

Diagnostic Agreement of a Structured Telehealth Adult Autism Assessment

Eileen Crehan

University of Massachusetts Medical School

Kelly Whaling

kelly.whaling@prosperhealth.io

Prosper Health <https://orcid.org/0000-0002-2300-1255>

Hannah Jones

Prosper Health

Jacqueline Shinall

Prosper Health

Rachel Loftin

Prosper Health

Research Article

Keywords: neurodiversity, neurodiversity-affirming, assessment, autism assessment, autism, adult diagnosis, diagnostic agreement, telehealth

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11166196/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: EC served as the blinded expert reviewer for this study. EC has no employment relationship with Prosper Health, has not provided clinical services for Prosper Health, and did not supervise any Prosper Health clinicians or cases; EC's expert ratings were completed independently. EC's time on this study was supported through a paid time-buyout arrangement between Prosper Health and the University of Massachusetts Chan Medical School; EC received no direct payment from Prosper Health. EC maintains a private clinical practice specializing in autism assessment for adolescents and adults and holds an academic appointment at the University of Massachusetts Chan Medical School. EC subsequently assumed first authorship after completion of the expert review; this dual role is discussed in the manuscript. KW, HJ, JS, and RL are current employees of Prosper Health, the telehealth organization whose assessment pathway was evaluated in this study. The assessment protocol evaluated in this study was developed and implemented within Prosper Health. The study was designed, conducted, and reported by authors with relationships to the organization and/or the clinical model under evaluation. Bias-reduction steps included preregistration, random case selection, de-identification, blinded expert review from a non-employee, and complete reporting of all primary metrics.

Version of Record: A version of this preprint was published at Journal of Autism and Developmental Disorders on October 12th, 2026. See the published version at <https://doi.org/10.1007/s10803-026-07571-2>.

Abstract

Purpose: This preregistered retrospective study evaluated diagnostic agreement under shared-information conditions for a structured, neurodiversity-affirming telehealth adult autism assessment protocol mapped to DSM-5-TR criteria.

Methods: The sample included 100 adults who primarily communicated through spoken language and was intentionally stratified to include at least 50% Black, Indigenous, and people of color (BIPOC; 51% achieved). Evaluations generally included two 90-minute structured interviews across ten DSM-5-TR-mapped domains, criterion-level decision rules, behavioral observation, questionnaires, and collateral information when available. A blinded expert reviewed de-identified assessment materials and independently classified cases as ASD or non-ASD.

Results: Overall agreement was 86.0% (95% CI [77.9%, 91.5%]), with Cohen's $\kappa = .684$ (95% CI [.531, .836]). The point estimate exceeded the preregistered .66 benchmark, although the confidence interval included lower values. ASD classification rates were 68% for original clinicians and 66% for expert review. Positive and negative percent agreement were 90.9% and 76.5%, respectively. Exploratory subgroup analyses suggested lower agreement among BIPOC participants, although estimates were imprecise.

Conclusion: Findings provide preliminary evidence that a structured, neurodiversity-affirming telehealth framework can yield concordant ASD/non-ASD classifications under shared-information conditions. Neurodiversity-affirming practice can coexist with structured, criterion-based diagnostic decision-making: affirming clients' experiences and reducing deficit-based assumptions does not require presuming an autism diagnosis. Because reviewers relied on substantially overlapping information, findings do not establish diagnostic accuracy or criterion validity. Independent prospective validation using separate evaluations and multiple blinded reviewers is needed.

Introduction

Autism assessment in adulthood presents distinct challenges because many adults seeking first-time evaluation have complex psychiatric histories and longstanding compensatory or masking strategies that can reduce observable autistic features during structured observation, particularly among adults who communicate primarily through spoken language (Belcher et al., 2023; Hull et al., 2017; Lai et al., 2017; Maddox et al., 2017; McQuaid et al., 2022). These characteristics complicate the use of pediatric or adolescent assessment instruments, necessitating further evaluation of whether existing tools perform equivalently across the demographic groups increasingly seeking adult evaluation (Christiansen & Pedersen, 2025; Conner et al., 2019; Kamp-Becker et al., 2018). Additionally, many autistic adults have expressed varying degrees of acceptability with traditional diagnostic instruments (Doherty et al., 2026).

Several adult autism assessment approaches address important clinical needs, including standardized observation, developmental-history interviews, dimensional clinical interviews, sensory-based interview frameworks, and adult-focused clinical interviews. However, many remain proprietary, oriented toward in-person administration, dependent on developmental informants, or supported by limited adult diagnostic-accuracy evidence (Berney et al., 2017; Coleman et al., 2026; Goodwin et al., 2026; Huang et al., 2020; Kamp-Becker et al., 2018; Lord et al., 1994, 2012; Maddox et al., 2017; Mandy et al., 2018; Monteiro, 2018; Wigham et al., 2019, 2020; Wing et al., 2002). These limitations are especially relevant for later-identified adults, for whom developmental informants may be unavailable and observable features may be reduced by masking or compensatory strategies. A detailed description of existing instruments and the present protocol is provided in Supplemental Table S1.

Adult autism diagnoses in the United States have risen substantially, with the largest increases among individuals aged 26 to 34 and among groups historically underdiagnosed in childhood (Goldblum et al., 2024; Grosvenor et al., 2024; Nayyar et al., 2025), and timely identification in adulthood can shape self-understanding, quality of life, and access to supports (Atherton et al., 2022). Telehealth approaches may address some access barriers by reducing geographic constraints and allowing assessment in familiar environments (Alfuraydan et al., 2020; Gibbs et al., 2021; Sutherland et al., 2018). Pediatric and mixed-age telehealth studies have reported moderate to high diagnostic agreement, often around 88% to 92% (Corona et al., 2024;

Gangi et al., 2025; Stavropoulos et al., 2022), but adult-specific evidence remains limited (Gibbs et al., 2021; Matthews et al., 2021).

A structured telehealth pathway could complement established instruments by preserving DSM-5-TR (American Psychiatric Association, 2022) criterion-level clinical reasoning while increasing feasibility for adult evaluation. However, few published studies have examined whether diagnostic classifications from structured, DSM-5-TR-mapped telehealth clinical interviews for adults who communicate primarily through spoken language are reproducible under blinded secondary review. Agreement between raters is an important component of evaluating the reproducibility of diagnostic classification.

Accordingly, the present preregistered study evaluated diagnostic agreement between original clinical classifications generated through a structured telehealth adult autism assessment pathway and blinded secondary classifications based on de-identified assessment records under shared-information conditions. The primary objective was to quantify agreement for dichotomous ASD/non-ASD classification using Cohen's kappa and percent agreement.

The secondary outcomes were directional agreement with blinded expert review and descriptive characterization of discordant cases. Exploratory analyses examined agreement across information quality, age, gender, and aggregate BIPOC versus non-BIPOC groups to identify possible measurement-equity signals warranting adequately powered follow-up.

Methods

Study Design

This preregistered retrospective shared-information diagnostic agreement study compared original clinician ASD/non-ASD classifications from a structured telehealth adult autism assessment pathway with classifications assigned by a blinded expert reviewer. The blinded expert reviewed de-identified assessment materials generated during the original clinical evaluation but did not interact directly with participants. Because the two classifications were derived under different assessment conditions but from substantially overlapping source information, the study is characterized as diagnostic agreement under shared-information conditions.

