J Autism Dev Disord
DOI 10.1007/s10803-014-2290-8

ORIGINAL PAPER

Facial Structure Analysis Separates Autism Spectrum Disorders into Meaningful Clinical Subgroups

Tayo Obafemi-Ajayi · Judith H. Miles · T. Nicole Takahashi · Wenchuan Qi · Kristina Aldridge · Minqi Zhang · Shi-Qing Xin · Ying He · Ye Duan

Abstract Varied cluster analysis were applied to facial surface measurements from 62 prepubertal boys with essential autism to determine whether facial morphology constitutes viable biomarker for delineation of discrete Autism Spectrum Disorders (ASD) subgroups. Earlier study indicated utility of facial morphology for autism subgrouping (Aldridge et al. in Mol Autism 2(1):15, 2011). Geodesic distances between standardized facial landmarks were measured from three-dimensional stereo-photogrammetric images. Subjects were evaluated for autism-related symptoms, neurologic, cognitive, familial, and phenotypic variants. The most compact cluster is clinically characterized by severe ASD, significant cognitive impairment and language regression. This verifies utility of facially-based ASD subtypes and validates Aldridge et al.'s severe ASD subgroup, notwithstanding different techniques. It suggests that language regression may define a unique ASD subgroup with potential etiologic differences.

Keywords Autism · Cluster analysis · Language regression · Facial phenotype · Biomarker · Outcome indicators

Introduction

Autism Spectrum Disorder (ASD) comprises a group of complex neuropsychiatric disorders of childhood, diagnosed on the basis of the behavioral phenotype. The ASD phenotype is characterized by social deficits, impaired

Some preliminary result from this work is being presented as a poster presentation at the 2014 International Meeting for Autism Research (IMFAR) in Atlanta, Georgia, USA on May 14, 2014.

Present Address:
T. Obafemi-Ajayi
Applied Computational Intelligence Lab, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, 301 W. 16th St, Rolla, MO 65409, USA

T. Obafemi-Ajayi · W. Qi · Y. Duan (✉)
Department of Computer Science, University of Missouri, 201 Engineering Building West, Columbia, MO 65211, USA
e-mail: duanye@missouri.edu

J. H. Miles · T. N. Takahashi
Thompson Center for Autism and Neurodevelopmental Disorders, University of Missouri, 205 Portland Street, Columbia, MO 65211, USA

J. H. Miles
Department of Child Health, University of Missouri School of Medicine, One Hospital Dr, N712, Columbia, MO 65212, USA

K. Aldridge
Department of Pathology and Anatomical Sciences, University of Missouri School of Medicine, One Hospital Dr, M309 Med Sci Bldg, Columbia, MO 65212, USA

M. Zhang · S.-Q. Xin · Y. He
School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

Present Address:
S.-Q. Xin
College of Information Science and Engineering, Ningbo University, 818 Fenghua Road, Ningbo, Zhejiang, China

Published online: 29 October 2014

communication, and restricted and repetitive behavior patterns (American Psychiatric Association 2013). A recent study by Aldridge et al. (2011) using 3D facial imaging discovered structural differences between faces of children with ASD and typically developing children. They suggested that differences in facial morphology may reflect alterations in embryologic brain development. Within ASD they identified two clinically discrete ASD subgroups, using cluster analysis of the facial measurements. This current study validates the previous findings by using an alternative distance measurement and multiple clustering techniques to verify the power and utility of facially based ASD subtypes. Geodesic (surface) rather than Euclidean (straight) measurements and four exceptionally robust clustering techniques were utilized to determine whether similar or additional subgroups would be identified. Extensive use of mathematical algorithms for data selection, multiple cluster analysis techniques, validity and classification models optimized the results.

Experimental results from cluster analysis based on facial morphology using surface distance features revealed that the ASD cohort studied could be separated into three clusters. Examination of clinical data using mean and correlation analysis revealed that each of the three clusters demonstrated relatively distinctive clinical and behavioral traits. One of the clusters (Cluster 2) exhibited clinical traits similar to those described by Aldridge et al. (2011) in their subgroup 1. If these facial groups identify etiologically discrete subsets of ASD, their identification may allow clinicians and researchers to identify precise etiologic bases of the ASD. This study demonstrates the generalization of facial phenotypes as a viable biomarker for identifying ASD subgroups. The similarity of the results obtained show that it is not dependent on measurement type (Euclidean vs. geodesic) or the cluster technique. This confirms that facial measurements provide a replicable and important biomarker in autism.

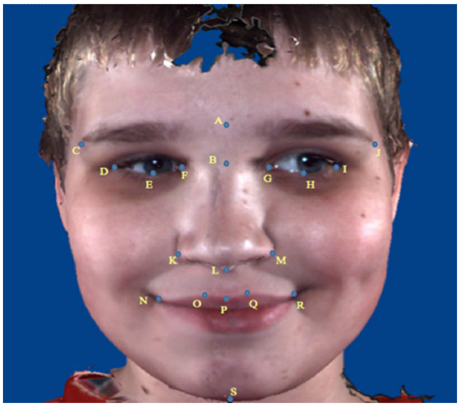


Fig. 1 Illustration of the 19 Farkas anthropometric landmark points used to derive facial surface distance features

Methods

Subjects

Sixty-two prepubertal Caucasian boys between 8 and 12 years of age, who had been diagnosed with ASD at the Thompson Center for Autism and Neurodevelopmental Disorders, were recruited for study. Forty-two subjects had also participated in the Simons Simplex Collection (SSC) and their clinical data were available for analysis. The remaining 22 subjects were recruited from the Thompson Center database, which contains similar clinical data. To ensure a homogeneous study set, all subjects were male, of Caucasian ancestry and old enough to have a mature facial and skull growth, but prepubertal to avert androgen surge effects on facial bone growth (Farkas and Posnick 1992) and classified as having essential autism. Boys with recognized genetic syndromes, including fragile X syndrome, chromosomal disorders, including copy number variants (CNV), generalized dysmorphology or gestational age less than 35 weeks were excluded. Generalized dysmorphology was assessed using the autism dysmorphology measure (ADM) (Miles et al. 2008). In addition, 83 % (52/63) of the ASD subjects overlap with the Aldridge cohort (Aldridge et al. 2011). A control group of 36 typically developing prepubertal Caucasian boys between 8 and 12 years of age were recruited from the community under the Thompson Center control subject recruitment protocol.

ASD Diagnosis

ASD diagnoses were made using the Thompson Center diagnostic protocol, which consists of complete clinical, medical, behavioral, and family histories, physical, neurologic and dysmorphology examinations, and autism diagnostic measures. Of the 62 boys with ASD, 42 had also completed the SSC protocol, which included the Autism Diagnostic Interview—Revised (ADI-R) (Lord et al. 1994) and Autism Diagnostic Observation Schedule (ADOS) (Lord et al. 2000). The 20 boys diagnosed exclusively through the Autism Medical Clinic were diagnosed on the basis of DSM-IV (American Psychiatric Association 2000) criteria (as appropriate during the time period of diagnosis) using a center-specific protocol based on the ADI-R, clinical observation and judgment of the clinician. Seventy-five percent also had an ADI-R or ADOS, which substantiated the Thompson Center diagnosis. ASD DSM-IV subtype diagnoses present within the study population were Autistic Disorder, Asperger Syndrome, and Pervasive Development Disorder-not otherwise specified (PDD-NOS).

This study was carried out under the guidelines and approval of the Health Sciences Institutional Review Board. The parents or legal guardians of all subjects provided written consent for participation in this study; each subject provided voluntary assent.

Data Acquisition

The 3DMD® Cranial system was used to reconstruct the 3D surface model (both the geometry and the co-registered texture image) of each subject, similar to previous work (Aldridge et al. 2011). We used the 3dMD software to obtain 3D coordinate data for a set of 19 anthropometric facial landmarks, as shown in Fig. 1, following (Farkas 1994). These landmark measurements were carried out by a rater (WQ) trained in use of the software program and verified by another rater (TO). Facial surface (geodesic) distances between all possible pairs of the 19 landmark coordinate points were computed to obtain a total of 171 facial distance features, as described in Fig. 1. For example, the distance from the midpoint between both eyes (B) to the midpoint of the chin (S) is designated as BS distance feature. Each subject's facial distance measurements were normalized by dividing them by the geometric mean of all the geodesic distances obtained for the subject.

