The Journal of Child Psychology and Psychiatry

The Association for Child and Adolescent Mental Health

Journal of Child Psychology and Psychiatry 65:5 (2024), pp 656–667 doi:10.1111/jcpp.13867

The influence of loss to follow-up in autism screening research: Taking stock and moving forward

R. Christopher Sheldrick,[1†] Jessica L. Hooker,[2†] Alice S. Carter,[3] Emily Feinberg,[4] Lisa A. Croen,[5] Jocelyn Kuhn,[4] Elizabeth Slate,[6] and Amy M. Wetherby[2]

[1]UMass Chan Medical School, Worcester, MA, USA; [2]College of Medicine, Florida State University, Tallahassee, FL, USA; [3]Department of Psychology, University of Massachusetts Boston, Boston, MA, USA; [4]School of Public Health, Brown University, Providence, MA, USA; [5]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA; [6]Department of Statistics, Florida State University, Tallahassee, FL, USA

Background: How best to improve the early detection of autism spectrum disorder (ASD) is the subject of significant controversy. Some argue that universal ASD screeners are highly accurate, whereas others argue that evidence for this claim is insufficient. Relatedly, there is no clear consensus as to the optimal role of screening for making referral decisions for evaluation and treatment. Published screening research can meaningfully inform these questions—but only through careful consideration of children who do not complete diagnostic follow-up. **Methods:** We developed two simulation models that re-analyze the results of a large-scale validation study of the M-CHAT-R/F by Robins et al. (2014, *Pediatrics*, **133**, 37). Model #1 re-analyzes screener accuracy across six scenarios, each reflecting different assumptions regarding loss to follow-up. Model #2 builds on this by closely examining differential attrition at each point of the multi-step detection process. **Results:** Estimates of sensitivity ranged from 40% to 94% across scenarios, demonstrating that estimates of accuracy depend on assumptions regarding the diagnostic status of children who were lost to follow-up. Across a range of plausible assumptions, data also suggest that children with undiagnosed ASD may be more likely to complete follow-up than children without ASD, highlighting the role of clinicians and caregivers in the detection process. **Conclusions:** Using simulation modeling as a quantitative method to examine potential bias in screening studies, analyses suggest that ASD screening tools may be less accurate than is often reported. Models also demonstrate the critical importance of every step in a detection process—including steps that determine whether children should complete an additional evaluation. We conclude that parent and clinician decision-making regarding follow-up may contribute more to detection than is widely assumed. **Keywords:** Screening; autism spectrum disorders; validity; methodology.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impairments in social interaction and communication and restrictive and repetitive behaviors and activities (American Psychiatric Association, 2013). For over a decade, the American Academy of Pediatrics (AAP) has recommended supplementing developmental surveillance with universal screening for ASD for all children at ages 18 and 24 months to facilitate early intervention (Hyman et al., 2020; Johnson & Myers, 2007). To assess the potential benefits of ASD screening, the United States Preventive Services Task Force (USPSTF) commissioned a systematic review to evaluate the evidence on ASD screening, including the accuracy, benefits, and potential harms of ASD screening instruments administered during routine primary care visits (McPheeters et al., 2016). The purpose of this paper is to reanalyze a primary source of evidence identified by the USPSTF and discuss implications for using screening tools.

In their recommendation statement, the USPSTF concluded, “several screening tools for ASD are available, but the … most applicable evidence is for the M-CHAT/F and M-CHAT-R/F (Robins et al., 2014), 2 versions of the same tool” (p. 694; Siu et al., 2016). In particular, the USPSTF noted that the original Modified Checklist for Autism in Toddlers (M-CHAT) and its revision (M-CHAT-R) with follow-up interviews (M-CHAT/F and M-CHAT-R/F, respectively), were supported by “two good- and four fair-quality studies,” but that “none of the studies followed either the complete sample of screen-negative children or a truly random sample in order to assess missed cases” (McPheeters et al., 2016). In short, the USPSTF raised concerns about the potential for ascertainment bias in diagnostic evaluations.

The consequences of these limitations have led to significant controversies regarding screening performance. The USPSTF concluded that “the methods used for following screen negatives do not permit calculations of sensitivity, specificity, or NPV” (McPheeters et al., 2016). Likewise, in their review of research on early ASD screening, Zwaigenbaum et al. (2015) did not include any reported estimates of sensitivity and specificity for the M-CHAT, stating “Estimates of sensitivity and specificity cannot be

†Joint first authors.

Conflict of interest statement: See Acknowledgements for full disclosures.

determined by this study" due to a lack of evaluation of screen-negative cases. Others clearly disagree—several of the original papers reviewed by the USPSTF report estimated sensitivity and specificity, and these estimates are also reported in two recent meta-analyses (Sanchez-Garcia, Galindo-Villardon, Nieto-Librero, Martin-Rodero, & Robins, 2019; Wieckowski, Williams, Rando, Lyall, & Robins, 2023) and in the AAP practice recommendation on ASD screening (Hyman et al., 2020). Moreover, recent evidence published by Guthrie et al. (2019) reported that the M-CHAT-R was much less accurate than prior studies suggest. To evaluate the accuracy of the M-CHAT-R, there is a critical need to resolve these discrepancies. We argue that much of the confusion and conflicting evidence on screening accuracy results from unstated and divergent assumptions about children who are lost to diagnostic follow-up. Building on prior studies (Kuntz et al., 2013; Sainfort et al., 2013), we present a simulation model that analyzes the potential influence of loss to follow-up on estimates of sensitivity and specificity informed by highly cited papers in the field. Because loss to follow-up has implications beyond screening accuracy, it is also important to understand the implications of attrition in the entire "screening-to-treatment chain" (Silverstein & Radesky, 2016). Therefore, we extend our analysis with a second simulation model that analyzes the potential for differential attrition between children with and without ASD. We conclude by discussing implications of loss to follow-up for the interpretation of screening studies published since the USPSTF recommendation, clinical implications for improving multi-stage screening processes, and opportunities for further research.

The impact of lack of follow-up

Sensitivity and specificity estimates compare an index test (e.g., screening tool) to a clinical reference such as a diagnostic evaluation (Cohen et al., 2016). A central challenge to making such comparisons is that most children who are screened for ASD (or any developmental-behavioral disorder) do not receive a diagnostic evaluation. This is particularly true following negative screens, which is a well-recognized challenge to diagnostic accuracy studies in general (Whiting et al., 2011) and a particular area of concern noted by the USPSTF (McPheeters et al., 2016). Yet it is equally true that many children who screen positive are not evaluated. Indeed, children can be lost to follow-up at any step of a screening process for many reasons, including their clinician's choice not to refer, their parents' choice not to pursue further assessment, a follow-up test discourages evaluation, or an unaddressed barrier (e.g., scheduling conflicts, financial constraints, or translation need). When follow-up is incomplete, analyses of the accuracy of screening tools hinge on assumptions about the diagnostic status of children who never completed a full diagnostic evaluation.

Further, the accuracy and efficiency of a screening process depends on every step (Kaminsky, Benneyan, & Mullins, 1997). The health care delivery process can break down at many points, thereby limiting the population impact of screening in primary care (Gardner, Bevans, & Kelleher, 2021). For clinical care, it is therefore critical to understand the impact of loss to follow-up in a multi-step identification process.

This paper considers results from the seminal study by Robins et al. (2014), one of the largest screening validation studies to date. In this large-scale study (n = 16,071), the authors implemented a multi-step screening process at 18- and 24-month well-child visits, followed by a diagnostic evaluation. They concluded that the M-CHAT-R (a parent-report screener) has a sensitivity of 91.1% and a specificity of 95.5%. When combined with the secondary follow-up (M-CHAT-R/F), they reported a higher specificity (99.3%) but lower sensitivity (85.4%)—likely because the follow-up interview ruled out some true positives. The USPSTF described this paper as "the most recent study in the United States of population-level screening for ASD" and determined it was a "good-quality U.S. study." Using results from this study, this paper explores implications of lack of follow-up for the accuracy of universal autism screening and participation in screening processes.

