eMethods.

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 is SAS® 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[3], 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[1]: 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=0.20]; 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 ware 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.