3D Geodesic Distance Computation

Geodesic distance is defined as the shortest distance between any pair of anatomical landmark points along the surface of the face. Computing geodesics on polyhedral surfaces has been a fundamental problem in digital geometry processing and has been extensively studied. Representative work includes the Mitchell, Mount and Papadimitriou algorithm (MMP) (Mitchell et al. 1987) and the Chen and Han algorithm (CH) (Han 1990), which both compute the exact geodesic distance on triangle meshes. It is well recognized that this geodesic suffers from topological and geometric changes due to its local nature. For example, a small shortcut or miss-measurement may result in a significantly large change of the geodesic path and distance. In this work, we apply a global approach for the robust computation of geodesics on polygonal meshes (Quynh et al. 2012). This method takes a completely different strategy to compute the geodesic in an iterative and global manner, in contrast to the MMP and CH algorithms, which propagate the window (a data structure which encodes the distance) from the source to the destination.

To compute the shortest distance along the surface, the first iteration is initialized using the Euclidean distance, which is able to bridge small holes and gaps. For each iteration, our method computes the vector field X which matches the gradient of the current distance field, and normalizes X. Then it finds the closest scalar potential d by minimizing $\int_M |\nabla d - X|^2 dA$ over the entire mesh M, which is equivalent to solve a Poisson equation $\Delta d = div(X)$. These procedures are repeated until the convergence. This algorithm for Defect-Tolerant Geodesic (DTG) distance works quite well for the 3D face model, as the computed geodesics are very resilient to small topological and geometric noises (Xin et al. 2012). Hence, no pre-processing is required for smoothing or noise removal.

Clinical Data Evaluation

Each of the boys was evaluated for characteristics of their ASD diagnosis (social function, verbal function, repetitive behavior and language level), behavioral problems (aggression, attention deficits and self-injurious behaviors), out-come measures (IQ, communication, daily living skills, socialization and Vineland Adaptive Behavior Scale composite scores), the clinical course of their disorder (presence of regression at onset), medical and neurological variables (seizures, electroencephalogram results) and physical morphology (head circumference and dysmorphology).

Measures administered include the ADI-R (Lord et al. 1994), ADOS (Lord et al. 1989), Vineland Adaptive Behavior Scale II (Sparrow et al. 1984), an age- and development-appropriate IQ test (Full Scale IQ (FSIQ), Verbal IQ (VIQ), Nonverbal IQ (NVIQ)), Social Responsiveness Scale (SRS) (Constantino and Gruber 2005), and Broad Autism Phenotype (BAPQ) (Sasson et al. 2013). Parental alcohol use data was obtained using a Parent Substance Use questionnaire, based on the CAGE Assessment (Ewing 1984), for families who participated in the SSC project. A similarly detailed questionnaire was completed by parents of subjects recruited the Autism Medical Clinic. Alcoholism was defined as excessive use of alcohol, tolerance to high amounts of alcohol consumption and/or negative consequences to family, jobs or health (Wade et al. 2014). In addition, a detailed family history was obtained for all subjects by the clinician with extensive experience in the family history method. Not all measures of IQ were available for a small number of boys. All participants received complete medical and neurological examinations, including assessment of growth and dysmorphology.

The ADI-R and ADOS, which are considered the gold standard diagnostic instruments, measure the amount of impairment for the three autism core symptom areas; 1. Social functioning, 2. Communication, both verbal and nonverbal and 3. Repetitive behaviors. Higher scores indicate greater symptom severity. For each area of impairment a numeric score specifies the cut-off above which an ASD diagnosis is indicated. SRS score, developed for children between 4 and 18 years, measures the severity of autism spectrum symptoms that occur in social settings. It assesses social awareness, social information processing, capacity for reciprocal social communication, social anxiety/avoidance, and autistic preoccupations and traits. BAPQ scores are used to assess relatives of the study subject for language and personality characteristics diagnostic of a broad autism phenotype. The three subscales quantitatively measure characteristics that correspond to the diagnosis of autism in the DSM—IV: social deficits, stereotyped-repetitive behaviors, and social language deficits.

Cluster Analysis

Cluster analysis is the identification of groups of observations that are cohesive and separated from other groups (Fraley and Raftery 2000). Our goal was to identify clusters of boys with similar facial morphological features within the ASD dataset that correlate with clinical and behavioral traits. Our hypothesis is derived from previous work (Aldridge et al. 2011) that suggested differences in facial morphology reflect alterations in embryologic brain development in children with ASD compared to typically developing children as well as suggesting potential etiologic differences. A variety of clustering algorithms can be used to separate a finite unlabeled data set, like ours, into a finite and discrete set of "natural," hidden data structures (Xu and Wunsch 2005). We chose 4 different clustering algorithms to apply to our dataset: expectation maximization (EM) (Fraley and Raftery 2000), self-organizing feature map (SOM) (Kohonen 1998), K-means (Hartigan and Wong 1979), and partitioning around medoids (PAM) (Kaufman and Rousseeuw 1990).

EM algorithm is a well-known general-purpose machine learning technique for clustering. It is a model-based method. EM assigns a probability distribution to each instance, which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation (as done in all our experiments), or you may specify a priori how many clusters to generate. We implemented the EM algorithm using the Weka datamining tool (Witten and Frank 2005). SOM also is a model-based clustering method and uses a neural network approach. It maps all the instances (points) of a given dataset in a high-dimensional source space into a 2 to 3-d target space, such that, the distance and proximity relationship among the examples in the dataset are preserved as much as possible. The objective of SOM is to represent high-dimensional input patterns with prototype vectors that can be visualized in a two-dimensional lattice structure (Xu and Wunsch 2005). Each unit in the lattice is called a neuron, and adjacent neurons are connected to each other, which provide clear topology of how the network fits itself to the input space. Input patterns are fully connected to all neurons via adaptable weights, and during the training process, neighboring input patterns are projected into the lattice, corresponding to adjacent neurons. The size of the lattice, i.e. the number of clusters (k), must be predefined. K-means is a very simple and widely used partition based clustering method (Jain 2010). K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. The goal of K-means is to minimize the sum of the squared error over all K clusters. K-means algorithm requires three user-specified parameters: number of clusters K, cluster initialization, and distance metric. The PAM clustering algorithm is also a partition based clustering method. PAM tries to avoid outlier sensitivity, a known fault of K-means, by using medoids (the most centrally located object in the cluster) as a reference point rather than the mean value of the objects in a cluster. Thus, PAM starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids, if it improves the total distance of the resulting clustering.

Different cluster configurations results were obtained by varying the input parameters of the four clustering algorithms when applied to the ASD data. However the question remains: how do we know which set of clusters is valid or best fit the data set and how many clusters actually do exist in the data? Cluster validity refers to formal procedures that evaluate the results of cluster analysis in a quantitative and objective fashion (Jain 2010). In cluster validation (Kovács et al. 2005), two measurement criteria have been proposed for evaluating and selecting an optimal clustering scheme: Compactness and Separateness. Compactness measures how close the members of each cluster are to each other. A typical measure of compactness is the variance. Separateness measures how separated the clusters are from each other. A good cluster algorithm result should yield clusters that are compact and well separated. The aim of cluster validation is to find the cluster partition set which is the most appropriate/optimal to the input dataset.

The cluster validity analysis platform (CVAP) Matlab tool (Wang et al. 2009) estimates the quality of the different clustering algorithms' results and attempts to determine statistically which set of clusters are optimal using multiple validity indices. There are different types of cluster validity indices that measure the quality of clustering results. Validation indices based on internal criteria assess the fit between the structure imposed by the clustering algorithm (clustering) and the data by itself. Thus, the clustering results are evaluated using the quantities and features inherent in the data set (Arbelaitz et al. 2013). For the evaluation of the multiple clustering results obtained on the ASD study population, we used four following internal criteria validation indices to measure the goodness of the clusters, since the underlying structure of the data is unknown.