Methods

Using Microsoft Excel, we constructed two simulation models that examined loss to follow-up in a multi-stage screening and diagnostic process.

Simulation model #1

Model #1 has three steps (Figure 1): (1) children seen during routine pediatric care screen positive or negative; (2) children either complete a diagnostic evaluation or not; (3) diagnostic outcome status—ASD or no ASD—is determined. This represents a simplification of the multi-stage screening process reported in Robins et al. Specifically, we do not include the secondary interview; therefore, no distinction is made between individuals who scored negative on the M-CHAT follow-up interview and those who did not complete evaluation for other reasons. Consequently, Model #1 simulates the accuracy of a stand-alone parent-report screener (Step 1).

Data were drawn from Robins et al. (2014). We analyzed six different scenarios—each with different assumptions regarding the proportion of children with ASD who did not complete evaluations. Table 1 describes assumptions and their rationale. For each scenario, we estimated the sensitivity and specificity of the M-CHAT-R. As a metric of overall accuracy, we estimated the diagnostic odds ratio (DOR; odds of an ASD diagnosis given a positive screen divided by the odds of an ASD diagnosis given a negative screen). If there is no accuracy (e.g., classification is random), the odds of diagnosis will not differ

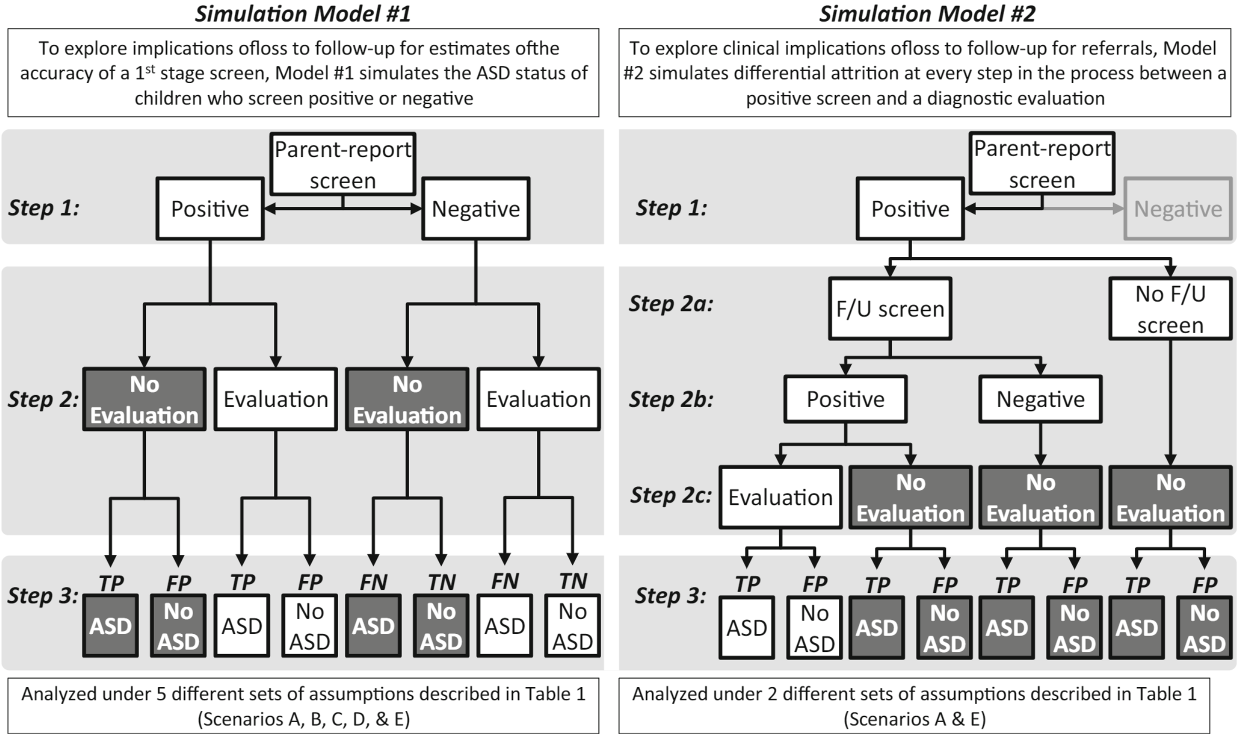

Simulation Model #1
To explore implications ofloss to follow-up for estimates ofthe accuracy of a 1st stage screen, Model #1 simulates the ASD status of children who screen positive or negative
Step 1:
Parent-report screen
Positive
Negative
Step 2:
No Evaluation
Evaluation
No Evaluation
Evaluation
Step 3:
TP
FP
TP
FP
FN
TN
FN
TN
ASD
No ASD
ASD
No ASD
ASD
No ASD
ASD
No ASD
Analyzed under 5 different sets of assumptions described in Table 1 (Scenarios A, B, C, D, & E)
Simulation Model #2
To explore clinical implications ofloss to follow-up for referrals, Model #2 simulates differential attrition at every step in the process between a positive screen and a diagnostic evaluation
Step 1:
Parent-report screen
Positive
Negative
Step 2a:
F/U screen
No F/U screen
Step 2b:
Positive
Negative
Step 2c:
Evaluation
No Evaluation
No Evaluation
No Evaluation
Step 3:
TP
FP
TP
FP
TP
FP
TP
FP
ASD
No ASD
ASD
No ASD
ASD
No ASD
ASD
No ASD
Analyzed under 2 different sets of assumptions described in Table 1 (Scenarios A & E)

Figure 1 Overview of simulation model #1 and simulation model #2. ASD, autism spectrum disorder; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

between positive and negative screens (DOR = 1). DOR >1 indicates some level of accuracy, whereas DOR <1 is worse than random.

Simulation model #2

Model #2 builds on Model #1 by evaluating the influence of loss to follow-up on the accuracy of the whole referral process (Figure 1). Step 1 (screen positive or not) is identical to Model #1. Unlike Model #1, Step 2 is separated into three parts: (2a) children who screened positive either complete a follow-up screen or not; (2b) children score positive or negative on the follow-up screen; (2c) children either complete a diagnostic evaluation or not. Step 3 (diagnostic results) is identical to Model #1.

For Model #2, we differentiate between steps that involve test results, which we characterize by their sensitivity and specificity, and steps that involve follow-up completion, which require new metrics. For the latter, we estimate true positives who complete follow-up (TPCFU) and true negatives with incomplete follow-up (TNIFU). While TPCFU and TNIFU are analogous to sensitivity and specificity (respectively), they are conceptually distinct as they involve factors independent of screening results (as described above) that can influence follow-up completion (Figure S1).

Figure 2 depicts two methods for estimating process sensitivity and specificity. The first method relies on familiar equations. The second method combines estimates of the accuracy of individual steps of the screening process.

This second equation allows for separate estimates of the combined accuracy of steps involving the accuracy of test results and steps involving the accuracy of follow-up completion, which combine to determine the accuracy of the overall process. Analyses of Model #2 focus on scenarios A and F (see Table 1).

