
Kondolot et. al. (2016)	2016	Turkey	Population Based/General Population		CARS	Single Clinician	*Adequate*
Magan-Maganto et. al. (2018)	2018	Spain	Population Based/General Population		ADOS, VABS	Single Clinician	*Adequate*
Matson et al. (2013)	2013	USA	High-Risk	Early intervention system	None	Single Clinician	*Adequate*
Nygren et al. (2012)	2012	Sweden	Population Based/General Population		ADOS, VABS	Multiple Clinicians	*Excellent*
Oien et al. (2018)	2018	Norway	Population Based/General Population		ADOS, ADI-R	Unsure - Unclearly Reported	*Poor*
Oner & Munir (2019)	2019	Turkey	Population Based/General Population		ADOS	Single Clinician	*Adequate*
Snow & Lecavalier (2008)	2008	USA	High-Risk	Specialty clinic referrals	ADOS, ADI-R	Multiple Clinicians	*Excellent*

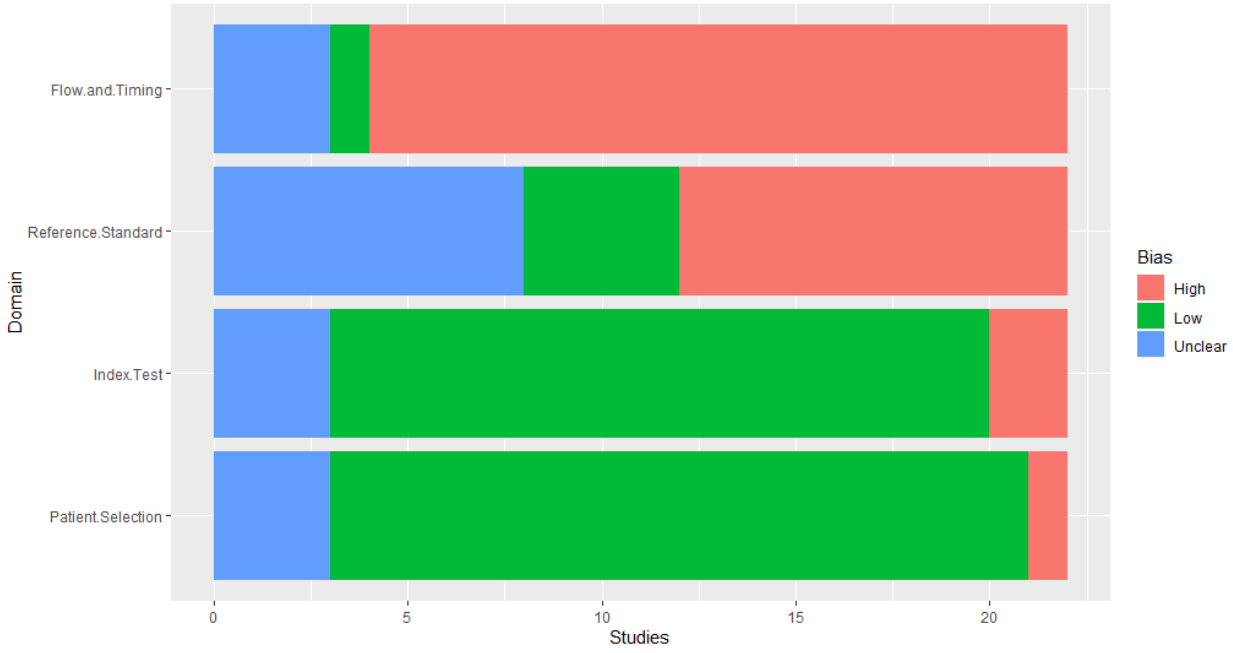
Table 1 (Continued)

Study	Year	Country	Study Type	Clinical Description	Diagnostic Measure	Diagnosis	Rank Based on Figure 4
Srisinghasongkram et al. (2016)	2016	Thailand	High-Risk	Identified with language delay	None	Multiple Clinicians	*Good*
Sturner et al. (2013)	2013	USA	High-Risk	Testing at an Autism Center	ADOS	Single Clinician	*Adequate*
Toh et al. (2018)	2018	Malaysia	Population Based/General Population		None	Single Clinician	*Adequate*
Topcu et al. (2018)	2018	Turkey	High-Risk	Social Pediatrics Department Ankara University	None	Single Clinician	*Adequate*
Tsai et al. (2019)	2019	Taiwan	High-Risk	Referred from Child Development Clinics	None	Multiple Clinicians	*Good*
Wong et al. (2018)	2018	Taiwan	High-Risk	Suspected developmental delays from home visits	ADOS	Single Clinician	*Adequate*

Figure 5 was created with QUADAS-2 bias assessments for each study's methodological quality. The "Flow and Timing" stands out as the category with a large majority (18/22, 81.8%) of studies with high bias for these studies with reference standard the second highest bias group with 10/22 (45.5%). The other two QUADAS-2 domains provide a large majority of studies matching low bias criteria for qualitative assessments. This qualitative assessment means that up to a quarter of studies are questionable for inclusion in the quantitative analysis.

Figure 5

QUADAS-2 Results



Diagnostic Accuracy of Screening Tools

The accuracy of screening tools was evaluated in 22 peer reviewed publications reporting data on 24 independent samples utilizing the M-CHAT for screening population based or high-risk samples. The full collected study metrics are shown in Table 2 below. As described in Table 3 the pooled sensitivity was 0.782 (95% CI 0.663-0.867) and the pooled specificity was 0.980 (95% CI 0.941-0.988). The LR+ was 41.826 (95% CI) and the DOR was 192.100 (95% CI

76.267 – 483.858).. The reported Se of each study varied between 0.18 and 1.00 and the Sp varied between 0.38 and 1.00.

Table 2

Reported Study Accuracy Metrics as Reported by Collected Studies

Study	Year	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+
Baduel et al. (2017)	**2017**	**12**	**1201**	**8**	**6**	**0.67**	**0.99**	**0.60**	**1.00**	**100.75**
Canal-Bedia et al. (2011)	2011	6	2024	25	0	1.00	0.99	0.19	1.00	81.96
Canal-Bedia et al. (2011)	2011	23	2394	63	0	1.00	0.97	0.27	1.00	39.00
Chlebowski et al. (2013)	2013	92	18269	79	6	0.94	1.00	0.34	1.00	218.03
Coelho-Medeiros et al. (2019)	2019	1	90	1	0	1.00	0.83			
Coelho-Medeiros et al. (2019)	2019	17	0	3	0	1.00	0.83			
Cuesta-gomez et al. (2016)	2016	1	402	1	0					
Hoang et al. (2019)	2019	129	17021	118	1	0.99	0.99	0.51	1.00	144.13
Kamio et al. (2014)	2014	20	1683	24	22	0.48	0.99	0.45	11.00	33.43
Kerub et al. (2018)	2018	7	1538	43	3	0.70	0.98	0.20	1.00	
Kondolot et al. (2016)	2016	2	2004	15	0	1.00	0.99	0.12	1.00	134.60
Magan-Maganto et al. (2018)	2018	9	3485	10	2	0.82	1.00	0.08	1.02	285.95
Matson, et al. (2013)	2013	150	151	150	101	0.60	0.50			1.20

Table 2 (Continued)

Study	Year	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	LR+
Oien et al. (2018)	2018	69	67969	1402	228					
Oner & Munir (2019)	2019	57	6388	95	0	1.00	0.91	0.09	1.00	
Snow & Lecavalier (2008)	2008	38	5	8	5	0.88	0.38	0.83	0.50	1.44
Srisinghasongkram et al. (2016)	2016	49	785	2	5	0.91	1.00	0.96	0.99	357.06
Sturner et al. (2013)	2013	23	4568	17	16	0.59	1.00	0.38	6.22	159.06
Topcu et al. (2018)	2018	3	465	15	0	1.00	0.92	0.07	1.00	
Tsai et al. (2019)	2019	19	273	17	3	0.88	0.94	0.61	0.99	
Wong et al. (2018)	2018	65	115	58	14	0.46	0.93		1.92	
Oien et al. (2018)	2018	69	67969	1402	228					
Oner & Munir (2019)	2019	57	6388	95	0	1.00	0.91	0.09	1.00	
Snow & Lecavalier (2008)	2008	38	5	8	5	0.88	0.38	0.83	0.50	1.44

The reported TP, TN, FP, and FN of table 2 were used to calculate sensitivity, *Sp*, and DOR for all studies which can be seen in Figures 6,7, and 8. The forest plots in Figure 6 show substantial variability in sensitivity with a minimum point estimate of 0.23 and a maximum of 0.99. For *Sp* many studies show highly specific estimate with fewer exhibiting large confidence intervals, the min. was 0.39 and max. was 0.99.

