Phenotypic differences between female and male individuals with suspicion of autism spectrum disorder

Sanna Stroth1*, Johannes Tauscher2, Nicole Wolff3, Charlotte Küpper4, Luise Poustka5, Stefan Roepke4, Veit Roessner3, Dominik Heider2 and Inge Kamp-Becker1

Abstract

Background: Although autism spectrum disorder (ASD) is a common developmental disorder, our knowledge about a behavioral and neurobiological female phenotype is still scarce. As the conceptualization and understanding of ASD are mainly based on the investigation of male individuals, females with ASD may not be adequately identified by routine clinical diagnostics. The present machine learning approach aimed to identify diagnostic information from the Autism Diagnostic Observation Schedule (ADOS) that discriminates best between ASD and non-ASD in females and males.

Methods: Random forests (RF) were used to discover patterns of symptoms in diagnostic data from the ADOS (modules 3 and 4) in 1057 participants with ASD (18.1% female) and 1230 participants with non-ASD (17.9% % female). Predictive performances of reduced feature models were explored and compared between females and males without intellectual disabilities.

Results: Reduced feature models relied on considerably fewer features from the ADOS in females compared to males, while still yielding similar classification performance (e.g., sensitivity, specificity).

Limitations: As in previous studies, the current sample of females with ASD is smaller than the male sample and thus, females may still be underrepresented, limiting the statistical power to detect small to moderate effects.

Conclusion: Our results do not suggest the need for new or altered diagnostic algorithms for females with ASD. Although we identified some phenotypic differences between females and males, the existing diagnostic tools seem to sufficiently capture the core autistic features in both groups.

Keywords: Female autism, Sex, ASD, ADOS, ADI-R, Diagnostics, Phenotype

Background

Autism spectrum disorder (ASD) is a common developmental disorder with an onset within the first years of life and early emerging atypicalities in social attention and reciprocity [1]. Since the early days of autism research, the condition has been understood to predominantly affect males. Most epidemiological studies report an approximately 4:1 male to female ratio [2, 3], which has recently shifted toward a 3:1 ratio [4]. It has consistently been shown that the sex imbalance in prevalence varies with cognitive ability, with a male to female ratio of 2:1 among individuals with low cognitive ability or co-occurring intellectual disability and a ratio as high as 9:1 among individuals with average to above-average IQ [2, 5]. Consequently, most research has involved males,
leading to a male-biased understanding and conceptualization of ASD. Furthermore, although the female ASD phenotype may present differently, current defining diagnostic criteria are mainly based on male characteristics, and diagnostic instruments may be biased toward detecting ASD among male individuals, with similar diagnostic thresholds for females and males [6, 7].

Previous work suggests that ASD may be more difficult to detect in females, as they tend to be diagnosed later than males [8, 9] and seem to require a more significant etiological load to manifest autistic behavioral characteristics and autistic symptoms, or concurrent impairments need to be more severe for the diagnosis to be given [10]. It is argued that phenotypic sex differences might lead to delayed or even missed diagnoses in girls and women with ASD [11].

In the past decades, a wealth of investigations have been conducted to examine the relationship between sex and clinical profiles of individuals with ASD. Research findings on differences between the sexes provide some insight into why females might be more difficult to detect and are diagnosed later in life than males. However, knowledge about differences between the sexes in terms of the phenotypic presentation of ASD symptoms is still lacking, as the available studies yielded inconsistent findings regarding symptom severity across different age groups and different levels of functioning. While some studies did not find sex differences during a behavioral observation, e.g., [12, 13], others did report some differences [14].

It has been argued that females and males might meet the diagnostic criteria for ASD differently, as a range of different behaviors can be mapped onto each broad criterion. For example, deficits in social-emotional reciprocity may be composed of impairments in spoken language, reduced joint attention, and reduced sharing of interest, emotions, and affect. To meet the social-emotional reciprocity criterion, an individual does not need to present with all of these behaviors—rather, the clinician needs to decide whether or not an individual meets a particular criterion based on the available information. Despite some work on a female autism phenotype, little is known about how females and males meet the diagnostic criteria.

Females with ASD appear to score lower than males on measures of restricted and repetitive behavior (RRB), they seem less likely to present with stereotyped use of objects and show different types of restricted interests than males [15]. Specific differences in social communication deficits have not been consistently observed. Some girls were more likely to show an ability to integrate non-verbal and verbal behaviors, maintain a reciprocal conversation, and be able to initiate, but not maintain friendships [15] others showed more impairment in communication [16] compared to boys. Overall, results remain inconsistent (for reviews, see [5, 17–19]. If there are indeed different symptom patterns in females and males, but diagnostic instruments are biased toward the male ASD phenotype, one solution to better recognize ASD in females would be to revise the diagnostic criteria and the diagnostic algorithms of standard diagnostic instruments.

The current study aimed to investigate an ASD specific behavioral observation tool and explore whether there are differences in how female and male individuals meet the actual criteria for ASD. As females with ASD without cognitive and language deficits are at risk of not being identified until later in life, we investigated a sample of individuals with fluent language and without profound intellectual disabilities. Diagnoses at an older age have been associated with increased comorbidity [20]. Moreover, the presentation of ASD symptoms can significantly overlap with other mental disorders [21, 22]. Therefore, it is essential to investigate sex differences in ASD symptom presentation not only in a sample of individuals with ASD but also in those with suspicion of ASD but no actual ASD diagnosis. The present study thus aimed to extend previous research on a female ASD phenotype, which focused on differences in ASD symptoms between females and males already diagnosed with ASD, by including a large clinical sample comprising individuals without a diagnosis of ASD but with a diagnosis of other mental disorders. We aimed to identify those symptoms, directly observed by trained specialists, which optimally discriminate between ASD and non-ASD within a female and a male sample, and to then compare these discriminative features between the sexes. We thus aim to facilitate the diagnostic identification of females by highlighting potential nuanced differences between the sexes. By using machine learning models, we sought to identify the particular contributions of individual pieces of diagnostic information (item codes from the ADOS) for the diagnosis of female and male children and young adolescents and for later diagnosis of older adolescents and adults.