Results

Simulation model #1

The top portion of Figure 3 presents data as published by Robins et al. (2014). The boxes with dark shading highlight unknown diagnostic outcomes for children who did not complete evaluations. The lower portion of Figure 3 presents assumptions from the six scenarios. Estimates of sensitivity, specificity, and ASD prevalence in the overall population appear to the right. Results of each scenario are as follows:

1. *Assumptions made by* Robins et al. (2014). Among the 123 children with ASD, 112 screened positive (sensitivity = 91.1%), and among the 15,948 children without ASD, 14905 screened negative (specificity = 93.5%). These assumptions imply 0.77% prevalence in the screened population. Note that while Figure 3 includes all children screened, Robins et al. excluded children if follow-ups were attempted but evaluations remained incomplete. This difference in assumptions results in a small difference between our estimate of specificity and the original estimate as published. Scenario A assumes that among children without ASD, 14,905 initially screened negative and 1,043 screened positive. In contrast, Robins et al. assumed that among children without ASD, 14,798 initially screened negative and 700 screened positive—hence, specificity = 95.5%.

Table 1 Scenarios considered in the simulation, their assumptions, and their rationale

Scenario	Rationale	Assumptions concerning loss to follow-up
A. Assumptions implicit in analyses published by Robins et al. (2014).	This paper counted only cases of ASD that were confirmed by a diagnostic evaluation, 112 of whom initially screened positive and 11 of whom initially screened negative (see table 3 of Robins et al. (2014)). Children were assumed not to have ASD if (a) they screened negative on the parent-report screener (M-CHAT-Revised) or at follow-up interview, and (b) there was no further attempt at follow-up. Based on this reasoning, sensitivity was reported as 91.1% (i.e., 112/123).	Consistent with this logic, scenario A assumes that none of the children lost to follow-up have ASD.
B. Scenario A modified with USPSTF assumptions regarding screen positive cases	The USPSTF questioned the assumptions of the original paper. For example, they noted that the true positive cases detected by screening represent 0.65% of the population and that “expected prevalence in the current sample would likely be lower because children identified early by parent or clinical concern, or lost to follow-up (29% of screen-positive cases were not followed up) were excluded. If one assumed no differential attrition, projected prevalence of screen-detected ASD would be 0.92 percent. . .”	Consistent with this logic, scenario B assumes that there were an additional 36 cases of ASD among screen positive children who did not complete evaluations (raising the “projected prevalence of screen-detected ASD” to [112 + 36]/16,071 = 0.92%). Total ASD cases: 159 (112 + 11 + 36).
C. Scenario B modified with USPSTF assumptions regarding population prevalence	The USPSTF also assumed that “the known population prevalence” was 1.47%. If this assumption were true, there must have been additional cases of ASD that were missed.	Building on scenario B, scenario C assumes that an additional 77 children who screened negative but were lost to follow-up have ASD. This assumption raises the total number of cases to 236 (112 + 11 + 36 + 77), implying a prevalence in the screened population of 1.47% (i.e., 236/16071).
D. Scenario C modified with more specific reasoning regarding the diagnostic status of the 920 children who initially screened positive but did not complete the diagnostic evaluation.	*Scenario D assumes that among screen positive children, prevalence varies across each of the following groups: (a) 584 children who completed the follow-up interview, screened negative, and did not complete evaluation; (b) 209 children who did not complete the follow-up interview, and (c) 127 children who screened positive on the follow-up interview but did not complete the diagnostic evaluation. Specific assumptions for each group are grounded as follows*:	
	Robins et al. (2014) reported that 7 of the 598 children[a] (1.2%) who scored positive on the parent-report screen but negative on the follow-up interview were diagnosed with ASD.	Similarly, we assumed that among the 584 children[a] who scored positive on the parent-report screen but negative on the follow-up interview and did not complete evaluation, 7 had ASD (1.2%).
	Robins et al. (2014) reported that 112 of the 1,155 children (10%) who scored positive on the parent-report screen were diagnosed with ASD.	Similarly, we assumed that among the 209 children who scored positive on the parent-report screen but did not complete the follow-up interview, 20 had ASD (10%).
	Robins et al. (2014) reported that 105 of the 348 children (30%) who scored positive on the parent-report screen and positive on the follow-up interview were diagnosed with ASD.	Similarly, we assumed that among the 127 children who scored positive on the parent-report screen and positive on the follow-up interview but did not complete a diagnostic evaluation, 38 had ASD (30%).
	In total, scenario D assumes that among children who initially screened positive but were lost to follow-up, 65 (7 + 20 + 38) have ASD (instead of the 36 in scenario B). Like scenario C, scenario D assumes a prevalence of 1.47%. This requires that 48 children who initially screened negative but were lost to follow-up also have ASD (instead of the 77 in scenario C). This brings the total number of cases to 236 (112 + 11 + 65 + 48).	
E. Scenario D modified with higher prevalence	Scenario E assumes that the overall population prevalence is higher than suggested by the USPSTF—i.e., 2.20% rather than 1.47%. This higher estimate is equivalent to the report by Guthrie et al. (2019) regarding the proportion of children who were screened with the original M-CHAT plus follow-up interview in primary care and later diagnosed with ASD, somewhat lower than parent report in the National Survey of Children’s Health (Kogan et al., 2018), and within the range of state-level estimates of prevalence reported by the Centers for Disease Control and Prevention (Sheldrick & Carter, 2018).	To align with this assumption, scenario E assumes that an additional 118 children who screened negative have ASD. This assumption raises the total number of cases to 354 (112 + 11 + 65 + 48 + 118), implying a prevalence in the screened population of 2.20% (i.e., 354/16071).

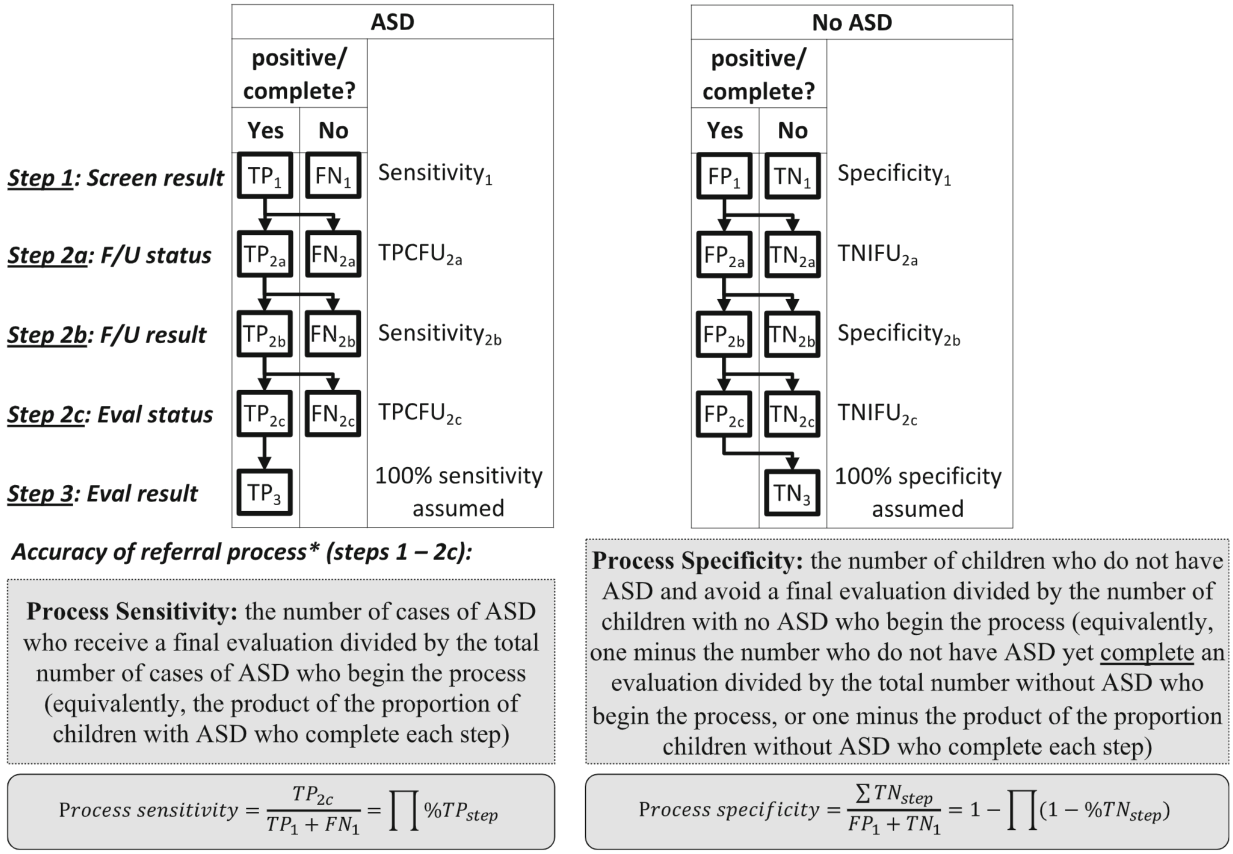
(continues)