Figure 6

Sensitivities of Included Studies Organized by Population Based or High-Risk

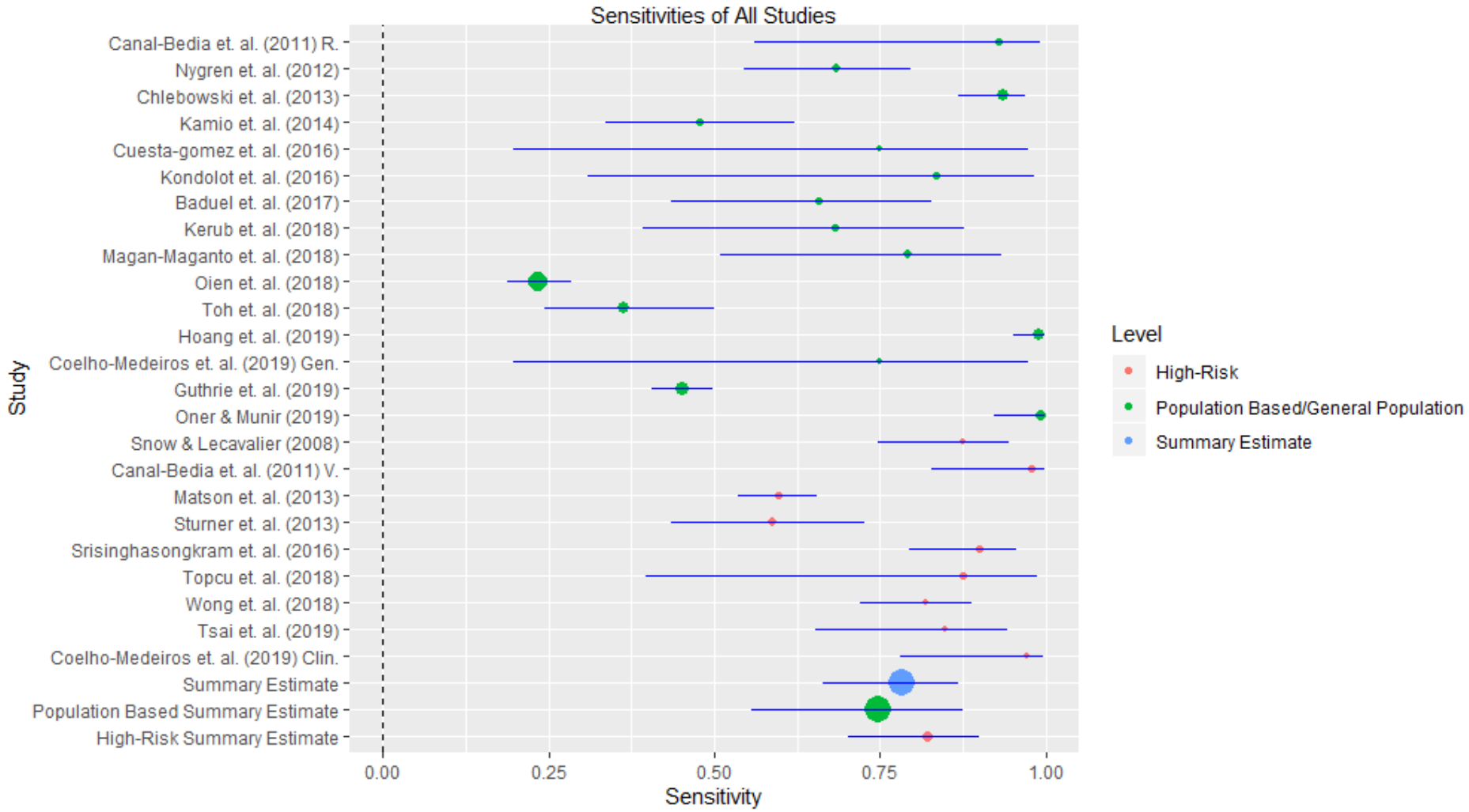


Figure 7

Specificities of Included Studies Organized by Population Based or High-Risk

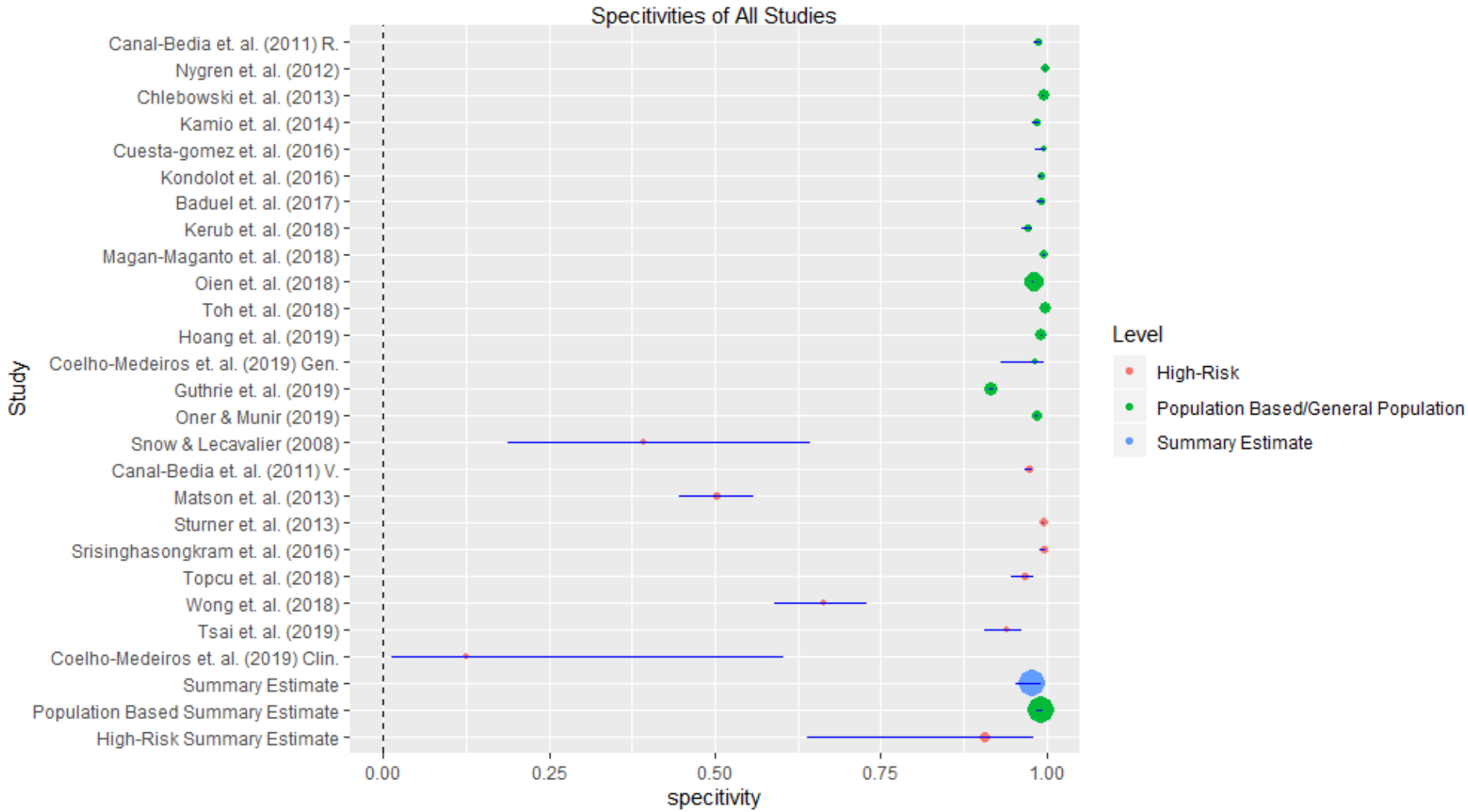


Figure 8

Natural LOG DOR of Included Studies Organized by Population Based or High-Risk

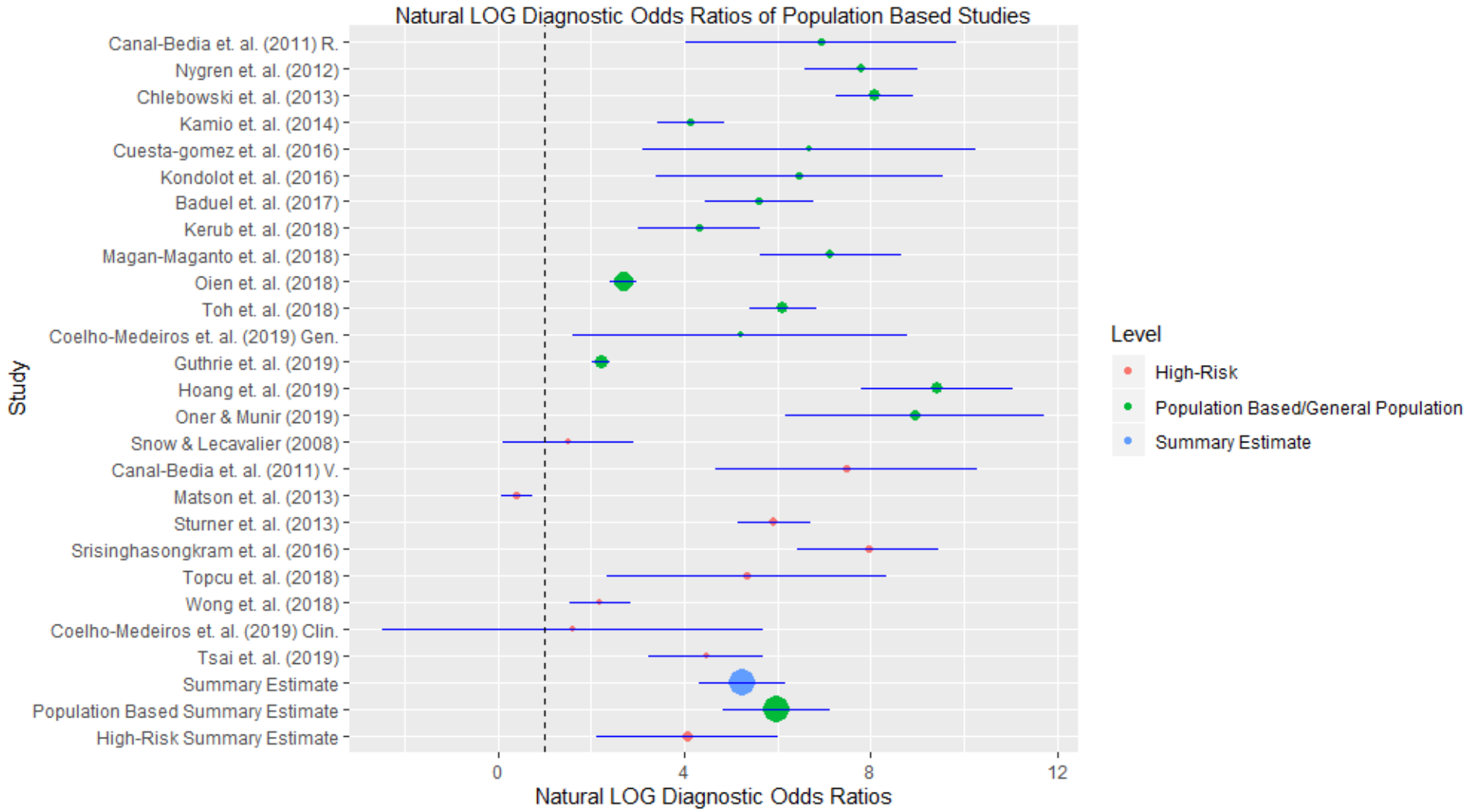


Table 3

Meta-Analysis Estimates All Studies

	Point Estimate	95% Lower Bound	95% Upper Bound
DOR	192.100	76.267	483.585
Pooled Sensitivity	0.782	0.663	0.867
Pooled Specificity	0.980	0.954	0.991

Exploration of Heterogeneity

Using the Cochran Q test, the DOR of 24 data sets were assessed for heterogeneity and the results were found to be 28.47 (df = 23, p = 0.198) which shows there is insubstantial heterogeneity in the sample. However, Study Type level analyses were planned *a priori* and thus conducted. In addition to the Cochran Q test, the spearman rho correlation between sensitivity and false positive rate (FPR) is 0.184 (-0.237, 0.547). This suggests there are no threshold effects between sensitivity and FPR in this analysis (though, as previously mentioned, there is low power).

Using all 24 sets of data an SROC curve was constructed displaying sensitivity and FPR. The studies were similar in that all M-CHAT scoring was based on suggested original cut-off values proposed by Robins (2001, 2014) except one study, Kamio et al. (2014). A variation between the studies collected was whether or not they used solely the M-CHAT's initial stage, the follow-up stage or a combination of the two. The prediction region shows a wide range for both FPR and sensitivity when considering all studies. Figure 9 demonstrates that studies vary in their results even though they all report on M-CHAT screening. Figure 10 shows the entire sample of studies sensitivity related to the FPR visualized as an SROC. The 95% interval is situated above 0.6 sensitivity and less than 0.1 FPR. There are four notable studies with low

FPR; it is noteworthy is that all four are high-risk studies (Coelho-Modeiros et al., 2019; Matson et al., 2013; Snow & Lecavlier, 2008; Wong et al., 2018).

Figure 9

ROC Ellipse and Confidence Intervals of all Studies

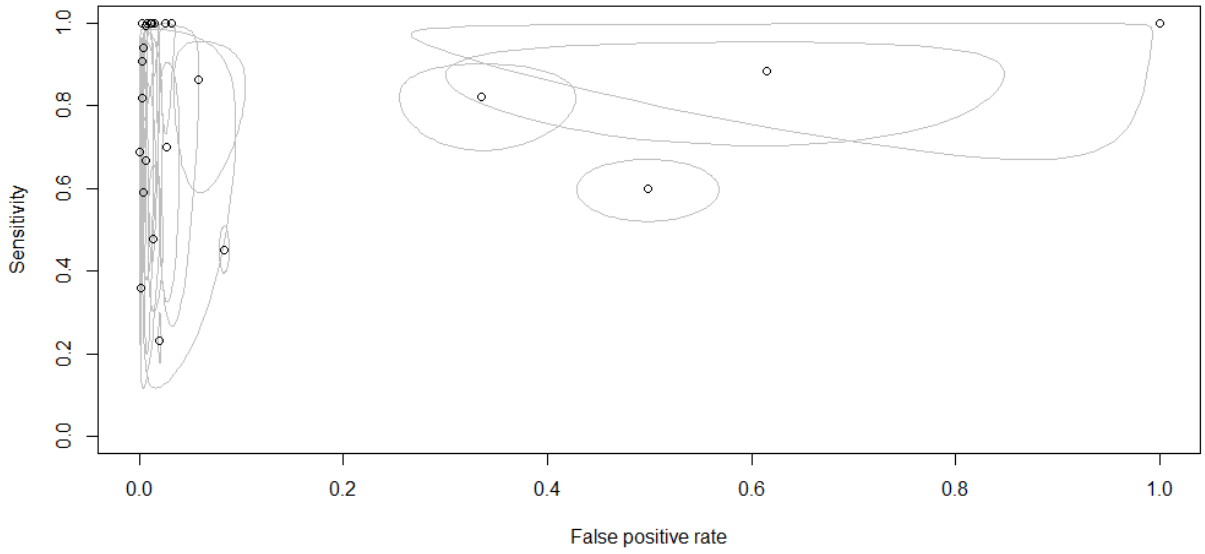
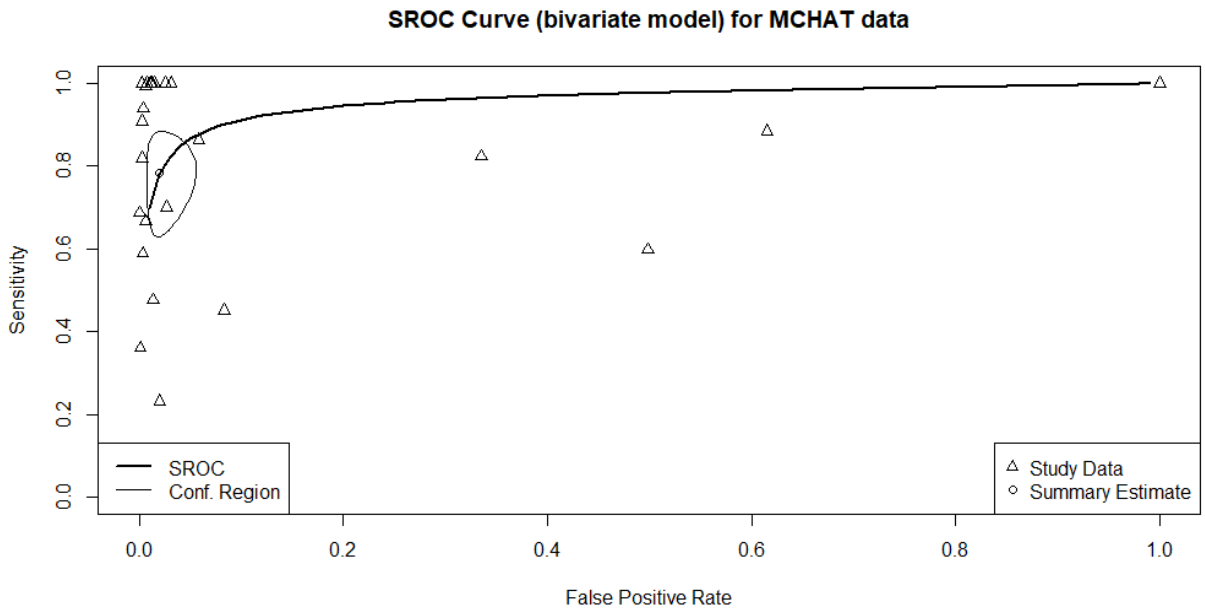


Figure 10

Bivariate Reitsma SROC of All Included Studies



Subgroup Comparisons

Population vs. Clinical Study Types

Even though the data was not found to be heterogeneous based on the Cochrane Q test, subgroup comparisons were performed between the studies examining general populations and those examining clinical populations. When examining Figures 6, 7, and 8 again the studies are organized first by Study Type, population based first and then high-risk, in order to visually describe similarities by Study Type. The clinical studies display substantially lower and different sensitivity and specificity than population based studies this is shown in Figure 6 and 7.

The forest plot of Figure 6 shows variability in the sensitivity of the studies as a whole, but assessing the sensitivity by population based or high-risk reduces the variation and partially explains some Study Type relationship. A similar visual description is seen in Figure 7 where

variability is largely contained by the high-risk studies, and Figure 8, where the DOR is high and significant for all population based studies but some high-risk show non-significant results. The population specificities have a tight range from 0.92 to 1.00 while the clinical has a much wider range of estimates. Splitting these by their subgroups has helped to clarify the differences between these two study types when describing their sensitivity and specificity. Examining the high-risk studies once again, in Figure 8, shows more variability where one study, Matson (2013), had a DOR of less than 1.0 with a confidence interval that did not cross 1.0 indicating a significant result.

When inspecting the numbers closer, Table 4 serves to show the specifics of the univariate characteristics of data sets when they are split. The summary estimates that are used in Figures 6, 7, and 8 are shown here more clearly with their confidence intervals (95% CI). For population based studies the pooled sensitivity was 0.746 (0.555, 0.874) and the pooled *Sp* was 0.992 (0.985, 0.995). The estimate of sensitivity for the high-risk study types (0.821 (0.701, 0.900) is higher than the population based counterpart, but has a lower estimate for specificity 0.906 (0.639, 0.981). The difference in DOR between the two is large as well with a population DOR of 396.756 (126.753, 1241.906) and high-risk of 58.389 (8.318, 409.866). Like with the full sample Cochran Q and Spearmans's rho is non-significant for each of these study samples.