1. Silhouette index (Rousseeuw 1987). This is a composite index that measures both the compactness (using the distance between all the points in the same cluster) and separation of clusters (based on the nearest neighbor distance). A larger average Silhouette index indicates a better overall quality of the clustering result.
2. Dunn index (Halkidi et al. 2001). A measure that maximizes the inter-cluster distances while minimizing the intra-cluster distances. A large value indicates the presence of compact and well-separated clusters. Thus, the maximum value is the optimal clustering result.
3. Davies-Bouldin (DB) index (Bolshakova and Azuaje 2003). This measures the average value of the similarity between each cluster and its most similar cluster. A lower DB index implies a better cluster configuration.
4. Calinski-Harabasz (CH) index (Dudoit and Fridlyand 2002). This measures between-cluster isolation and within-cluster coherence. Its maximum value determines the optimal clustering configuration.

Another approach to validating the number of clusters present in a dataset is to view clustering as a supervised classification problem, in which we must also estimate the "true" class labels (Tibshirani and Walther 2005). Given the output labels of a given clustering algorithm, we apply it to train and build classification models (classifiers). Our goal is to see how well the models can predict the labels using the output of the clustering algorithms. The basic idea is that 'true' class labels will improve the prediction strength of the classification models. Hence, the resulting "prediction strength" measure assesses the quality of the clustering results. We applied three different classification models (support vector machines (SVM) (Burges 1998), neural networks multilayer perceptron (MLP) (Jain et al. 1996), and random forest (RF) (Breiman 2001)) to the clustering results.

An essential aspect of all cluster analysis is feature selection/extraction. Using a large number of features (171 in our case) increases the likelihood of feature redundancy. The goal of feature selection is to remove irrelevant/redundant features by finding the minimal feature subset necessary and sufficient to support the target concept (Dash and Liu 1997). The feature subset should improve and not degrade prediction accuracy and be a fairly accurate representation of the original feature distribution. To determine which facial features were significant and discriminant among the 171 features, we applied three feature selection methods. parallel scatter search algorithm (García López et al. 2006), best first search (Xu et al. 1988), and linear forward selection (Gutlein et al. 2009). We validated the significance of the features by reapplying the classification models. We expected that the discriminant features would improve the prediction strength of the models or at least not degrade performance of the classifiers.

Table 1 Evaluation of clustering algorithms using internal criteria cluster validation measures

Clustering algorithm (no. of clusters)[a]	Cluster validation measures (index scores)			
	Silhouette	Davies-Bouldin	Calinski-Harabasz	Dunn
K-means (3)	0.12	**1.65**	**12.42**	0.84
K-means (4)	**0.13**	1.80	11.26	0.85
Expectation maximization (3)	0.12	1.91	12.21	0.81
Self-organizing feature map (4)	0.11	1.71	10.24	**0.87**
Self-organizing feature map (3)	**0.13**	1.88	12.41	0.85
Partitioning around medoids (3)	0.10	1.73	11.41	0.77

Best method according to each index is highlighted in bold

For all validation measures except Davies-Bouldin (DB), a higher score indicates better cluster configuration. For the DB index, a lower score implies better cluster configuration

[a] Number of clusters in algorithm output result. For example, K-Means (3) = 3 cluster K-means result

Statistical Comparisons

To determine significance of results obtained for the facially defined clusters, we evaluated the statistical differences between the clusters using the univariate one-way analysis of variance (ANOVA) test along with the Student's *t* test for continuous variables, χ^2 and test for categorical variables. The ANOVA test generalizes the Student's *t* test for between comparisons for multiple groups. Hence, in addition, we performed the student's *t* test for each distance measure for comparisons between each pair of clusters to gain insight into the significance of difference, where needed. The *t* test informs us on which pairs of clusters are actually statistically different since the ANOVA's *p* value only indicates that at least one cluster is statistically different from another.

Results

Choosing Optimal Set of Clusters

The EM algorithm (unlike K-means, SOM and PAM) can decide how many clusters to create by cross validation based on resampling (Fraley and Raftery 2000). Thus we ran the EM algorithm initially in this manner and it

Table 2 Evaluation of cluster separateness using all 171 facial distance features versus minimal 31 feature set in three classification models

Classification models	Classification accuracy (%)		Overall precision/ recall		Precision/recall per cluster					
					Cluster 1		Cluster 2		Cluster 3	
	All	31	All	31	All	31	All	31	All	31
Support vector machine (SVM)	91.94	95.16	0.92/0.92	0.96/0.95	0.90/0.93	0.91/1.0	0.92/0.86	1.0/1.0	0.94/0.94	1.0/0.83
Neural networks multilayer perceptron (MLP)	93.55	93.55	0.94/0.94	0.94/0.94	0.91/0.97	0.91/0.97	1.0/0.86	1.0/1.0	0.94/0.94	0.93/0.83
Random forest	88.71	91.94	0.91/0.89	0.92/0.92	0.81/1.0	0.88/0.97	1.0/0.79	1.0/0.86	1.0/0.78	0.94/0.89

Overall performance of the classfication models improved when trained and tested using minimal set of 31 features rather than all 171 features, except in the case of cluster 3 for SVM and MLP

estimated 3 clusters within the dataset. We also reran the EM algorithm with different values of k (number of clusters) from 2 to 7. (We did not go beyond 7 due to the limited size of the ASD data.) The best EM result (as determined by cluster validity indices) was for $k = 3$. Based on this, for the remaining three cluster algorithms we varied k from 3 to 7. The best EM result was compared to the 21 outputs from the other three algorithms (K-means, SOM, PAM). In Table 1, we compare the top six best results. Based on the internal criteria cluster validation indices, we selected the K-means output with $k = 3$ as the optimal cluster configuration. Those 3 clusters identified within the 62 subjects ASD dataset were designated Cluster 1 (29 %, 18 boys), Cluster 2 (23 %, 14 boys), and Cluster 3 (48 %, 30 boys).

Machine learning techniques and evaluation metrics were employed to verify the distinctness of the clusters by training and testing three different classification models (SVM, MLP and RF). Models were trained to classify using the entire set of 171 facial geodesic distance measures. Given the limited size of the dataset, a threefold cross-validation approach, that splits the dataset into 3 groups, was used. Thus, we train on two-thirds of the data and test on the reminder third. Results are average of the three separate runs (folds). Evaluation metrics used were Classification Accuracy, Precision (Positive Predictive Value), and Recall (Sensitivity). Classification accuracy is defined as the percentage of test set samples that are correctly classified by the model. Precision (exactness) measures the proportion of actual positives that are correctly identified by the model. Recall (completeness) is the ratio of correctly classified samples to total number of samples for a given class. We report both metrics to present a complete depiction of the overall performance of the models in terms of how precisely and completely it correctly identified each cluster on the average.

A minimal set of 31 features was derived by taking the mathematical union of output results from three feature selection algorithms. Using this minimal set of 31 facial distance measures resulted in improved performance for two models (SVM, RF), and equal performance for MLP (Table 2). This validates that 31 features provide a robust and discriminant representation of the entire 171 facial distance measures.

To obtain a visual description of cluster separation, we performed a Principal Component Analysis (PCA) on the 31 significant features. The distribution of the clusters using the first two principal component axes is shown in Fig. 2a. An illustration of the data using a dendrogram based on mean linkage is shown in Fig. 2b. Note that the dendrogram is only for visualization not interpretation of data, as hierarchical clustering methods were not applied to the data. Figure 3 shows the distribution of the clusters plus the control group. The control group overlaps strongly with the cluster 3, partially with cluster 1and not at all with cluster 2.

Facial Features Selection

The discriminant set of 31 facial geodesic features is illustrated in Fig. 4. Each feature is denoted as the distance from one anthropometric facial landmark to another. For instance, BS feature indicates the distance measure from the nasion (i.e. the midpoint of the forehead) to the gnathion (i.e. the chin point).