Table 1 (continued)

Scenario	Rationale	Assumptions concerning loss to follow-up
F. Scenario E modified with higher prevalence	Scenario F assumes that the overall population prevalence is higher than suggested by the USPSTF or Guthrie et al. (2019)—i.e., 2.78 rather than 1.47% (Scenario D) or 2.20% (Scenario E). This higher estimate is the prevalence estimate reported by the Autism and Developmental Disabilities Monitoring Network in 2023. It is the highest estimate of prevalence to date.	To align with this assumption, scenario F assumes that an additional 93 children who screened negative have ASD. This assumption raises the total number of cases to 447 (112 + 11 + 65 + 48 + 118 + 93), implying a prevalence in the screened population of 2.78% (i.e., 447/16071).

ASD, autism spectrum disorder; M-CHAT, Modified Checklist for Autism in Toddlers; USPSTF, United States Preventive Services Task Force.

[a]Discrepancy between $n = 584$ and $n = 598$ is explained in the caption to Figure 3.



Process Sensitivity: the number of cases of ASD who receive a final evaluation divided by the total number of cases of ASD who begin the process (equivalently, the product of the proportion of children with ASD who complete each step)

$$\textit{Process sensitivity} = \frac{TP_{2c}}{TP_1 + FN_1} = \prod \%TP_{step}$$

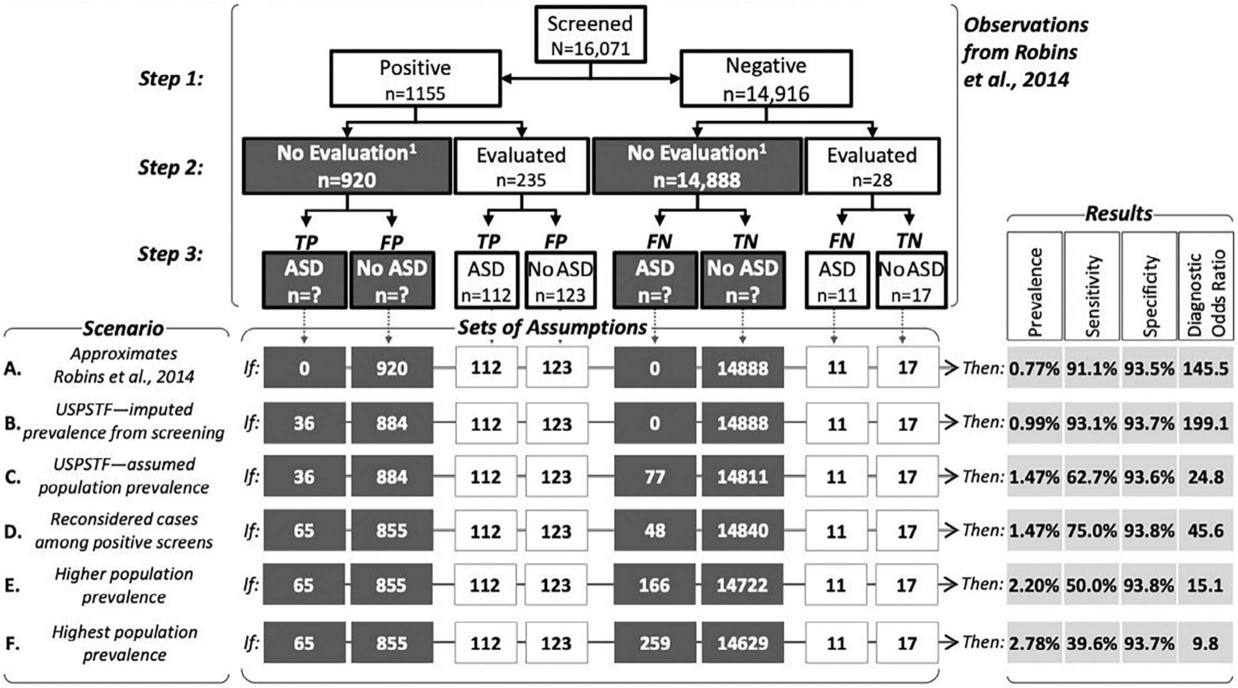
Process Specificity: the number of children who do not have ASD and avoid a final evaluation divided by the number of children with no ASD who begin the process (equivalently, one minus the number who do not have ASD yet complete an evaluation divided by the total number without ASD who begin the process, or one minus the product of the proportion children without ASD who complete each step)

$$\textit{Process specificity} = \frac{\sum TN_{step}}{FP_1 + TN_1} = 1 - \prod (1 - \%TN_{step})$$

Figure 2 Analyzing the accuracy of a screening and referral process. ASD, autism spectrum disorder; FN, false negative; FP, false positive; TN, true negative; TNIFU, True negative incomplete follow-up; TP, true positive; TPCFU, True positive complete follow-up. If the specificity of any step = 100%, process specificity = 100%. Becauase the evaluations is assumed to be perfectly accurate, we therefore exclude the evaluation result (step 3) to focus on the referral process (steps 1–2c)

2 *Modified with USPSTF assumptions regarding screen positive cases*. Following USPSTF assumptions regarding cases of ASD among screen-positive children who were lost to follow-up, scenario B suggests that of the 1,155 children who screened positive, 148 have ASD (sensitivity = 93.1%). Specificity was similar to scenario A (14,905/15912 = 93.7%). These assumptions imply 0.99% prevalence in the screened population (and 0.92% "prevalence of screen-detected ASD"). In summary, assuming that there are at least some cases of ASD among screen-positive children who are not evaluated increases the number of true positives, thereby increasing estimates of sensitivity and prevalence.

3 *Modified with USPSTF assumptions regarding population prevalence*. Following the USPSTF assumption that "the known population prevalence" is 1.47%, scenario C results in an estimate of sensitivity that is much lower (62.7%) than scenario B and a roughly equivalent specificity (93.6%). In summary, assuming that there are at



Scenario	TP ASD	FP No ASD	TP ASD	FP No ASD	FN ASD	TN No ASD	FN ASD	TN No ASD	Prevalence	Sensitivity	Specificity	Diagnostic Odds Ratio
A. Approximates Robins et al., 2014	0	920	112	123	0	14888	11	17	0.77%	91.1%	93.5%	145.5
B. USPSTF—imputed prevalence from screening	36	884	112	123	0	14888	11	17	0.99%	93.1%	93.7%	199.1
C. USPSTF—assumed population prevalence	36	884	112	123	77	14811	11	17	1.47%	62.7%	93.6%	24.8
D. Reconsidered cases among positive screens	65	855	112	123	48	14840	11	17	1.47%	75.0%	93.8%	45.6
E. Higher population prevalence	65	855	112	123	166	14722	11	17	2.20%	50.0%	93.8%	15.1
F. Highest population prevalence	65	855	112	123	259	14629	11	17	2.78%	39.6%	93.7%	9.8

Figure 3 Influence of missing data assumptions on analyses of screener accuracy (Model #1). ASD, autism spectrum disorder; FN, false negative; FP, false positive; TN, true negative; TP, true positive; USPSTF, United States Preventive Services Task Force. [1]Robins et al. report that 221 evaluations were completed with children who scored positive on both the first and second screens, but it is unclear how many of the remaining 42 evaluations were conducted with children who also screened positive stage 1 (see Robins et al. Figure 1). Given that seven cases wer diagnosed among children who initially screened positive but scored negative at the second stage, the minimum is 7. Given that the only pathway to receive an evaluation for children who scored negative at the second stage was to be "invited to complete the Screening Tool for Autism in Two-Year-Olds (STAT)" and that "of 375 children who completed the STAT, 20 were evaluated on the basis of a screen positive STAT," the maximum is 20. For the purposes of this paper, we assumed that 14 children who screened initially screened positive but scored negative at the second stage completed evaluations and 28 children initially screened negative but completed an evaluation

least some cases of ASD among screen-negative children who are not evaluated increases the number of false negatives, thereby decreasing estimates of sensitivity.