Table 4

Univariate Statistical Measures of Included Studies by Study Type

	Population	df	p-value	Clinical	df	p-value
k	15			9		
Equality of Sensitivities	364.986	14	<0.001	56.8563	8	<0.001
Equality of Specificities	4586.487	14	<0.001	2621.484	8	<0.001
Rho (95% CI)	-0.388 (-0.725, 0.211)			0.114 (-0.595, 0.723)		
DOR (95% CI)	396.756 (126.753, 1241.906)			58.389 (8.318, 409.866)		
Cochran's Q	13.488	14	0.489	6.95	14	0.542
Tau of DOR (95% CI)	2.055 (0.000, 2.551)			2.811 (0.000, 4.388)		
Tau^2 of DOR (95% CI)	4.221 (0.000, 6.508)			7.901 (0.000, 19.258)		
Sensitivity (95% CI)	0.746 (0.555, 0.874)			0.821 (0.701, 0.900)		
Specificity (95% CI)	0.992 (0.985, 0.995)			0.906 (0.639, 0.981)		

The following tables represent the meta-regression analysis. Table 5 reports the log likelihood of each model and the comparison to the null model. The log likelihood of the model

incorporating Study Type (population based and high-risk) is 64.075. The Chi-squared analysis of the model is significant at p = 0.004 showing that Study Type model explains some of the heterogeneity between included studies. Finally Table 6 shows the meta regression table for Study Type. For the Study Type model the regression coefficient for the FPR is significant, indicating that the population based studies result in a better FPR than high-risk studies.

Table 5

Regression Model Performance Comparison Characteristics

	Study Type	
	Model	Null Model
Log Likelihood	64.075	57.850
df	7	5
k	2	2
AIC	-114.149	-105.699
BIC	-101.051	-96.343
CHI-SQ (df, p-value)	10.84 (2, 0.004*)	

Note. Significant p-value set at α=0.10

Table 6

Bivariate Meta-Regression Coefficients

	Estimate	2.5% CI	97.5% CI	p-value
	Study Type Model			
Sensitivity (Intercept)	1.673	0.712	2.634	0.001*
Sensitivity Population Based	-0.647	-1.871	0.577	0.300
FPR (Intercept)	-2.343	-3.508	-1.179	0.000*
FPR Population Based	-2.443	-3.904	-0.982	0.001*

Note. The control group for the Study Type model is High-Risk samples.

Multi-Clinician vs Other Reference Standard

Figures 11, 12, and 13 show the sensitivity, specificity, and DOR with summary estimates when organized by reference standard level. These forest plots are very similar to Figures 6, 7, and 8 for use describing the breakdown of these two groups visually. When observing these studies organized by Multi-Clinician and Other there appears to be variability in both levels of reference standard quality. Sensitivity shows high variability throughout the studies in Figure 11 while Specificity shows less variability overall in Figure 12. Unlike the Study Type breakdown, where most variability was seen in the high-risk studies, the variability of Reference Standard does not appear to be represented by one level more than the other. Using Table 7 The pooled sensitivity estimate of Multi-Clinician reference standard in 0.874 and the pooled estimate of Other is 0.711. The CI regions of these estimates overlap but the point estimate difference is 0.163 for sensitivity. The specificities on the other hand are very similar, 0.983 and 0.978 for Multi-Clinician and Other respectively. Finally, the DORs are 460.927 and 121.330 for Multi-Clinician and Other respectively show a large difference between the two point estimates and a much higher DOR for Multi-Clinician.

Descriptive characteristics of these groupings of data are the Cochran Q test and Spearman's Rho. The Cochran's Q shows a non-significant result indicating no heterogeneity within the groupings of *excellent* and *good* or the grouping of *adequate, poor,* and *unacceptable*. Spearman's rho for these sets of data are 0.102 (-0.649, 0.753) and 0.247 (-0.283, 0.662) both non-significant results and indicating no threshold effects between sensitivity and FPR in either of these sets of data. These results can be seen in Table 4.

Figure 11

Sensitivities of Included Studies Organized by Multi-Clinician or Other Reference Standard

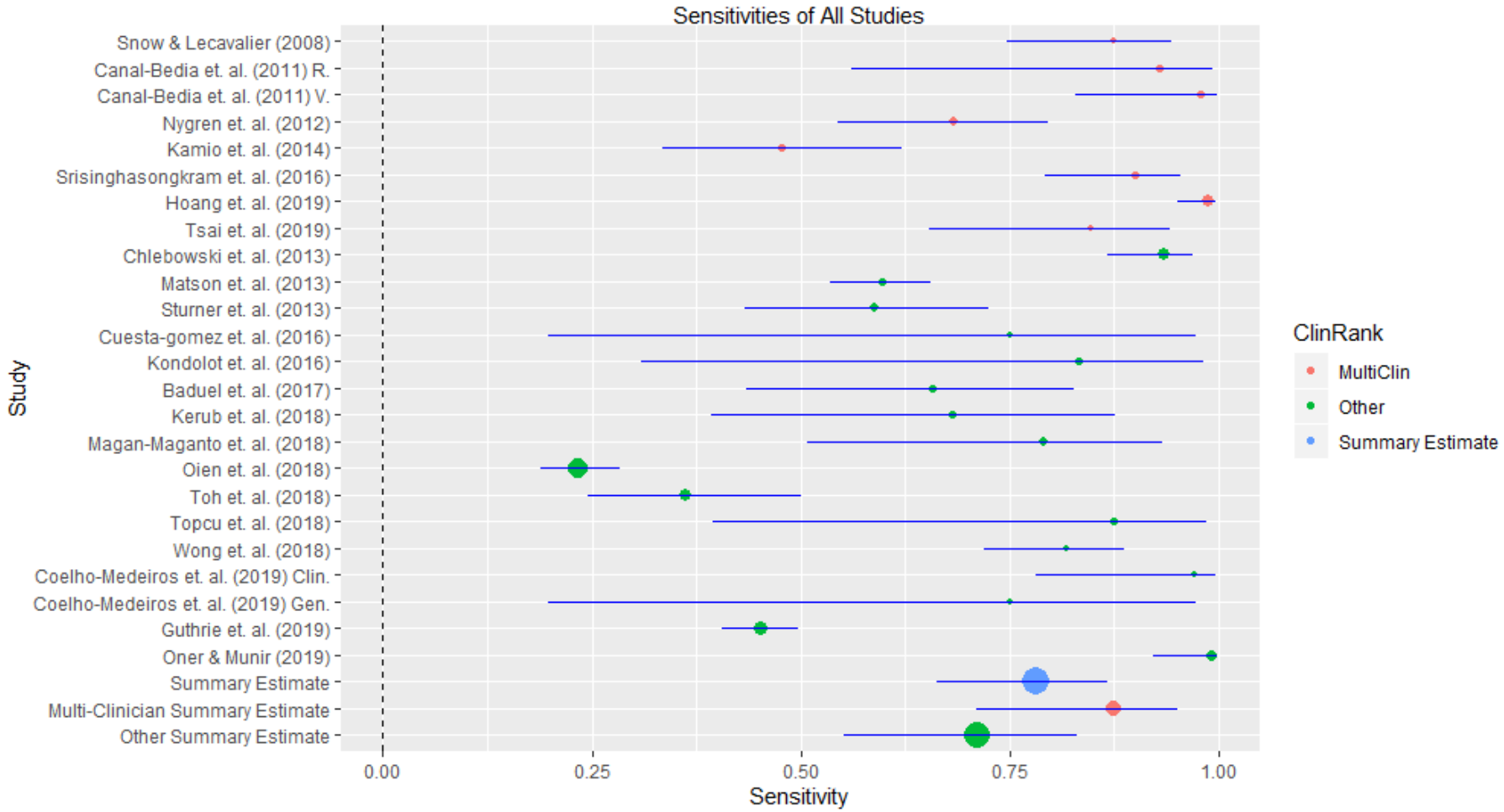


Figure 12

Specificities of Included Studies Organized by Multi-Clinician or Other Reference Standard

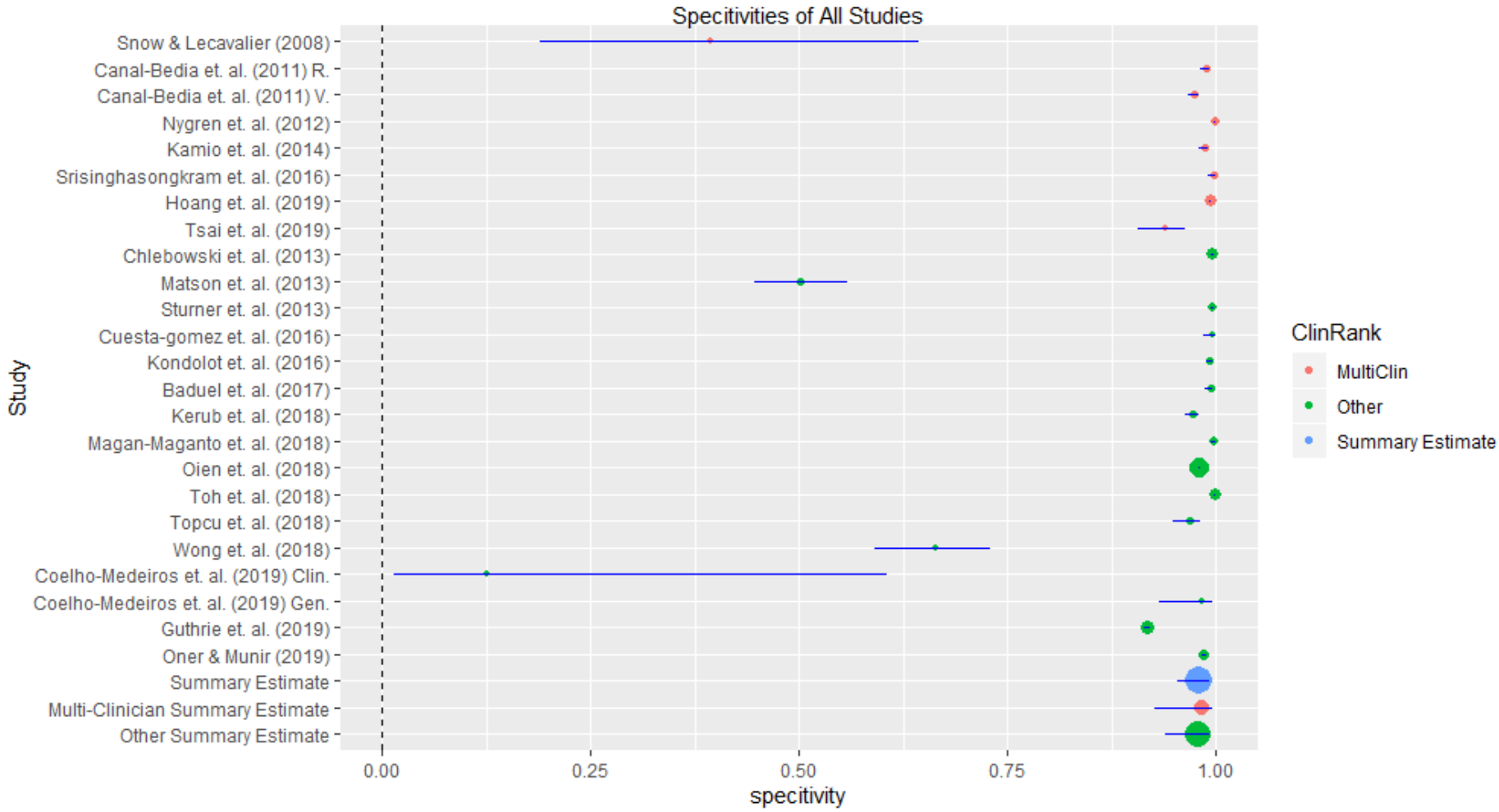


Figure 13

Natural LOG DOR of Included Studies Organized by Multi-Clinician or Other Reference Standard

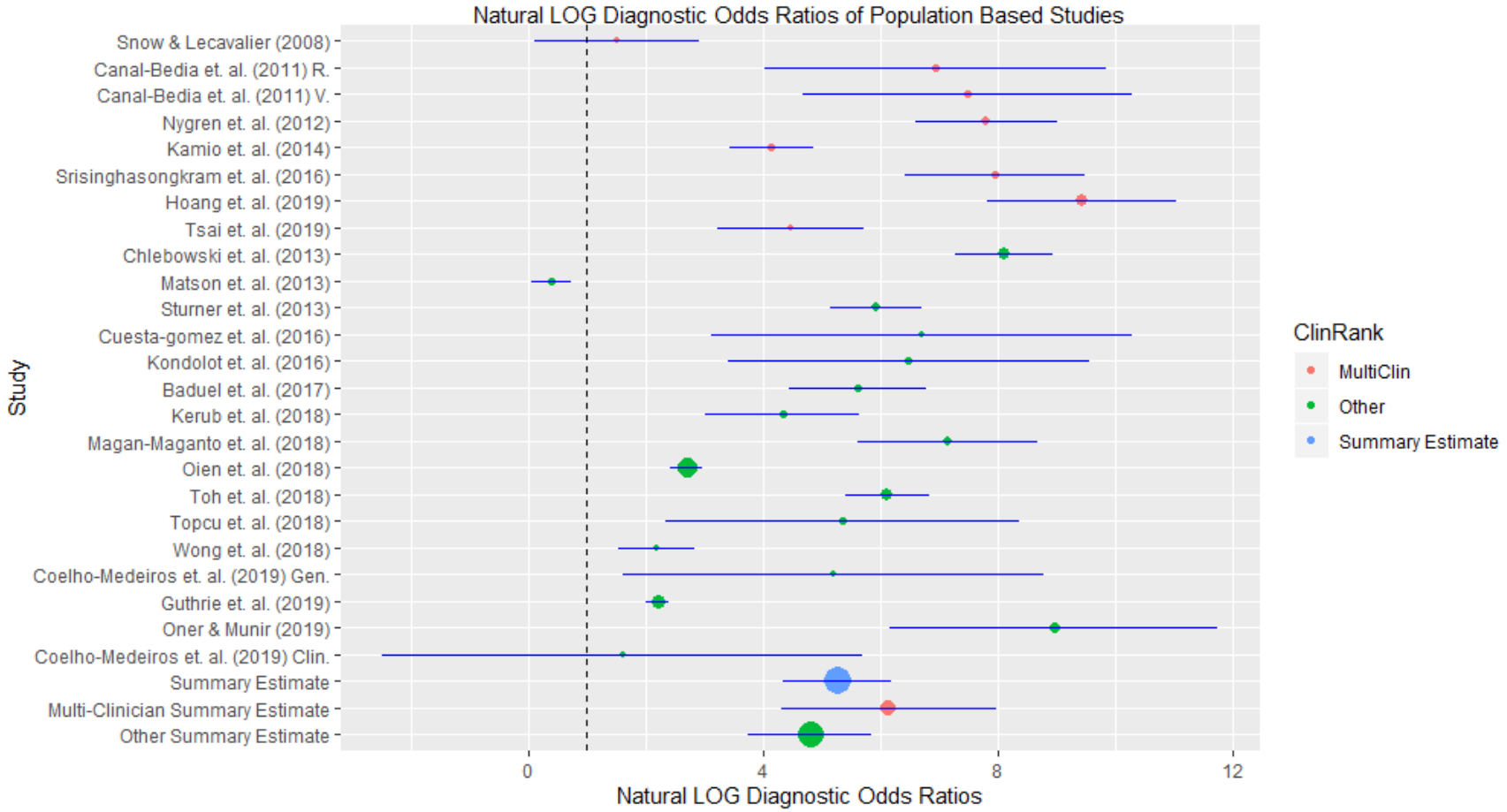


Table 7 shows the univariate characteristics of data sets when they are split by Reference Standard level. The summary estimates that are used in Figures 11, 12, and 13 are shown here more clearly with their confidence intervals. For Multi-Clinician studies the pooled sensitivity was 0.874 (0.711, 0.951) and the pooled *Sp* was 0.983 (0.927, 0.996) which is higher than the estimate of sensitivity and specificity for the Other reference standard study types, 0.711 (0.552, 0.831) and 0.978 (0.939, 0.992). The difference in DOR between the two is large as well with a Multi-Clinician DOR of 460.927 (73.658, 2884.340) and Other DOR of 121.330 (42.312, 347.913). Finally, Cochran's Q and Spearman's rho are non-significant for both of these sets of studies.

