Supplementary Online Content


eAppendix. eMethods and eResults.
eReferences
eFigure 1. Site density plots and tolerance bands for selected diagnostic and developmental variables.
eFigure 2. CART.1 for diagnoses with only diagnostic scales as predictors.
eFigure 3. CART.3 for diagnoses with diagnostic scales, site and demographic and behavior characteristics as predictors.
eFigure 4. CART for diagnoses by individual sites using all predictors.
eTable. Summary of CART models for individual sites.

This supplementary material has been provided by the authors to give readers additional information about their work.
Procedures for establishing and maintaining inter-rater reliability for use of the ADOS and ADI-R in the SSC sample

Each site was required to have two clinical supervisors who meet research reliability standards on the ADI–R and ADOS in order to insure that clinicians were available who had experience making diagnostic decisions using both instruments. Over the course of the study (which occurred over three years), there were 29 clinical supervisors, 56 senior diagnosticians (i.e., clinicians making BEC diagnostic decisions), 82 examiners who administered the ADOS, and 81 examiners who administered the ADI-R across the 12 sites. Twenty-three of the 56 senior diagnosticians were also clinical supervisors; the remaining senior diagnosticians were reliable on either the ADI-R or ADOS (28) or neither (5).

Initially, clinical supervisors were required to meet research reliability standards with the study consultants (at any one time, there were 5-7 consultants, for a total of 9 different people), each of whom had acted as a trainer for both instruments on numerous occasions and established reliability with other research reliable examiners from UMACC and from other centers, as well as with standard reliability videos. Trainings and maintenance reliability checks then took place on a quarterly basis, with semi-annual visits to centers by the consultants and semi-annual training sessions for all examiners held at central locations. In addition, each consultant reviewed all data from each child at each center every quarter and asked to view videos of ADI-Rs and ADOSes of questionable cases. Once the study was underway, it was clear that reliability with supervisors was not equivalent to reliability with the consultants and a new procedure was put into place where each examiner submitted a reliability tape to the consultant every 3 months. If reliability fell below 85% exact agreement on an item level for the ADI-R or 75% for the ADOS (modules 1 or 2 and for modules 3 or 4), examiners were asked to submit an additional tape. This process occurred twice at which point, examiners were required to have a “reliable” examiner present during all further participation. Examiners were also given specific feedback regarding the fidelity of the administrations.

In general, inter-rater reliability was very good with averages for independent, consultant-coded ADOS tapes at 85% exact item agreement and for the ADI-R, 90% exact item agreement (estimated tapes scored per year, n=41 ADOSes and 39 ADIs). Presented another way, over the course of checking, 94% of the clinical supervisors met or exceeded minimum goals of reliability on ADOS protocols and 88% met the minimum goals for reliability for the ADI-R protocols. 87% of the examiners (not including supervisors) met or exceeded minimum goals for reliability on the ADOS and for the ADI-R. The remainder followed the procedure described above of submitting additional tapes until reliability was established.

Test-retest reliability was not assessed as part of this study. Data for test-retest of the instruments is reported in the manuals for each respective measure.¹ ²

Sites were required to include the standard set of measures in their assessments, but could differ in the context and order of their administration. For example, while most sites administered the ADI-R to a caregiver while the proband was given the cognitive tests and ADOS, both in a clinic, some did home-based assessments and others did various components sequentially. Thus, opportunities for observation or reporting of different information could have been available to different sites in addition to the standard instrument data. When sites were asked to indicate what information their senior clinicians used in making BEC diagnoses, all reported only the standard
information, but it is possible that, beyond the inclusion of site as a variable, implicit differences in information could account for differences across centers.

**Interclass correlations (ICCs)**

The ICC for each of the continuous core features of ASD and developmental scores in Table 2, is computed as the ratio of the between-site variance divided by the sum of the within-site and between-site variances, i.e. the total variance. The variances are estimated based on a mixed effect regression, modeling the mean of the measure as a function of a fixed overall mean term, plus a random site effect, plus a random error term. The model-based estimate of the variance of the random site effects estimates the between-site variance; the estimated variance of the random error estimates the between-site variance. PROC MIXED, an SAS® program, was used to fit the mixed effects models.