The study was approved by an institutional review board. Full ethics approval information, preregistration details (including the registry DOI), funding, data availability, and conflicts of interest are provided on the separately submitted title page to preserve double-anonymized review. The study was reported using STARD-informed elements (Bossuyt et al., 2015) adapted for a shared-information diagnostic agreement design. A reporting checklist is provided in Supplemental Table S2. Original assessments occurred during a five-month eligibility window, and case selection and blinded expert review proceeded on a rolling, batch-wise basis during a subsequent 13-week review period. Exact case-level elapsed time between the original assessment and expert review could not be calculated because review-completion dates were not retained in the analytic dataset.

Rater Information Conditions

The original evaluating clinician made the diagnostic determination after direct synchronous video interaction with the participant across the assessment sessions. The clinician therefore had access to live behavioral observations, the direct clinical interview, the participant's clinical record, and questionnaires, collateral information, and other assessment materials when available. The blinded expert did not interact with participants and instead reviewed de-identified materials generated during the original assessment, including clinician session notes, available verbatim transcripts, client-completed questionnaires and intake forms, available standardized measure data, and available collateral information. The expert packet excluded the original diagnostic conclusion, billing diagnosis, feedback-letter diagnosis, and any documentation field that directly stated the clinician's final ASD/non-ASD determination. Cohen's kappa was used to quantify agreement between the two classifications.

Setting and Eligibility

Assessments were conducted in a U.S.-based specialty telehealth clinic providing adult autism evaluations by licensed psychologists. The pathway was designed for adults seeking first-time or confirmatory autism evaluation whose communication and cognitive profile was appropriate for an extended spoken-language telehealth interview. No standardized language, verbal-fluency, intellectual-functioning, or cognitive assessment instruments were administered. Eligibility was determined clinically during a structured orientation period at the beginning of the initial session. Clinicians considered communication during conversation, educational and occupational history, and brief orientation and reasoning questions to screen for possible intellectual disability or cognitive impairment and to inform judgments about communication and cognitive functioning relevant to telehealth assessment. This screening was not used to estimate IQ, diagnose cognitive impairment, or assign autism support level. Clients for whom the orientation raised concern for intellectual disability, significant cognitive impairment, or other needs requiring formal cognitive assessment were referred for a more comprehensive in-person evaluation.

The same orientation process was used to evaluate broader clinical appropriateness for a remote, autism-only assessment. Clinicians screened for significant traumatic brain injury or other neurological conditions likely to substantially affect diagnostic interpretation, active psychosis, florid mania, acute suicidality, intensive psychiatric treatment, and active substance use disorders, and deferred evaluation or referred clients for in-person or higher-acuity evaluation when needed.

The clinical approach emphasized structured integration of DSM-5-TR criterion-level evidence, developmental history, masking and compensatory strategies, sensory experience, functional impact, differential diagnosis, and contextual factors relevant to adult presentations. For inclusion in the study evaluation pool, cases also had to be completed by clinicians with at least six months of clinic tenure and no past or current quality-assurance concerns regarding documentation or diagnostic decision-making. Study inclusion further required age 19 years or older, completion of all telehealth evaluation sessions, documentation meeting post-standardization quality standards, and a completed diagnostic determination. These criteria yielded a sample of cases evaluated under the clinic's established training, documentation, and quality-assurance procedures.

Assessment Protocol

Each evaluation used a standardized multi-method clinical protocol comprising direct clinical interview, DSM-5-TR criterion-level clinical judgment, behavioral observation during synchronous video sessions, and integration of collateral information, questionnaires, and transcripts when those materials were available. Questionnaire scores were reviewed as part of the broader diagnostic evidence base and were not used as standalone diagnostic determinants or primary study outcomes.

Measures available in the broader clinical record in some cases included the Social Communication Questionnaire-Lifetime Version (SCQ; Rutter et al., 2003), the Social Responsiveness Scale, Second Edition (SRS-2; Constantino & Gruber, 2012), and the Adaptive Behavior Assessment System, Third Edition (ABAS-3; Harrison & Oakland, 2015). Because questionnaire availability was not extracted as a structured variable, the study could not determine how often standardized measures contributed to expert review or whether measure availability differed across cases.

The structured clinical interview was adapted from the Royal College of Psychiatrists diagnostic interview guide for autism in adults (Berney et al., 2017) and reorganized around ten DSM-5-TR-mapped assessment domains: developmental history, the three Criterion A social-communication domains, the four Criterion B restricted and repetitive behavior domains, functional impact, and differential diagnosis. For each domain, clinicians followed structured probes, documented specific examples from adulthood and development when the record supported them, and applied predefined decision rules for rating criterion-level evidence.

Masking, camouflaging, sensory experience, adult functional impact, and differential diagnosis were treated as explicit assessment targets. Interviews generally consisted of two 90-minute sessions. After the interview, clinicians completed a standardized DSM-5-TR decision form embedded in the clinical record. The form required criterion-level ratings across the three Criterion A social-communication domains and the four Criterion B restricted and repetitive behavior domains.

For an autism diagnosis to proceed, social-emotional reciprocity had to be rated clearly evident, the remaining Criterion A domains had to be rated at least possibly evident, and at least two Criterion B domains had to be rated clearly evident. The form also required clinicians to document childhood signs of autism and current functional differences, including whether each was supported by self-report, collateral report, or both.

The present study evaluates a structured assessment framework and replicable components are the ten-domain interview structure, DSM-5-TR criterion-level mapping, behavioral observation during synchronous clinical interview, differential-diagnosis documentation, and integration of multiple evidence sources (see Supplemental Table S3). The protocol was designed as a clinical protocol that does not require specialized certification, proprietary materials, or controlled access to administration procedures.

Interview Guide Development and Adaptation

The interview guide was adapted from the Royal College of Psychiatrists' Interview Guide for the Diagnostic Assessment of Able Adults with Autism Spectrum Disorder (Berney et al., 2017). The source guide provided structured coverage of developmental history, social communication, restricted and repetitive behavior, sensory features, and functional impact. For the present pathway, these areas were reorganized around DSM-5-TR criterion-level decision-making and adapted for two synchronous telehealth sessions. The language and clinical framing were also revised to reflect a neurodiversity-affirming approach, with greater attention to the client's own account of their experience, developmental and environmental context, functional impact, strengths, and support needs.

Protocol development used two parallel consultative processes and was not designed as a formal content-validation study. Seven autistic adults reviewed draft interview materials using cognitive-pretesting principles and provided feedback focused on wording and interpretability, emotional valence, potentially leading language, deficit-based framing, and stereotyped assumptions. Meanwhile, five autism-specialist clinical psychologists and psychiatrists separately reviewed the protocol for diagnostic specificity, developmental inquiry, masking and compensatory strategies, co-occurring psychiatric presentations, and differential diagnosis. Feedback from both consultation streams informed revisions to wording, structured probes, developmental-history procedures, and clinical framing. The consultation was not organized according to a prespecified sequence of review rounds and did not use formal consensus or adjudication rules or quantitative content-validity indices.

These consultations led to changes in both the wording of the interview and the way information was gathered and interpreted. Greater emphasis was placed on clients' internal experiences and autonomy, with less reliance on externally observable behavior alone. Both autistic consultants and autism specialists also raised the difficulty of obtaining collateral developmental information in adult assessment. Developmental history procedures were therefore adapted to treat clients as the primary informants, with collateral information incorporated when available but not required. Clinicians used age-anchored prompts (e.g., "What did you play when you were around 5?") and revisited specific experiences with clear and concrete prompts across developmental periods to support recall and contextualize current presentations.