4 *Modified with more specific reasoning regarding the diagnostic status of the 920 children who initially screened positive but did not complete evaluations*. Scenario D simulates a somewhat more complex line of reasoning regarding the 920 screen-positive children who did not complete evaluations. While scenarios C and D are consistent with the USPSTF assumption that prevalence equals 1.47%, scenario D differs from scenario C in that it assumes proportionally fewer cases of ASD among children who screen negative (i.e., 48 among those lost to follow-up). Reflecting these assumptions, scenario D assumes that of the 1,155 children who screened positive, 177 (65 + 112) have ASD (sensitivity = 75.0%), and of the 14,916 who screened negative, only 59 have ASD (specificity = 93.8%).

5 *Modified with higher prevalence*. Scenario E builds on the assumptions of scenario D regarding screen-positive children but assumes that the overall population prevalence is 2.2%. Because its assumptions imply a greater number of false negative cases, the estimate of sensitivity is much lower (50.0%) while specificity is equivalent to scenario D (93.8%).

6 *Highest prevalence*. Scenario F assumes an even greater number of false negative cases, resulting in a population prevalence is 2.78%. This scenario results in the lowest estimate of sensitivity (39.6%), with similar specificity (93.7%).

Simulation model #2

Figure 4 estimates the contribution of each step of the screening process under scenarios A and F described above. Because the final diagnostic evaluation is assumed to be perfectly accurate, we focus only on the prior referral process (Steps 1–2c). The left-hand panel displays observed results (reflecting Figure 1 in Robins et al., 2014). The middle panel models how children with and without ASD proceed through each step under Scenario A. The right-hand panel does the same for Scenario F.

Consistent with Robins et al. (2014), scenario A assumes that among 123 children with ASD, 112 screen positive. All of these children also complete the follow-up screen, so Step 2a's TPCFU =100%. Of

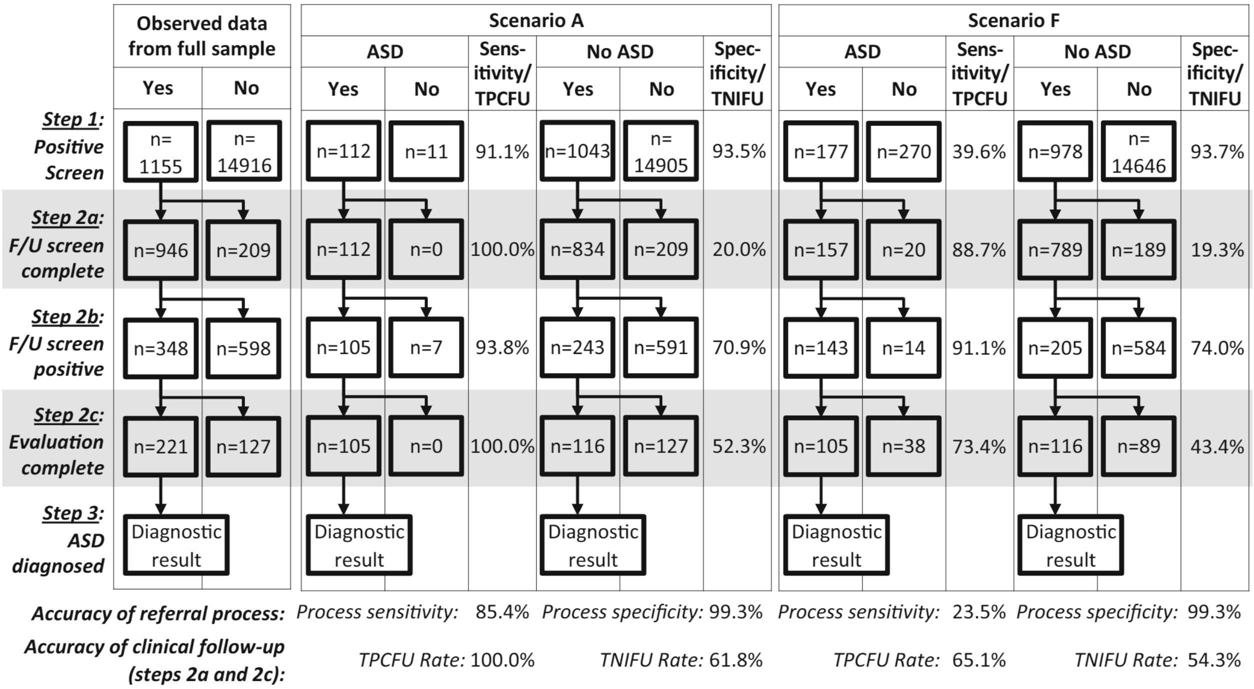


Figure 4 Sensitivity and specificity at each step following a positive initial screen in Scenarios A and F (Model #2). ASD, autism spectrum disorder; TPCFU rate, True positive complete follow-up rate (see Figure S1; estimated based on steps 2a and 2c using the equation for sensitivity in Figure 2); TNIFU Rate, True negative incomplete follow-up rate (see Figure S1; estimated based on steps 2a and 2c using the equation for specificity in Figure 2)

these, 105 score positive on the follow-up screen; thus, Step 2b's sensitivity = 93.8%. All 105 children complete a diagnostic evaluation (TPCFU = 100%) and all are diagnosed with ASD. Combining Steps 2a and 2c, which involve follow-up completion, the TPCFU rate is 100%. However, the sensitivity of the overall process is 85.4% (equivalent to the estimate for the M-CHAT-R/F in Robins et al., 2014) because some children with ASD incorrectly screen negative in Steps 1 and 2b.

Scenario A (Figure 4) assumes that 15,948 children do not have ASD. Of these, 14,905 correctly screen negative; thus, Step 1's specificity = 93.5%. Of the 1,043 children who incorrectly screen positive, 209 do not complete the follow-up screen, so Step 2a's TNIFU = 20%. Of the 834 children who complete a follow-up screen, 591 score negative, thus, Step 2b's specificity = 70.9%. Of the 243 children who incorrectly screen positive at Step 2b, 127 do not complete an evaluation (TNIFU = 52.3%). Combining Steps 2a and 2c, the TNIFU rate is 61.8%. However, the specificity of the overall process is 99.3% (equivalent to the estimate for the M-CHAT-R/F in Robins et al., 2014) because test results in Steps 1 and 2b rule out a high proportion of children with no ASD (specificity$_{1,2b}$ = 98.1%). For the entire process, DOR = 796.1—that is, the odds that a child with autism receives a clinical diagnosis is approximately 796 times as high as the odds that a child without autism receives a clinical diagnosis. In comparison, a value of 290 was the largest odds ratio identified in a review of epidemiological studies, which classified any value over 6.7 as large (Chen, Cohen, & Chen, 2010). For the clinical steps, DOR is infinitely high (because TPCFU rate = 100%—i.e., follow through on referrals is assumed to be perfect for children with autism).