**CART**

*Classification and Regression Tree (CART)* is a non-parametric statistical modeling tool that builds classification and regression trees for predicting continuous (regression) or categorical (classification) outcomes. In this research, the outcome is the categorical variable best estimate clinical (BEC) diagnosis, therefore, the models are classification trees. CART is particularly well suited when there is no coherent a priori set of predictions regarding which variables are related to the outcome and how they interact with each other to affect the outcome. This is true for this paper, where the question is how BEC diagnosis is made in different sites. In such situations, CART methods can reveal simple relationships between variables that could go unnoticed using other analytic techniques. In comparison to linear parametric methods for modeling and prediction, such as logistic or normal linear regression, CART does not require pre-specifying interactions between the predictors. Furthermore, there is no implicit assumption that the underlying relationships between the predictors and the outcome are linear, or that they follow some specific non-linear form, or even that they are monotonic in nature. CART can also handle cases with incomplete data in a sophisticated way, thus maximizing the use of available information.

CART algorithms aim at achieving the best possible predictive accuracy. Specifically, the most accurate prediction is defined as the prediction with the minimum costs. Minimizing costs (rather than simply the proportion of misclassified cases) allows for the situation where some wrong predictions are more catastrophic than others, or where some wrong predictions occur more frequently than others; in the former case, different costs can be assigned to different types of misclassification; in the latter case, different prior distributions (priors) for the levels of the categorical outcome can be specified. In the current investigation, equal costs were assigned to all misclassification types. The rates of BEC diagnoses (AUT, PDD-NOS and ASP) in the full sample of n=2102 probands was 0.70, 0.21 and 0.09, respectively, and varied during the accumulation of the data around 0.7, 0.2 and 0.1. These proportions were more variable across sites (AUT ranged from 47% to 100%, PDD-NOS ranged from 0% to 45% and ASP ranged from 0% to 21%). All CART models (for the whole sample and individually for each site) were fit with a prior equal to the observed proportion in the respective sample (e.g. site specific prior) and the overall sample converging prior (0.7, 0.2, 0.1).

Decision trees contain a binary question with a “yes” or “no” answer about some feature at each node in the tree; for example “Is the ADOS-S+C total < 12?” Cases in a node are assigned to either a left or a right branch according to the answer. CART algorithms are based on minimizing the “impurity” (i.e. heterogeneity) of the nodes. In the case of classification, the impurity parallels
misclassification. At each node, a split based on one of the predictors (e.g., an autism-specific diagnostic measure, such as the ADOS-S+C total) is identified; the predictors are ordered by the reduction in the impurity measure that they achieve. The predictor giving maximum reduction is chosen at each node. In addition to the predictor achieving maximum reduction of impurity, information regarding the next few best predictors can be informative and is stored in the results. Another set of predictors that are related to the chosen variable are used to classify cases for which the selected predictor is missing – they are called surrogates. Different measures of impurity exist. Here we use the Gini index, which takes value 0 if the node contains elements from only one category of the outcome (best classification) and takes value 1 if the cases in the node are distributed equally between the categories of the outcome (worst classification). The splitting rule is applied recursively at each branch until some stopping criterion is achieved (e.g., purity or a minimum number of cases in the node). The leaves, or terminal nodes, of the tree contain the cases that satisfy the conditions of the branches that lead to them and assign prediction for the outcome for those cases, for example, “AUT” as a BEC diagnosis.

As with the more familiar parametric regression models, larger trees have smaller misclassification rates (or sums of squared errors in the case of continuous outcomes); however, larger trees often have poor properties in terms of prediction of new observations, due, in essence, to over-fitting of the existing data. To control the problem, CART estimates cross-validation error-rates (misclassification rate for data not used in fitting the tree). CART models for the whole sample used 10-fold cross-validation. The final model is selected to optimize the prediction error (cross-validation error-rate) using the procedures for pruning the full grown trees suggested by Breiman and colleagues: it selects the size of the tree to be that of the simplest tree, which is within one SD of the cross-validation error rate from the tree that minimizes the mean cross-validation error rate.