All 31 distance measures were also verified to be significant (p value less than 0.05) by doing a between comparison among all the clusters using the ANOVA test. In addition, we performed the student's t test for each distance measure for comparison between each pair of clusters. We were interested in identifying which means were statistically different among all possible pairs of the three clusters (clusters 1:2; clusters 2:3, clusters 3:1). 12 facial distance measures were ascertained as statistically significant among all three clusters, based on which features had a p value of less than 0.05 for the pair-wise student's t tests. We describe these 12 facial distances, which were informative for all three clusters, in detail by mean and standard

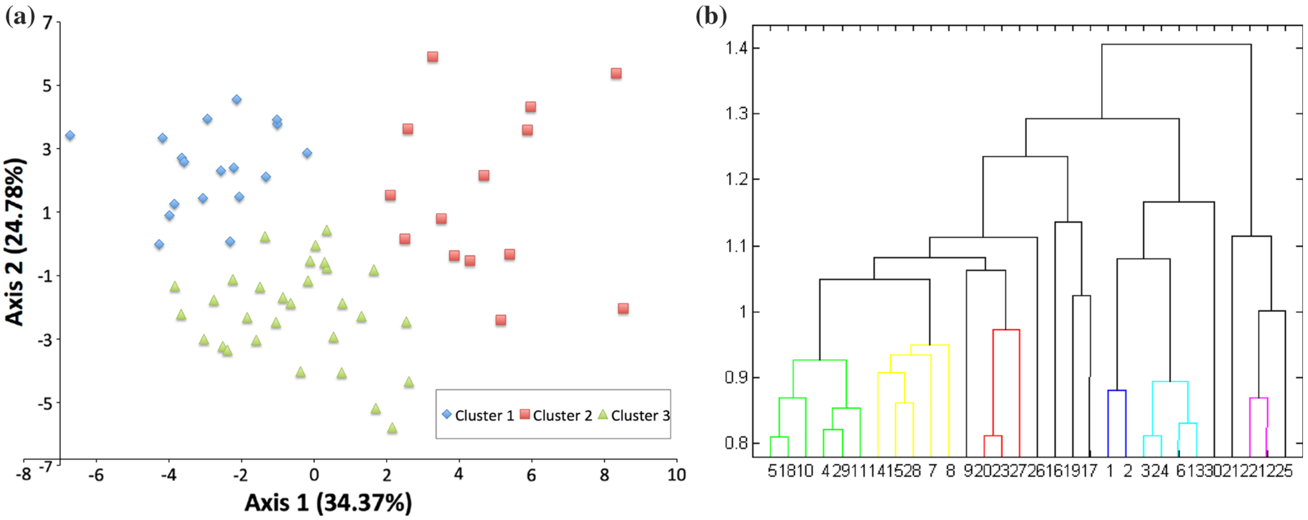
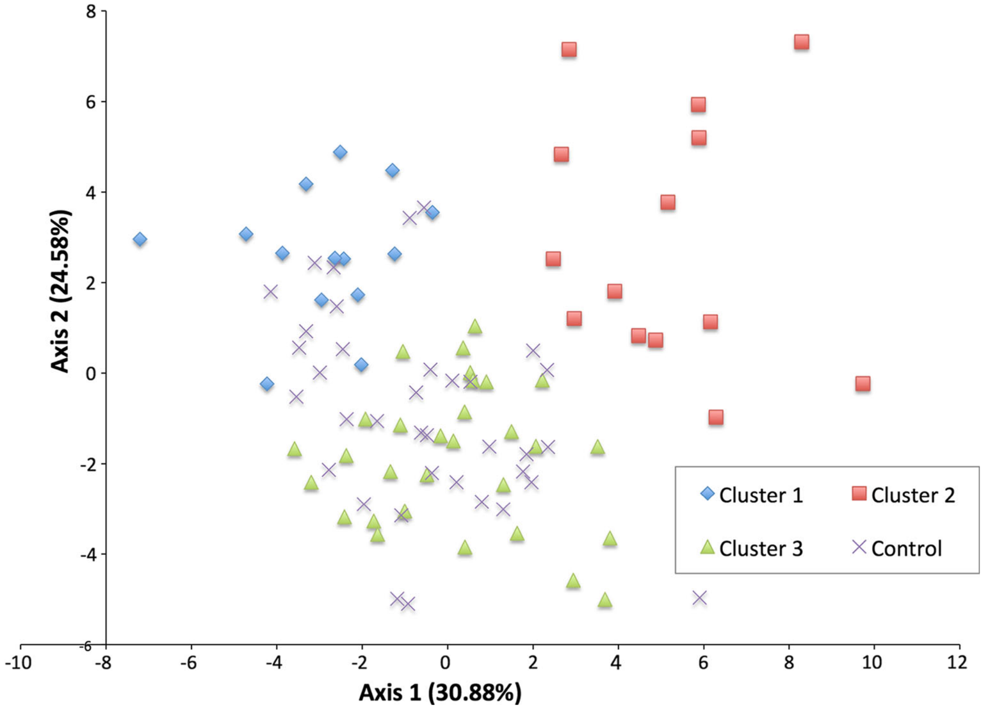


Fig. 2 Visualization of ASD clusters using **a** principal coordinates analysis plot of eigenscores for the first two principal axes. Axis 1 accounts for 34.37 % of the variance within the entire sample, and axis 2 accounts for 24.78 % of the variance. **b** Dendrogram based on mean linkage. *Note*: this is only for visualization not interpretation of data, as hierarchical clustering methods were not applied to the data

Fig. 3 Visualization of ASD clusters overlapped with the control group of 36 boys using principal coordinates analysis plot of eigenscores for the first two principal axes. Axis 1 accounts for 30.88 % of the variance within the entire sample, and axis 2 accounts for 24.58 % of the variance



deviation values (Table 3). Clustering results presented in Fig. 3 along with Table 3 validates cluster 2 group as a very compact and separate group among the ASD study population and the typically developing boys (the control group) using facial geodesic distance measures.

For each cluster, we identified which set of features were discriminant and useful in describing each cluster facially (Fig. 5). Cluster 1 is described by overall decreased surface facial heights (BS, ES, GS, HS, LS, MS, QS), combined with a broader maxillary midface from the temporal landmark to the lower nose landmarks (JK, JL). However, these individuals, demonstrate some overlap with the typically developing controls (Fig. 3). Interestingly, cluster 1 has the lowest standard deviation values for 7 of the 12 discriminant facial distances measurements (Table 3). This further verifies its compactness as the relatively low standard deviation values imply facial distances measured do not vary widely among the group (Fig. 2a). Cluster 2 subjects are facially defined by overall increased surface facial heights (BS, ES, GS, HS, LS, MS, QS), a decreased mid-face height (HS), and longest

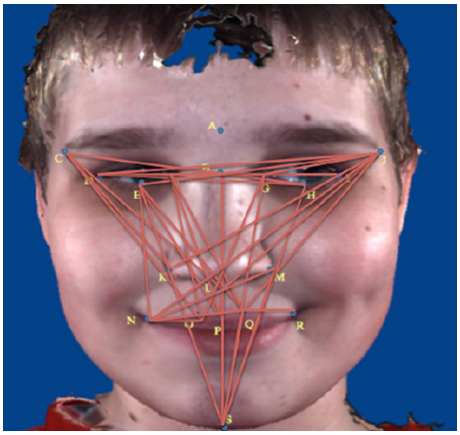


Fig. 4 Illustration of the 31 discriminant facial distance features for ASD clusters. 31 minimal discriminant features set from feature selection phase is illustrated on the face (*BS, EJ, GJ, LS, CH, EN, GO, MN, CK, EP, GS, MS, CL, EQ, HS, NO, CO, ES, JK, NQ, DH, FO, JL, QR, DI, FP, JN, QS, DJ, FQ,* and *JQ*). Though these distances are described using *straight lines*, they are not straight but rather the *shortest lines* along the surface from one landmark point to the other

mouth widths (NO and NQ). They also show no overlap with the control boys (Fig. 3). Cluster 2 is characterized by the most exaggerated facial features (Table 3; Fig. 5) among the ASD study population. For 11 of the 12 facial distance measures in Table 3, cluster 2 subjects either have the maximum or the minimum distance among the three clusters. Cluster 3 appears to be in between clusters 1 and 2, based on facial morphological features. They also have the smallest NO and NQ surface distances. Similar to cluster 1, cluster 3 individuals demonstrate considerable overlap with along with the typically developing boys. Thus, we observe that majority of the boys with ASD cluster with the typically developing controls, as also demonstrated by Aldridge et al. (2011).