**Methods**

**Participants**

Datasets were drawn from medical records (retrospective chart review of the period between 2000 and 2019) from five specialized autism centers that were part of the ASD-Net, a research consortium funded by the German Federal Ministry of Education and Research (BMBF) [22] and coordinated by the authors. Experienced clinicians with continuous ADOS coding experience and research
reliable ADOS experts for supervision at each site applied the current diagnostic gold standard.

All data were analyzed anonymously, with approval from the local ethics committee (Az. 92/20). Due to the retrospective nature of data collection and analysis based on anonymized data, the need for informed consent was waived by the ethics committee. All methods were applied following relevant institutional and international research guidelines and regulations. Sex is defined as biological sex as assessed by caregivers or the participants themselves. If sex was unknown or not reported, data were excluded (N = 4).

The dataset included 2287 individuals who underwent a complete clinical examination after an initial suspicion of ASD. 46.2% received a diagnosis of ASD (n = 866 male; n = 191 female). The remaining participants (53.8%) did not receive a diagnosis of ASD, and were diagnosed with other mental or developmental disorders or no disorders (n = 1010 male; n = 220 female). The non-ASD group represents a well-balanced data set comprising differential disorders with some traits or symptoms of ASD (leading to the suspicion of ASD). The ratio of males to females with ASD was 4.5:1. The sample was separated into subsamples, as participants were administered different measures (ADOS modules) depending on age and language ability. The subsamples are henceforth labeled “children and young adolescents” (examined with ADOS module 3) and “older adolescents and adults” (examined with ADOS module 4).

Child and young adolescent sample

Of the children and young adolescents with ASD, 51% had further co-occurring diagnoses (most commonly attention deficit hyperactivity disorder (ADHD, F90 according to ICD-10) and developmental disorders (F80 and F82). The most common non-ASD diagnoses were ADHD (23%) and conduct disorders (11%). 79% had further co-occurring diagnoses. In contrast to this high multi-comorbidity, 22% of non-ASD cases had no clinical mental disorder but did have some autistic traits that had led to the suspicion of ASD. Further details can be found in Additional file 1: Tables S1 and S2.

On average, the non-ASD group was younger than the ASD group (T(152) = 2.24, p = 0.027 for females, T(1196) = 2.77, p = 0.006) for males), with small to moderate effect sizes (Cohen’s d = 0.38 and 0.16) (see Table 1). Comparing only those female and male individuals with an ASD diagnosis (see Table 2), females were older than males (T(545) = 2.8, p = 0.005; mean = 11.4, median = 11, SD = 3.2 in females and mean = 10.3, median = 10, SD = 2.8 in males), with a moderate effect size (d = 0.40). There were no differences concerning IQ.

Older adolescent and adult sample

In older adolescents and adults with ASD, 51% had co-occurring diagnoses (most commonly depressive disorders, in 25% of the sample). The most common non-ASD diagnoses were affective disorders (21%) and personality disorders (20.5%). 72% had further co-occurring diagnoses. Again, in contrast to this high multi-comorbidity, 35% of non-ASD cases had no clinical mental disorder. Further details can be found in Additional file 1: Tables S1 and S2.

There were no differences between the ASD and non-ASD groups with regard to age or IQ (see Table 1). Comparing only those female and male individuals with an ASD diagnosis (see Table 2), again, females were older than males on average (T(507) = 4.58, p = 0.000; mean = 29.8, median = 28, SD = 11.5 in females and mean = 24.9, median = 22, SD = 10.6 in males, d = 0.44), and females had a slightly higher full IQ (T(452) = 2.0, p = 0.045, d = 0.20).

Measures

The Autism Diagnostic Observation Schedule (ADOS-G/ADOS-2) [23, 24] is a standardized instrument that assesses social interaction, communication, and imagination during a semi-structured interaction with an examiner. It is an internationally used diagnostic instrument that consists of a module for toddlers and four additional modules to be administered based on the individual’s level of expressive language, chronological age, and appropriateness of the respective assessment materials. ADOS codes indicate symptom severity by coding increasing severity with codes of 0, 1, 2, and 3. Specific ADOS codes additionally contain information about peculiar or abnormal behavior using codes 7 or 8. There are 29 behavioral aspects (very specific aspects such as eye contact and broader aspects such as quality of social overtures) that have to be observed and coded in Module 3, and 31 behavioral codes in Module 4, of which 14 are entered into the respective classification algorithm. The ADOS provides diagnostic cut-offs for “no autism,” “autism spectrum,” and “autism” and metrics of ASD symptom severity (“comparison score,” CSS) [23].

For the best estimate clinical diagnosis (BEC), the ADOS needs to be complemented by the Autism Diagnostic Interview—Revised (ADI-R [25], a structured clinical caregiver interview that mostly focuses on ASD-related symptoms at the age of 4.0–5.0 years. The scoring of the ADI-R is organized into three behavioral domains: qualitative abnormalities in reciprocal social interaction (A); qualitative abnormalities in communication (B); and restricted and repetitive behavior (C). Furthermore, a careful differential diagnostic...
examination, physical examination, medical history-taking, and assessment of intellectual abilities are required for BEC [26] and were undertaken in the present study using standard diagnostic instruments. All diagnoses in the current study were built on a thorough clinical characterization of all individuals, leading to BEC diagnoses that did not always correspond to the classification according to the ADOS diagnostic cut-offs (see fourfold Table 3 for details).

### Data analysis

To explore differences in demographic characteristics between females and males, t-tests were conducted. To explore which ADOS items discriminate best between ASD and non-ASD within the two groups of females and males, we trained a random forest (RF) algorithm. Twenty-eight items of the ADOS module 3 and 31 items of the ADOS module 4 entered as predictors (item Amount of Social Overtures/ Maintenance of Attention
to Examiner of ADOS-2 was excluded because it is only used in ADOS-2 and was not available for cases examined with ADOS-G). Following the ADOS manual instructions, we remapped codes of 7 and 8 to 0, and codes of 3 were recoded to 2. The ASD best estimate clinical diagnosis was the classification criterion.