Classification trees, like regression models, can be characterized with respect to how well they fit the data. In the case of normal regression models, “percent of variance explained by the model” is often reported as a measure of how important the variables in the model are; the percent is with respect to the total variance in the outcome, without any predictors. In classification models, the misclassification error rate without any predictors is the equivalent of total variance, and the models can be described in terms of how much the model decreases this error rate (i.e., explains the error). In this investigation, the proportion of cases in the BEC diagnostic categories (AUT, PDD-NOS, ASP) in the total sample is (0.7, 0.21, 0.09) and the best classification of a new participant without considering any predictors would be “AUT”; this makes the total misclassification error rate 30% (i.e., 100% - 70% of children identified with AUT = 30%). If a model decreases this misclassification error to 24%, this would constitute a reduction of 20% of the total error [(30-24)/30=.20]; i.e. the model explains 20% of the error. If a model with more predictors has a misclassification error of 20%, the total error is reduced by 33% [(30-20)/30=.33], i.e., 33% explained error. In this case, the second model constitutes 13% improvement over the first one (.33-.20=.13); i.e., the additional predictors explained 13% of the total error.

An important limitation of this analytic technique is that CART is not based on a probabilistic model. Thus, significance testing (e.g. for importance of variables) cannot be performed and confidence intervals (e.g., for cut-off points in a split) cannot be constructed. Models based on small samples can be quite unreliable and this is particularly relevant for the models for individual sites. For this reason, CART analyses can be difficult to replicate exactly, although the reporting of predictors that are ranked second and third after the selected ones in terms of improving the model, can be informative. It is worth noting that the analyses in this paper were initially run for the first half of the sample (n=933), and the resulting trees were used to predict the newer data (cases 934 to 2102). The misclassification error rates for the new cases were very close to the misclassification error rate for the cases on which the tree models were built. For example,
the CART.2 based on the first 933 participants had a misclassification rate of 0.23; using the same model to predict BEC diagnosis on the new children resulted in a misclassification rate of 0.25, which is very good.
eResults

CART analysis

CART.1 Relating BEC diagnosis to diagnostic scales

The CART model with prior equal to the sample proportions for AUT, PDD-NOS and ASP (i.e., 0.70, 0.21 and 0.09) is shown on eFigure 2. The first selected predictor was ADOS-S+C. ADOS-RRB and ADOS-CSS were ranked 2nd and 3rd best predictors, respectively, for the 1st node; these do not appear on the figures, but are reported because they reduced the impurity measure almost as much as the selected variable. Children with moderate to severe raw social communication scores (ADOS-S+C \( \geq 12 \)) on the clinician-administered standardized observation and interview (\( n_{\text{Total}}=1277, n_{\text{AUT}}=1087, n_{\text{PDD-NOS}}=136, n_{\text{ASP}}=54 \)) ended in a single terminal node which was relatively pure (85.1% of participants in this node had AUT compared to 69.9% of the full sample). Thus, the CART model would predict that all of these children had AUT diagnoses, yielding a misclassification error of 15%.

In contrast, for children with relatively mild social-communication scores (ADOS-S+C < 12; \( n_{\text{Total}}=825, n_{\text{AUT}}=383, n_{\text{PDD-NOS}}=299, n_{\text{ASP}}=143 \)), there were many further nodes, indicating interactions among predictors. The second split was on the ADOS-CSS, a measure that calibrates ADOS algorithm scores, including RRBs, across ages and language levels for Modules 1 - 3. Cases with ADOS-CSS \( \geq 6 \), indicating moderate to severe autism severity as observed during the ADOS (left branch, \( n_{\text{Total}}=576, n_{\text{AUT}}=313, n_{\text{PDD-NOS}}=172, n_{\text{ASP}}=91 \)) were further split based on ADI-R Social domain. The ADI-R Social domain differs from the ADOS in that it is a parent-report measure that includes history as well as current functioning and separates social scores from nonverbal and verbal communication. Participants split with the right branch, where ADI-R Social < 13 (indicating quite mild social deficits), resulted in a terminal node. Of the 82 cases here, 50 (61%) were given a BEC diagnosis of PDD-NOS. CART.1 would classify all children in this terminal node as PDD-NOS with 39% misclassification error.