Clinical Results

The goal of this study section is to determine whether the ASD subgroups defined by the cluster analysis are clinically distinctive. The clinical phenotype associated with each cluster is described based on five clinical areas. ASD diagnostic measures (ASD subsets, ADI-R and ADOS scores), outcome indicators (IQ, Adaptive Behavior, language), neurologic indicators (head size, seizures, electroencephalogram and brain Magnetic Resonance Imaging (MRI) results), family history (alcoholism, ASD symptoms), and clinical course (regression).

ASD Core Symptoms

ADI-R and ADOS scores, which indicate greater impairment with higher scores, were above ASD diagnostic cutoffs for subjects in each cluster affirming their autism diagnoses (Table 4). Social dysfunction, measured by the ADI-R, was most impaired in cluster 2, and significantly so compared with cluster 3. Cluster 2 also contained the highest percentage of nonverbal subjects; verbal subjects were more impaired in clusters 1 and 2 than in cluster 3. Repetitive behaviors were highest in clusters 1 and 2, with cluster 1 having a statistical significance over cluster 3. Consistent with the ADI-R, ADOS calculated severity scores were higher for clusters 1 and 2. Overall, individuals in cluster 2 were most impaired, though often not significantly from cluster 1. Cluster 3 was less symptomatic generally and with a wider range of scores suggesting a more heterogeneous subset of individuals.

Intelligence and Adaptive Behavior Scores

Though long-term functional outcomes are difficult to predict in ASD, IQ scores, language development and adaptive functioning provide some direction (Table 5). All intelligence scores (NVIQ, VIQ and FSIQ) indicate that boys in cluster 2 have significantly lower intelligence than those in either cluster 1 or 3. Cluster 1 presented the highest scores throughout though differences were not significantly different from those in cluster 3. Wide ranges and high standard deviations indicate significant heterogeneity in IQ and adaptive functioning in the 3 clusters. The Vineland II adaptive scores did not discriminate between the three clusters to the same degree as IQ (Table 6). Vineland Adaptive Scores were similar for the three groups with the exception of lower communication scores for cluster 2.

Clinical Course

A history of language regression at the onset of ASD symptoms in the first 3 years occurred in cluster 2 subjects more than twice as often as in clusters 1 or 3 (57.1 vs. 16.7 and 20 %). (Table 7). There was no significant difference in language regression between clusters 1 and 3. When regression history was compared with IQ there was a significant inverse association for all intelligence scores such that individuals whose ASD presented with regression had the lowest IQ scores (Table 8).

ASD Behavioral Subtype Diagnoses

Though Autism behavioral subtype diagnoses are no longer considered valid diagnostic indicators (Lord et al. 2012), these data are available and do convey some information about what the diagnosing clinicians thought about the subjects. Cluster 2 boys consisted of 79 % Autistic Disorder, 14 % as PDD-NOS, and 7 % as Asperger Syndrome. A key finding is that individuals in cluster 2 were

Table 3 Statistically significant facial distance measurements across clusters

Landmark	Indicates	Facial description	Cluster 1		Cluster 2		Cluster 3	
			Mean	SD	Mean	SD	Mean	SD
BS Nasion–gnathion	Facial height	Mid nasal bridge to chin point	↓2.05	0.05	↑2.25	0.08	2.14	0.07
ES Palpebrale inferius–gnathion	Facial height	Rt Mid eye to chin point	↓1.81	0.06	↑2.00	0.08	1.92	0.08
GS Endocanthion–gnathion	Facial height	Lf Inner canthus to chin point	↓1.87	0.05	↑2.05	0.08	1.98	0.07
HS Palpebrale inferius–gnathion	Facial height	Lf Mid eye to chin point	↓1.83	0.06	↑2.01	0.09	1.93	0.07
FP Endocanthion–labiale superius	Mid Face height	Rt Inner canthus to mid upper lip	1.15	0.04	↓1.13	0.04	↑1.21	0.03
JK Frontotemporale–alare	Mid Face breadth	Lf Lateral eye brow to Rt nasal edge	↑1.89	0.05	1.79	0.05	↓1.83	0.05
JL Frontotemporale–pronasale	Mid Face breadth	Lf lateral eye brow to nose septum	↑1.31	0.05	↓1.21	0.04	1.27	0.05
LS Pronasale–gnathion	Lower Face height	Nose septum to chin point	↓1.30	0.06	↑1.50	0.07	1.38	0.08
MS Alare–gnathion	Lower Face height	Lf lateral nose to chin point	↓1.19	0.07	↑1.42	0.09	1.26	0.08
QS Crista philtri–gnathion	Lower face height	Lf Cupids bow to chin point	↓0.90	0.07	↑1.15	0.09	0.96	0.08
NO Cheilion–crista philtri	Mouth width	Rt lateral mouth to Rt cupids bow	0.47	0.05	↑0.53	0.06	↓0.43	0.04
NQ Cheilion–crista philtri	Mouth width	Rt lateral mouth to Lf cupids bow	0.65	0.05	↑0.72	0.10	↓0.61	0.05

Significance of means of facial distances determined by univariate ANOVA test between the three clusters along with pairwise student *t* test ($p < 0.001$)

SD standard deviation

diagnosed primarily with Autistic Disorder (78.6 %) whereas cluster 1 (50 % Autistic Disorder, 44 % Asperger Syndrome, 6 % PDD-NOS) and cluster 3 (47 % Autistic Disorder, 33 % Asperger Syndrome, 20 % PDD-NOS) consist of a distribution of subtypes reflective of the total study population (55 % Autistic Disorder, 31 % Asperger Syndrome, 15 % PDD-NOS). Also, the Asperger diagnosis is closely correlated with IQ measurements, especially verbal IQ and verbal functioning. Separation of patients proposed in this paper provides subsets of patients based on a physical biomarker—facial morphology. Facially defined clusters reflect separation between more severely autistic children (previously grouped under Autistic Disorder) and less severe (previously grouped in Asperger Syndrome and PDD).

Neurologic Indicators

Complete data on neurologic indicators were available for seizures and head circumference. Seizures were more common in cluster 2 (28.6 %) than in clusters 1 (22.2 %) or 3 (10.0 %) though differences were not statistically different. This may reflect the small number and young age of the subjects. Head size measured by orbital occipital circumference and converted to Z scores for analysis, revealed no significant differences. Cluster 1 had the highest mean Z score (1.21) which was not statistically different from clusters 2 (0.87) and 3 (0.70). This indicated that the facial phenotypes were not driven by differences in head size. Head size groups' results for the facial distance defined clusters also showed that clusters are not related to macrocephaly, as the percentage of macrocephalic subjects in each cluster were similar (cluster 1–28 %, cluster 2–29 %, and cluster 3–20 %, all–24 %) and not statistically significant.