Table 7

Univariate Statistical Measures of Included Studies by Reference Standard

	Multi-Clinician	df	p-value	Other	df	p-value
k	8			16		
Equality of Sensitivities	83.797	7	<0.001	291.9054	15	<0.001
Equality of Specificities	748.783	7	<0.001	7423.318	15	<0.001
Rho (95% CI)	0.102 (-0.649, 0.753)			0.247 (-0.283, 0.662)		
DOR (95% CI)	460.927 (73.658, 2884.340)			121.330 (42.312, 347.913)		
Cochran's Q	7.008	7	0.428	20.253	15	0.162
Tau (95% CI)	2.491 (0.000, 4.617)			1.929 (0.000, 3147)		
Tau^2 (95% CI)	6.204 (0.000, 21.314)			3.720 (0.000, 9.902)		
Sensitivity (95% CI)	0.874 (0.711, 0.951)			0.711 (0.552, 0.831)		
Specificity (95% CI)	0.983 (0.927, 0.996)			0.978 (0.939, 0.992)		

Table 8 shows the comparison of the Reference Standard model to the null model. The CHI-SQ analysis is not significant for this model, ($p = 0.240$). This model regressed on Reference Standard does not substantially explain the heterogeneity of these studies more than the null model. Table 9 shows the results of the meta-regression as a regression table for Reference Standard model. For the Reference Standard model neither regression coefficient is significant indicating neither reference standard ranking of Multi-Clinician or Other offers significantly better sensitivity or FPR. However, these findings should be judged in light of the low power of the study and that we considered $p < .10$ to indicate potentially meaningful relationships; the Reference Standard analysis indicates a trend may be present wherein studies with Multi-Clinician reference standards have higher sensitivities than those without.

Table 8

Regression Model Performance Comparison Characteristics

	Reference Standard	
	Model	Null Model
Log Likelihood	60.466	57.850
df	7	5
k	2	2
AIC	-106.932	-105.699
BIC	-93.833	-96.343
CHI-SQ (df, p-value)	2.855 (2, 0.240)	

Table 9

Bivariate Meta-Regression Coefficients

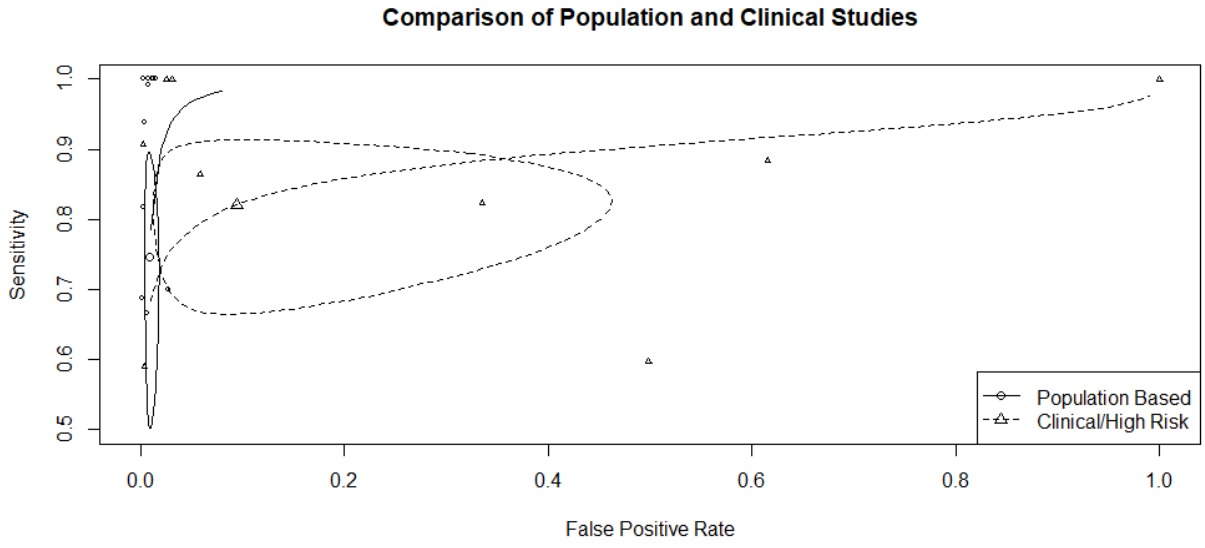
	Estimate	2.5% CI	97.5% CI	p-value
	Reference Standard Model			
Sensitivity (Intercept)	1.913	0.930	2.897	0.000*
Sensitivity Other	-0.999	-2.213	0.214	0.107
FPR (Intercept)	-4.034	-5.523	-2.545	0.000*
FPR Other	0.243	-1.585	2.071	0.794

Note. The control group for the Reference Standard Model is Multi-Clinician reference standard.

Diagnostic accuracy metrics differ between population and clinical studies as illustrated by Table 4 and in the bivariate Reitsma illustrated in Figure 14. The population based and high-risk study samples can further be broken up by reference standard quality categories. This further analysis can be observed in Appendix D where Supplemental Figure 1 and 2 show SROC plots of population based and high-risk studies by reference standard criteria. When assessing the population-based studies by sensitivity and FPR (1-specificity), they are all grouped near 0 FPR, but widely variable when assessed by sensitivity. The high-risk studies are variable in both sensitivity and false positive rate indicating these types of studies may vary on more between them in terms of study variance.

Figure 14

Comparison of Bivariate Reitsma models for Population Based and High-Risk Studies Through SROC

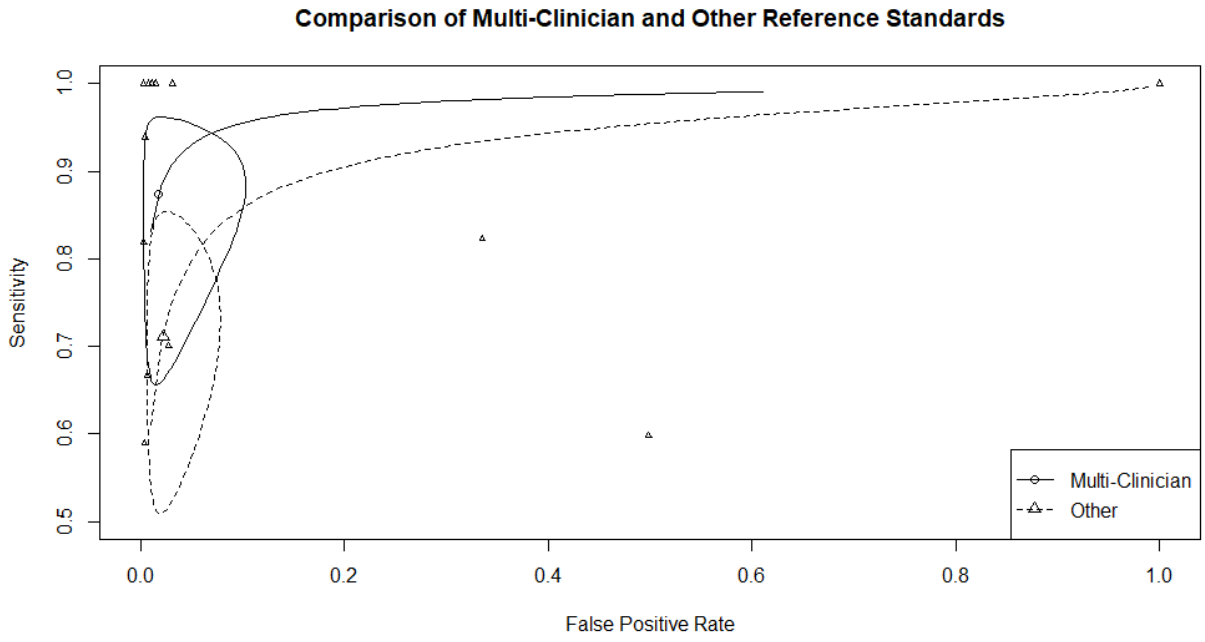


Finally, the last subgroup comparison, Reference Standard, is shown in Figure 15 where the studies have been split into groups based on their strength of reference standard. While the proposed framework of ranking screening accuracy studies has 6 levels, the studies of this

systematic review only identified reference 5 levels of reference standard: *Excellent*, *Good*, *Adequate*, *Poor* and *Unclear*. These were split into Multi-Clinician and Other categories. The SROC ellipse shows Multi-Clinician and Other overlap however the Multi-Clinician group appears to be better than the Other category.

Figure 15

Comparison of Bivariate Reitsma models for Multi-Clinician and Other Reference Standard Levels Through SROC



Discussion

This study identified 24 samples across 22 peer-reviewed publications reporting data on TP, TN, FP, and FN from which we conducted a MADA. The combined accuracy was summarized by the bivariate Reitsma model. The results of the full combined study sample analysis shows a summary estimate of 192.100 (95% CI 76.267, 483.858) indicating that the odds of a positive result on the M-CHAT is 192 times more likely on a patient with ASD than on a patient without ASD. Other than DOR the LR+ was 37.368 indicating positive cases were likely to be indicated by a positive test result and the LR- was 0.337 indicating that positive cases were not likely to be indicated by a negative test result. Despite the fact that our heterogeneity scores did not indicate any significant between-study heterogeneity, the decision was made a priori to split the studies by the level at which their populations were collected, population-based vs high-risk. This knowledge was informed by Tipton et al. (2019) who recommend best practice of reducing confounding by separating moderators which may have an association with one another; in this case study level of population based and high-risk samples. The meta-regression showed that population-based samples have significantly lower FPR than high-risk studies. More generally, when compared to clinical studies, population-based studies appear to show much higher DOR, 396.756 (126.753, 1241.906) than high-risk studies, 58.389 (8.318, 409.866) and pooled specificity, 0.992 (0.985, 0.995) to 0.906 (0.639, 0.981) but a lower pooled sensitivity 0.746 (0.555, 0.874) than the high-risk studies 0.821 (0.701, 0.900).

While population based vs high-risk was an expected sample difference, the main focus of this study was clinical reference standard. We conducted an additional meta-regression of reference standard showing that while we hypothesized an inverse relationship between reference standard, this was not fully explored due to the low power of this universe. Instead, we

conducted a meta-regression between Multi-Clinician reference standards and Other which yielded interesting data. The Multi-Clinician reference standard showed a higher DOR 460.927 (73.658, 2884.340), sensitivity 0.874 (0.711, 0.951), and specificity 0.983 (0.927, 0.996) than the Other group's results of 121.330 (42.312, 347.913), 0.711 (0.552, 0.831), 0.978 (0.939, 0.992), respectively. The DOR and sensitivity differences appear to be large, however the specificity differences are quite small. This binary split shows a positive relationship between multi-clinician outcomes and reference standard quality, however this result was not significant when performing a meta-regression. Ideally the number of studies included would give sufficient power to every reference standard category. This study has too few studies per group and so a grouping technique, Mutli-Clincian and Other, was employed to increase the size of the categories.Even when employing this technique the power remains low. Collectively, these results indicate that there is not a significant difference between Multi-Clinician reference standard and Other when determining sensitivity and FPR, though meta-regression suggested a trend.

A secondary analysis was attempted when studies were split by Study Type. These results are displayed in Appendix D. A meta-regression was not performed on this data, due to low power. If this work is analyzed only visually, the relationships apparent show a small difference in the sensitivity between population based samples who employ Multi-Clinician reference standards over Other but not much difference in FPR. For High-Risk samples the sensitivity once again is favored by Multi-Clinician, however FPR is significantly different between these two. The high-risk sample included 9 studies total which may bias these results. The conclusion reached from this analysis is that these results can be viewed as preliminary. The relationship between Reference Standard within population based or high-risk studies should be pursued by a systematic review and meta-analysis with more inclusion criteria, possibly through including the grey literature.

This study is unique in the universe of screener reviews in a number of different ways. First, this study includes both a study that Sanchez-Garcia missed, Magan-Maganto and a sample that Yuen missed. Yeun also notes Stenberg et al. (2014), but our study includes this sample in Oien et al. (2018). Furthermore, of the existing quantitative reviews none explore the effect of high-risk vs population studies or reference standard effects on diagnostic outcomes. Despite its non-significance, this analysis helps underscore that studies utilizing different reference standards might impact diagnostic outcomes, and should be considered for other MADA studies.