The left branch, with mild to severe ADI-R Social \( \geq 13 \) (\( n_{\text{Total}}=494, n_{\text{AUT}}=291, n_{\text{PDD-NOS}}=122, n_{\text{ASP}}=81 \)) had further splits based on the characteristics of the senior clinician. Diagnoses of children seen by a clinician with more than 10 years of experience (\( n=301 \)) fell in another terminal node of AUT with a misclassification error, compared to the actual clinical diagnoses, of 23%. Classifications based on diagnoses made by clinicians with less experience (\( n=193 \)) were further improved by ADI-R-RRB (parent-reported histories of RRBs). Ten children with lower RRB scores ended up in a pure terminal node with a classification of PDD-NOS. The remaining 192 children with higher RRB scores had further split on ADOS module (separating children with phrase speech from those with complex expressive language), such that children with less language (Modules 1 or 2) were never diagnosed with ASP and children with more language (Modules 3 and 4) were again differentiated by ADI-RRB. At that node, children with higher RRB scores were classified as AUT (even though BEC diagnoses were heterogeneous) and children with lower RRB scores were classified as PDD (with a majority of these children receiving BEC diagnoses of PDD-NOS and ASP). Misclassification errors for these final nodes were all at about 50%, reflecting considerable heterogeneity even with the multiple interactions.

The mildest cases in terms of clinician standardized observations, those with ADOS-S+C < 12 and ADOS-CSS < 6 (\( n_{\text{Total}}=249, n_{\text{AUT}}=70, n_{\text{PDD-NOS}}=127, n_{\text{ASP}}=52 \)) were split further first on parent-reported history of abnormalities in verbal communication (ADI-VC domain) such that children with more severe scores almost all received AUT diagnoses, even though they had less severe ADOS scores. For children with milder verbal communication abnormalities, interactions occurred with ADOS module (no child who had received a Module 1, indicating single words or...
less on the ADOS, received an ASP diagnosis). For children with phrases or more complex language (ADOS Modules 2, 3 or 4), the type of degree of the most senior clinician predicted PDD-NOS diagnoses (proportionately more commonly made by Ph.D.-level diagnosticians) versus ASP diagnoses (more commonly made by M.D.- and master-level diagnosticians). As in the previous branch, final nodes here were also very heterogeneous, with errors of 44% and 52%, respectively.

At the root node, everyone would be classified as AUT and thus, the total misclassification error rate (without predictors) was 30%. Thus, altogether, the CART.1 model reduced the misclassification error to 24%, (20% improvement, or explained error). The model with prior equal to the overall converging proportions (0.7, 0.2, 0.1) gave identical results.

**CART.2 Relating BEC diagnosis to diagnostic scales and site**

The classification tree modeling BEC diagnosis using site indicators in addition to the diagnostic scales as predictors, assuming a prior equal to the sample proportions in the overall sample, is shown in Figure 2 in the main text.

After the root node, the first split was made based on ADOS-S+C with the same cut-off as in CART.1. Ranked second after ADOS-S+C in reducing the impurity was ADOS-RRB, followed closely by ADOS-CSS, ADOS module and site. Unlike CART.1, the 1277 cases with ADOS-S+C ≥ 12 were further split based on site, with 9 of the sites classifying all individuals as AUT (with only 9% misclassification) and 3 other sites making further decisions based on ADOS-RRB. Diagnosis of cases with ADOS-RRB ≥ 5 (nTotal =161, nAUT=138, nPDD-NOS=17, nASP=6) was further refined based on ADI-Social. Those with ADI-Social ≥ 13 were primarily diagnosed with AUT, whereas those with ADI-Social < 13 were diagnosed with PDD-NOS; diagnosis of cases with ADOS-RRB < 5 (nTotal =149, nAUT=68, nPDD-NOS=57, nASP=24) seemed to further take into account ADOS module, ADOS-RRB and ADI-R Verbal Communication, to separate AUT and PDD-NOS.