Potential biases due to site effects were tested by including site as a predictive feature in the RF. As it was of no importance, the final RF included only ADOS items. RFs are ensemble classifiers based on several decision trees aggregated by majority voting. Each decision tree yields a class prediction considering a random subset of features, and a majority vote of all the trees (“the forest”) forms the final classification [27]. Figure 1 summarizes the steps during training, testing, and validating the random forest.

We implemented guards against overfitting by splitting the total sample into a 70% RF (training and test) set and a 30% hold out set for validation of the final models. The RF set was again split into a 75% training set for model building and hyperparameter estimation and a 25% test set for model evaluation. The procedure consists of four consecutive steps, uses the R package randomForest [28] and is described in detail in [29].

The first step was a feature selection that gave us a hierarchy of all features regarding their importance for predicting class membership (i.e., ASD versus non-ASD). In a second step, we stepwise reduced the number of features that entered the training of the RF according to their importance rank. Reduced models were trained

### Table 2 Characteristics of the ASD sample for males and females separately

|               | Female |        | Male |        | T   | df  | p    | ES  |
|---------------|--------|--------|------|--------|-----|-----|------|-----|
| **Age**       | 11.5 (3.2) | 10.3 (2.8) | 495 | 545 | 2.8 | .005 | 0.40 |
| **IQ**        | 93.8 (17.9) | 97.7 (18.2) | 408 | 453 | 1.4 | .167 | 0.22 |
| **ADOS-SA**   | 8.8 (4.2) | 9.9 (4.1) | 495 | 545 | 1.9 | .054 | 0.27 |
| **ADOS-RRB**  | 1.2 (1.4) | 1.5 (1.4) | 495 | 545 | 1.4 | .054 | 0.27 |
| **ADOS-CSS**  | 5.8 (2.5) | 6.6 (2.4) | 495 | 545 | 2.4 | .016 | 0.33 |
| **ADI_A**     | 14.6 (6.2) | 16.4 (6.0) | 339 | 370 | 1.7 | .096 | 0.30 |
| **ADI_B**     | 10.3 (4.9) | 12.7 (5.2) | 339 | 370 | 2.5 | .011 | 0.48 |
| **ADI_C**     | 3.3 (1.9) | 4.3 (2.4) | 339 | 370 | 2.2 | .026 | 0.46 |

**Older adolescents and adults**

|               | Female |        | Male |        | T   | df  | p    | ES  |
|---------------|--------|--------|------|--------|-----|-----|------|-----|
| **Age**       | 29.8 (11.5) | 24.9 (10.6) | 371 | 507 | 4.5 | .000 | 0.44 |
| **IQ**        | 106.8 (14.4) | 103.6 (17) | 326 | 452 | 2.0 | .046 | 0.20 |
| **ADOS-SA**   | 9.3 (4.2) | 10.11 (4.2) | 371 | 508 | 1.8 | .048 | 0.19 |
| **ADOS-RRB**  | 1.3 (1.2) | 1.6 (1.4) | 371 | 508 | 2.3 | .023 | 0.23 |
| **ADOS-CSS**  | 5.7 (2.6) | 6.2 (2.5) | 371 | 508 | 2.3 | .024 | 0.20 |
| **ADI_A**     | 13.2 (7.0) | 14.0 (6.1) | 205 | 234 | 0.6 | .550 | 0.12 |
| **ADI_B**     | 8.3 (5.0) | 10.0 (4.9) | 205 | 233 | 2.4 | .016 | 0.34 |
| **ADI_C**     | 3.0 (2.4) | 3.1 (2.1) | 205 | 40  | 1.8 | .076 | 0.04 |

**Table 3 Fourfold table of ASD/non-ASD versus ADOS cut-offs**

|               | Non-autism | Spectrum | Autism |
|---------------|------------|----------|--------|
| **Female**    |            |          |        |
| BEC           |            |          |        |
| Non-ASD       | 86 (84.3%) | 5 (4.9%) | 11 (10.8%) |
| ASD           | 10 (19.2%) | 12 (23.1%) | 30 (57.7%) |
| **Male**      |            |          |        |
| BEC           |            |          |        |
| Non-ASD       | 571 (81.2%) | 64 (9.1%) | 68 (9.7%) |
| ASD           | 61 (12.3%) | 74 (14.9%) | 360 (72.7%) |
| **Older adolescents and adults** | | | |
| Female        |            |          |        |
| BEC           |            |          |        |
| Non-ASD       | 103 (87.3%) | 6 (5.1%) | 9 (7.6%) |
| ASD           | 25 (18.0%) | 24 (17.3%) | 90 (67.7%) |
| Male          |            |          |        |
| BEC           |            |          |        |
| Non-ASD       | 220 (71.7%) | 34 (11.1%) | 53 (17.3%) |
| ASD           | 55 (14.8%) | 43 (11.6%) | 273 (73.6%) |

**BEC best estimate clinical diagnoses, ASD autism spectrum disorder**
with 20-fold cross-validation using 95% of the data for training and 5% for testing. Subsequently, the reduced feature classifiers were evaluated on the previously held out and unseen validation data set, and the “optimal model” was determined by calculating a weighted ratio of accuracy and complexity (number of variables) for each model, with the choice of weights favoring simpler models in a 2:1 ratio (i.e., $w_1 \times \text{AUC} + w_2 \times \text{complexity}$ where $w_1 = 0.35$ and $w_2 = 0.65$). Each model’s accuracy (ACC), sensitivity, and specificity are presented as indices of model quality. In the final step, the optimal model’s predictive performance (accuracy) was statistically tested against the full features model using the McNemar test.