The particular focus on the M-CHAT is important as recent research from Guthrie et al. (2019) indicates that the current published screener studies have identified too few cases and long term follow-up shows the M-CHAT has a lower rate of accuracy than initially thought. Specifically, their calculated sensitivity was 0.388 and specificity was 0.949, which is substantively lower than sensitivity of .911 and specificity of .955 as reported by Robins et al. (2014). A preliminary exploratory analysis in Appendix E showed that using a population prevalence baseline estimate to adjust FN and TN metrics can mimic the effect of complete follow-up. Supplemental Table 1 shows the original reported study values, population prevalence estimate matched by country and year and adjusted study metrics. Supplemental Figure 3, 4, and 5 show sensitivity, specificity, and DOR that can be compared to original forest plots in figure 6, 7, and 8 in the population based study section. This strategy was an idea taken from Barbaro et al. (2010) who used Australian prevalence estimate rates for ASD instead of following-up all negatives. After using their idea of population prevalence baseline, our largest study, Chlebowski, had adjusted sensitivity of 0.34 which closely resembles the follow-up adjusted metrics of the Guthrie study. More information on the preliminary exploratory analysis can be

observed in Appendix E. Guthrie leaves readers with the task of identifying new methods to detect a greater proportion of children with ASD.

The studies included in this review were selected because the authors reported all TP, TN, FN, and FP data points. The main focus was to use clear psychometric data to calculate MADA while fitting the analysis into the timeline of a thesis. While these studies served their purpose of psychometric data reporting, some studies were found to lack positive qualities from an ASD identification process focus, based on Figure 1. Certain studies were noted during the QUADAS-2 process for not even including a flow chart, specifically Matson et al. (2013), Oien et al. (2018), Snow& Lecavalier (2008), and Toh et al. (2018) making it difficult to ensure numbers reported could be correctly replicated or when trying to understand the flow of their studies visually. QUADAS-2 drastically helped to pinpoint which domains this literature base needed improvement on; specifically Domain 4: "Flow and Timing" needed improvement because many studies resulted in high bias estimates (18/22, 81.8%) due to the amount of questions that had to be marked "unclear". The QUADAS-2 covers all categories that need to be reviewed for collecting and reviewing studies for a meta-analysis. Besides Flow and Timing, results for the Domains: 1 Patient Selection and 2 Index Test were generally low bias while Domain 3: Reference Standard has a large number of unclear results as marked by this study. The QUADAS-2 review likely found more unclear results due to the ranking system modeled previously by Barger and redesigned for this study. Because of the strict guidelines of information to collection our focus on reference standard categories resulted in numerous high bias estimates as opposed to similar reviews that did not use strict methods.

This study highlights a main area where research on the early identification process can improve: Clarity of reference standard reporting in all literature, particularly peer-reviewed

publications. Screening tools are an important aspect of early identification, but are most useful when predicting in a system with high quality diagnosis and long term ability to continue monitoring potential cases (Barger et al., 2018). Improvements to reference standard reporting in peer-reviewed articles can be seen in this study's diagnostic framework which extends the QUADAS-2 by improving the categorical analysis in Domain 3: Reference Standard. This is done by creating specific and reproducible levels for bias estimates instead of general "low" or "high" bias labels. Thus, this framework might be useful for future meta-analysis studies or for the QUADAS-2 Appendix F to guide new bias categories this may reduce the amount of "unclear" answers for studies when performing a QUADAS-2 review for Domain 3.

Systematic reviews and meta-analyses have tremendously helpful guidelines to follow (Pigott & Polanin, 2020; Polanin et al., 2019). The guidelines help keep the focus on the review question and therefore more able to answer the question. This study adds to systematic review methods by creating robust and defendable clinical reference standard categories that accounts for important insights about reference standard norms from the field of autism research. There may be more room for improvement since the original categories had to be modified to fit the data of this study, but disambiguating single clinician diagnosis from multi-clinician diagnosis is a valid change to the original format. Future direction should seek to use frameworks, such as seen in **Figure 4** *Redesigned Framework for Categorical Assessment of Reference Standards*. It's important for Categorical analysis that the meaning of a category is well defined which is not the case with QUADAS-2 Domain 3: Reference Standard questions. A more focused reference standard review can lead to better comparison of MADA results from other studies by accounting for reference standard quality in the study collection phase.

Limitations

This study required a number of important methodological decisions, each with a potential limitation. First, M-CHAT studies vary in their choice cut-off rules and following up with positive and negative screens, and within text descriptions are often opaque. When available, we prioritized using flow charts to best determine identification procedures and interpret in-text descriptions; however, alternative approaches could be developed. In this review there are two studies that report two sample sets, Canal-Bedia et al. (2011) and Coelho-Medeiros et. al. (2019), because within-publication samples were collected independent of one another they are both reported and not expected to compromise findings. One major advantage of this work is that all studies used recommended M-CHAT cut-off scores (except Kamio) which likely helped control for a source of heterogeneity (Doebler & Bohning, 2010). The implications of including only studies that report all of their own metrics are that they can be misleading. Sensitivity and specificity from partial data across more studies can be back calculated which was not performed by this study to confirm the reported psychometric data reported. This would fill gaps in the current literature because this level of confirmation is not present in current MADA studies.

The second important limitation is related to the number of primary studies that do not follow up all negative screens in order to determine the absolute number of false negatives that exist in their sample. In addition to not following up all negative screens, not all studies report sample demographics which may explain screening differences because individuals of different races have been shown to screen differently with the M-CHAT (Khowaja et al., 2015). This is an important methodological choice for study inclusion because this will positively bias psychometric results due to the likely under-identified number of cases reported in any given

study that does not follow-up completely. During the meta-regression there is indication that population based studies have significantly higher FPR which would the factor most affected by not following up on a sample for TN.

References

AAP Schedule of Well-Child Care Visits. (n.d.). Retrieved June 25, 2019, from https://www.healthychildren.org/English/family-life/health-management/Pages/Well-Child-Care-A-Check-Up-for-Success.aspx

Adak, B., & Halder, S. (2017). Systematic review on prevalence for autism spectrum disorder with respect to gender and socio-economic status. *J Ment Disord Treat*, *3*(133), 2.

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American Psychiatric Association.

American Psychiatric Association: **Cautionary statement.** *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). 2000.

Armstrong, R., Hall, B. J., Doyle, J., & Waters, E. (2011). 'Scoping the scope'of a cochrane review. *Journal of Public Health*, *33*(1), 147-150.

Arnold, C. L., Davis, T. C., Frempong, J. O., Humiston, S. G., Bocchini, A., Kennen, E. M., & Lloyd-Puryear, M. (2006). Assessment of newborn screening parent education materials. *Pediatrics*, *117*(Supplement 3), S320-S325.

Baduel, S., Guillon, Q., Afzali, M. H., Foudon, N., Kruck, J., & Rogé, B. (2017). 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*, *47*(2), 297-304.*

Baio, J. (2014). Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010.

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., Kurzius-Spencer, M., Zahorodny, W., Rosenberg, C. R., White, T., Durkin, M. S., Imm, P., Nikolauo, L., Yeargin-Allsopp, M., Lee, L., Harrington, R., Lopez, M., Fitzgerald, R. T., Hewitt, … Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries*, *67*(6), 1.

Barbaro, J., & Dissanayake, C. (2010). Prospective identification of autism spectrum disorders in infancy and toddlerhood using developmental surveillance: the social attention and communication study. Journal of Developmental & Behavioral Pediatrics, 31(5), 376-385.

Barger, B., Rice, C. & Roach, A. (2018a). Socioemotional developmental surveillance in young children: Monitoring and screening best identify young children that require mental health treatment. *Child and Adolescent Mental Health 23(3),* 206 – 213.

Barger, B., Rice, C., & Roach, A. (2018b). Commentary: Response to Foreman's commentary on detecting unmet mental health needs in preschool children (2018). *Child and Adolescent Mental Health*, *23*(3), 217-219.

Barger, B., Rice, C., Simmons, C. A. & Wolf, R. (2018c). A systematic review of Part C early identification studies. *Topics in Early Childhood Special Education, 38(1),* 4-16.

Barger, B., Rice, C., Wolf, B. & Roach, A. (2018d). Better together: Developmental screening and monitoring best predict Part C early intervention receipt. *Disability and Health Journal 11(3) 420-426.*

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., & De Vet, H. C. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Radiology*, *226*(1), 24-28.

Brennan, L., Fein, D., Como, A., Rathwell, I. C., & Chen, C. M. (2016). Use of the modified checklist for autism, revised with follow up-Albanian to screen for ASD in Albania. *Journal of autism and developmental disorders*, *46*(11), 3392-3407.

Bricker, D., Macy, M., Squires, J., & Marks, K. (2013). Developmental screening in your community. *Baltimore, MD: Paul H. Brookes*.

Bright Futures Steering Committee, & Medical Home Initiatives for Children With Special Needs Project Advisory Committee. (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics*, *118*(1), 405-420.

Bright Futures Steering Committee, & Medical Home Initiatives for Children With Special Needs Project Advisory Committee. (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. Pediatrics, 118(1), 405-420.

Canal-Bedia, R., García-Primo, P., Martín-Cilleros, M. V., Santos-Borbujo, J., Guisuraga-Fernández, Z., Herráez-García, L., Herráez-García, M., Voada-Munoz, L., Fuentes-Biggi, J., & Posada-de La Paz, M. (2011). Modified checklist for autism in toddlers: cross-cultural adaptation and validation in Spain. *Journal of autism and developmental disorders*, *41*(10), 1342-1351.*

Carcani-Rathwell, I., Rabe-Hasketh, S., & Santosh, P. J. (2006). Repetitive and stereotyped behaviours in pervasive developmental disorders. *Journal of Child Psychology and Psychiatry*, *47*(6), 573-581.

Charman, T., & Gotham, K. (2013). Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders–lessons from research and practise. Child and adolescent mental health, 18(1), 52-63.

Chawarska, K., Klin, A., Paul, R., & Volkmar, F. (2007). Autism spectrum disorder in the second year: Stability and change in syndrome expression. *Journal of Child Psychology and Psychiatry*, *48*(2), 128-138.

Chlebowski, C., Robins, D. L., Barton, M. L., & Fein, D. (2013). Large-scale use of the modified checklist for autism in low-risk toddlers. *Pediatrics*, *131*(4), e1121-e1127.*

Coelho-Medeiros, M. E., Bronstein, J., Aedo, K., Pereira, J. A., Arraño, V., Perez, C. A., Valenzuela, P. M., Moore, R., Garrido, I., & Bedregal, P. (2019). M-CHAT-R/F Validation as a screening tool for early detection in children with autism spectrum disorder. *Revista chilena de pediatria*, *90*(5), 492.*

Committee on Children with Disabilities. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics*, *108*(1), 192-195.

Cuesta-Gómez, J. L., Andrea Manzone, L., & Posada-De-La-Paz, M. (2016). Modified checklist for autism in toddlers cross-cultural adaptation for Argentina. *International journal of Developmental Disabilities*, *62*(2), 117-123.*

Data & Statistics on Autism Spectrum Disorder | CDC. (n.d.). Retrieved October 5, 2019, from https://www.cdc.gov/ncbddd/autism/data.html.

Data & Statistics on Autism Spectrum Disorder | CDC. (n.d.). Retrieved October 5, 2019, from https://www.cdc.gov/ncbddd/autism/data.html.

Davidovitch, M., Hemo, B., Manning-Courtney, P., & Fombonne, E. (2013). Prevalence and incidence of autism spectrum disorder in an Israeli population. *Journal of autism and developmental disorders*, *43*(4), 785-793.

Dendukuri, N., Schiller, I., Joseph, L., & Pai, M. (2012). Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*, *68*(4), 1285-1293.

Dendukuri, N., Schiller, I., Joseph, L., & Pai, M. (2012). Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*, *68*(4), 1285-1293.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics*, *26*, 979-1030.

Doebler, P., & Holling, H. (2015). Meta-analysis of diagnostic accuracy with mada. *Reterieved at: https://cran. rproject. org/web/packages/mada/vignettes/mada. pdf*.

Doebler, P., & Holling, H. (2015). Meta-analysis of diagnostic accuracy with mada. *Reterieved at: https://cran. rproject. org/web/packages/mada/vignettes/mada. pdf*.

Durkin, M. S., Elsabbagh, M., Barbaro, J., Gladstone, M., Happe, F., Hoekstra, R. A., … Shih, A. (2015). Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. *Autism research : official journal of the International Society for Autism Research*, *8*(5), 473–476. 10.1002/aur.1575

Durkin, M. S., Elsabbagh, M., Barbaro, J., Gladstone, M., Happe, F., Hoekstra, R. A., … Shih, A. (2015). Autism screening and diagnosis in low resource settings: Challenges and opportunities to enhance research and services worldwide. *Autism research : official journal of the International Society for Autism Research*, *8*(5), 473–476. 10.1002/aur.1575

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2009). Meta-analysis of early intensive behavioral intervention for children with autism. *Journal of Clinical Child & Adolescent Psychology*, *38*(3), 439-450.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2009). Meta-analysis of early intensive behavioral intervention for children with autism. *Journal of Clinical Child & Adolescent Psychology*, *38*(3), 439-450.

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcín, C., Monteil-Nava, C., Patel, V., Paula, C. S., Wang, C., Yasamy, M. T., & Fombonne, E. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism research*, *5*(3), 160-179.

Fernell, E., & Gillberg, C. (2010). Autism spectrum disorder diagnoses in Stockholm preschoolers. *Research in developmental disabilities*, *31*(3), 680-685.

Filipek, P. A., Accardo, P. J., Baranek, G. T., Cook, E. H., Dawson, G., Gordon, B., ... & Minshew, N. J. (1999). The screening and diagnosis of autistic spectrum disorders. *Journal of autism and developmental disorders*, *29*(6), 439-484.

Foley-Nicpon, M., Fosenburg, S. L., Wurster, K. G., & Assouline, S. G. (2017). Identifying high ability children with DSM-5 autism spectrum or social communication disorder: performance on autism diagnostic instruments. *Journal of autism and developmental disorders*, *47*(2), 460-471.

Gatsonis, C., & Paliwal, P. (2006). Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *American Journal of Roentgenology*, *187*(2), 271-281.

Goin-Kochel, R. P., Mackintosh, V. H., & Myers, B. J. (2006). How many doctors does it take to make an autism spectrum diagnosis?. *Autism*, *10*(5), 439-451

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *Journal of autism and developmental disorders*, *37*(4), 613.

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91-108.

Green, J., Pickles, A., Pasco, G., Bedford, R., Wan, M. W., Elsabbagh, M., Slonims, V., Gliga, T., Jones, E., Cheung, C., Charman, T., Johnson, M., & The British Autism Study of Infant Siblings (BASIS) Team. (2017). Randomised trial of a parent-mediated intervention for infants at high risk for autism: longitudinal outcomes to age 3 years. *Journal of Child Psychology and Psychiatry*, *58*(12), 1330-1340.