Diagnosis of the n=825 milder cases with ADOS-S+C < 12 strongly depended on site; 5 of the sites predominantly classified those cases as AUT and the other 7 sites most often diagnosed PDD-NOS. One of the 5 sites diagnosed only AUT in all cases, whereas the other 4 sites sometimes diagnosed ASP, apparently based on ADOS-CSS and ADI-R Verbal Communication. The cases with ADOS-S+C < 12 in the 7 sites that were most likely to diagnose PDD-NOS (nTotal =471, nAUT=146, nPDD-NOS=244, nASP=81), were further split based on ADOS-CSS (overall severity, calibrated by age and language level and including RRBs) with milder cases of ADOS-CSS < 6 ending in a terminal node; 103 of the 149 cases received a BEC PDD-NOS diagnosis. For cases with moderate to severe ADOS-CSS ≥ 6 (nTotal =322, nAUT=123, nPDD-NOS=141, nASP=58), further consideration of ADI-Social was taken into account; for those with ADI-R Social > 12, depending on characteristics of the senior clinician and ADI-RRB, AUT or PDD-NOS was diagnosed.

This model reduced the misclassification rate from total (without predictors) error rate of 30% to 21%, i.e. an improvement of 29%. The error rate in CART.2, when site was included, was lower than in CART.1 by 9%, i.e., sites explained 9% more error than CART.1. The model with prior equal to the overall converging proportions (0.7, 0.2, 0.1) gave identical results.

**CART.3 Relating BEC diagnosis to diagnostic scales, site and demographic/behavioral characteristics**

The model tree with a prior distribution equal to the observed overall prior is presented on eFigure3. The first split of ADOC-S+C was the same as in CART.1 and CART.2. The split of cases with ADOS-S+C ≥ 12, however, was not based on site (as in CART.2) but on Verbal IQ, a developmental characteristic, rather than a diagnostic scale. Within the group of participants with higher ADOS social-communication scores, cases with Verbal IQ ≤ 85 and ADOS-RRB ≥ 1 were almost all classified as AUT (nTotal = 822, © 2011 American Medical Association. All rights reserved.
only 12 of these cases had Verbal IQ $\leq 85$ and no RRBs on the ADOS, 8 of whom were classified as PDD-NOS and the others with AUT.

The 443 cases with ADOS-S+C $\geq 12$ and Verbal IQ $> 85$ were differentially classified depending most importantly on site. Within 6 of the 12 sites, diagnoses were based further on ADOS-RRB (in this case, differentiated at a score of $> 4$, which was quite high), together with other diagnostic and developmental scores (ADI-R Social, Nonverbal IQ, ADI-R Verbal Communication, ADOS-CSS and ADI-R Nonverbal Communication). The classification of the 825 cases with ADOS-S+C $< 12$ was very similar to the one in CART.2, but further improvement of the model was achieved by including developmental and behavioral characteristics such as Verbal IQ, Vineland Composite, child’s chronological age and hyperactivity.

The misclassification error rate of CART.3 was 17%, or improvement over no model by 43% (43% explained error by the model). The demographic and behavioral characteristics contributed 14% to the error explained by diagnostics scales and site. Together site and demographic/behavioral explained 23% more of the error than diagnostic scales alone. The model with prior equal to the overall converging proportions (0.7, 0.2, 0.1) gave identical results.

**CART models for individual sites**

CART models were fit for all but one of the sites, where AUT BEC diagnosis was given to all participants. The site-specific tree models were fit using (i) only diagnostic scales and clinicians’ characteristics and (ii) diagnostic scales, clinicians’ characteristics and demographic/behavioral variables. When models (i) were fit with priors equal to the distribution of BEC diagnosis in the individual sites, the most frequent predictor used as a first split was ADOS-S+C, followed by ADOS-CSS, ADOS SA and ADI-R Social; ADI-R verbal and nonverbal communication were used at the branches following the first split. The results of CART models (i) are not shown here.