### Results

#### Differences in symptom severity

In the child and young adolescent sample with ASD, no differences between females and males were observed concerning the social affect and RRB domains of the ADOS (see Table 2). However, the Calibrated Severity Score (CSS) was higher in males than in females ($T(545) = 2.4, p = 0.016$), with a small effect size ($d = 0.33$). In the ADI-R, we found differences in the Communication domain ($T(370) = 2.5, p = 0.011$) and the RRB domain ($T(370) = 2.2, p = 0.026$), with moderate effect sizes ($d = 0.48$ and 0.46).

In older adolescents and adults with ASD, differences were observed between females and males concerning
the RRB domain of the ADOS ($T(508)=2.3, \ p=0.023$, $\chi^2=0.23$) (see Table 2), but this difference did not emerge in the anamnestic interview (ADI-R). Males showed slightly more deficits in the Communication domain ($T(233)=2.4, \ p=0.016$), with only a small effect size ($d=0.34$). In particular, adult females with ASD had significantly more comorbidities (e.g., depression, social phobia), whereas males with ASD had more attention deficit hyperactivity disorders. In the non-ASD samples, the females and males had a similar number of further diagnoses, but again, females were more likely to have anxiety disorders (F40-48 of ICD-10) and males were more likely to have attention deficit hyperactivity disorders.

**Diagnostic threshold of the ADOS**

In the child and young adolescent sample, 80.8% of females and 87.6% of males diagnosed with ASD met the Autism Spectrum cut-off of the ADOS-2 classification algorithm. On the other hand, 15.7% of females and 18.8% of males without an ASD diagnosis also met the ADOS-2 Autism Spectrum cut-off. In the older adolescent and adult sample, 82.7% of females and 88.4% of males diagnosed with ASD met the cut-off, but so too did 12.7% of females and 28.4% males without an ASD diagnosis. The fourfold Table 3 presents the ADOS scores and diagnostic groups.

**Random forest (RF) analyses**

The endorsement of ADOS items and their importance for the diagnostic classification (ASD versus non-ASD) was explored through a random forest approach for females and males separately. The first step of the analysis focused on identifying the latent feature importance ranking. Figure 2a, b shows the average rank of each feature from the cross-validation procedure in a heatmap—comparing the ranking of features between sexes in children and young adolescents (Fig. 2a) and older adolescents and adults (Fig. 2b).

Model performance indices from the RF models are computed, and yielded an AUC of 0.91 with 1.00 sensitivity and 0.88 specificity. For independent validation of the classifier, its performance on the validation set was computed and yielded an AUC of 0.86 with 0.63 sensitivity and 0.81 specificity (see also Table 4 for an overview). A model including five features yielded optimal results in the validation set, with an AUC of 0.83 with 0.81 sensitivity and 0.87 specificity. The optimal model comprised the following features: Quality of Social Overtures (QSOV), Facial Expressions Directed to Examiner (EXPE), Conversation (CONV), Shared Enjoyment in Interaction (EN), Descriptive, Conventional, Instrumental, or Informational Gestures (DGES). A comparison of the models’ performance via McNemar’s test for differences in classification error rates showed no advantage of the full-feature model (28 features) over the weighted optimal model with five features ($\chi^2=0.266$, $p=0.60$).

In males, the model output from the test set, including all 28 variables, showed an AUC of 0.93 with 0.93 sensitivity and 0.86 specificity. For independent validation of the classifier, its performance on the validation set was computed, and yielded an AUC of 0.79 with 0.85 sensitivity and 0.88 specificity. A model including eight features yielded optimal results in the validation set, with an AUC of 0.81 and 0.85 sensitivity and 0.85 specificity. The following eight features were included in the optimal model: Speech Abnormalities Associated With Autism (SPAB), Conversation (CONV), Quality of Social Overtures (QSOV), Insight Into Typical Social Situations and Relationships (INS), Descriptive, Conventional, Instrumental, or Informational Gestures (DGES), Amount of Reciprocal Social Communication (ARSC), Stereotyped/Idiosyncratic Use of Words or Phrases (STER), Unusual Eye Contact (EYE). A comparison of the models’ performance via McNemar’s test for differences in classification error rates showed no advantage of the full-feature model (28 features) over the weighted optimal model with eight features ($\chi^2=0.209, \ p=0.14$).
| Feature | Female | Male |
|---------|--------|------|
| NESL    | 23     | 27   |
| SPAB    | 18     | 1    |
| IECO    | 22     | 22   |
| STER    | 18     | 6    |
| OINF    | 23     | 13   |
| ASK     | 10     | 12   |
| REPT    | 22     | 15   |
| CONV    | 3      | 2    |
| DGES    | 5      | 5    |
| EYE     | 11     | 8    |
| EXPE    | 3      | 9    |
| LLNC    | 15     | 14   |
| ENJ     | 4      | 10   |
| EMO     | 6      | 11   |
| INS     | 12     | 4    |
| QSOV    | 1      | 3    |
| QSR     | 15     | 16   |
| ARSC    | 6      | 7    |
| OQR     | 17     | 17   |
| IMAG    | 16     | 19   |
| SINT    | 21     | 23   |
| MAN     | 20     | 20   |
| INJ     | 23     | 25   |
| XINT    | 26     | 18   |
| RITÖ    | 9      | 21   |
| OACT    | 22     | 24   |
| TAN     | 24     | 23   |
| ANX     | 11     | 26   |

| Feature | Female | Male |
|---------|--------|------|
| NESL    | 29     | 29   |
| SPAB    | 5      | 19   |
| IECO    | 31     | 25   |
| STER    | 9      | 20   |
| OINF    | 27     | 26   |
| ASK     | 25     | 22   |
| REPT    | 19     | 13   |
| CONV    | 14     | 6    |
| DGES    | 4      | 8    |
| EGES    | 10     | 22   |
| EYE     | 1      | 3    |
| EXPE    | 3      | 5    |
| LLNC    | 13     | 14   |
| ENJ     | 12     | 15   |
| AFF     | 17     | 13   |
| EMO     | 2      | 10   |
| INS     | 7      | 11   |
| RESP    | 26     | 23   |
| QSOV    | 8      | 6    |
| QSR     | 11     | 1    |
| ARSC    | 6      | 3    |
| OQR     | 21     | 4    |
| IMAG    | 23     | 30   |
| SINT    | 20     | 25   |
| MAN     | 30     | 25   |
| INJ     | 26     | 25   |
| XINT    | 24     | 16   |
| RITL    | 18     | 20   |
| OACT    | 22     | 25   |
| TAN     | 16     | 16   |
| ANX     | 15     | 13   |