Scenario F (Figure 4) assumes that 447 children have ASD. Of these, 177 screen positive; thus, Step 1's sensitivity = 39.6%. Only 157 children complete follow-up, so Step 2a's TPCFU = 88.7%. Of these, 143 screen positive at follow-up; thus, Step 2b's sensitivity = 91.1%. Only 105 children complete a diagnostic evaluation (TPCFU = 73.4%), and all are diagnosed with ASD. Combining Steps 2a and 2c, the TPCFU rate is 65.1%. However, the sensitivity of the overall process is only 23.5% because some children with ASD incorrectly screen negative in Steps 1 and 2b (sensitivity$_{1,2b}$ = 45.5%).

Scenario F (Figure 4) assumes that 15,624 children do not have ASD. Of these, 14,739 correctly screen negative; thus, Step 1's specificity = 93.7%. Of the 978 children who incorrectly screen positive, 189 do not complete follow-up, so Step 2a's TNIFU = 19.3%. Of the 789 children who complete follow-up, 584 screen negative, thus, Step 2b's specificity = 74.0%. Of the 205 children who incorrectly screen positive, 89 do not complete an evaluation (TNIFU = 43.4%). Combining Steps 2a and 2c, the TNIFU rate is 54.3%. However, the specificity of the overall process is 99.3% because test results in Steps 1 and 2b rule out a high proportion of children with no ASD (specificity$_{1,2b}$ = 98.4%). The DOR for the entire process is 41.0. For the clinical steps, DOR = 2.2.

Discussion

This paper presents six empirically-derived scenarios that demonstrate the degree to which estimates of screening accuracy can be influenced by assumptions regarding the diagnostic status of children who do not complete evaluations. Analyzing precisely the same data, implicit assumptions adopted by the USPSTF and by Robins et al. (2014) result in sensitivity estimates that range from 62.7% to 93.1%. Assuming greater prevalence resulted in an even greater range, with sensitivity estimates as low as 39.6%. Different assumptions also have implications for differential attrition in each step of the "screening-to-treatment chain" (Silverstein & Radesky, 2016). For example, the scenario that most closely reflects the assumptions of Robins et al. (2014) implies that the M-CHAT-R's sensitivity = 91.1%, but also that all children with ASD who screened positive completed follow-up at every step (i.e., TPCFU = 100%) and only children without ASD failed to do so (TNIFU = 61.8%). In contrast, our final scenario suggests that the M-CHAT-R's sensitivity = 39.6%. For this to be true, only a portion of children with ASD who screened positive must have completed follow-up (i.e., TPCFU = 65.1%), while many children without ASD also failed to do so (TNIFU = 54.3%).

On a technical level, these scenarios also demonstrate the close linkage between loss to follow-up and prevalence. Assumptions about the diagnostic status of children lost to follow-up (as in Scenarios A and B) have implications for prevalence. Conversely, assumptions about prevalence (as in scenarios C–F) have implications for the number of cases that must have been lost to follow-up. Either way, simulation models constrain the two sets of assumptions to be consistent. Also notable is the apparent stability of specificity across scenarios. However, recall that the numerator for specificity is a count of true negatives, whereas the numerator for sensitivity is a count of true positives. Across scenarios, the number of true negatives who received no evaluation varied between 14,629 and 14,888 (range = 259). The number of true positives who received no evaluation varied between zero and 65 (range = 65). Thus, the numerator for specificity varied by more than the numerator for sensitivity. However, the denominator for specificity is very large (i.e., ranging from 15,624 to 15,948 children without ASD across scenarios) compared to the denominator for sensitivity (i.e., ranging from 123 to 447 children with across scenarios). Thus, differences in specificity obscure differences in the number of true negatives while sensitivity accentuates differences in the number of true positives.

Updates published since the USPSTF recommendation

In 2019, Guthrie et al. published an analysis of 25,999 children screened with the M-CHAT-R (many of whom received the follow-up screen) that is among the most influential papers on autism screening published since the last USPSTF review. The methodology differed from Robins et al. (2014). Instead of prospectively following screened children, the authors examined electronic medical records to retrospectively determine whether children diagnosed with ASD through at least 4 years of age had previously screened positive on the M-CHAT-R. The paper estimated much higher prevalence (2.2%) and lower sensitivity (38.8%; CI: 34.3%–43.3%) and specificity (94.9%; CI: 94.5%–95.2%; Guthrie et al., 2019). Notably, a similar longitudinal study from Norway that analyzed 54,463 children screened with the M-CHAT reported similar sensitivity (31.2%), yet lower prevalence (0.62%; Schjolberg et al., 2021).

By applying higher prevalence estimates (2.2%, 2.78%) from Guthrie et al. (2019) and from recent surveillance (Maenner et al., 2023) to the observations reported by Robins et al. (2014), our final two scenarios offer a potential explanation for the discrepancy in sensitivity estimates. Assuming that a significant proportion of children lost to follow-up initially screened negative, observations reported by Robins et al. (2014) are consistent with lower estimates of accuracy (sensitivity = 39.6%–50%, specificity = 93.7%–93.8%) that are closer to Guthrie et al. (sensitivity = 34.3%; specificity = 94.9%) than to the original paper (sensitivity = 91.1%; specificity = 95.5%). Notably, the assumptions inherent in the report by Robins et al. (2014) imply that families, clinicians, and the system in which they work are very successful at ensuring that children who have ASD (but are not yet diagnosed) receive evaluations while children without ASD avoid them (TPCFU = 100%; TNIFU = 61.8%). In contrast, scenario F, which assumes that the overall population prevalence is 2.78%, implies a more modest level of accurate clinical follow-up (TPCFU = 65.1%; TNIFU = 54.3%).

Given differences in prevalence and sample ages between the two studies, it is reasonable to ask whether developmental trajectories play a role. As Robins (2020) noted, "all ASD cases likely are not detectable at 18–24 months, based on the trajectories of emerging symptoms, severity of symptoms, and the child's ability to employ compensatory mechanisms, which can mask impairments". Indeed, a body of evidence suggests that many children who meet criteria for ASD at an older age do not meet criteria when evaluated at a younger age. For example, many younger siblings of children with ASD who are not diagnosed at one age are found to meet diagnostic criteria at a later age (Brian et al., 2016; Ozonoff et al., 2015; Zwaigenbaum et al., 2016). Likewise, reviews of health and school records that form the basis of US prevalence estimates find higher prevalence of ASD at age 8 than at age 4 (Christensen et al., 2015, 2016). Finally, studies of diagnostic stability consistently find that more children transition into an ASD diagnosis as they age than transition away from one (Guthrie, Swineford, Nottke, & Wetherby, 2013; Pierce et al., 2019; Woolfenden, Sarkozy, Ridley, & Williams, 2012). Thus, one might expect that some proportion of children who were found (or assumed) not to have ASD at 18–24 months would, if evaluated at a later age, be diagnosed with ASD. However, there are two types of undiagnosed cases: those who initially screened positive and those who initially screened negative. The M-CHAT-R's sensitivity can only be high if screen positive children were later diagnosed with ASD. However, Guthrie et al.'s findings suggest that the majority of children diagnosed with ASD by age 4 years originally screened negative. Therefore, developmental trajectories are not likely to explain the difference in findings.