Genetic Indicators

Genetic indicators are those data that may provide insight into the genetic basis of ASD. These may include gender

Table 4 ASD core symptoms distribution by cluster

Diagnostic measures	Cluster 1 (18)	Cluster 2 (14)	Cluster 3 (30)
Social (ADI-R A) (cutoff = 10)			
Mean (SD)	23.27 (5.57)	25.58 (4.14)	19.80 (7.05)
Range	9–30	18–30	5–30
Three cluster comparison (p value)	**0.02**		
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	0.23	**<0.01**	0.09
Verbal scores (ADI-R B) (cutoff = 8)			
Mean (SD)	18.80 (3.19)	18.70 (2.50)	15.91 (5.06)
Range	14–24	16–23	7–23
Three cluster comparison (p value)	0.07		
Nonverbal scores (ADI-R B) (cutoff = 7)			
Percent of group measured by Nonverbal criteria	0.00 % (0)	14.29 % (2)	6.67 % (2)
Repetitive behavior (ADI-R C) (cutoff = 3)			
Mean (SD)	8.87 (2.47)	7.75 (1.71)	6.80 (2.47)
Range	4–12	5–10	2–12
Three cluster comparison (p-value)	**0.03**		
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	0.18	0.18	**0.02**
ADOS calculated severity scores			
Mean (SD)	7.47 (1.88)	7.55 (1.51)	6.52 (1.60)
Range	5–10	6–10	4–9
Three cluster comparison (p value)	0.15		

Statistically significant p-values are highlighted in bold

ADI-R data was available for 83 % of both clusters 1 and 3 and 86 % of cluster 2 while ADOS data was available for 83, 79 and 70 % of data for clusters 1- 3 respectively

Significance figure derived using univariate ANOVA test between the three clusters

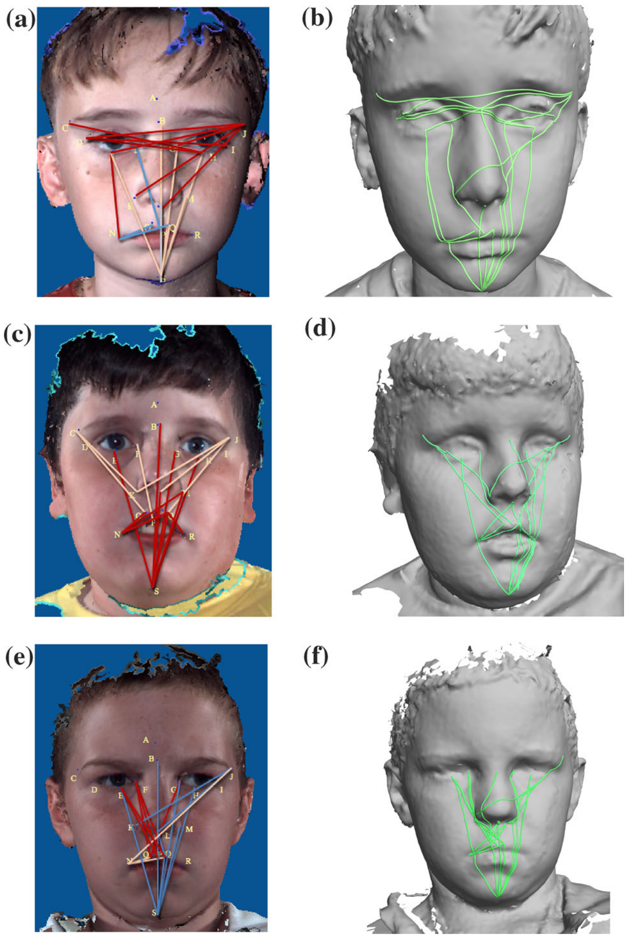


Fig. 5 Illustration of statistically significant facial distance measurements per cluster. **a** Cluster 1: 2D representation. **b** Cluster 1: 3D facial surface distance description. **c** Cluster 2: 2D representation. **d** Cluster 2: 3D facial surface distance description. **e** Cluster 3: 2D representation. **f** Cluster 3: 3D facial surface distance description. *Note*: Facial surface distance features are compared among the 3 clusters. *Red lines* indicate maximum, orange are minimum distances while *blue* imply distance is neither maximum nor minimum among the 3 clusters (Color figure online)

and family history of autism and related neuropsychiatric disorders. Social Responsiveness Scale (SRS) and Broad Autism Phenotype Questionnaire (BAPQ) scores which are designed to assess the number of autism symptoms in the parents of individuals with ASD were analyzed for the 42 SSC project boys. Though no significant differences were found between the clusters (Table 9), it is noted that in each of the three measures (SRS, BAPQ-Autism, BAPQ-Traits), mothers of boys in cluster 2 had somewhat higher scores, indicating possible genetic or epigenetic predisposition to develop an ASD. The portion of parents with alcoholism, which is known to be significantly higher than in families identified through an ASD (Miles et al. 2003), did not assort by cluster. Consistent with previously published data (Constantino and Gruber 2005), (Sasson et al. 2013), the paternal BAPQ scores for the ASD study population were significantly higher compared to the maternal scores (Table 9). The relationship to gender and other neuropsychiatric disorders could not be measured since all subjects were male and the SSC were precluded families with significant histories of ASD or major neuropsychiatric diagnoses. None of the subjects had a history of chromosomal or other autism related disorders.

Discussion

Children with ASD diagnoses comprise a heterogeneous population with a wide range in type, number and severity

Table 5 Intelligence scores by cluster

Outcome indicators	Cluster 1 (18)	Cluster 2 (14)	Cluster 3 (30)
Full Scale IQ[a]			
Mean (SD)	95.1 (18.60)	69.8 (25.98)	86.5 (21.58)
Range	68–127	31–112	38–130
FSIQ < 70	5.6 % (1)	42.9 % (6)	16.7 % (5)
FSIQ ≥ 70	72.2 % (13)	50.0 % (7)	63.3 % (19)
Three cluster comparison (p value)	**0.02**		
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	**0.01**	0.06	0.21
Verbal IQ[b]			
Mean (SD)	93.9 (20.68)	66.0 (29.66)	84.4 (26.50)
Range	65–121	13–112	23–126
VIQ < 70	11.1 % (2)	42.9 % (6)	20.0 % (6)
VIQ ≥ 70	66.7 % (12)	50.0 % (7)	56.7 % (17)
Three cluster comparison (p value)	**0.02**		
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	**0.01**	0.08	0.23
Non verbal IQ[c]			
Mean (SD)	94.8 (16.13)	73.7 (26.66)	92.3 (18.66)
Range	70–129	33–119	53–129
NVIQ < 70	0.0 % (0)	42.9 % (6)	10.0 % (3)
NVIQ ≥ 70	94.4 % (17)	50.0 % (7)	73.3 % (22)
Three cluster comparison (p value)	**0.01**		
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	**0.02**	**0.04**	0.64

Statistically significant p-values are highlighted in bold

Significance figure derived using univariate ANOVA test between the three clusters

[a] FSIQ scores were available for 78, 93, and 80 % of clusters 1–3 respectively

[b] VIQ scores were available for 78, 93, and 77 % of clusters 1–3 respectively

[c] NVIQ scores were available for 94, 93, and 83 % of clusters 1–3 respectively

Table 6 Vineland adaptive scores by cluster

Vineland II Scores	Cluster 1 (18)	Cluster 2 (14)	Cluster 3 (30)
Vineland Composite Score			
Mean (SD)	73.8 (11.6)	71.0 (8.10)	77.2 (9.98)
Range	57–95	56–84	56 -100
Three cluster comparison (p value)	0.31		
Communication			
Mean (SD)	77.9 (10.79)	70.2 (8.83)	80.3 (10.95)
Range	57–98	54–81	57–103
Three cluster comparison (p value)	**0.04**		
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	0.07	**0.01**	0.55
Daily living skills			
Mean (SD)	77.8 (14.29)	78.5 (13.90)	81.4 (13.66)
Range	59–101	62–109	58–117
Three cluster comparison (p value)	0.73		
Socialization			
Mean (SD)	69.7 (12.83)	69.3 (8.27)	73.9 (10.19)
Range	50–91	48–80	54–96
Three cluster comparison (p value)	0.39		

Statistically significant p-values are highlighted in bold

Significance figure derived using univariate ANOVA test between the three clusters

Vineland II scores were available for 67, 79, and 73 % of clusters 1–3 respectively. Vineland Composite scores were available for 67, 79, and 67 % of clusters 1–3 respectively

Table 7 Language regression by cluster

Language regression	Cluster 1 (18)	Cluster 2 (14)	Cluster 3 (30)	Total (62)
Language regression % (#)	16.7 % (3)	57.1 % (8)	20.0 % (6)	27.4 % (17)
Three cluster comparison (p value)	**0.02**			
Pairwise comparisons (Clusters 1:2, 2:3, 3:1) (p value)	**0.02**	**0.02**	0.24	

Statistically significant p-values are highlighted in bold

Regression data was available for all subjects

Significance figure derived using χ^2 test between the three clusters

of social deficits, behavior, communication, and cognitive difficulties which undoubtedly reflect multiple etiologic origins (Eaves et al. 1994). An initial step in search for etiologically discrete autism subgroups is discovery of phenotypic features that are present in some but not all ASD subjects, relatively discrete, quantifiable and pathophysiologically relevant (Miles 2011). We proposed that facial morphology, assessed by Euclidean and Geodesic distances between anatomical landmarks, could be used to reveal biologic homogeneity within ASD. Aldridge et al. (2011) showed that young boys diagnosed with ASD project a distinctive facial phenotype compared to typical controls. The ASD face was characterized by increased breadth of the upper face, orbits and mouth, a flattener nasal bridge and reduced height of the philtrum and maxillary region. Moreover, their data suggested biologic subsets that correlated with ASD severity.