Fig. 2 (See legend on previous page.)
Table 4: The performance of the ADOS diagnostic cut-off (ADOS), performance of the RF models on the test set (test) and the previously unseen validation data set (val) for nonverbal children and young adolescents as well as older adolescents and adults.

| Module                        | No. of features | AUC ADOS | Sens. ADOS | Spec. ADOS | AUC test | ACC test | Sens. test | Spec. test | J | p McNe | AUC val | ACC val | Sens. val | Spec. val |
|-------------------------------|-----------------|----------|------------|------------|----------|----------|------------|------------|---|-------|----------|---------|-----------|-----------|
| Children and young adolescents | Female          | ADOS algorithm | 0.83       | 0.81       | 0.84     | 0.91     | 0.94       | 1          | 0.88 | 0.86  | 0.94     | 0.86    | 0.92      | 0.86      | 0.72 | 0.63  | 0.81    |
|                               |                 | All 28 features | 0.92       | 0.93       | 0.97     | 0.91     | 0.43       | 0.60       | 0.83 | 0.84  | 0.60     | 0.84    | 0.81      | 0.87      |
|                               |                 | 5 features (optimal model) | 0.92       | 0.88       | 0.91     | 0.86     | 0.40       | 0.14       | 0.81 | 0.85  | 0.14     | 0.85    | 0.85      | 0.85      |
|                               | Male            | ADOS algorithm | 0.87       | 0.88       | 0.81     | 0.93     | 0.89       | 0.93       | 0.86 | 0.79  | 0.89     | 0.79    | 0.85      | 0.85      |
|                               |                 | All 28 features | 0.92       | 0.88       | 0.91     | 0.86     | 0.40       | 0.14       | 0.81 | 0.85  | 0.14     | 0.85    | 0.85      | 0.85      |
|                               |                 | 7 features (optimal model) | 0.92       | 0.88       | 0.91     | 0.86     | 0.40       | 0.14       | 0.81 | 0.85  | 0.14     | 0.85    | 0.85      | 0.85      |
|                               | Older adolescents and adults | Female          | ADOS algorithm | 0.89       | 0.82       | 0.88     | 0.83     | 0.88       | 0.91       | 0.82 | 0.83  | 0.83     | 0.93    | 0.93      | 0.93      | 0.72 |
|                               |                 | All 31 features | 0.88       | 0.88       | 0.92     | 0.85     | 0.42       | 0.18       | 0.86 | 0.78  | 0.18     | 0.84    | 0.84      | 0.84      | 0.72 |
|                               |                 | 5 features (optimal model) | 0.88       | 0.88       | 0.92     | 0.85     | 0.42       | 0.18       | 0.86 | 0.78  | 0.18     | 0.84    | 0.84      | 0.84      | 0.72 |
|                               | Male            | ADOS algorithm | 0.85       | 0.85       | 0.72     | 0.82     | 0.82       | 0.83       | 0.81 | 0.55  | 0.82     | 0.87    | 0.79      | 0.80      | 0.77 |
|                               |                 | All 31 features | 0.82       | 0.80       | 0.84     | 0.76     | 0.48       | 0.43       | 0.82 | 0.76  | 0.43     | 0.82    | 0.76      | 0.81      | 0.71 |

AUC: area under the curve, ACC: accuracy, Sens.: sensitivity, Spec.: specificity, J: Youden's index, McNemar: level of significance— each model tested against the full-feature sets of available features.
optimal model comprised the following features: Unusual Eye Contact (EYE), Comments on Others’ Emotions/Empathy (EMO), Facial Expressions Directed to Examiner (EXPE), Descriptive, Conventional, Instrumental, or Informational Gestures (DGES), Speech Abnormalities Associated with Autism (SPAB). A comparison of the models’ performance via McNemar’s test for differences in classification error rates showed no advantage of the full-feature model (31 features) over the weighted optimal model with five features ($\chi^2 = 1.76, p = 0.18$).

In males, the model output from the test set, including all 31 variables, showed an AUC of 0.82, with 0.83 sensitivity and 0.81 specificity. For independent validation of the classifier, its performance on the validation set was computed, and yielded an AUC of 0.87, with 0.80 sensitivity and 0.77 specificity. A model including eight features yielded optimal results in the validation set, with an AUC of 0.87 and 0.80 sensitivity and 0.77 specificity. The optimal model comprised the following features: Quality of Social Responses (QSR), Amount of Reciprocal Social Communication (ARSC), Unusual Eye Contact (EYE), Overall Quality of Rapport (OQR), Facial Expressions Directed to Examiner (EXPE), Quality of Social Overtures (QSOV), Conversation (CONV), Descriptive, Conventional, Instrumental, or Informational Gestures (DGES). A comparison of the models’ performance via McNemar’s test for differences in classification error rates showed no advantage of the full-feature model (31 features) over the weighted optimal model with eight features ($\chi^2 = 0.622, p = 0.43$).

**Discussion**

The present study aimed to explore potential differences in how female and male individuals meet the diagnostic criteria for ASD assessed by the ADOS. We aimed to identify a potential female phenotype from behavioral observations in a well-characterized clinical population of children, adolescents, and adults. Using a random forest approach, we compared subsets of diagnostic features of the ADOS that were most indicative of an ASD diagnosis between females and males.