The results from CART models (ii) for the 11 sites are summarized in eTable 1 and shown in eFigure 4. eTable 1 shows the predictors on which the first and second splits in the individual sites were made (these can also be seen in eFigure4). In addition to the selected predictor, the four next best variables are also shown in eTable 1. Verbal IQ appeared as a selected variable on the first split in 5 of the 11 sites and was among next best variables in 5 other sites. The cut-offs for the splits, however, were widely different between sites – from Verbal IQs of 62 to 127 (shown in eFigure 4). The second most often selected variable at the first split was ADOS-S+C (4 sites) and it was among the next best predictors at 7 other sites; the cut-off values for this variable did not vary much – from 11 to 13 in sites where it was the first split. ADOS RRB was selected on the first split two times with cut-offs at 3 and 5 respectively. The CART models for individual sites provided further evidence for idiosyncrasies and different criteria for making BEC diagnoses in these 12 sites.

In addition to the variables used in the splits and the next highest ranking ones, eTable 1 also reports the misclassification error resulting from the CART models as well as how much improvement in misclassification was due to the model. Notice that because the proportions of different BEC diagnostic categories were different across sites, the initial misclassification errors (with no model) were also different. For example, in a site where 60% of the cases are AUT, the initial misclassification rate is 40%, and if the model improves this error by 50%, the model misclassification rate will be 20%. On the other hand, in a site where 80% of the cases are AUT, the initial misclassification rate is 20% and a model that does not improve the classification will also have a misclassification error of 20%. Thus, for site “a,” for example, the improvement in misclassification rate was 58% compared with no predictors and this resulted in a model error rate.
of 12% [from here we can compute the initial misclassification rate in site “a” as 0.29=0.12/(1-0.58)].

References

Table 1. Summary of CART models for individual sites

<table>
<thead>
<tr>
<th>Site</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Verbal IQ</td>
<td>ADOS S+C</td>
<td>ADOS S+C</td>
<td>Verbal IQ</td>
<td>ADOS S+C</td>
<td>Verbal IQ</td>
<td>ADOS RRB</td>
<td>ADOS S+C</td>
<td>Verbal IQ</td>
<td>ADOS S+C</td>
<td>ADOS RRB</td>
</tr>
<tr>
<td>1st split</td>
<td>ADOS module</td>
<td>Vineland Composi te</td>
<td>CSS</td>
<td>Vineland Composi te</td>
<td>CSS</td>
<td>ADI-R NV Comm</td>
<td>ADOS S+C</td>
<td>CSS</td>
<td>ADOS S+C</td>
<td>ADOS S+C</td>
<td>ADOS S+C</td>
</tr>
<tr>
<td></td>
<td>ADOS S+C</td>
<td>ADI-R Social</td>
<td>Verbal IQ</td>
<td>ADOS S+C</td>
<td>Nonverb al IQ</td>
<td>ADOS RRB</td>
<td>Maternal education</td>
<td>Verbal IQ</td>
<td>Nonverb al IQ</td>
<td>ADOS module</td>
<td>Verbal IQ</td>
</tr>
<tr>
<td></td>
<td>ADOS RRB</td>
<td>Verbal IQ</td>
<td>ADOS RRB</td>
<td>CSS</td>
<td>Verbal IQ</td>
<td>ADOS S+C</td>
<td>CSS</td>
<td>ADOS module</td>
<td>CSS</td>
<td>Nonverb al IQ</td>
<td>ADOS module</td>
</tr>
<tr>
<td></td>
<td>Nonverb al IQ</td>
<td>Nonverb al IQ</td>
<td>ADI-R RRB</td>
<td>ADOS RRB</td>
<td>ADOS RRB</td>
<td>Nonverb al IQ</td>
<td>ADOS module</td>
<td>ADOS RRB</td>
<td>ADOS RRB</td>
<td>Vineland ABC</td>
<td>CSS</td>
</tr>
<tr>
<td>2nd split</td>
<td>Age</td>
<td>ADOS</td>
<td>Vineland Composi te</td>
<td>Nonverb al IQ</td>
<td>CSS</td>
<td>ADOS RRB</td>
<td>ADI-R Social</td>
<td>Age</td>
<td>ADOS</td>
<td>Verbal IQ</td>
<td>ADOS S+C</td>
</tr>
<tr>
<td></td>
<td>Vineland Composi te</td>
<td>Verbal IQ</td>
<td>CSS</td>
<td>Vineland Composi te</td>
<td>Verbal IQ</td>
<td>ADI-R NV Comm</td>
<td>Verbal IQ</td>
<td>CSS</td>
<td>CSS</td>
<td>Verbal IQ</td>
<td>Nonverb al IQ</td>
</tr>
<tr>
<td></td>
<td>ADOS S+C</td>
<td>ADI-R Social</td>
<td>ADOS S+C</td>
<td>ADOS S+C</td>
<td>ABC Hyperact</td>
<td>Age</td>
<td>ADOS</td>
<td>ABC Hyperact</td>
<td>ADOS module</td>
<td>Verbal IQ</td>
<td>ADOS RRB</td>
</tr>
<tr>
<td></td>
<td>ABC Hyperact</td>
<td>ADOS module</td>
<td>Verbal IQ</td>
<td>ADI-R Social</td>
<td>ADI-R NV Comm</td>
<td>ADOS S+C</td>
<td>Nonverba l IQ</td>
<td>RACE</td>
<td>Nonverb al IQ</td>
<td>ADOS module</td>
<td>Verbal IQ</td>
</tr>
<tr>
<td></td>
<td>ADI-R Social</td>
<td>ADI-R NV Comm</td>
<td>ADI-R RRB</td>
<td>ADOS RRB</td>
<td>ADI-R Social</td>
<td>ABC Irritabilit y</td>
<td>ABC Irritabilit y</td>
<td>ADOS S+C</td>
<td>Age</td>
<td>ADOS</td>
<td>Nonverb al IQ</td>
</tr>
<tr>
<td>n1</td>
<td>140</td>
<td>207</td>
<td>213</td>
<td>170</td>
<td>144</td>
<td>145</td>
<td>97</td>
<td>229</td>
<td>204</td>
<td>193</td>
<td>184</td>
</tr>
</tbody>
</table>