Other changes based on consultant feedback included that observable social behavior was described using more behaviorally specific language. For example, earlier guidance included descriptors such as speech sounding "normal," eye contact being "natural and expressive," facial expression being "varied, communicative and vivacious," and behavior being "unusual" or "eccentric." The revised protocol shifted toward describing what was directly observed and then contextualizing those observations using the client's own experience of nonverbal communication, including effort, learned strategies, masking, and compensatory behavior. A client might therefore appear highly expressive, warm, and socially engaged during the assessment while also reporting that producing those behaviors requires substantial conscious effort or feels unnatural and exhausting. Similarly, limited eye contact or speech with relatively unvarying pitch was documented as an objective behavioral observation that may be consistent with autism, without characterizing the behavior itself as eccentric, unusual, or inherently abnormal. Supplemental Table S3 summarizes the domain structure and DSM-5-TR mapping of the assessment framework, with one domain (social-emotional reciprocity) shown in full detail, including its example structured probe, documentation requirement, rating anchor, and decision rule. The remaining domains are intentionally not specified to this level of detail as the protocol is designed as an adaptable clinical framework and not a fixed, standardized instrument. Clinicians or researchers implementing

this approach are expected to develop probes, documentation practices, and decision rules suited to their own setting, following the structure illustrated here.

Clinician Training and Quality Assurance

All evaluating clinicians were licensed psychologists who completed clinic-specific onboarding before conducting independent assessments. Onboarding included review of the ten-domain interview structure, DSM-5-TR decision rules, documentation expectations, and differential-diagnosis considerations in adult autism assessment. Ongoing internal quality-improvement activities focused on documentation quality, application of decision rules, and diagnostic consistency. Quality assurance included review of completed assessment documentation, with feedback on domain coverage, differential-diagnosis documentation, criterion-level rating consistency, and alignment between documented evidence and final classification. Clinicians whose documentation did not meet quality standards received additional supervision and, when indicated, retraining.

For the present study, the sampling frame was restricted to cases completed by clinicians with at least six months of clinic tenure and no past or current quality-assurance concerns regarding documentation or diagnostic decision-making. Although the clinic used ongoing quality monitoring and diagnostic-consistency review, formal research calibration procedures, such as prospective group rating of shared case vignettes or quantified monitoring of rater drift, were not conducted during the study period.

Sampling and Participant Flow

Assessment data were stored in a structured clinical data warehouse populated from the electronic health record. Personnel who were not involved in expert review, data analysis, or outcome interpretation queried the warehouse to identify completed adult autism assessment cases from the eligible date range. After clinician-eligibility and case-completeness criteria were applied, 2,130 cases remained in the sampling frame. The final analytic sample comprised 100 cases completed by 52 clinicians.

Per-clinician contribution was generally sparse. The median number of cases per clinician was 1 (mean = 1.92; range = 1-6; IQR = 1-3). Thirty-three clinicians (63.5%) contributed a single case. The highest-contributing clinicians each contributed 6 cases (6.0% of the analytic sample). Five clinicians contributed 26 of 100 cases (26.0%), and the ten highest-contributing clinicians contributed 45 of 100 cases (45.0%). No single clinician contributed more than 6% of the analytic sample. Formal clinician-level multilevel modeling was not conducted because most clinicians contributed too few cases for stable variance-component estimation.

Before case selection, the research team defined sampling strata that were implemented by the data engineering team when they generated the case pool. Cases were selected using a computer-generated random number algorithm within strata and released to the expert reviewer in sequential batches across the 13-week review period, with each subsequent batch provided after the expert had completed ratings for the preceding batch. Sampling was stratified to achieve a target of at least 50% representation among participants identifying as Black, Indigenous, and people of color (BIPOC). Original diagnostic outcomes were not disclosed to the study analyst or expert reviewer during case selection or packet preparation. A de-identification protocol removed names, dates, geographic information, and other direct identifiers from all case materials before expert review. A separate re-identification key was inaccessible to the reviewer. Three non-study practice cases were used to pilot the rating system and were not included in the analytic sample. Participant selection and case flow are summarized in Supplemental Table S4.

Blinded Expert Review

Blinded expert review was conducted by a single licensed clinician-researcher with extensive postdoctoral experience in autism assessment across the lifespan. The expert reviewer is an author on this manuscript. The reviewer had no involvement in the original evaluations, case selection, de-identification, packet preparation, or outcome analysis. The reviewer also had no clinical, supervisory, or employment relationship with the evaluating clinicians or the organization in which the data were collected and received no direct payment from that organization. During review, the expert was blinded to the original diagnostic outcomes.

Expert-review packets contained the shared assessment materials described above but omitted all fields that directly disclosed the original clinician's diagnostic conclusion.

Several safeguards were used to reduce the influence of diagnostic framing. Records were de-identified, direct diagnostic conclusions were removed, and the expert was blinded to the original outcome. Review was guided by a structured DSM-5-TR-based rating form, and the expert completed the classification before it was compared with the original clinician classification. Because the expert reviewed notes and transcripts generated during the original assessment, some implicit diagnostic framing may still have remained in the source material.

For each case, the reviewer completed a structured DSM-5-TR rating form. The reviewer rated A1-A3 and B1-B4, two childhood-evidence items, and two functional-impact items using a three-category scale of Not evident, Maybe, or Clearly evident. The reviewer then selected an autism diagnostic classification and assigned a support level when autism was diagnosed. The rating form also included an optional item assessing the extent to which masking or camouflaging appeared to contribute to the presentation; this item was not completed for all cases. The reviewer also documented areas of diagnostic ambiguity or concern and cultural or contextual considerations in open-text fields and rated the overall quality of information available for making a confident diagnostic determination as Excellent, Good, Fair, or Poor.

Open-text responses describing diagnostic ambiguity or concern and cultural or contextual considerations were subsequently recoded into non-mutually exclusive descriptive themes for summary reporting. Rating-form domains, response options, and codebook anchors are presented in Supplemental Table S5. Criterion-level ratings informed the expert determination but were not analyzed as standalone outcomes because comparable clinician-side ratings were unavailable in the extracted analytic dataset.

Statistical Analysis

Analyses were conducted in R version 4.5.3. The complete analytic code is publicly available in an online repository. The study hypotheses, sampling plan, and analysis plan were preregistered on the Open Science Framework before data access and analysis. Deviations from the preregistered analysis plan and their rationale are summarized in Supplemental Table S6. Sample size was determined for the primary Cohen's kappa agreement analysis. As preregistered, the power analysis indicated that approximately 90 to 100 cases would provide 80% power to detect $\kappa \geq .66$, assuming an expected ASD prevalence of approximately 60%; the target sample was therefore set at $N = 100$. The preregistered interpretive benchmark of $\kappa \geq .66$ was specified because it falls above the conventional .61 threshold for substantial agreement (Landis & Koch, 1977) and was consistent with kappa values reported in prior autism assessment research (Lord et al., 2012; Rice et al., 2022; Zander et al., 2017).

Diagnostic agreement between assessing clinicians and the blinded expert reviewer was measured using Cohen's kappa (Cohen, 1960) with 95% confidence intervals for the dichotomous classification, with percent agreement reported concurrently. Because kappa is sensitive to prevalence and marginal asymmetry (Byrt et al., 1993; Feinstein & Cicchetti, 1990), prevalence and bias indices, prevalence-adjusted bias-adjusted kappa (PABAK), Gwet's AC1 (Gwet, 2008), symmetric positive and negative agreement, and Bangdiwala's B (Bangdiwala & Shankar, 2013) were computed as supplementary agreement metrics. Kappa, percent agreement, positive percent agreement, and negative percent agreement are emphasized in the main text; expanded metrics are presented in Supplemental Table S7.