Instead, Robins (2020) argued that "the purpose of screening toddlers is to refer to expert diagnosticians as many cases as possible in the first 2 years of life so that children can start treatment as early as possible to maximize outcomes." Thus, "we should not be using ASD diagnosis at age 4, 8, or 10 as the primary metric to evaluate the success of the screening program. Rather, we should rigorously apply concurrent case confirmation approaches to measure the performance of the screening tool against the gold standard assessment at that target age." The validity of this argument rests on the definition of the word "case." Does "case" refer only to children who meet diagnostic criteria right now (as reflected by point prevalence)? Or does the word also refer to children with "prodromal" presentations who exhibit symptoms but do not (yet) meet diagnostic criteria for ASD? Is it only the former who should "start treatment as early as possible to maximize outcomes," or would the latter group benefit as well? Evidence that "red flags" for ASD observed by 12 months of age are predictive of later ASD diagnoses suggests the potential utility of offering intervention to children with a broader range of presentations (Dow, Day, Kutta, Nottke, & Wetherby, 2020; Pileggi et al., 2021).

Screening as a process

Simulations demonstrate that when screening is conceptualized as a process, every step matters—including steps that determine whether or not children complete an additional evaluation (Gardner et al., 2021). Research on family navigation recognizes this fact (Broder-Fingert et al., 2020; Feinberg et al., 2021). However, as the National Academy of Medicine (NAM) notes in *Improving Diagnosis in Healthcare*, most research on assessment tools focuses on accuracy (National Academies of Science, Engineering, and Medicine, 2015), not their clinical application. As the NAM highlights, accurate results are not sufficient by themselves to improve care—results must also influence case conceptualization and decision-making among both clinicians and caregivers. Indeed, accurate screening may improve care by highlighting unrecognized signs of autism. But screening may improve care in other ways, for example by prioritizing attention to child development and facilitating shared decision-making. For instance, the results of a screening questionnaire may not be unexpected to families, yet still facilitate building consensus on interpretation and seeking care (Mackie et al., 2021). Thus, accuracy may not be the only important attribute of a screening tool.

Notably, in the scenarios we considered, children were more likely to proceed to the next step of the screening process if they truly had ASD than if they did not (see Figure S1). That is, not only were children more likely to screen positive, they were also more likely to complete follow-up. Thus, higher attrition among children without ASD may enhance the specificity—and therefore the overall accuracy—of screening processes. As suggested above, the mechanisms by which this occurs are likely to involve decisions by clinicians and parents. On reflection, the hypothesis that decision-making can (under some conditions) contribute to the accuracy of a screening process is not all that implausible. Indeed, Guthrie et al. (2019) reported that a majority of children with ASD were detected by means other than screening and diagnosed before 4 years of age. Likewise, in a recent implementation study that documented substantial increases in the diagnosis of ASD while reducing disparities (Sheldrick et al., 2022), an important pathway to diagnosis involved completing an evaluation based on parent and/or clinician concerns despite a negative screen (Sheldrick et al., 2019). In this study, considering the possibility of misclassification was conceptualized as part of the screening process (Sheldrick et al., 2015; Sheldrick & Garfinkel, 2017) and was therefore incorporated in clinicians' training. Designed well, a screening process that enhances shared decision-making can help to mitigate false positive and/or false negative errors that should be expected from any single screening tool.

We note some limitations. Each analysis depended on multiple assumptions, and not all plausible assumptions were considered. For example, our models do not consider possible moderators of screening accuracy, such as child age and cognitive level or race, ethnicity, culture, and language. This is an important topic for future research as the Guthrie et al. (2019) study reported differential performance of the M-CHAT-R in children who were racially and economically diverse. Analyses focused on missing or incomplete data and did not address other potential methodological biases, such as the likelihood of imperfect diagnostic evaluations (see

Appendix S1 for detail). Moreover, we did not analyze reasons for loss to follow-up, such as stigma and social determinants of health, that may be important for improving care.

Conclusions

In this paper, we reconsidered inferences regarding the accuracy of the M-CHAT-R reported by Robins et al. (2014), while highlighting implications for every step in the screening process. On one hand, we find that ASD screening tools may be less accurate than is often reported. On the other hand, parent and clinician decision-making regarding follow-up may contribute more to detection than is widely assumed. We pose two questions for future research. First, what is known versus what is assumed regarding the accuracy and clinical effectiveness of screening tools? As Zwaigenbaum and Maguire (2019) argue, "Ultimately, the potential added value of ASD screening must be considered relative to what would occur in its absence." When evaluating a screening tool, it is critical to consider assumptions underlying estimates of sensitivity and specificity. Second, how can entire screening processes be engineered and monitored to maximize effectiveness? A screening tool is only one element in a larger system of care. Attention to other elements of system design is also warranted, such as how best to mitigate false positive and false negative errors (which inevitably result from any screener [Sheldrick et al., 2015]) and how to support shared decision-making with families who are diverse with respect to race, ethnicity, culture, and language in a way that facilitates timely detection and diagnosis for all children who can benefit from early intervention services.

Data availability statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Figure S1. Overview of key terms.

Appendix S1. Commentary on diagnostic uncertainty.

Acknowledgements

R.C.S. and A.S.C. contributed to the development of a freely-available universal screening questionnaire for autism known as the Parents' Observations of Social Interactions (POSI). A.M.W. contributed to the development of the Early Screening for Autism and Communication Disorders, which is freely-available within Baby Navigator™. She also owns Autism Navigator, LLC, which distributes Baby Navigator™, but does not receive any salary. The remaining authors have no conflicts of interest to disclose.

This research was supported in part by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01HD093055). The content is solely the responsibility of the authors and does not necessarily represent the official views of these federal agencies.

Correspondence

R. Christopher Sheldrick, Chan Medical School, University of Massachusetts, Worcester, MA, USA; Email: radley.sheldrick@umassmed.edu

Key points

- Several past studies conclude that autism screening tools are highly accurate. However, more recent studies suggest lower accuracy. In addition, the National Academy of Medicine emphasizes that accurate screening is not sufficient to improve care—results must also influence decision-making among both clinicians and caregivers.
- Our results demonstrate that estimates of accuracy depend on assumptions regarding the diagnostic status of children who were lost to follow-up. We conclude that ASD screening tools may be less accurate than is often reported.
- However, simulations also demonstrate how every step in a screening process matters—including steps that determine whether children complete an additional evaluation. We conclude that parent and clinician decision-making regarding follow-up may contribute more to detection than is widely assumed.

References

American Psychiatric Association. (2013). *The diagnostic and statistical manual of mental disorders (DSM-5)* (5th edn). Washington, DC: Author.

Brian, J., Bryson, S.E., Smith, I.M., Roberts, W., Roncadin, C., Szatmari, P., & Zwaigenbaum, L. (2016). Stability and change in autism spectrum disorder diagnosis from age 3 to middle childhood in a high-risk sibling cohort. *Autism*, *20*, 888–892.

Broder-Fingert, S., Stadnick, N.A., Hickey, E., Goupil, J., Diaz Lindhart, Y., & Feinberg, E. (2020). Defining the core components of family navigation for autism spectrum disorder. *Autism*, *24*, 526–530.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics: Simulation and Computation*, *39*, 860–864.

Christensen, D.L., Bilder, D.A., Zahorodny, W., Pettygrove, S., Durkin, M.S., Fitzgerald, R.T., … Yeargin-Allsopp, M. (2015). Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network. *Journal of Developmental and Behavioral Pediatrics*, *37*, 1–8.

Christensen, D.L., Van Naarden Braun, K., Baio, J., Bilder, D., Charles, J., Constantino, J.N., … Yeargin-Allsopp, M. (2016). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2012. *MMWR Surveillance Summaries*, *65*, 1–24.

Cohen, J.F., Korevaar, D.A., Altman, D.G., Bruns, D.E., Gatsonis, C.A., Hooft, L., … Bossuyt, P.M.M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. *BMJ Open*, *6*, e012799.

Dow, D., Day, T.N., Kutta, T.J., Nottke, C., & Wetherby, A.M. (2020). Screening for autism spectrum disorder in a naturalistic home setting using the systematic observation of red flags (SORF) at 18–24 months. *Autism Research*, *13*, 122–133.