**Overall results**

The results revealed similar classifier performances in the female and male samples, but relying on slightly different features for classification. Concentrating on a few core behavioral aspects for female and male samples led to classification performances that were equally as good as those based on information from the complete examination. For an optimal performance, the classifiers needed fewer features in the female sample than in the male sample in both age groups. It has been argued that since the defining diagnostic criteria are historically based on the

---

**Table 5** The ADOS codes of behaviors identified as the optimal feature subset in descending order of importance for males and females

| Female | Male |
|---|---|
| **Children and young adolescents** | **Male** |
| (1) Quality of Social Overtures (QSOV) | (1) Speech Abnormalities Associated with Autism (SPAB) |
| (2) Facial Expressions Directed to Examiner (EXPE) | (2) Conversations (CONV) |
| (3) Conversation (CONV) | (3) Quality of Social Overtures (QSOV) |
| (4) Shared Enjoyment in Interaction (ENJ) | (4) Insight into Typical Social Situations and Relationships (INS) |
| (5) Descriptive, Conventional, Instrumental or Informational Gestures (DGES) | (5) Descriptive, Conventional, Instrumental or Informational Gestures (DGES) |
| **Older adolescents and adults** | **Male** |
| (1) Unusual Eye Contact (EYE) | (1) Quality of Social Responses (QSR) |
| (2) Comments on others’ emotions/empathy (EMO) | (2) Amount of Reciprocal Social Communication (ARSC) |
| (3) Facial Expressions Directed to Examiner (EXPE) | (3) Unusual Eye Contact (EYE) |
| (4) Descriptive, Conventional, Instrumental or Informational Gestures (DGES) | (4) Overall Quality of Rapport (OQR) |
| (5) Speech Abnormalities Associated with Autism (SPAB) | (5) Facial Expressions Directed to Examiner (EXPE) |
| **Quality of Social Over-| (6) Quality of Social Overtures (QSOV) |
| tures** | (7) Conversation (CONV) |
| **Conversa-| (8) Descriptive, Conventional, Instrumental or Informational Gestures (DGES) |
| **Descriptive, Conventional, Instrumental or Informational Gestures** | |
| (1) Speech Abnormalities Associated with Autism (SPAB) | |
| (2) Conversations (CONV) | |
| (3) Quality of Social Overtures (QSOV) | |
| (4) Insight into Typical Social Situations and Relationships (INS) | |
| (5) Descriptive, Conventional, Instrumental or Informational Gestures (DGES) | |
| (6) Amount of Reciprocal Social Communication (ARSC) | |
| (7) Stereotyped/Idiosyncratic Use of Words or Phrases (STER) | |
| (8) Unusual Eye Contact (EYE) | |

Bold font indicates items are comprised in the diagnostic algorithms (ADOS-2). Overlap between the sexes is written in italics.
male phenotype and the diagnostic thresholds are similar, a female phenotype may be missed if it presents differently, even if these females present with a substantial clinical burden [7]. However, the current study demonstrates that although slightly different features were most discriminative, classification in females was just as good as in males.

**Differences in symptom severity**

In the current study, females were older at the time of the diagnostic appointment—an effect that was pronounced in the *older adolescent and adult sample*. In the *young adolescent and adult group*, males with ASD scored higher in the RRB domain of the behavior observation than females with ASD, but the effect size was small. We observed no differences in social affect between the sexes in the ASD samples of either age group, but males scored higher on overall symptom severity. These findings are in line with a meta-analysis that reported few differences in communication and social behavior between males and females and only in the RRB domain did girls show fewer symptoms than boys [18].

The present findings indicate, however, that as ASD symptoms present differently across development, the developmental aspect might be important with respect to sex differences: In the *older adolescent and adult sample*, we found fewer symptoms of RRB and lower overall ASD severity (ADOS CSS total) in females than in males. From the parental perspective (anamnestic data from the ADI-R), females showed fewer symptoms in the communication domain. In the *child and young adolescent sample*, more parent-reported RBB were observed in males compared to females, with moderate effect sizes. Classification accuracy of the RF models was similar to the diagnostic accuracy of the ADOS-2 algorithm in females as well as males. Interestingly, we found more females than males who were diagnosed with ASD while scoring below the ADOS autism spectrum diagnostic cut-off (18.6% females vs. 13.5% males, i.e., false negative ADOS classifications). This suggests that information from outside the standardized behavioral observation may be of greater importance for the diagnostic decision in females than in males, giving rise to the question of which particular additional information clinicians rely on in order to classify a female as autistic. On the other hand, more males than females did not receive an ASD diagnosis despite exceeding the ADOS diagnostic threshold (6.4% females vs. 14.2% males, i.e., false positive ADOS classifications). This suggests that autistic traits in males may be present during the behavioral observation but are attributed to other underlying conditions or symptoms of a differential diagnosis. However, our female sample had more comorbid diagnoses (e.g., depression, social phobia), and particularly in females with ASD, there is evidence that the presence of depression and anxiety is associated with enhanced ASD symptoms [30–35]. The considerable symptom overlap of ASD with depressive and anxiety disorders entails the risk of false-positive evaluations in females. Although the ADOS-2 shows high sensitivity (0.91; [23], p. 243) for detecting autism versus non-spectrum cases, emerging research shows that it may be less accurate in detecting ASD in individuals with complex psychiatric presentations [36]. Moreover, the observation in the current sample that the prevalence of ASD diagnoses increases with age (45.9% of all adolescent/adult females, but only 33.8% of the younger sample, received an ASD diagnosis) underlines the need to carefully consider differential, potentially overlapping diagnoses during the diagnostic process.