Using the blinded expert review classification as the reference classification for directional agreement only, a 2 x 2 table was constructed and the following indices were computed: positive percent agreement (PPA), negative percent agreement (NPA), clinician-positive concordance with expert review, clinician-negative concordance with expert review, clinician-positive/expert-negative discordance, and clinician-negative/expert-positive discordance. Balanced agreement was calculated as $(PPA + NPA) / 2$, thereby giving equal weight to agreement among expert-positive and expert-negative classifications regardless of their relative prevalence. Confidence intervals for percent agreement, PPA, NPA, clinician-positive concordance, and clinician-negative concordance used the Wilson score method. Intervals for kappa used analytic asymptotic standard errors. Intervals

for discordance rates, balanced agreement, and supplementary metrics lacking closed-form variance estimators used nonparametric percentile bootstrap procedures with 10,000 resamples (Efron & Tibshirani, 1993).

Exploratory subgroup analyses computed agreement indices separately by information quality, age group, analytic gender group, and BIPOC versus non-BIPOC status. These analyses were preregistered as exploratory and were interpreted as hypothesis-generating because several subgroup cell sizes were small. For exploratory subgroup analyses, 95% confidence intervals for subgroup-specific kappa estimates were obtained using nonparametric percentile bootstrap procedures with 10,000 case-level resamples within each stratum. Bootstrap replicates in which kappa was undefined because of absent marginal variation were excluded from percentile estimation.

For the age 19-24 subgroup, perfect observed agreement yielded a degenerate distribution of defined bootstrap estimates; therefore, a kappa confidence interval was not reported. Race and ethnicity were self-reported at intake and analyzed descriptively. The aggregate BIPOC contrast was used only as a preliminary equity screen and should not be interpreted as evidence about any specific racial or ethnic group; disaggregated race and ethnicity agreement estimates were not interpreted because small cell sizes made subgroup estimates unstable. For descriptive reporting, gender was summarized as a mutually exclusive participant-centered category using the final coded gender group. For exploratory agreement analyses, we created a three-category analytic gender variable: cisgender woman, cisgender man, and gender-minoritized. The gender-minoritized category included participants who indicated they identified as transgender, nonbinary/gender expansive, multiple gender identities, or questioning/unsure. Participants who met criteria for the gender-minoritized indicator were assigned to the gender-minoritized analytic category, even if their mutually exclusive descriptive gender group was woman or man. Open-text expert-review fields, including diagnostic concerns and cultural or contextual considerations, were recoded into non-mutually exclusive descriptive categories for summary reporting.

Results

Sample Characteristics

Sample characteristics, expert information-quality ratings, and diagnostic outcome distributions are shown in Table 1. The analytic sample included 100 adults. Participants were distributed across age, gender, and race/ethnicity groups. Descriptively, 49 participants were coded as White, 29 as multiracial, 12 as Black, 6 as Hispanic/Latine, 2 as Asian, and 2 as Native American/American Indian. For aggregate exploratory analyses, 51 participants were categorized as BIPOC and 49 as non-BIPOC. Expert review classified 66 participants as ASD and 34 as non-ASD; original clinician determinations classified 68 participants as ASD and 32 as non-ASD. No case received a Poor information-quality rating.

Primary Outcome: Diagnostic Agreement

Of 100 cases, 86 received concordant classifications: 60 were classified as ASD by both the original evaluating clinician and the expert reviewer, and 26 were classified as non-ASD by both raters. Fourteen cases were discordant, comprising eight classified as ASD by the original clinician but non-ASD by the expert reviewer and six classified as non-ASD by the original clinician but ASD by the expert reviewer (Table 2).

Overall percent agreement was 86.0% (95% Wilson CI [77.9%, 91.5%]). Cohen's kappa was .684 (95% CI [.531, .836]). The point estimate exceeded the preregistered kappa agreement benchmark of .66, but the confidence interval included values below that threshold. Marginal distributions indicated a moderate prevalence effect (prevalence index = .34) and minimal rater bias (bias index = .02). Gwet's AC1 was .749. Main agreement metrics are shown in Table 3; expanded metrics, including symmetric positive and negative agreement, are presented in Supplemental Table S7.

Secondary Outcome: Directional Agreement With Blinded Expert Review

PPA with expert-positive classifications was 90.9% (95% Wilson CI [81.6%, 95.8%]). NPA with expert-negative classifications was 76.5% (95% Wilson CI [60.0%, 87.6%]). Clinician-positive concordance with expert review was 88.2% (95% Wilson CI [78.5%,

93.9%]), and clinician-negative concordance with expert review was 81.2% (95% Wilson CI [64.7%, 91.1%]). The clinician-positive/expert-negative discordance rate was 23.5% of expert-negative cases, and the clinician-negative/expert-positive discordance rate was 9.1% of expert-positive cases. These directional metrics describe agreement with blinded expert review and are not estimates of diagnostic accuracy.

Exploratory Analyses

All subgroup analyses reported below were preregistered as exploratory analyses. Within-stratum sample sizes were small for several comparisons, resulting in wide confidence intervals around some subgroup kappa estimates (Table 4), and no between-stratum bootstrap confidence interval for kappa differences excluded zero. Accordingly, these exploratory analyses are interpreted as descriptive, hypothesis-generating findings rather than evidence of subgroup differences.

Agreement was numerically higher across higher information-quality strata. Percent agreement and kappa were 94.9% and .804 in the Excellent stratum, 81.4% and .612 in the Good stratum, and 77.8% and .556 in the Fair stratum (Table 4). Because information quality and reference classification were assigned by the same reviewer, these estimates should be interpreted as descriptive associations, not independent evidence about documentation quality.

Discordant and concordant cases did not differ detectably by age or gender. BIPOC participants represented nine of 14 discordant cases (64.3%) and 42 of 86 concordant cases (48.8%), compared with 51.0% of the full sample. Agreement was 89.8% among non-BIPOC participants ($n = 49$; $\kappa = .787$) and 82.4% among BIPOC participants ($n = 51$; $\kappa = .547$), a raw difference of 7.4 percentage points and a kappa difference of $-.240$ with a bootstrap CI crossing zero (Table 5). The BIPOC subgroup kappa estimate fell below the preregistered applied benchmark used for the full sample, whereas the non-BIPOC estimate exceeded it; however, the between-group bootstrap confidence interval crossed zero. This aggregate contrast reflects the analytic approach described above and should not be interpreted as evidence about any specific racial or ethnic group. Expert-rated ASD prevalence was numerically higher among BIPOC participants (72.5%) than among non-BIPOC participants (59.2%), yielding a higher prevalence index in the BIPOC subgroup (.471 vs. .204); bias indices were equivalent across groups (.020 in both strata).

The 14 discordant cases were summarized from structured expert-review ratings and recoded expert open-text notes (Supplemental Table S8). All eight clinician-positive/expert-negative cases had at least one differential consideration documented in open text, most often anxiety-spectrum considerations or trauma. Among the six clinician-negative/expert-positive cases, four had no competing alternative diagnosis documented in open text. Across the remaining two cases, documented competing considerations included social communication disorder and social anxiety. Diagnostic considerations were documented in 43 of 86 concordant cases (50.0%) and 10 of 14 discordant cases (71.4%; Supplemental Table S9). These patterns are descriptive because the number of discordant cases was small.