Misclas sification error2

| Improv | 12% | 18% | 13% | 17% | 18% | 18% | 13% | 11% | 19% | 24% | 18% |

Improv | 58% | 49% | 15% | 67% | 48% | 37% | 28% | 28% | 36% | 52% | 66% |
Note. Results from the first two splits in each site are given. The highlighted rows correspond to the actual selected variables; the next 4 variables within each split are the 2nd, 3rd, 4th and 5th best variables for this split, i.e., variables that provide the next best improvement in the model. ¹ Sample size for that site; ² Misclassification error of the CART model for that site; ³ Reduction of the error rate compared to no predictors; ADOS S+C=ADOS Social+Communication Domain Total; ADOS RRB=Repetitive Behavior Domain Total; CSS=ADOS Calibrated Severity Score; ADI-R NV Comm=ADI-R Nonverbal Communication Domain Total; ABC Hyperact=Aberrant Behavior Checklist Hyperactivity Total
**Figure 1.** Density and Tolerance Band Plots by Site and Overall

![Density plots and tolerance bands for selected diagnostic and developmental variables.](image)

Aut = Autism, PDQ-CD = Parental Developmental Disorder = Not Otherwise Specified, AD = Asperger's.
Figure 2. CART 1 for diagnoses with only diagnostic scales as predictors. 

Legend: AUT, Autism spectrum diagnosis; ADOS, Autism Diagnostic Observation Schedule; ADOS-Mod, Autism Diagnostic Observation Schedule-MODIFIED; ADOS-C, Autism Diagnostic Observation Schedule-Communication Domain Total; ADOS-COM, Autism Diagnostic Observation Schedule-Commentary; ADOS-RRB, Autism Diagnostic Observation Schedule-Repetitive Behavior Scale; Yr Exp, Years of Experience; Highest Degree, Patient's Highest Degree of Education.
Figure 3. CARS for diagnosis with diagnostic scales, site and demographic and behavioral characteristics as predictors.