Guthrie, W., Wallis, K., Bennett, A., Brooks, E., Dudley, J., Gerdes, M., ... & Miller, J. S. (2019). Accuracy of autism screening in a large pediatric network. *Pediatrics*, *144*(4), e20183963.*

Hagan, J. F., Shaw, J. S., & Duncan, P. M. (2007). *Bright futures: Guidelines for health supervision of infants, children, and adolescents*. American Academy of Pediatrics.

Harbord, R. M., & Whiting, P. (2009). Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. *The Stata Journal*, *9*(2), 211-229.

Harbord, R. M., & Whiting, P. (2009). Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. *The Stata Journal*, *9*(2), 211-229.

Harris, S. R. (2017). Early motor delays as diagnostic clues in autism spectrum disorder. *European journal of pediatrics*, *176*(9), 1259-1262.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Hirai AH, Kogan MD, Kandasamy V, Reuland C, Bethell C. Prevalence and Variation of Developmental Screening and Surveillance in Early Childhood. *JAMA Pediatr.* 2018;172(9):857–866. 10.1001/jamapediatrics.2018.1524

Hirai AH, Kogan MD, Kandasamy V, Reuland C, Bethell C. Prevalence and Variation of Developmental Screening and Surveillance in Early Childhood. *JAMA Pediatr.* 2018;172(9):857–866. 10.1001/jamapediatrics.2018.1524

Hoang, V. M., Le, T. V., Chu, T. T. Q., Le, B. N., Duong, M. D., Thanh, N. M., ... & Bui, T. T. H. (2019). Prevalence of autism spectrum disorders and their relation to selected socio-demographic factors among children aged 18-30 months in northern Vietnam, 2017. *International journal of mental health systems*, *13*, 29-29.*

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004)

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: potential and promise. *Statistics in medicine*, *30*(20), 2481-2498.

Johnson, C. P., & Myers, S. M. (2007). Identification and evaluation of children with autism spectrum disorders. *Pediatrics*, *120*(5), 1183-1215.

Kamio, Y., Inada, N., Koyama, T., Inokuchi, E., Tsuchiya, K., & Kuroda, M. (2014). Effectiveness of using the Modified Checklist for Autism in Toddlers in two-stage screening of autism spectrum

disorder at the 18-month health check-up in Japan. *Journal of autism and developmental disorders*, *44*(1), 194-203.*

Kasari, C., Gulsrud, A., Freeman, S., Paparella, T., & Hellemann, G. (2012). Longitudinal follow-up of children with autism receiving targeted interventions on joint attention and play. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(5), 487-495.

Kasari, C., Gulsrud, A., Paparella, T., Hellemann, G., & Berry, K. (2015). Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of consulting and clinical psychology*, *83*(3), 554.

Kasari, C., Gulsrud, A., Paparella, T., Hellemann, G., & Berry, K. (2015). Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of consulting and clinical psychology*, *83*(3), 554.

Kawamura, Y., Takahashi, O., & Ishii, T. (2008). Reevaluating the incidence of pervasive developmental disorders: impact of elevated rates of detection through implementation of an integrated system of screening in Toyota, Japan. *Psychiatry and clinical neurosciences*, *62*(2), 152-159.

Kerub, O., Haas, E. J., Meiri, G., Davidovitch, N., & Menashe, I. (2018). A comparison between two screening approaches for ASD among toddlers in Israel. *Journal of autism and developmental disorders*, 1-8.*

Khowaja, M. K., Hazzard, A. P., & Robins, D. L. (2015). Sociodemographic barriers to early detection of autism: screening and evaluation using the M-CHAT, M-CHAT-R, and follow-up. *Journal of autism and developmental disorders*, *45*(6), 1797-1808.

Kleinman, J. M., Ventola, P. E., Pandey, J., Verbalis, A. D., Barton, M., Hodgson, S., ... & Fein, D. (2008). Diagnostic stability in very young children with autism spectrum disorders. *Journal of autism and developmental disorders*, *38*(4), 606-615.

Kleinman, J. M., Ventola, P. E., Pandey, J., Verbalis, A. D., Barton, M., Hodgson, S., ... & Fein, D. (2008). Diagnostic stability in very young children with autism spectrum disorders. *Journal of autism and developmental disorders*, *38*(4), 606-615.

Kleinman, J. M., Ventola, P. E., Pandey, J., Verbalis, A. D., Barton, M., Hodgson, S., ... & Fein, D. (2008). Diagnostic stability in very young children with autism spectrum disorders. *Journal of autism and developmental disorders*, *38*(4), 606-615.

Klin, A., & Volkmar, F. R. (1995). Asperger's Syndrome: guidelines for assessment and diagnosis. *Learning Disabilities Association of America. Available online at http://www. aspennj. org/guide. html*.

Klin, A., Saulnier, C., Tsatsanis, K., & Volkmar, F. R. (2005). Clinical evaluation in autism spectrum disorders: Psychological assessment within a transdisciplinary framework. *Handbook of autism and pervasive developmental disorders*, *2*, 772-798.

Kondolot, M., Özmert, E. N., Öztop, D. B., Mazıcıoğlu, M. M., Gümüş, H., & Elmalı, F. (2016). The modified checklist for autism in Turkish toddlers: A different cultural adaptation sample. *Research in Autism Spectrum Disorders*, *21*, 121-127.*

Le Couteur, A., Haden, G., Hammal, D., & McConachie, H. (2008). Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: the ADI-R and the ADOS. *Journal of autism and developmental disorders*, *38*(2), 362-372.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of general psychiatry*, *63*(6), 694-701.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), 659-685.

Magán-Maganto, M., Canal-Bedia, R., Hernández-Fabián, A., Bejarano-Martín, Á., Fernández-Álvarez, C. J., Martínez-Velarte, M., ... & de la Paz, M. P. (2018). Spanish Cultural Validation of the Modified Checklist for Autism in Toddlers, Revised. *Journal of autism and developmental disorders*, 1-12.*

Mandell, D. S., Listerud, J., Levy, S. E., & Pinto-Martin, J. A. (2002). Race differences in the age at diagnosis among Medicaid-eligible children with autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(12), 1447-1453.

Mandell, D. S., Listerud, J., Levy, S. E., & Pinto-Martin, J. A. (2002). Race differences in the age at diagnosis among Medicaid-eligible children with autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(12), 1447-1453.

Matson, J. L., Kozlowski, A. M., Fitzgerald, M. E., & Sipes, M. (2013). True versus false positives and negatives on the modified checklist for autism in toddlers. *Research in Autism Spectrum Disorders*, *7*(1), 17-22.*

Mazefsky, C. A., McPartland, J. C., Gastgeb, H. Z., & Minshew, N. J. (2013). Brief report: Comparability of DSM-IV and DSM-5 ASD research samples. *Journal of autism and developmental disorders*, *43*(5), 1236-1242.

McPheeters, M. L., Weitlauf, A., Vehorn, A., Taylor, C., Sathe, N. A., Krishnaswami, S., ... & Warren, Z. E. (2016). Screening for Autism Spectrum Disorder in Young Children.

National Research Council. (2001). *Educating children with autism*. National Academies Press.

National Research Council. (2001). *Educating children with autism*. National Academies Press.

Norris, M., & Lecavalier, L. (2010). Screening accuracy of level 2 autism spectrum disorder rating scales: A review of selected instruments. *Autism*, *14*(4), 263-284.

Nygren, G., Sandberg, E., Gillstedt, F., Ekeroth, G., Arvidsson, T., & Gillberg, C. (2012). A new screening programme for autism in a general population of Swedish toddlers. *Research in developmental disabilities*, *33*(4), 1200-1210.*

Office of Disease Prevention and Health Promotion, US Department of Health and Human Services. Healthy People 2020 maternal, infant, and child health objectives. https://www.healthypeople.gov/2020/topics-objectives/topic/maternal-infant-and-child-health/objectives

Øien, R. A., Schjølberg, S., Volkmar, F. R., Shic, F., Cicchetti, D. V., Nordahl-Hansen, A., ... & Ventola, P. (2018). Clinical features of children with autism who passed 18-month screening. *Pediatrics*, *141*(6).*

Oner, O., & Munir, K. M. (2019). 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*, 1-8.*

Pierce, K., Carter, C., Weinfeld, M., Desmond, J., Hazin, R., Bjork, R., & Gallagher, N. (2011). Detecting, studying, and treating autism early: the one-year well-baby check-up approach. *The Journal of pediatrics*, *159*(3), 458-465.

Pierce, K., Courchesne, E., & Bacon, E. (2016). To screen or not to screen universally for autism is not the question: why the task force got it wrong. *The Journal of pediatrics*, *176*, 182-194.

Pierce, K., Courchesne, E., & Bacon, E. (2016). To screen or not to screen universally for autism is not the question: why the task force got it wrong. *The Journal of pediatrics*, *176*, 182-194.

Pigott, T. (2012). *Advances in meta-analysis*. Springer Science & Business Media.

Pigott, T. D., & Polanin, J. R. (2019). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, 0034654319877153.

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*(3), 330-342.

Posserud, M., Lundervold, A. J., Lie, S. A., & Gillberg, C. (2010). The prevalence of autism spectrum disorders: impact of diagnostic instrument and non-response bias. *Social psychiatry and psychiatric epidemiology*, *45*(3), 319-327.

Recommendations & Guidelines. (2019, August 27). Retrieved from https://www.cdc.gov/ncbddd/autism/hcp-recommendations.htmlhttps://www.cdc.gov/ncbddd/autism/hcp-screening.html#Screening

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, *58*(10), 982-990.

Robins, D. (n.d.). he Modified Checklist for Autism in Toddlers –Revised/Follow-up (M-CHAT –R/F). Retrieved October 5, 2019, from http://www.amchp.org/programsandtopics/CYSHCN/projects/spharc/peer-to-peer-exchange/Documents/M-CHAT.pdf. Committee on Children with Disabilities. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics*, *108*(1), 192-195.

Robins, D. (n.d.). he Modified Checklist for Autism in Toddlers –Revised/Follow-up (M-CHAT –R/F). Retrieved October 5, 2019, from http://www.amchp.org/programsandtopics/CYSHCN/projects/spharc/peer-to-peer-

exchange/Documents/M-CHAT.pdf. Committee on Children with Disabilities. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics*, *108*(1), 192-195.

Robins, D. L. (2008). Screening for autism spectrum disorders in primary care settings. *Autism*, *12*(5), 537-556.

Robins, D. L., Casagrande, K., Barton, M., Chen, C. M. A., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics*, *133*(1), 37-45.

Robins, D. L., Fein, D., Barton, M. L., & Green, J. A. (2001). The Modified Checklist for Autism in Toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of autism and developmental disorders*, *31*(2), 131-144.

Rogers, S. J. (1996). Brief report: Early intervention in autism. *Journal of autism and developmental disorders*, *26*(2), 243-246.

Rogers, S. J. (1996). Brief report: Early intervention in autism. *Journal of autism and developmental disorders*, *26*(2), 243-246.

Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford university press.

Sanchez-Garcia, A. B., Galindo-Villardon, P., Nieto-Librero, A. B., Martin-Rodero, H., & Robins, D. L. (2019). Toddler screening for autism spectrum disorder: A meta-analysis of diagnostic accuracy. *Journal of autism and developmental disorders*, *49*(5), 1837-1852.

Sánchez-García, A. B., Galindo-Villardón, P., Nieto-Librero, A. B., Martín-Rodero, H., & Robins, D. L. (2019). Toddler Screening for Autism Spectrum Disorder: A Meta-Analysis of Diagnostic Accuracy. *Journal of autism and developmental disorders*, *49*(5), 1837-1852.

Sánchez-García, A. B., Galindo-Villardón, P., Nieto-Librero, A. B., Martín-Rodero, H., & Robins, D. L. (2019). Toddler Screening for Autism Spectrum Disorder: A Meta-Analysis of Diagnostic Accuracy. *Journal of autism and developmental disorders*, *49*(5), 1837-1852.

Sand, N., Silverstein, M., Glascoe, F. P., Gupta, V. B., Tonniges, T. P., & O'Connor, K. G. (2005). Pediatricians' reported practices regarding developmental screening: do guidelines work? Do they help?. *Pediatrics*, *116*(1), 174-179.

Screening and Diagnosis of Autism Spectrum Disorder for Healthcare Providers. (2020, February 11). https://www.cdc.gov/ncbddd/autism/hcp-screening.html#Screening

Sheldrick, R. C., Benneyan, J. C., Kiss, I. G., Briggs-Gowan, M. J., Copeland, W., & Carter, A. S. (2015). Thresholds and accuracy in screening tools for early detection of psychopathology. *Journal of Child Psychology and Psychiatry*, *56*(9), 936-948.

Sheldrick, R. C., Breuer, D. J., Hassan, R., Chan, K., Polk, D. E., & Benneyan, J. (2016). A system dynamics model of clinical decision thresholds for the detection of developmental-behavioral disorders. *Implementation Science*, *11*(1), 156.

Sheldrick, R. C., Merchant, S., & Perrin, E. C. (2011). Identification of developmental-behavioral problems in primary care: a systematic review. *Pediatrics*, *128*(2), 356-363.

Sheldrick, R. C., Schlichting, L. E., Berger, B., Clyne, A., Ni, P., Perrin, E. C., & Vivier, P. M. (2019). Establishing new norms for developmental milestones. *Pediatrics*, *144*(6).

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, *70*, 747-770.

Siu AL, and the US Preventive Services Task Force (USPSTF). Screening for Autism Spectrum Disorder in Young Children: US Preventive Services Task Force Recommendation Statement. *JAMA.* 2016;315(7):691–696. 10.1001/jama.2016.0018

Siu AL, and the US Preventive Services Task Force (USPSTF). Screening for Autism Spectrum Disorder in Young Children: US Preventive Services Task Force Recommendation Statement. *JAMA.* 2016;315(7):691–696. 10.1001/jama.2016.0018

Smith, I. C., Reichow, B., & Volkmar, F. R. (2015). The effects of DSM-5 criteria on number of individuals diagnosed with autism spectrum disorder: A systematic review. *Journal of Autism and Developmental Disorders*, *45*(8), 2541-2552.

Snow, A. V., & Lecavalier, L. (2008). Sensitivity and specificity of the Modified Checklist for Autism in Toddlers and the Social Communication Questionnaire in preschoolers suspected of having pervasive developmental disorders. *Autism*, *12*(6), 627-644.*

Soto, S., Linas, K., Jacobstein, D., Biel, M., Migdal, T., & Anthony, B. J. (2015). A review of cultural adaptations of screening tools for autism spectrum disorders. *Autism*, *19*(6), 646-661.