Discussion

Clinical and Assessment Implications

This preregistered retrospective shared-information agreement study found 86.0% agreement and a Cohen's kappa of .684 between original clinician diagnoses from a structured, DSM-5-TR-mapped telehealth adult autism assessment protocol and blinded expert record review. The point estimate exceeded the preregistered agreement benchmark of .66, although the lower bound of the 95% CI (.531) did not. Directional agreement was stronger when the expert classified autism as present than when the expert classified autism as absent (see Directional Agreement and Differential Diagnosis below). These findings provide preliminary evidence that the pathway can yield reproducible classifications under shared-information review conditions.

The kappa estimate should be interpreted within a literature in which agreement statistics arise from materially different populations and designs. For context, the DSM-5 field trial reported a test-retest kappa of .69 for autism spectrum disorder (Regier et al., 2013), and ADOS-2 Module 4 studies have reported inter-rater reliability estimates ranging from approximately .57

to .82 in adult clinical samples (Christiansen & Pedersen, 2025; Kamp-Becker et al., 2018; Lord et al., 2012). These values are not psychometric benchmarks for the present study as the field trial involved repeated clinical evaluations, ADOS-2 estimates arise from standardized observational scoring, and the present study involved retrospective expert review of substantially shared assessment information. As noted above, kappa is also sensitive to classification prevalence and marginal distributions, further limiting direct numerical comparison across studies.

The 95% CI width (.531 to .836) reflects the precision achievable at N = 100 with a 66% expert-rated ASD base rate. The lower bound falls below the conventional lower boundary of substantial agreement (.61; Landis & Koch, 1977), meaning the data remain compatible with moderate agreement; the interval should therefore be considered alongside the point estimate.

For clinicians, these findings suggest that neurodiversity-affirming, telehealth-based adult autism evaluations can support reproducible clinical classification under trained-use conditions when conducted by clinicians with adult autism assessment training and embedded within structured documentation and quality-review procedures. The results should not be generalized to abbreviated video visits, unstructured consultations, record-only determinations, or evaluations conducted without specialized training. For institutions, payers, and disability offices, the findings underscore the potential relevance of documentation standards that emphasize assessment comprehensiveness, criterion-level evidence, differential diagnostic reasoning, and functional-impact documentation rather than treating any single proprietary instrument as equivalent to the full diagnostic process.

Directional Agreement and Differential Diagnosis

Directional agreement was asymmetric. PPA with expert-positive classifications was 90.9%, whereas NPA with expert-negative classifications was 76.5%. The clinician-positive/expert-negative discordance rate exceeded the clinician-negative/expert-positive discordance rate. This pattern indicates stronger concordance for expert-positive than expert-negative classifications. Because expert review was not an independent diagnostic gold standard, the asymmetry should not be interpreted as sensitivity, specificity, or evidence that either rater was correct. The discordant-case summaries provide more specific descriptive context for where disagreement occurred. Discordant-case review showed that competing diagnostic considerations were documented far more often in one direction of disagreement than the other (see Results), but the small number of discordant cases requires caution. This pattern supports a descriptive association between disagreement and expert-noted differential considerations in some cases but does not establish that co-occurring conditions caused disagreement.

Information Quality and Documentation Standards

Agreement was numerically highest among cases rated Excellent and lowest among cases rated Fair, but these estimates are descriptive because the same expert reviewer assigned both information-quality ratings and reference classifications. The pattern demonstrates the value of auditable records that document criterion-level examples, developmental evidence, functional impact, masking and compensation, and differential diagnostic reasoning. Such documentation makes diagnostic decisions more visible for external review, calibration, and quality assurance.

Measurement Equity

Agreement was lower among BIPOC participants than non-BIPOC participants, and BIPOC participants were mildly overrepresented among discordant cases. Because the study intentionally stratified sampling to achieve at least 50% BIPOC representation, this equity screen was a planned design feature. However, the aggregate BIPOC category combined participants across heterogeneous identities, and cell sizes precluded stable disaggregated estimates. These findings should therefore be interpreted as a measurement-equity signal, not as evidence about any specific racial or ethnic group.

The expert-rated ASD prevalence was higher among BIPOC participants than non-BIPOC participants, producing a larger prevalence index. This pattern is consistent with referral-based samples in the broader literature. Although ASD was historically underidentified among racially and ethnically minoritized groups, recent CDC surveillance data show that identified prevalence

among Black, Hispanic, and Asian/Pacific Islander children now meets or exceeds that of White children (Maenner et al., 2023; Shaw et al., 2025). Clinical samples also suggest that BIPOC individuals who reach autism evaluation may have higher or more readily recognized support needs, potentially reflecting a higher threshold for referral, recognition, or access to diagnostic assessment for people of color (Martin et al., 2024). Because kappa is sensitive to prevalence asymmetry, the lower BIPOC kappa may partly reflect this statistical property. At the same time, documentation quality, cultural presentation differences, differential masking patterns, or rater interpretive frameworks may also have contributed. Autism identification has a history of racial and ethnic disparities (Goldblum et al., 2024; Mandell et al., 2009), and whether adult diagnostic systems show equitable agreement across groups remains an open empirical question.

Limitations

Reference classification and reviewer design. The study used a shared-information blinded-review design rather than an independent reference-standard evaluation. The reference classification was based on de-identified materials produced during the original assessment, so shared information may have increased agreement relative to fully independent evaluations. Clinical notes and transcripts may also have contained implicit diagnostic framing even after direct diagnostic conclusions were removed. Agreement indices should therefore be interpreted as agreement with blinded expert review, not diagnostic accuracy or criterion validity against an independent reference standard. A single expert reviewer also precluded evaluation of agreement across multiple independent expert reviewers, and the observed agreement reflects the relation between the clinician pool and one reviewer. The expert reviewer is an author on this manuscript. Although the reviewer's role was limited as described above, shared professional context with the research team could have influenced interpretation of ambiguous documentation, masking evidence, differential-diagnosis boundaries, or threshold judgments.

Protocol development. The interview guide was developed through a consultative process intended to improve clarity, clinical relevance, acceptability, and alignment with adult autistic presentations. Feedback from autistic adults and autism-specialist clinicians informed revisions to wording, structured probes, developmental-history procedures, and clinical framing; however, this process was intended for iterative protocol refinement and was not designed as a formal content-validation study. The development process did not use prespecified adjudication procedures, quantitative content-validity indices, or other psychometric methods for establishing the relevance or representativeness of individual interview domains or items. Accordingly, the present study evaluates diagnostic agreement for the implemented assessment pathway and should not be interpreted as establishing the formal content validity of the interview guide or its individual components.

Setting, clinicians, and implementation conditions. The study was conducted in a single specialty telehealth clinic. Because eligibility required completion of the full pathway and documentation meeting clinic quality standards, the findings reflect agreement under trained, quality-controlled implementation conditions and may not generalize to settings with less specialized training, less structured documentation, or fewer quality-assurance procedures. The 100 cases were drawn from 52 unique clinicians, with a median of one case per clinician and a range of one to six cases. The relatively diffuse distribution of cases across clinicians reduces the likelihood that the overall agreement estimate was driven by a single high-volume clinician. However, some clinicians contributed more than one case, so observations were not fully independent. Clinician-level clustering could not be modeled reliably because most clinicians contributed very few cases.

Population scope. The study applies only to adults whose communication profile fit the spoken-language telehealth interview format and who completed the full assessment pathway. No standardized language or verbal-fluency assessment was administered; communication eligibility was determined clinically during orientation and intake. The pathway-eligibility description is therefore operational and reflects the clinician's determination that a client's communication and cognitive profile was appropriate for an extended spoken-language telehealth interview. It does not represent a standardized language, verbal-fluency, intellectual-functioning, or cognitive classification.