**Differences in diagnostic features of the ADOS**

To the best of our knowledge, this is the first study to explore sex differences in an ASD and a non-ASD sample with the aim of identifying those symptoms that are most important for the classification and subsequently comparing these discriminative features between females and males. The most discriminative features all stem from the social communication domain of the ADOS, whereas only speech items (Speech Abnormalities Associated with Autism, and Stereotyped and Idiosyncratic Use of Words or Phrases) of the RBB domain are included in the optimal feature models. This may be due to the rather short time span of the ADOS (45–60 min duration of administration), which limits the time for observations of repetitive behaviors and/or the overall more verbal character of the ADOS modules 3 and 4. Furthermore, although males showed more RBBs than females, the pattern in male and female non-ASD cases seemed similar thus not providing the RF classifier information relevant for the distinction of ASD and non-ASD cases within each group. The effect may also be attributed to basic sex differences in the occurrence of RRB in the diagnostic situation elicited in boys by a male-biased toy selection. As has been pointed out, the restricted and repetitive interests among females may be more “random” and more difficult to categorize and thus to “identify as atypical” [15, p. 1391].

Our optimal models include mainly ADOS items mapping onto “Basic Social Communication Skills.” According to Bishop and colleagues [37], social communication deficits captured by the ADOS can be divided into “Basic Social Communication Skills” (including Gestures, Eye Contact, Facial Expressions, and Shared Enjoyment) and “Interaction Quality” (including Conversation, Amount of Reciprocal Social Communication, and other Quality items). These ‘basic’ impairments seem to be specific for
ASD regardless of sex, age, and intelligence [37, 38]. In our models, these basic impairments appear, overall, to be sufficient in order to discriminate females with ASD from those with other mental disorders when flanked by the two additional items of “Interaction Quality,” with good specificity and sensitivity. Moreover, in contrast to the findings for males, they are not correlated with age and IQ (see Additional file 1: Table S3). Some previous studies found that females with ASD exhibit less severe impairments in social communication behaviors [39, 40], although we and others [7, 14] cannot confirm this for the behavior observation. Nevertheless, these items do seem to be essential for the differentiation of ASD from other mental disorders, particularly in females.

In the child and young adolescent sample, we found similarities between females and males concerning the following items: Quality of Social Overtures, Conversation, and Gestures. Differences were especially evident in the communication domain. Speech abnormalities were also relevant for the differentiation from other mental disorders. Such speech abnormalities are important for females: For the female group, all items are algorithm items, whereas for the male group, six items are part of the algorithm and two additional items are needed (Speech Abnormalities and Insight) for the model to reach optimal classification performance. In the older adolescent and adult sample, similarities were only found concerning the basic skills Eye Contact, Facial Expressions, and Gestures. However, for the differentiation from other mental disorders in males, many aspects of the Quality of Interaction are additionally needed; in females, only Empathy and Speech Abnormalities are relevant.

In the male, the most discriminative ADOS items all stem from the classification algorithm plus the item Descriptive, Conventional, Instrumental or Informational Gestures (DGES). In the female sample, though a smaller number of features seem to suffice for an optimal classification, only 3 out of 5 items stem from the ADOS classification algorithm. Particularly, the item Comments on Others’ Emotions/Empathy that is linked to cognitive empathy, a construct often impaired in ASD [41], was of prime importance in the optimal model.

Overall, the optimal models of our RF approach yielded slightly different distinctive features for females and males but did not outperform the ADOS-2 classification algorithm (grouping the autism spectrum and autism cases together). These results do not suggest an adaptation of the ADOS-2 classification algorithm for a female phenotype.

Future aim of the present work is to break down these most discriminative subsets of diagnostic items into their underlying mechanisms or processes and translate them into research on biomarkers in order to identify the behaviorally observed differences between females and males on a molecular level. This needs to be the next step on the way to the identification of a female phenotype as both measures—ADOS and ADI-R—cannot simply be abbreviated, as, e.g., ADOS codes are attained throughout the observation session and are not strictly tied to single subtasks [24] and thus items cannot be observed independently and the impact of each item for the diagnostic decision is difficult to extract.

**Strengths and limitations**

The observation of differences between the sexes not only regarding the most discriminating diagnostic features but also across the age groups leads to the assumption that gender associated symptom presentation changes during development. Future studies therefore need to evaluate sex differences in younger age groups and ideally in longitudinal studies in children at risk, who are eventually diagnosed with ASD or other developmental or clinical conditions. Only longitudinal data can clarify “age differences in how ASD manifests in boys vs. girls, from other phenotypic differences” [14], p. 102).

A particular strength of the present study lies in the composition of the sample. Previous research on sex/gender effects only included individuals with a confirmed ASD diagnosis, and may therefore have missed females with different symptom profiles (the “female phenotype”). By contrast, the present study investigated a broader clinical sample that also included individuals with suspicion of ASD. This had the advantage that we were able to evaluate sensitivity and specificity, and did not merely treat scores as indices of symptom severity, as was the case in previous studies [7, 14]. Thus, it was possible to evaluate the utility of standard instruments also among individuals with autistic traits but with other mental disorders. In turn, this enabled us to identify symptom profiles in females that led to a diagnostic decision and to compare them to symptom profiles in males. Furthermore, diagnoses in the present study were best estimate clinical diagnoses (BEC) and did not solely rely on the diagnostic thresholds of the “gold standard” instruments (ADOS and ADI-R). The sample thus included individuals (female and male alike) who did not meet the ADOS/ADI-R cut-offs but were nevertheless diagnosed with ASD, or conversely, individuals who were not diagnosed with ASD despite their scores exceeding the diagnostic threshold.

As was the case in previous studies, our sample of females with ASD was smaller than the male sample. Therefore, females may still be underrepresented, limiting the statistical power to detect small to moderate effects. A further limitation concerns our study design:
Although ASD diagnoses in the current study were BEC diagnoses and did not rely solely on ADOS scores, these scores were nevertheless employed as part of the diagnostic assessment, leading to a certain degree of circularity. This is also associated to the limitation that behaviors captured by the ADOS might already be male-biased because the development and validation of the instrument were undertaken with predominantly male cases. We tackled this by relying on BEC diagnoses that included multiple sources of information a mere classification based on ADOS (and ADI-R) cut-off scores. We thus have individuals in the sample that scored beyond cut-off but were nevertheless diagnosed with ASD and individuals that exceeded the cut-off but were not diagnosed with ASD. In order to approach this limitation, future studies need to extend the methodological approach to data-driven analyses. Previous studies have pursued subgroups within the autism spectrum and were able to identify subgroups based on social interaction and communication, intelligence, and morphological abnormalities. However, behavioral subgroups have not yet been replicated [42] and sex or gender has not yet been taken into account.