Srisinghasongkram, P., Pruksananonda, C., & Chonchaiya, W. (2016). Two-step screening of the modified checklist for autism in toddlers in Thai children with language delay and typically developing children. *Journal of autism and developmental disorders*, *46*(10), 3317-3329.*

Sturner, R., Howard, B., Bergmann, P., Morrel, T., Andon, L., Marks, D., ... & Landa, R. (2016). Autism screening with online decision support by primary care pediatricians aided by M-CHAT/F. *Pediatrics*, *138*(3), e20153036.*

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research synthesis methods*, *10*(2), 180-194.

Toh, T. H., Tan, V. W. Y., Lau, P. S. T., & Kiyu, A. (2018). Accuracy of Modified Checklist for Autism in Toddlers (M-CHAT) in detecting autism and other developmental disorders in community clinics. *Journal of autism and developmental disorders*, *48*(1), 28-35.*

Topçu, S., Ulukol, B., Öner, Ö., Şimşek Orhon, F., & Başkan, S. (2018). Comparison of tidos with m-chat for screening autism spectrum disorder. *Psychiatry and Clinical Psychopharmacology*, *28*(4), 416-422.*

Tsai, J. M., Lu, L., Jeng, S. F., Cheong, P. L., Gau, S. S. F., Huang, Y. H., & Wu, Y. T. (2019). Validation of the modified checklist for autism in toddlers, revised with follow-up in Taiwanese toddlers. *Research in developmental disabilities*, *85*, 205-216.*

Uman, L. S. (2011). Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *20*(1), 57.

US Department of Health and Human Services. (2014). Birth to 5: Watch me thrive. *A compendium of screening measures for young children. Washington, DC: US Department of Health and Human Services*.

van Bakel, M. M. E., Delobel-Ayoub, M., Cans, C., Assouline, B., Jouk, P. S., Raynaud, J. P., & Arnaud, C. (2015). Low but increasing prevalence of autism spectrum disorders in a French area from register-based data. *Journal of autism and developmental disorders*, *45*(10), 3255-3261.

Van Cong, T., Weiss, B., Toan, K. N., Le Thu, T. T., Trang, N. T. N., Hoa, N. T. K., & Thuy, D. T. T. (2015). Early identification and intervention services for children with autism in Vietnam. *Health psychology report*, *3*(3), 191.

Ventola, P. E., Kleinman, J., Pandey, J., Barton, M., Allen, S., Green, J., ... & Fein, D. (2006). Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. *Journal of autism and developmental disorders*, *36*(7), 839-847.

Virués-Ortega, J. (2010). Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical psychology review*, *30*(4), 387-399.

Virués-Ortega, J. (2010). Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical psychology review*, *30*(4), 387-399.

Volkmar, F. R., Booth, L. L., McPartland, J. C., & A. Wiesner, L. (2014a). Clinical evaluation in multidisciplinary settings. *Handbook of Autism and Pervasive Developmental Disorders, Fourth Edition*.

Volkmar, F., Siegel, M., Woodbury-Smith, M., King, B., McCracken, J., & State, M. (2014b). Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(2), 237-257.

Walker, E., Hernandez, A. V., & Kattan, M. W. (2008). Meta-analysis: Its strengths and limitations. *Cleveland Clinic journal of medicine*, *75*(6), 431.

Walter, S. D., & Jadad, A. R. (1999). Meta-analysis of screening data: a survey of the literature. *Statistics in medicine*, *18*(24), 3409-3424.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... & Bossuyt, P. M. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, *155*(8), 529-536.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... & Bossuyt, P. M. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, *155*(8), 529-536.

Wong, Y. S., Yang, C. C., Stewart, L., Chiang, C. H., Wu, C. C., & Iao, L. S. (2018). Use of the Chinese version modified checklist for autism in toddlers in a high-risk sample in Taiwan. *Research in Autism Spectrum Disorders*, *49*, 56-64.*

World Health Organisation. Mental Disorders: A Glossary And Guide To Their Classification In Accordance With The 10th Revision Of The International Classification Of Disease – Research Diagnostic Criteria: ICD- 10. Geneva: WHO, 1993.

World Health Organisation. Mental Disorders: A Glossary And Guide To Their Classification In Accordance With The 10th Revision Of The International Classification Of Disease – Research Diagnostic Criteria: ICD- 10. Geneva: WHO, 1993.

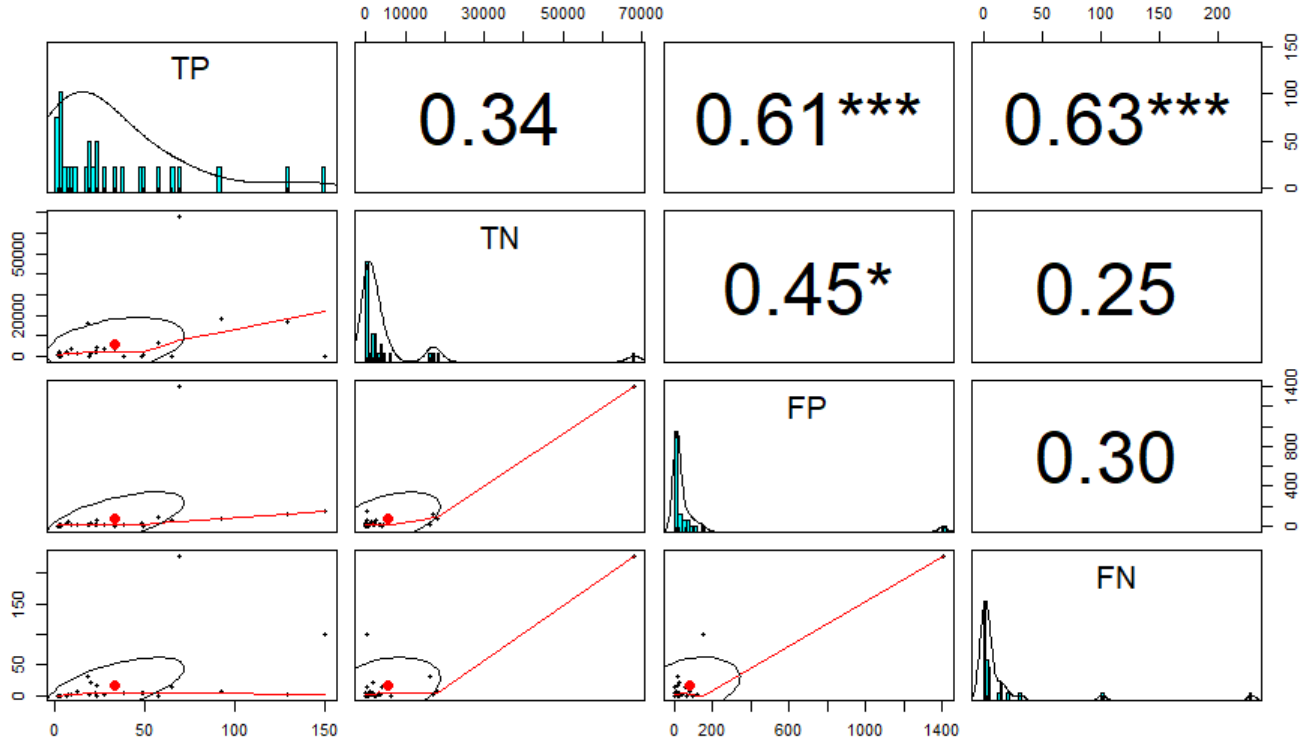
Yuen, T., Penner, M., Carter, M. T., Szatmari, P., & Ungar, W. J. (2018). Assessing the accuracy of the Modified Checklist for Autism in Toddlers: a systematic review and meta-analysis. *Developmental Medicine & Child Neurology*, *60*(11), 1093-1100.

Zwaigenbaum, L., Bauman, M. L., Fein, D., Pierce, K., Buie, T., Davis, P. A., ... & Kasari, C. (2015). Early screening of autism spectrum disorder: recommendations for practice and research. *Pediatrics*, *136*(Supplement 1), S41-S59.

Appendix A

Cross table and list of presumed outliers in the

data



TP/TN outlier: Matson (150, 151) Hoang (129, 17021) Chlebowski (92, 18269) Oien (69, 67969)

TP/FP outlier: Matson (150, 150) Hoang (129, 118) Chlebowski (92, 79) Oien (69, 1402)

TP/FN outlier: Matson (150, 101) Hoang (129, 1) Chlebowski (92, 6) Oien (69, 228)

TN/FP outlier: Oien (67969, 1402)

TN/FN outlier: Oien (67969, 228) Matson (150, 101)

FP/FN outlier: Oien (1402, 228) Matson (150, 101)

Appendix B

Modified-Checklist for Autism in Toddlers

M-CHAT

Please fill out the following about how your child usually is. Please try to answer every question. If the behavior is rare (e.g., you've seen it once or twice), please answer as if the child does not do it.

1.Does your child enjoy being swung, bounced on your knee, etc.? Yes No

2.Does your child take an interest in other children? Yes No

3.Does your child like climbing on things, such as up stairs? Yes No

4.Does your child enjoy playing peek-a-boo/hide-and-seek? Yes No

5.Does your child ever pretend, for example, to talk on the phone or take care of a doll or pretend other things? Yes No

6.Does your child ever use his/her index finger to point, to ask for something? Yes No

7.Does your child ever use his/her index finger to point, to indicate interest in something? Yes No

8.Can your child play properly with small toys (e.g. cars or blocks) without just mouthing, fiddling, or dropping them? Yes No

9.Does your child ever bring objects over to you (parent) to show you something? Yes No

10.Does your child look you in the eye for more than a second or two? Yes No

11.Does your child ever seem oversensitive to noise? (e.g., plugging ears) Yes No

12.Does your child smile in response to your face or your smile? Yes No

13.Does your child imitate you? (e.g., you make a face-will your child imitate it?) Yes No

14.Does your child respond to his/her name when you call? Yes No

15.If you point at a toy across the room, does your child look at it? Yes No

16.Does your child walk? Yes No

17.Does your child look at things you are looking at? Yes No

18.Does your child make unusual finger movements near his/her face? Yes No

19.Does your child try to attract your attention to his/her own activity? Yes No

20.Have you ever wondered if your child is deaf? Yes No

21.Does your child understand what people say? Yes No

22.Does your child sometimes stare at nothing or wander with no purpose?

Yes No

23.Does your child look at your face to check your reaction when faced with something unfamiliar? Yes No

Appendix C

QUADAS-2 Appendix F Questions

Domain 1: Patient selection

A. Risk of bias	
Describe methods of patient selection:	
Was a consecutive or random sample of patients enrolled?	**Yes / No / Unclear**
Was a case-control design avoided?	**Yes / No / Unclear**
Did the study avoid inappropriate exclusions?	**Yes / No / Unclear**
Could the selection of patients have introduced bias? **Risk: Low / High / Unclear**	
B. Concerns regarding applicability	
Describe included patients (prior testing, presentation, intended use of index test and setting): . . .	
Is there concern that the included patients do not match the review question? **Concern: Low / High / Unclear**	

Domain 2: Index test(s)

A. Risk of bias
Describe the index test and how it was conducted and interpreted:

. . . .
Were the index test results interpreted without knowledge of the results of the reference standard? **Yes / No / Unclear**
If a threshold was used, was it pre-specified? **Yes / No / Unclear**
Could the conduct or interpretation of the index test have introduced bias? **Risk: Low / High / Unclear**
B. Concerns regarding applicability
Is there concern that the index test, its conduct, or interpretation differ from the review question? **Concern: Low / High / Unclear**

Domain 3: Reference standard

A. Risk of bias
Describe the reference standard and how it was conducted and interpreted: . . .
Is the reference standard likely to correctly classify the target condition? **Yes / No / Unclear**
Were the reference standard results interpreted without knowledge of the results of the index test? **Yes / No / Unclear**
Could the reference standard, its conduct, or its interpretation have introduced bias? **Risk: Low / High / Unclear**
B. Concerns regarding applicability
Is there concern that the target condition as defined by the reference standard does not

match the review question? **Concern: Low / High / Unclear**

Domain 4: Flow and timing

A. Risk of bias	
Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram): . . . Describe the time interval and any interventions between index test(s) and reference standard:	
Was there an appropriate interval between index test(s) and reference standard?	**Yes / No / Unclear**
Did all patients receive a reference standard?	**Yes / No / Unclear**
Did patients receive the same reference standard?	**Yes / No / Unclear**
Were all patients included in the analysis?	**Yes / No / Unclear**
Could the patient flow have introduced bias? **Risk: Low / High / Unclear**	

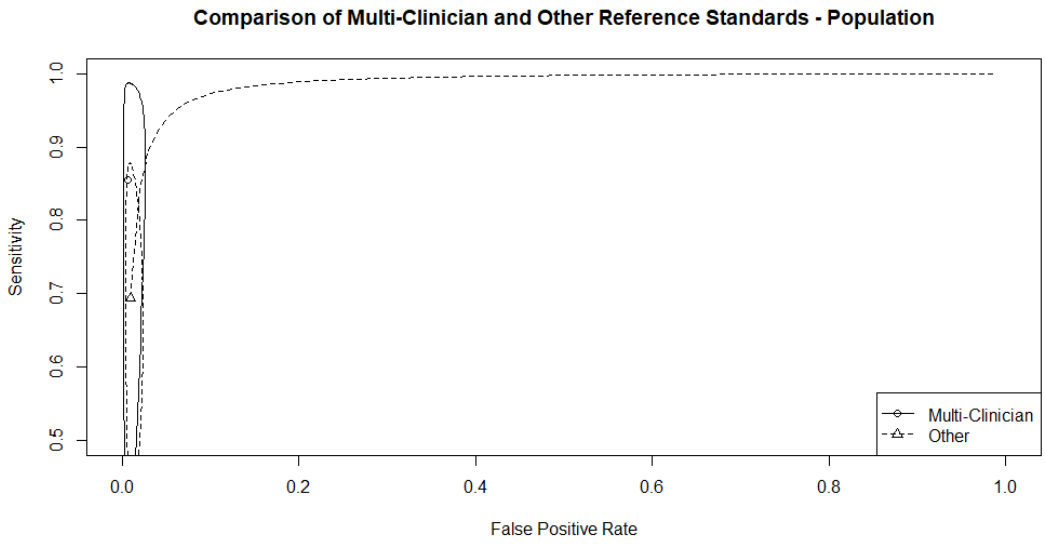
Appendix D

Population based and high-risk studies broken into reference standard categories

Supplemental Figures 1 and 2 are the bivariate Reitsma illustrated by an SROC of subgroup analysis and clinical reference criteria. In Figure 17 Multi-Clinician reference standard shows a higher sensitivity than Other and neither are significantly different in FPR. In Figure 18 Multi-Clinician reference standard has a higher sensitivity but Other appears to have a significantly better FPR.