Legend: 
- Predominantly Autism
- Predominantly POD-NOS
- Predominantly Asperger's

Legend for Scales: 
- ADOS: ADI-DAP (ADL, Social, Communication, Total) 
- ASP: ADOS-GS (ADL, Social, Communication, Total) 
- POD: POD-CAS (Communication, Social, Total) 
- PDD: PDD-CAS (Communication, Social, Total) 
- VIG: Vineland Adaptive Behavior Scales 
- VIG: Vineland Adaptive Behavior Scales

Legend for Behaviors: 
- ADOS: ADOS-CAS (Communication, Social, Total) 
- ADOS: ADOS-CAS (Communication, Social, Total) 
- POD: POD-CAS (Communication, Social, Total) 
- POD: POD-CAS (Communication, Social, Total) 
- VIG: Vineland Adaptive Behavior Scales

Legend for Ages: 
- Age (Chronological Age in Years)

Legend for Scales: 
- VIG: Vineland Adaptive Behavior Scales

Legend for Hyperactivity: 
- Hyperactivity Total

Legend for Site: 
- Site

Legend for Sex: 
- Sex

Legend for Race: 
- Race

Legend for Ethnicity: 
- Ethnicity

Legend for Disability: 
- Disability

Legend for Education: 
- Education

Legend for Other: 
- Other

Legend for Diagnoses: 
- Diagnoses

Legend for Other: 
- Other
Figure 4. CART for diagnoses by individual sites using all predictors.

N=1229 (excluding 1 site where all probands received AUT diagnosis, 1213) ADOS-SCQ+COM+ADOS Total + Communication Domain Total. ADOS-C3+COM+COM-C3 Severity Score: ADOS-NOS=ADOS NOS; ADOS-C3=ADOS Total; ADOS-C3+COM+COM-C3 Total. ADOS-C3+COM+COM-C3 Total = ADOS+COM+COM-C3 Total + ADOS-C3+COM+COM-C3 Total. ADOS-C3+COM+COM-C3 Total = ADOS-C3+COM+COM-C3 Total + ADOS-C3+COM+COM-C3 Total. ADOS-C3+COM+COM-C3 Total = ADOS-C3+COM+COM-C3 Total + ADOS-C3+COM+COM-C3 Total. ADOS-C3+COM+COM-C3 Total = ADOS-C3+COM+COM-C3 Total + ADOS-C3+COM+COM-C3 Total. ADOS-C3+COM+COM-C3 Total = ADOS-C3+COM+COM-C3 Total + ADOS-C3+COM+COM-C3 Total.
eFigure 4 (continued). CART for diagnoses by individual sites using all predictors.
N=1925 (emerging, e.g., where all gradates received AUT Diagnosis). ADOS-C: ADOS-C Clinical + Communication Domain Total; ADOS-CG: ADOS-C Global Severity Score; ADOS-Mod: ADOS-Module; ADI-Q: ADI-Q Social + Communication Total; ADI-NQ-C: ADI-Q Nonverbal Communication Total; ADI-RQ: ADI-Q Restricted and Repetitive Behaviors Total; VIQ: Verbal Intelligence Quotient; PIQ: Performance Intelligence Quotient; FSIQ: Full Scale Intelligence Quotient; ASD: Autism Spectrum Disorder; EA: Early Autism; PDD: Pervasive Developmental Disorder; NOS: Not Otherwise Specified; M.1: Mental Age; N.1: Nondelayed Intellectual Quotient; P.1: Psychoeducational Quotient; O.1: Operational Quotient; E.1: Executive Quotient; M.2: Mental Age Equated; N.2: Nondelayed Intellectual Quotient Equated; P.2: Psychoeducational Quotient Equated; O.2: Operational Quotient Equated; E.2: Executive Quotient Equated; Verbal Communication: The ability to communicate in a verbal manner; Nonverbal Communication: The ability to communicate without using language; Restricted Repetitive Behaviors: The ability to engage in repetitive and restricted behaviors; Social: The ability to engage in social interactions; Communication: The ability to communicate effectively; Academic: The ability to perform academically; Adaptive: The ability to adapt to the environment; Intellectual: The ability to think and reason; Motor: The ability to perform physical tasks; Language: The ability to use language effectively.