Feinberg, E., Augustyn, M., Broder-Fingert, S., Bennett, A., Weitzman, C., Kuhn, J., … Blum, N.J. (2021). Effect of family navigation on diagnostic ascertainment among children at risk for autism: A randomized clinical trial from DBPNet. *JAMA Pediatrics*, *175*, 243–250.

Gardner, W., Bevans, K., & Kelleher, K.J. (2021). The potential for improving the population health effectiveness of screening: A simulation study. *Pediatrics*, *148*(Supplement 1), s3–s10.

Guthrie, W., Swineford, L.B., Nottke, C., & Wetherby, A.M. (2013). Early diagnosis of autism spectrum disorder: Stability and change in clinical diagnosis and symptom presentation. *Journal of Child Psychology and Psychiatry*, *54*, 582–590.

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*, e20183963.

Hyman, S.L., Levy, S.E., Myers, S.M., & AAP Council on Children with Disabilities, Section on Developmental and Behavioral Pediatrics. (2020). Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics*, *145*, e20193447.

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

Kaminsky, F.C., Benneyan, J.C., & Mullins, D.L. (1997). Automated rescreening in cervical cytology. Mathematical models for evaluating overall process sensitivity, specificity and cost. *Acta Cytologica*, *41*, 209–223.

Kogan, M.D., Vladutiu, C.J., Schieve, L.A., Ghandour, R.M., Blumberg, S.J., Zablotsky, B., … Lu, M.C. (2018). The prevalence of parent-reported autism spectrum disorder among US children. *Pediatrics*, *142*, e20174161.

Kuntz, K., Sainfort, F., Butler, M., Taylor, B., Kulasingam, S., Gregory, S., … Kane, R.L. (2013). *Decision and simulation modeling in systematic reviews. Methods research report. (prepared by the University of Minnesota Evidence-based Practice Center under contract No. 290–2007-10064-I.) AHRQ Publication No. 11(13)-EHC037-EF*. Rockville, MD: Agency for Healthcare Research and Quality.

Mackie, T.I., Schaefer, A.J., Ramella, L., Carter, A.S., Eisenhower, A., Jimenez, M.E., … Sheldrick, R.C. (2021). Understanding how parents make meaning of their child's behaviors during screening for autism spectrum disorders: A longitudinal qualitative investigation. *Journal of Autism and Developmental Disorders*, *51*, 906–921.

Maenner, M.J., Warren, Z., Williams, A.R., Amoakohene, E., Bakian, A.V., Bilder, D.A., … Shaw, K.A. (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2020. *MMWR Surveillance Summaries*, *72*, 1–14.

McPheeters, M.L., Weitlauf, A.S., Vehorn, A., Taylor, C., Sathe, N.A., Krishnaswami, S., … Warren, Z.E. (2016). *Screening for autism Spectrum disorder in Young children: A systematic evidence review for the U.S. preventive services task force. Evidence synthesis No. 129. AHRQ publication No. 13-05185-EF-1*. Rockville, MD: Agency for Healthcare Research and Quality.

National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care*. Washington, DC: National Academies Press.

Ozonoff, S., Young, G.S., Landa, R.J., Brian, J., Bryson, S., Charman, T., … Zwaigenbaum, L. (2015). Diagnostic stability in young children at risk for autism spectrum disorder: A baby siblings research consortium study. *Journal of Child Psychology and Psychiatry*, *56*, 988–998.

Pierce, K., Gazestani, V.H., Bacon, E., Barnes, C.C., Cha, D., Nalabolu, S., … Courchesne, E. (2019). Evaluation of the diagnostic stability of the early autism spectrum disorder phenotype in the general population starting at 12 months. *JAMA Pediatrics*, *173*, 578–587.

Pileggi, M.L., Brane, N., Bradshaw, J., Delehanty, A., Day, T., McCracken, C., … Wetherby, A.M. (2021). Early observation of red flags in 12-month-old infant siblings later diagnosed with autism spectrum disorder. *American Journal of Speech-Language Pathology*, *30*, 1846–1855.

Robins, D.L. (2020). How do we determine the utility of screening tools? *Autism*, *24*, 271–273.

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*, 37–45.

Sainfort, F., Kuntz, K.M., Gregory, S., Butler, M., Taylor, B.C., Kulasingam, S., & Kane, R.L. (2013). Adding decision models to systematic reviews: Informing a framework for deciding when and how to do so. *Value in Health*, *16*, 133–139.

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*, 1837–1852.

Schjolberg, S., Shic, F., Volkmar, F.R., Nordahl-Hansen, A., Stenberg, N., Torske, T., … Oien, R.A. (2021). What are we optimizing for in autism screening? Examination of algorithmic changes in the M-CHAT. *Autism Research*, *15*, 296–304.

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*, 936–948.

Sheldrick, R.C., & Carter, A.S. (2018). State-level trends in the prevalence of autism Spectrum disorder (ASD) from 2000 to 2012: A reanalysis of findings from the autism and developmental disabilities network. *Journal of Autism and Developmental Disorders*, *48*, 3086–3092.

Sheldrick, R.C., Carter, A.S., Eisenhower, A., Mackie, T.I., Cole, M.B., Hoch, N., … Pedraza, F.M. (2022). Effectiveness of screening in early intervention settings to improve diagnosis of autism and reduce health disparities. *JAMA Pediatrics*, *176*, 262–269.

Sheldrick, R.C., Frenette, E., Vera, J.D., Mackie, T.I., Martinez-Pedraza, F., Hoch, N., … Carter, A.S. (2019). What drives detection and diagnosis of autism spectrum disorder? Looking under the hood of a multi-stage screening process in early intervention. *Journal of Autism and Developmental Disorders*, *49*, 2304–2319.

Sheldrick, R.C., & Garfinkel, D. (2017). Is a positive developmental-behavioral screening score sufficient to justify referral? A review of evidence and theory. *Academic Pediatrics*, *17*, 464–470.

Silverstein, M., & Radesky, J. (2016). Embrace the complexity: The US preventive services task force recommendation on screening for autism spectrum disorder. *JAMA*, *315*, 661–662.

Siu, A.L., Bibbins-Domingo, K., Grossman, D.C., Baumann, L.C., Davidson, K.W., Ebell, M., … US Preventive Services Task Force. (2016). Screening for autism spectrum disorder in young children: US preventive services task force recommendation statement. *JAMA*, *315*, 691–696.

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*, 529–536.

Wieckowski, A.T., Williams, L.N., Rando, J., Lyall, K., & Robins, D.L. (2023). Sensitivity and specificity of the modified checklist for autism in toddlers (original and revised): A systematic review and meta-analysis. *JAMA Pediatrics*, *177*, 373–383.

Woolfenden, S., Sarkozy, V., Ridley, G., & Williams, K. (2012). A systematic review of the diagnostic stability of autism spectrum disorder. *Research in Autism Spectrum Disorders*, *6*, 345–354.

Zwaigenbaum, L., Bauman, M.L., Fein, D., Pierce, K., Buie, T., Davis, P.A., … Wagner, S. (2015). Early screening of autism spectrum disorder: Recommendations for practice and research. *Pediatrics*, *136*(Supplement_1), S41–S59.

Zwaigenbaum, L., Bryson, S.E., Brian, J., Smith, I.M., Roberts, W., Szatmari, P., … Vaillancourt, T. (2016). Stability of diagnostic assessment for autism spectrum disorder between 18 and 36 months in a high-risk cohort. *Autism Research*, *9*, 790–800.

Zwaigenbaum, L., & Maguire, J. (2019). Autism screening: Where do we go from here? *Pediatrics*, *144*(4), e20190925.

Accepted for publication: 27 June 2023