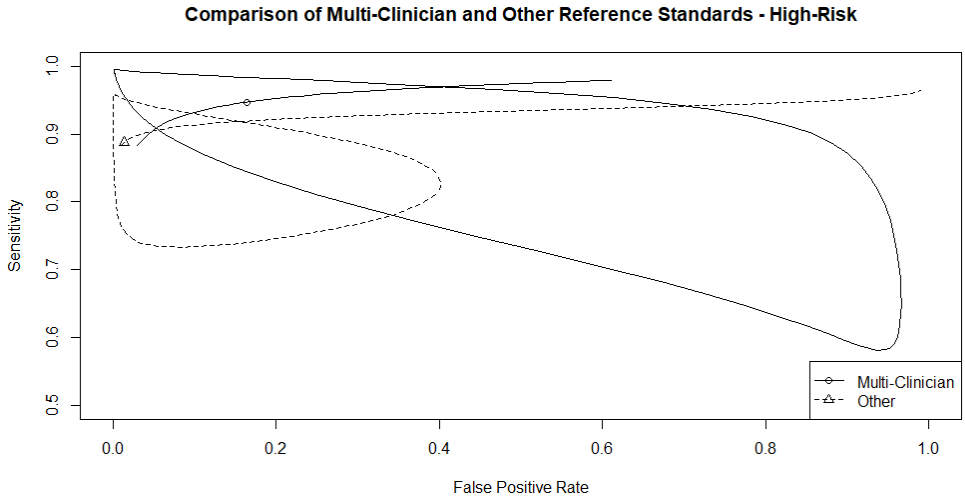
Supplemental Figure 1

SROC of Population by Reference Standard



Supplemental Figure 2

SROC of Clinical by Reference Standard



Appendix E

Exploratory Analysis of Population Based Estimate

In figure 17 the original sensitivity metrics are compared to the adjusted sensitivity metrics. Between these two forest plots there are six studies whose metrics were adjusted: Chlebowski, Hoang, Kondolot, Oien, Oner, and Toh. Due to the nature of the adjustments all of these resulted in the sensitivity estimate being lowered, some of these were lowered significantly like Chlebowski from 0.93 to 0.34. Figure 22, shows similar changes but with specificity, these comparisons are interesting because the six studies where adjustments are made, no point estimates changed from the original study to the adjusted version. Figure 17, resembles more of figure 21, where the 6 studies whose metrics were adjusted changed. None were reduced to the point of being non-significant. The adjusted *Se* became 0.522 from the original of 0.746 and the adjusted *Sp* did not change from the original of 0.992.

Supplemental Table 1

Adjusted identification metrics of population-based studies

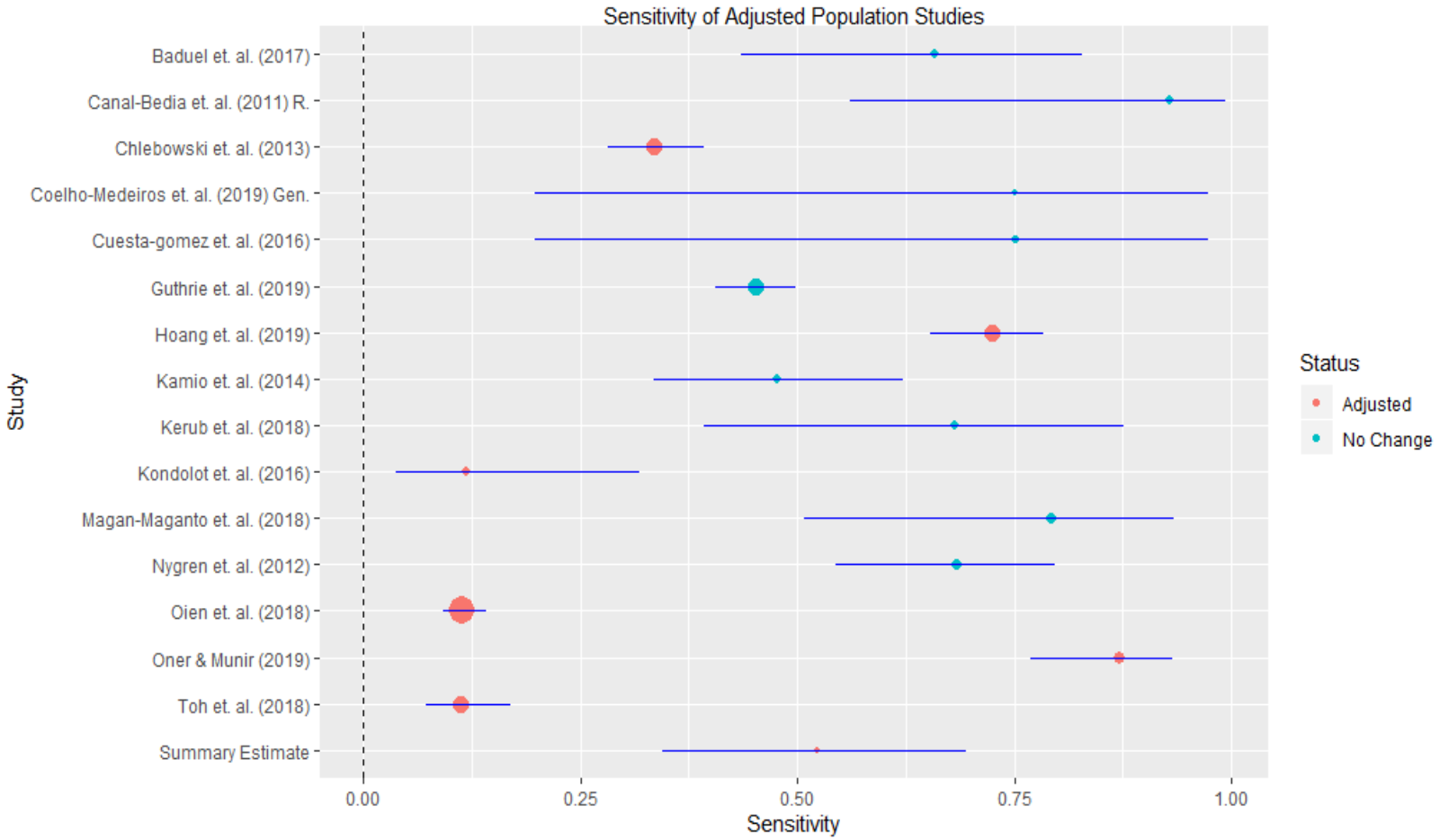
Author	Country	Pop. Size	TP	FN	FP	TN	Identified	Prev. Est. year	Prev. Est. Author	Prev. Est.	Expected	Difference	ADJ FN	Change	ADJ TN	ADJ Sensitivity	ADJ Specificity
Baduel et al. (2017)	France	1250	12	6	8	1201	18	2015	Van Bakel et al. (2017)	0.00365	5	-13	6	No	1201	0.6667	0.993383
Canal-Bedia et al. (2011)	Spain	2055	23	0	25	2024	23	2007	Adak & Halder (2017)	0.001297	3	-20	0	No	2024	1.00	0.987799
Chlebowski, Robins, Barton, & Fein (2013)	USA	18989	92	79	79	1826 9	171	2012	Baio et al. (2018)	0.014493	275	104	183	Yes	18165	0.3343	0.99567
Coelho-Medeiros et al. (2019)	Chile	100	1	90	1	0	17	2013	Van Cong et al. (2015)	0.01	1	-16	0	No	90	1.00	0.989011
Cuesta-gomez, Manzone, Posada-De-La-Paz (2016)	Argentina	420	1	0	1	402	1	2008	Elsabbagh et al. (2012)	0.00131	1	0	0	No	402	1.00	0.997519
Guthrie et al. (2019)	USA	20375	205	249	1658	1826 3	454	2014	Baio et al. (2018)	0.016949	345	-109	249	No	18263	0.4515	0.916771
Hoang et al. (2019)	Vietnam	17754	129	1	118	1702 1	130	2013	Van Cong et al. (2015)	0.01	178	48	49	Yes	16973	0.7266	0.993096
Kamio et al (2014)	Japan	1851	20	22	24	1661	42	2008	Kawamura et al. (2008)	0.01811	34	-8	22	No	1661	0.4762	0.985757
Kerub et al (2018)	Isreal	1591	7	3	43	1538	10	2001	Davidovitch et al. ((2013)	0.001	2	-8	3	No	1538	0.700	0.972802

Supplemental Table 1 (Continued)

Author	Country	Pop. Size	TP	FN	FP	TN	Identified	Prev. Est. year	Prev. Est. Author	Prev. Est.	Expected	Difference	ADJ FN	Change	ADJ TN	ADJ Sensitivity	ADJ Specificity
Kondolot et al 2016	Turkey	2021	2	0	15	2004	2	2013	Van Cong et al. (2015)	0.01	20	18	18	Yes	1986	0.0989	0.992503
Magan-Maganto et al (2018)	Spain	3529	9	2	10	3485	11	2007	Adak & Halder (2017)	0.001297	5	-6	2	No	3485	0.8182	0.997139
Nygren et al (2012)	Sweden	3985	33	15	3	3939	48	2010	Fernell & Gillberg (2010)	0.0062	25	-23	15	No	3939	0.6875	0.999239
Oien et al (2018)	Norway	69668	69	228	1402	67969	297	2010	Posserud et al. (2010)	0.0087	606	309	537	Yes	67660	0.1138	0.979699
Oner & Munir (2019)	Turkey	6540	57	0	95	6388	57	2013	Van Cong et al. (2015)	0.01	65	8	8	Yes	6380	0.8716	0.985327
Toh, Tan, Lau, & Kiyu (2018)	Malaysia	16297	18	32	20	16227	50	2013	Van Cong et al. (2015)	0.01	163	113	145	Yes	16114	0..1105	0.99876

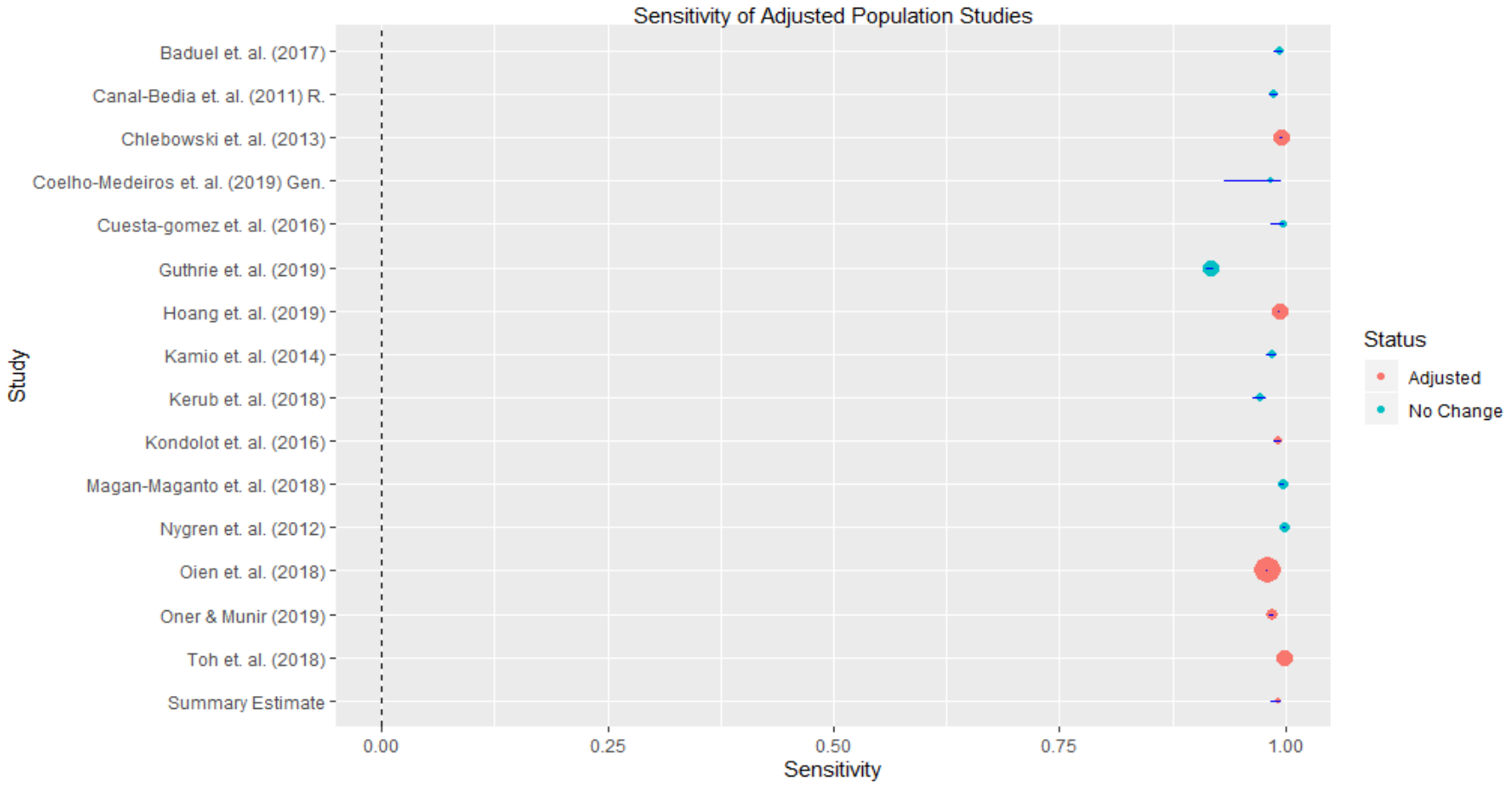
Supplemental Figure 3

Adjusted Sensitivity Metrics of Population Based Studies by Population Prevalence Estimates



Supplemental Figure 4

Adjusted Specificity Metrics of Population Based Studies by Population Prevalence Estimates



Supplemental Figure 5

Adjusted DOR Metrics of Population Based Studies by Population Prevalence Estimates