Eligibility criteria restrict generalizability as the present sample does not represent adult autism evaluation for individuals with intellectual disability, nonspeaking autistic adults, or adults who require substantial augmentative or alternative communication support. In clinical practice, the pathway has occasionally been used with clients who are minimally speaking, primarily

communicate through chat, attend with a parent or caregiver involved in the evaluation process, or are ultimately diagnosed with Level 2 autism. These presentations are uncommon, however, and none were represented in the present random sample.

Routine-care data constraints. The retrospective routine-care data structure limited available variables. The analytic dataset lacked consistently extracted indicators of collateral availability, questionnaire scores, questionnaire availability, clinician-side criterion-level ratings, clinician-side masking or camouflaging documentation, education, employment, prior psychiatric diagnoses, medication use, and parallel clinician-side differential-diagnosis fields. These omissions precluded analyses of measure-level performance, diagnostic complexity, presentation modifiers, collateral effects, and criterion-level agreement. This also meant the study could not determine whether cases with standardized measure data differed systematically in agreement or expert-rated information quality from cases without those data. This limitation reflects the constraints of secondary analysis of routine-care clinical records, where clinically available materials vary across cases and are not always captured as analyzable structured fields. As a result, the psychometric contribution is limited to dichotomous classification agreement and does not address item-level, domain-level, or criterion-level reproducibility. The disagreement discussion around anxiety and differential diagnosis is therefore based on expert open-text notes and expert-review fields, not parallel structured clinician-side differential-diagnosis variables.

Timing and base-rate considerations. Exact case-level elapsed time between the original assessment and expert review could not be calculated because review-completion dates were not retained. The study therefore could not evaluate whether elapsed time, documentation timing, or review-period effects influenced agreement. In addition, the expert-rated ASD base rate was 66%, which is plausible for a specialty adult autism clinic but may not generalize to lower-prevalence settings. Agreement indices and the practical meaning of positive and negative classifications may differ in settings with lower autism base rates.

Future Assessment Directions

Future work should address the design features that remain unresolved in this study. The most pressing need is for prospective studies that compare the telehealth pathway with fully independent adult autism evaluations conducted by clinicians who do not share source assessment materials with the original evaluator; independent parallel-assessment designs, such as the pediatric model used by Penner et al. (2023), illustrate the stronger separation of information sources needed for criterion-validity research, although adult-specific designs are required. Such studies should also use multiple blinded reviewers, standardized decision rubrics, matched or explicitly documented information access, and consensus procedures to reduce implicit diagnostic framing and quantify reviewer-to-reviewer agreement. Communication profiles should be characterized with standardized language or communication measures, and research should intentionally include nonspeaking adults and adults who use augmentative or alternative communication supports. Criterion-level agreement and documentation quality should likewise be examined alongside dichotomous classification, including whether structured decision rules reduce threshold variability. Future research should also formalize and evaluate standardized training and fidelity procedures, including calibration exercises and prospective monitoring of adherence to criterion-level decision rules. Measurement-equity analyses should be powered for disaggregated racial and ethnic groups and should evaluate documentation quality, criterion-level functioning, and rater-client demographic effects. Downstream outcomes such as accommodation access, care planning, therapy engagement, self-understanding, and client-reported impact also warrant evaluation. Community-engaged designs involving autistic adults, including BIPOC autistic adults and adults with diverse communication profiles, would strengthen protocol development, rater training, outcome selection, and interpretation of discordance.

Conclusion

In this preregistered study, a structured, neurodiversity-affirming, DSM-5-TR-mapped telehealth pathway yielded agreement between original clinician classifications and blinded expert review of shared assessment information, with a kappa point estimate in the substantial range. The findings support an adult-focused approach that integrates developmental history, masking and compensatory strategies, lived experience, functional impact, differential diagnosis, and direct behavioral observation within a transparent criterion-level decision framework. Although the shared-information design does not establish diagnostic accuracy or criterion validity, the results provide preliminary evidence that comprehensive adult autism assessment can be delivered through telehealth with reproducible diagnostic classification under trained, quality-controlled conditions.

Independent prospective studies with separate evaluations, multiple blinded reviewers, and broader representation across communication and support needs are needed to establish validity and generalizability.

Declarations

Participants provided informed consent for their de-identified clinical data to be used for research purposes, including secondary analyses. The study was reviewed by the University of Massachusetts Chan Medical School Institutional Review Board and was determined to be exempt.

References

1. Alfuraydan M, Croxall J, Hurt L, Kerr M, Brophy S (2020) Use of telehealth for facilitating the diagnostic assessment of Autism Spectrum Disorder (ASD): A scoping review. *PLoS ONE*, 15(7), e0236415
2. American Psychiatric Association (2022) Diagnostic and statistical manual of mental disorders (5th ed., text rev.). American Psychiatric Association. <https://doi.org/10.1176/appi.books.9780890425787>
3. Atherton G, Edisbury E, Piovesan A, Cross L (2022) Autism through the ages: A mixed methods approach to understanding how age and age of diagnosis affect quality of life. *J Autism Dev Disord* 52(8):3639–3654
4. Bangdiwala SI, Shankar V (2013) The agreement chart. *BMC Med Res Methodol* 13., Article 97. <https://doi.org/10.1186/1471-2288-13-97>
5. Belcher HL, Morein-Zamir S, Stagg SD, Ford RM (2023) Shining a light on a hidden population: Social functioning and mental health in women reporting autistic traits but lacking diagnosis. *J Autism Dev Disord* 53(8):3118–3132
6. Berney T, Brugha T, Carpenter P (2017) Interview guide for the diagnostic assessment of able adults with autism spectrum disorder (ASD) (Rev. ed.). Royal College of Psychiatrists
7. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF (2015) STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 351 Article h5527. <https://doi.org/10.1136/bmj.h5527>
8. Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. *J Clin Epidemiol* 46(5):423–429. [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
9. Christiansen J, Pedersen L (2025) ADOS-2 Module 4: Psychometric Properties and Diagnostic Performance at an autism-specialized clinic. *J Autism Dev Disord* 55(11):3900–3913
10. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46. <https://doi.org/10.1177/001316446002000104>
11. Coleman R, Weitlauf A, Hartnett H, Castillo-Martinez G, Stone C, Swanson A, Foster T, Hundley R, Kehl L, Vehorn A, Warren Z (2026) Evaluating the acceptability of an early adoption of the Tele-ASD Diagnostic Assessment Tool for Adults (TADA). *Journal of Autism and Developmental Disorders*. <https://doi.org/10.1007/s10803-026-07392-3>. Advance online publication
12. Conner CM, Cramer RD, McGonigle JJ (2019) Examining the diagnostic validity of autism measures among adults in an outpatient clinic sample. *Autism Adulthood* 1(1):60–68. <https://doi.org/10.1089/aut.2018.0023>
13. Constantino JN, Gruber CP (2012) Social Responsiveness Scale, Second Edition (SRS-2). Western Psychological Services
14. Corona LL, Wagner L, Hooper M, Weitlauf A, Foster TE, Hine J, Warren Z (2024) A randomized trial of the accuracy of novel telehealth instruments for the assessment of autism in toddlers. *J Autism Dev Disord* 54(6):2069–2080
15. Doherty M, Shaw SC, Dakin C, Milton D, Morton H (2026) The ADOS-2 rapport item: an unexamined assumption. *Lancet Psychiatry* 13(7):545–546
16. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall/CRC
17. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 43(6):543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)