Conclusion
Altogether, we found similarities and some differences between females and males with ASD. The reduced feature models in females relied on considerably fewer features from the ADOS than those in males, while still yielding similar classification performances. Although we identified some phenotypic differences between females and males with ASD, the existing diagnostic ADOS algorithm seems to be sufficient to capture the core diagnostic criteria in females and males. These results lead to the conclusion that the available standardized behavior observation (ADOS) should remain a substantial part of the diagnostic procedure and that clinicians need to be aware of potential differential diagnoses, particularly in females.

Acknowledgements
The authors would like to thank Friederike Hellbig, Gerti Gerber, Henrike Schmidt, Imke Garten, Marie Kollarzcyk, Miriam-Sophie Petasch, and Svenja Köne for their assistance in the conduct of this research.

Authors’ contributions
SS and IKB drafted the manuscript, NW, SST, and CK managed the data base, SST, JT and DH carried out analyses, LP, SR, VR and IKB designed the study; all authors interpreted the results, critically reviewed the manuscript drafts, and approved the final manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was funded by the German Federal Ministry of Education and Research (BMBF, Grant Number: FKZ 01EE1409A). Funding period: 2015–2022.

Availability of data and materials
The data are not publicly available due to medical confidentiality but are available from the first author on request pending the approval of the coauthors.

Declarations
Ethics approval and consent to participate
All data were analyzed anonymously, with approval from the local ethics committee (Az. 92/20). Due to the retrospective nature of data collection and analysis based on anonymized data, the need for informed consent was waived by the ethics committee. All methods were applied following relevant institutional and international research guidelines and regulations.

Consent for publication
Institutional consent forms were used.

Competing interests
Prof. Dr. Poustka has received conference attendance support or speaking fees from Shire. She receives research funding from the BMBF, DFG and EU and royalties from Hogrefe, Kohlhammer and Schattauer. Prof. Dr. Roessner has received payment for consulting and writing activities from Lilly, Novartis, Shire Pharmaceuticals, lecture fees from Lilly, Novartis, Shire Pharmaceuticals, and Medice Pharma, and support for research from Shire Pharmaceuticals and Novartis. He has carried out clinical trials in cooperation with the Novartis, Shire, Servier and Otsuka companies. The remaining authors declare no potential conflict of interest.

Author details
1 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Faculty of Human Medicine, Philippus University Marburg, Hans-Sachs Str. 6, 36037 Marburg, Germany. 2 Department of Mathematics and Computer Science, Philippus University Marburg, Hans-Meerwein-Straße 6, 35043 Marburg, Germany. 3 Department of Child and Adolescent Psychiatry, TU University Dresden, Fetscherstr. 74, 01307 Dresden, Germany. 4 Department of Psychiatry, Charité – Universitätsmedizin Berlin, Campus Benjamin Franklin, Hindenburgdamm 30, 12203 Berlin, Germany. 5 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Faculty of Human Medicine, Philippus University Marburg, Hans-Sachs Str. 6, 36037 Marburg, Germany. 6 Department of Child and Adolescent Psychiatry, TU University Dresden, Fetscherstr. 74, 01307 Dresden, Germany. 7 Department of Psychiatry, Charité – Universitätsmedizin Berlin, Campus Benjamin Franklin, Hindenburgdamm 30, 12203 Berlin, Germany. 8 Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Faculty of Human Medicine, Philippus University Marburg, Hans-Sachs Str. 6, 36037 Marburg, Germany.
Hartley SL, Sikora DM. Sex differences in autism spectrum disorder: an examination of developmental functioning, autistic symptoms, and coexisting behavior problems in toddlers. J Autism Dev Disord. 2009;39:1715–22. https://doi.org/10.1007/s10803-009-0819-8.

Lai M-C, Szatmari P. Sex and gender impacts on the behavioural presentation and recognition of autism. Curr Opin Psychiatry. 2020;33:117–23. https://doi.org/10.1097/YCO.0000000000000575.

Received: 13 September 2021   Accepted: 22 February 2022

Published online: 07 March 2022
positives. Clin Neuropsychol. 2021. https://doi.org/10.1080/13854046.2021.1942220.
37. Bishop SL, Havdahl KA, Huerta M, Lord C. Subdimensions of social-communication impairment in autism spectrum disorder. J Child Psychol Psychiatry. 2016;57:909–16. https://doi.org/10.1111/jcpp.12510.
38. Zheng S, Kaat A, Farmer C, Kanne S, Georgiades S, Lord C, et al. Extracting latent subdimensions of social communication: a cross-measure factor analysis. J Am Acad Child Adolesc Psychiatry. 2021;60:768-782.e6. https://doi.org/10.1016/j.jaac.2020.08.444.
39. Lai M-C, Lombardo MV, Ruigrok AN, Chakrabarti B, Auyeung B, Szatmari P, et al. Quantifying and exploring camouflaging in men and women with autism. Autism. 2017;21:690–702. https://doi.org/10.1177/1362361316671012.
40. Sedgewick F, Hill V, Yates R, Pickering L, Pellicano E. Gender differences in the social motivation and friendship experiences of autistic and non-autistic adolescents. J Autism Dev Disord. 2016;46:1297–306. https://doi.org/10.1007/s10803-015-2669-1.
41. Andreou M, Skrimpa V. Theory of mind deficits and neurophysiological operations in autism spectrum disorders: a review. Brain Sci. 2020. https://doi.org/10.3390/brainsci10060393.
42. Eapen V, Cmčec R, Walter A. Exploring links between genotypes, phenotypes, and clinical predictors of response to early intensive behavioral intervention in autism spectrum disorder. Front Hum Neurosci. 2013;7:567. https://doi.org/10.3389/fnhum.2013.00567.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.