Table 8 Correlation between language regression and IQ

	VIQ	NVIQ	FSIQ
Language regression (27.4 %, 17)	47.0	67.5	56.0
No language regression (72.6 %, 45)	90.0	93.9	91.6
p value*	<0.01	<0.01	<0.001

* p value reported in each column is based on using χ^2 test to compare mean IQ scores of subjects with no language regression to those that have

Table 9 Parental history of ASD symptoms and alcohol abuse by cluster

	Cluster1 (18)	Cluster2 (14)	Cluster3 (30)	*p* value*
Social Responsiveness Scale (SRS)				
Mother [mean (SD)]	27.7 (18.5)	38.6 (14.0)	30.1 (19.3)	0.31
Father [mean (SD)]	34.4 (31.0)	29.1 (20.0)	27.0 (20.3)	0.70
Broad autism phenotype (BAPQ)—autism				
Mother [mean (SD)]	2.3 (0.9)	2.6 (0.9)	2.2 (0.9)	0.62
Father [mean (SD)]	2.7 (0.7)	2.7 (0.9)	2.6 (1.1)	0.94
Broad autism phenotype (BAPQ)—traits				
Mother [mean (SD)]	81.3 (36.6)	88.5 (38.1)	81.0 (32.3)	0.84
Father [mean (SD)]	95.8 (23.4)	98.8 (28.9)	95.2 (34.9)	0.95
Alcoholism				
Mother alcoholic	44.4 % (8)	42.9 % (6)	40.0 % (12)	
Father alcoholic	50.0 % (9)	50.0 % (7)	50.0 % (15)	

68 % of the subjects were enrolled in the Simons Simplex Collection, thus they had no history of autism among 1st or 2nd degree relatives and no close relatives with major neuropsychiatric disorders

Raw SRS and BAP scores were available for 67, 79, and 73 % of Clusters 1–3 respectively

Alcoholism traits (parental history of alcohol abuse) data was available for all except 1 boy in cluster 3

* Significance figure derived using univariate ANOVA test between the three clusters

Our goals were to validate the Aldridge results and identify mathematically stronger clusters using additional statistical approaches. Identification of biologically valid, clinically distinctive subgroups is expected to expedite the search for autism genes and treatments. To minimize ASD's inherent heterogeneity, subjects were limited to Caucasian prepubertal boys, aged 8 to 12 with no significant dysmorphology or microcephaly. Facial distances were measured and mapped from three-dimensional stereophotogrammetric images of these boys. Each of the subjects was comprehensively evaluated for autism related symptoms, neurologic, cognitive, familial and phenotypic variants.

Three ASD subgroups were identified by cluster analysis based on geodesic distances between facial landmarks (Farkas 1994). Geodesic distance, defined as the shortest surface distance between anatomical landmarks, has been suggested as better suited to capture geometric structure of 3D models than Euclidean distance (Hamza and Krim 2006, Gilani et al. 2013). Our interpretation of the strength of the cluster analysis was based on four well-known internal criteria cluster validation indices (Silhouette, Dunn, Davies-Bouldin, Calinski-Harabasz) (Table 1). Cluster compactness is reflected by standard deviations (Table 3; Fig. 2), and separation of the clusters from each other is measured by prediction strength (as reflected by classification accuracy, sensitivity, and positive predictive value) of three classification models (Support Vector Machine, Neural Networks Multilayer Perceptron, and Random Forest) (Table 2; Fig. 3). Feature selection was also performed using established techniques (parallel scatter search algorithm, best first search, and linear forward selection) to select a subset of 31 geodesic distances that result in better classification and clustering of the data.

The three ASD subgroups, delineated by clusters 1, 2 and 3, have distinctive, though subtle, facial measurements. Cluster 1 is described by a reduction in facial height measures, combined with broader maxillary midface defined by temporal to lower nose landmarks conveys a shorter broader face. Cluster 1 faces are well separated from clusters 2 and 3 (as illustrated by the Principal Component Analysis—Fig. 2a); however, there is considerable overlap with typically developing subjects (Fig. 3). Features that describe cluster 3 faces include a shorter mid-face breadth, quantified by left lateral eye brow to right nasal edge, smaller mouth width and a decreased distance from the temporal area on the left to the outer edge of the right nasal alae, all of which portray a narrow face. Cluster 3 also has some overlap with the typical developing subjects.

Cluster 2 is mathematically the most distinctive and well-defined cluster (Tables 2, 3; Fig. 3). The faces are best described by an increased facial height measurements along the surface, with the exception of a shorter midface. Mouth widths are also wider. (Tables 2, 3; Fig. 3). Three supervised learning models (Support Vector Machine, Neural Networks Multilayer Perceptron, Random Forest) were used to verify the classification accuracy of the three clusters (Table 2). Using these models, we were able to almost perfectly train Support Vector Machine Classifier and Multilayer Neural Network Perceptron to identify cluster 2 correctly from the minimal set of 31 facial measurements. An F-measure of 1.0 indicates perfect classification. Cluster 2 also does not overlap with the control boys (Fig. 3). Thus, cluster 2 subjects not only show substantial cluster strength based on compactness and separateness criteria within the ASD population but also is distinct from the typically developing matched control group.

To determine whether these facial morphology based clusters would identify analogous clinical or behavioral subsets within the ASD diagnosis population, individuals in each cluster were assessed, using standard measures for ASD core symptoms, cognitive, adaptive, and language skills, ASD subtype diagnoses, type of ASD onset and parental autism broad phenotype indicators. *Cluster 2* subjects demonstrate the most coherent clinical phenotype with 79 % (11/14) described as Autistic Disorder, 14 % (2/14) as PDD-NOS, and 7 % (1/14) as Asperger Syndrome. They are clinically defined by significantly higher ADI-R A (Social) scores, (which implies a severe social diagnosis), severe verbal scores and an overall highest ADI CSS score. They also have the highest occurrence of non-verbal patients (14 %), the lowest IQ and Vineland II adaptive scores (except for daily living skills) in all categories. In addition to greater severity on autism measures, cluster 2 boys had more seizures (28 %) than boys in clusters 1 (22 %) or 3 (10 %). Moreover, this subgroup reported a likelihood of early language regression of 57 %, which is more than twice the frequency reported for clusters 1 (17 %) and 3 (20 %). The association of frequent language regression with cluster 2 and overall severity of Autistic Disorder diagnosis of these subjects provide additional evidence in line with Stefanotos' prognosis (2008). Stefanotos' findings suggest that the regressive subgroup of children with ASD may differ from the congenital form of the disorder in severity of behavioral symptoms and long-term prognosis, although he argues that more evidence is needed to justify them as a distinct subgroup with a distinguishable set of etiological considerations. Though sibling data was not available, severity of the maternal SRS, BAPQ—Autism, and BAPQ—Traits scores indicates an underlying genetic etiology for the individuals in cluster 2. Additional indicators of possible genetic differences between the clusters, including gender, autism and other psychiatric disorders in siblings and family members was not available because 68 % of ASD study population was from the Simons Simplex Collection (SSC). The SSC project recruited ASD patients based on exclusion of multiplex autism families and families with psychiatric disorders in close family members.