18. Gangi DN, Corona L, Wagner L, Weitlauf A, Warren Z, Ozonoff S (2025) In-home tele-assessment for autism in toddlers: Validity, reliability, and caregiver satisfaction with the TELE-ASD-PEDS. *J Dev Behav Pediatr* 46(3):e261–e268. <https://doi.org/10.1097/DBP.0000000000001358>
19. Gibbs V, Cai RY, Aldridge F, Wong M (2021) Autism assessment via telehealth during the COVID-19 pandemic: Experiences and perspectives of autistic adults, parents/carers and clinicians. *Res Autism Spectr Disorders* 88., Article 101859. <https://doi.org/10.1016/j.rasd.2021.101859>
20. Goldblum J, McFayden T, Bristol S, Putnam O, Wylie A, Harrop C (2024) Autism prevalence and the intersectionality of assigned sex at birth, race, and ethnicity on age of diagnosis. *J Autism Dev Disord* 54(10):3777–3791. <https://doi.org/10.1007/s10803-023-06104-5>
21. Goodwin J, Wigham S, Ingham B, Herrington R, Le Couteur A, Parr JR (2026) A survey of professionals' experiences of using the Autism Clinical Interview for Adults (ACIA) in diagnostic practice. *Res Autism* 132:202804
22. Grosvenor LP, Croen LA, Lynch FL, Marafino BJ, Maye M, Penfold RB, Simon GE, Ames JL (2024) Autism diagnosis among US children and adults, 2011–2022. *JAMA Netw Open* 7(10) Article e2442218. <https://doi.org/10.1001/jamanetworkopen.2024.42218>
23. Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 61(1):29–48. <https://doi.org/10.1348/000711006X126600>
24. Harrison PL, Oakland T (2015) *Adaptive Behavior Assessment System, Third Edition (ABAS-3)*. Western Psychological Services
25. Huang Y, Arnold SRC, Foley K-R, Trollor JN (2020) Diagnosis of autism in adulthood: A scoping review. *Autism* 24(6):1311–1327. <https://doi.org/10.1177/1362361320903128>
26. Hull L, Petrides KV, Allison C, Smith P, Baron-Cohen S, Lai M-C, Mandy W (2017) Putting on my best normal: Social camouflaging in adults with autism spectrum conditions. *J Autism Dev Disord* 47(8):2519–2534. <https://doi.org/10.1007/s10803-017-3166-5>
27. Kamp-Becker I, Albertowski K, Becker J, Ghahreman M, Langmann A, Mingebach T, Poustka L, Weber L, Schmidt H, Smidt J, Stehr T, Roessner V, Kucharczyk K, Wolff N, Stroth S (2018) Diagnostic accuracy of the ADOS and ADOS-2 in clinical practice. *Eur Child Adolesc Psychiatry* 27(9):1193–1207. <https://doi.org/10.1007/s00787-018-1143-y>
28. Lai M-C, Lombardo MV, Ruigrok ANV, Chakrabarti B, Auyeung B, Szatmari P, Happe F, Baron-Cohen S (2017) Quantifying and exploring camouflaging in men and women with autism. *Autism* 21(6):690–702. <https://doi.org/10.1177/1362361316671012>
29. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174. <https://doi.org/10.2307/2529310>
30. 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. *J Autism Dev Disord* 24(5):659–685
31. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop SL (2012) *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. Western Psychological Services
32. Maddox BB, Brodtkin ES, Calkins ME, Shea K, Mullan K, Hostager J, Mandell DS, Miller JS (2017) The accuracy of the ADOS-2 in identifying autism among adults with complex psychiatric conditions. *J Autism Dev Disord* 47(9):2703–2709. <https://doi.org/10.1007/s10803-017-3188-z>
33. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, Durkin MS, Fitzgerald RT, Furnier SM, Hughes MM, Ladd-Acosta CM, McArthur D, Pas ET, Salinas A, Vehorn A, Williams S, Esler A, Grzybowski A, Hall-Lande J, 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(2):1–14. <https://doi.org/10.15585/mmwr.ss7202a1>
34. Mandell DS, Wiggins LD, Carpenter LA, Daniels J, DiGuseppi C, Durkin MS, Giarelli E, Morrier MJ, Nicholas JS, Pinto-Martin JA, Shattuck PT, Thomas KC, Yeargin-Allsopp M, Kirby RS (2009) Racial/ethnic disparities in the identification of children with autism spectrum disorders. *Am J Public Health* 99(3):493–498. <https://doi.org/10.2105/AJPH.2007.131243>

35. Mandy W, Clarke K, McKenney M, Strydom A, Crabtree J, Lai M-C, Skuse D (2018) Assessing autism in adults: An evaluation of the Developmental, Dimensional and Diagnostic Interview-Adult Version (3Di-Adult). *J Autism Dev Disord* 48(2):549–560. <https://doi.org/10.1007/s10803-017-3321-z>
36. Martin AM, Keehn B, Paxton A, Ciccarelli MR, Keehn RM (2024) Associations among race, ethnicity, and clinical profiles of young children evaluated for autism in the primary care setting. *J Dev Behav pediatrics: JDBP*, 45(5), e414
37. Matthews NL, Skepnek E, Mammen MA, James JS, Malligo A, Lyon A, Smith CJ (2021) Feasibility and acceptability of a telehealth model for autism diagnostic evaluations in children, adolescents, and adults. *Autism Res* 14(12):2564–2579. <https://doi.org/10.1002/aur.2591>
38. McQuaid GA, Lee NR, Wallace GL (2022) Camouflaging in autism spectrum disorder: Examining the roles of sex, gender identity, and diagnostic timing. *Autism* 26(2):552–559. <https://doi.org/10.1177/13623613211042131>
39. Monteiro MJ (2018) Monteiro Interview Guidelines for Diagnosing the Autism Spectrum, Second Edition (MIGDAS-2). Western Psychological Services
40. Nayyar JM, Stapleton AV, Guerin S, O'Connor C (2025) Exploring lived experiences of receiving a diagnosis of autism in adulthood: A systematic review. *Autism Adulthood* 7(1):1–12. <https://doi.org/10.1089/aut.2023.0152>
41. Penner M, Senman L, Andoni L, Dupuis A, Anagnostou E, Kao S, Brian J (2023) Concordance of diagnosis of autism spectrum disorder made by pediatricians vs a multidisciplinary specialist team. *JAMA Netw Open*, 6(1), e2252879
42. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ (2013) DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry* 170(1):59–70
43. Rice CE, Carpenter LA, Morrier MJ, Lord C, DiRienzo M, Boan A, Wiggins LD (2022) Defining in detail and evaluating reliability of DSM-5 criteria for autism spectrum disorder (ASD) among children. *J Autism Dev Disord* 52(12):5308–5320
44. Rutter M, Bailey A, Lord C (2003) Social Communication Questionnaire (SCQ). Western Psychological Services
45. Shaw KA, Williams S, Patrick ME et al Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years – Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022. *MMWR Surveill Summ* 2025;74(No. SS-2):1–22. <http://dx.doi.org/10.15585/mmwr.ss7402a1>
46. Stavropoulos KKM, Bolourian Y, Blacher J (2022) A scoping review of telehealth diagnosis of autism spectrum disorder. *PLoS ONE* 17(2) Article e0263062. <https://doi.org/10.1371/journal.pone.0263062>
47. Sutherland R, Trembath D, Roberts J (2018) Telehealth and autism: A systematic search and review of the literature. *Int J Speech Lang Pathol* 20(3):324–336. <https://doi.org/10.1080/17549507.2018.1465123>
48. Wigham S, Ingham B, Le Couteur A, Berney T, Ensum I, Parr JR (2020) Development and initial utility of the Autism Clinical Interview for Adults: A new adult autism diagnostic measure. *Autism Adulthood* 2(1):42–47. <https://doi.org/10.1089/aut.2019.0052>
49. Wigham S, Rodgers J, Berney T, Le Couteur A, Ingham B, Parr JR (2019) Psychometric properties of questionnaires and diagnostic measures for autism spectrum disorders in adults: A systematic review. *Autism* 23(2):287–305. <https://doi.org/10.1177/1362361317748245>
50. Wing L, Leekam SR, Libby SJ, Gould J, Larcombe M (2002) The Diagnostic Interview for Social and Communication Disorders: Background, inter-rater reliability and clinical use. *J Child Psychol Psychiatry* 43(3):307–325. <https://doi.org/10.1111/1469-7610.00023>
51. Zander E, Willfors C, Berggren S, Coco C, Holm A, Jifält I, Bölte S (2017) The interrater reliability of the autism diagnostic interview-revised (ADI-R) in clinical settings. *Psychopathology* 50(3):219–227

Tables

Table 1
Sample Characteristics, Clinical Features, and Diagnostic Outcome Distributions

Characteristic	n (%) or value
Total analytic sample	100
Age 19–24	12 (12.0%)
Age 25–34	44 (44.0%)
Age 35–44	26 (26.0%)
Age 45+	18 (18.0%)
Woman	59 (59.0%)
Man	21 (21.0%)
Multiple gender identities	8 (8.0%)
Nonbinary/gender expansive	6 (6.0%)
Transgender man	3 (3.0%)
Transgender woman	3 (3.0%)
White	49 (49.0%)
Multiracial	29 (29.0%)
Black	12 (12.0%)
Hispanic/Latine	6 (6.0%)
Asian	2 (2.0%)
Native American/American Indian	2 (2.0%)
BIPOC analytic group	51 (51.0%)
Non-BIPOC analytic group	49 (49.0%)
Expert information quality: Excellent	39 (39.0%)
Expert information quality: Good	43 (43.0%)
Expert information quality: Fair	18 (18.0%)
Expert ASD	66 (66.0%)
Expert non-ASD	34 (34.0%)
Clinician ASD	68 (68.0%)
Clinician non-ASD	32 (32.0%)
Diagnostic agreement	86 (86.0%)
Diagnostic disagreement	14 (14.0%)
Note. ASD = autism spectrum disorder; BIPOC = Black, Indigenous, and people of color. Race/ethnicity categories are participant-centered descriptive categories. The aggregate BIPOC contrast was used only as a preliminary equity screen and should not be interpreted as evidence about any specific racial or ethnic group. Gender categories are mutually exclusive descriptive categories; no separate questioning/unsure category is shown. Participants who indicated a questioning/unsure identity are represented within their final coded descriptive gender category and were counted in the gender-minoritized analytic category (Table 4).	

Table 2
Two-by-Two Clinician Diagnosis by Blinded Expert Review Agreement Table

Clinician ASD	Expert ASD	Expert non-ASD	Total
	60	8	68
Clinician non-ASD	6	26	32
Total	66	34	100
Note. ASD = autism spectrum disorder. Clinician diagnosis was compared with blinded expert review as the reference classification for directional agreement only; the expert review was not an independent diagnostic gold standard.			

Table 3
Primary Diagnostic Agreement and Directional Agreement Metrics

Metric	Estimate	95% CI	CI Method
Percent agreement	86.0%	77.9%, 91.5%	Wilson
Cohen's kappa	.684	.531, .836	Analytic/asymptotic
Positive percent agreement (PPA)	90.9%	81.6%, 95.8%	Wilson
Negative percent agreement (NPA)	76.5%	60.0%, 87.6%	Wilson
Clinician-positive concordance with expert review	88.2%	78.5%, 93.9%	Wilson
Clinician-negative concordance with expert review	81.2%	64.7%, 91.1%	Wilson
Clinician-positive/expert-negative discordance rate	23.5%	10.0%, 38.2%	Bootstrap
Clinician-negative/expert-positive discordance rate	9.1%	2.9%, 16.7%	Bootstrap
Balanced agreement	83.7%	75.5%, 91.2%	Bootstrap
Note. CI = confidence interval. Cohen's kappa is interpreted as an agreement statistic. Directional agreement metrics use blinded expert review as the reference classification for descriptive agreement only and are not interpreted as criterion validity or diagnostic accuracy against an independent reference standard.			

Table 4
Exploratory Subgroup Agreement Estimates for Information Quality, Age, and Gender

Domain	Subgroup	n	Percent Agreement	Cohen's Kappa	95% CI
Information quality	Excellent	39	94.9%	.804	.458, 1.000
Information quality	Good	43	81.4%	.612	.345, .844
Information quality	Fair	18	77.8%	.556	.153, .889
Age	19–24	12	100.0%	1.000	Not estimable
Age	25–34	44	86.4%	.676	.415, .891
Age	35–44	26	76.9%	.530	.164, .842
Age	45+	18	88.9%	.766	.400, 1.000
Gender	Woman	59	84.7%	.671	.458, .855
Gender	Man	20	85.0%	.694	.306, 1.000
Gender	Gender-minoritized	21	90.5%	.611	– .068, 1.000

Note. CI = confidence interval. Kappa confidence intervals were estimated using nonparametric percentile bootstrap procedures with 10,000 case-level resamples within each subgroup. The confidence interval for the age 19–24 subgroup was not reported because perfect observed agreement produced a degenerate bootstrap distribution and kappa was undefined in a subset of bootstrap resamples. Subgroup analyses were exploratory and should be interpreted descriptively because several strata were small and confidence intervals were correspondingly wide. Table 1 reports mutually exclusive descriptive gender categories. Table 4 uses a derived three-category analytic gender variable. The gender-minoritized analytic category included participants coded as transgender, nonbinary/gender expansive, multiple gender identities, or questioning/unsure. One participant categorized as Man in Table 1 also met criteria for the gender-minoritized indicator and was reassigned to the gender-minoritized analytic subgroup, yielding n = 20 for the Man analytic category and n = 21 for the gender-minoritized analytic category.

Table 5
BIPOC Versus Non-BIPOC Exploratory Agreement Estimates

Subgroup	n	Expert ASD, n (%)	Percent Agreement	Cohen's Kappa	95% CI	Prevalence Index	Bias Index	Gwet's AC1	Positive Agreement	Negative Agreement
BIPOC	51	37 (72.5%)	82.4%	.547	.338, .730	.471	.020	.711	.880	.667
Non-BIPOC	49	29 (59.2%)	89.8%	.787	.626, .914	.204	.020	.804	.915	.872

Note. ASD = autism spectrum disorder; BIPOC = Black, Indigenous, and people of color; CI = confidence interval. BIPOC included all participants not identifying as exclusively White. The aggregate BIPOC contrast was used only as a preliminary equity screen and should not be interpreted as evidence about any specific racial or ethnic group. Disaggregated race and ethnicity agreement estimates are not reported because cell sizes were too small for stable interpretation. Delta kappa for BIPOC versus non-BIPOC was – .240, and the bootstrap CI crossed zero (–.572, .068).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DiagAgreePREPRINTSupplement.docx](#)