Clinical phenotype of *Cluster 1* subjects is described by 50 % (9/18) Autistic Disorder, 44 % (8/18) with Asperger Syndrome, and 6 % (1/18) with PDD-NOS. They are clinically defined by significantly higher (indicating greater severity) ADI-R C (Repetitive Behavior) scores. They have no occurrence of non-verbal patients along with the highest IQ and are the least likely group to experience language regression. Interestingly, this group also has the lowest Vineland II adaptive daily living skills scores. *Cluster 3* appears to represent the broad composition of children diagnosed with ASD. This is the largest subgroup (48 % (30/62)) with 47 % (14/30) of the boys described as Autistic Disorder, 33 % (10/30) as Asperger Syndrome, and 20 % as PDD-NOS. Clinically, they are defined by the lowest ADI-R scores in all the categories, which implies that this group has the least severe diagnosis socially, verbally, and repetitive behavior wise. This group also has the best Vineland II adaptive scores in all categories. However, this group has lower IQ scores compared to cluster 1 subjects, though much higher than cluster 2 subjects. This may be due to the presence of 2 (6.7 %) non-verbal boys in this group. There is a 20 % occurrence of language regression in this group. It is important to remember that both clusters 1 and 3 overlap with the control boys. These two clusters are clinically distinct from each other by their ADI-R scores with cluster 3 having better scores than cluster 1. Table 10 provides a clinical summary of each cluster in terms of the indicators/symptoms (autism core symptoms, cognitive, outcome, associated neurological symptoms and regression).

Table 10 Summary of clinical and behavioral severity levels for each cluster

Clinical/behavioral phenotypes	Cluster 1	Cluster 2	Cluster 3
Subjects			
Social competency (ADI-R)*	Severe	Most severe	Least severe
Verbal/communication (ADI-R, Vineland II)*	Severe	Most severe	Least severe
Repetitive behavior (ADI-R)*	Most severe	Severe	Least severe
ASD severity (ADOS)	Severe	Most severe	Least severe
ASD diagnostic subgroup (DSM-IV)	Asperger	Autistic disorder	PDD-NOS
Cognitive Level (VIQ, NVIQ, FSIQ)*	Highest	Lowest	High
Language regression (< year 3)*	Least frequent	Most frequent	Frequent
Parents			
SRS—Mother	Least severe	Most severe	Severe
BAPQ (autism)—Mother	Severe	Most severe	Least severe
BAPQ (traits)—Mother	Severe	Most severe	Least severe

* Comparison is significant, as determined by univariate ANOVA test between the three clusters

Our results are complimentary to previous study by Aldridge et al. (2011) performed on a similar, overlapping dataset (52 out of the 63 used previously in addition to 10 new boys) but with different research methodology. Key methodology differences are geodesic rather than Euclidean distance measurements, multiple clustering techniques versus principal component analysis; and two additional landmark points that further define measurements of facial height. In this study, we base our cluster separation decision solely on the ASD group in contrast to Aldridge et al., which includes separation from the control group as part of the cluster decision process. This report provides further evidence that the cluster results are strong, with a high degree of compactness and separateness not as easily appreciated as in the initial study. It is gratifying that both studies identified basically the same severe autism subgroup (Cluster 2 or Subgroup 1); characterized by severe ADI-R scores, low cognitive and functional IQ scores, highest maternal SRS scores and significant language regression. It is interesting to note that only 6 of the 12 boys identified by Aldridge et al. as belonging to the severe autism group (Subgroup 1) were included in our current study population. Based on our cluster analysis results, 5 of these 6 boys were included in our severe autism group (Cluster 2). Hence, both studies indicate that boys with ASD have altered development of their facial structure. In terms of Euclidean distance measurements, Aldridge et al. describes the severe autism subgroup with a decreased height of the facial midline and increased breadth of the mouth as well as the length and height of the chin. It is known that distance along the surface between two landmark points is not equivalent to the Euclidean distance between these points. Our findings indicate that cluster 2, our severe autism cluster, is characterized by an overall increased facial surface height measurements (with the exception of decreased mid-face height), and larger mouth widths compared to the measurements in individuals in clusters 1 and 3. This describes a longer face along the surface. Both studies indicate that distance measurements that describe decreased height of facial midline and long mouth widths are key biological traits for the severe autism group. This study demonstrates the generalization of facial phenotypes as a viable biomarker for identifying ASD subgroups, independent of measurement type (Euclidean vs. Geodesic) or cluster technique.

Our findings also indicate a strong association between language regression and cognitive performance, as indicated by IQ scores of cluster 2 as well as the entire study population. According to Table 7, 27 % of our study population has experienced language regression, which is consistent with the composition studied in literature about language regression in ASD (Jones and Campbell 2010). A pairwise comparison of the mean IQ scores in all three categories (VIQ, NVIQ, and FSIQ) between the regressed group and the non-regressed group (Table 8) reveals that the regressed group has significantly much lower IQ scores. Regression is a relatively common phenomenon in many pediatric neurologic disorders and has been linked to genetic diagnoses (Miles 2011). Though several reports have suggested that the eventual outcome in children with regression is that of a lower language level, lower IQ and lower adaptive level compared with those who do not regress, other studies have found no difference in outcome (Baird et al. 2008). Baird et al. found children with broad ASD diagnoses showed greater symptom severity in the presence of some language regression versus no regression. The outcomes from our study provide additional substantiation in support of a statistical correlation between language regression and cognitive performance.

Our findings also provide additional evidence that macrocephaly is an independent autism specific feature of autism. Head size results presented confirms that clusters are not related to macrocephaly, which is a relatively nonspecific finding in autism (Miles et al. 2000). Importantly, lack of association of head size with the clusters clearly indicates that brain and head growth are not the cause of the facial phenotypes.

The primary limitation of this study was the relatively small size and lack of some clinical data. Primary strength was the participation of mathematical and statistical scientists who designed a statistical approach that confirmed the validity of the cluster methodology.

Conclusion

Using comprehensive cluster analysis techniques, facial surface measurements were investigated in a cohort of 62 eight to twelve year old boys with essential ASD. Our results validate and extend the work of Aldridge et al. (2011) which showed for the first time that facial morphology differed significantly between groups of boys with ASD and matched controls and that subsets with distinctive facial morphology could be identified. Moreover, by using similar but different clustering methods, we also identified a comparable subset of boys with a classical autistic disorder phenotype characterized by lower IQs and Vineland Adaptive behavior scores, severe autism symptoms measured by gold standard autism diagnostic measures (ADI-R and ADOS), and more than twice likelihood of early language regression.

Based on these two studies, we assert that facial structure, based on 31 geodesic facial distances, should be considered a potentially useful biomarker to separate out a biologically discrete and homogeneous ASD subset for further study. This may help predict disorder severity and regression and has translational relevance as this ASD subset may represent genetically distinct individuals for whom specific treatment options may be tailored. Three dimensional facial imaging, which can be acquired with commercially available 3 dimensional systems already located in many university based tertiary care hospitals, should become a feasible autism biomarker with which to delineate homogeneous populations.

References

Aldridge, K., George, I., Cole, K., Austin, J., Takahashi, T. N., Duan, Y., et al. (2011). Facial phenotypes in subgroups of pre-pubertal boys with autism spectrum disorders are correlated with clinical phenotypes. *Molecular Autism, 2*(1), 15.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: APA.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243–256.

Baird, G., Charman, T., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., et al. (2008). Regression, developmental trajectory and associated problems in disorders in the autism spectrum: The SNAP study. *Journal of Autism Development Disorders, 38*(10), 1827–1836.

Bolshakova, N., & Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Processing, 83*(4), 825–833.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121–167.

Constantino, J. N., & Gruber, C. P. (2005). *Social Responsiveness Scale*. Los Angeles, CA: Western Psychological Services.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis, 1*(3), 131–156.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology, 3*(7), 1–21.

Eaves, L. C., Ho, H. H., & Eaves, D. M. (1994). Subtypes of autism by cluster analysis. *Journal of Autism and Developmental Disorders, 24*(1), 3–22.

Ewing, J. A. (1984). Detecting alcoholism: The CAGE questionnaire. *JAMA: Journal of the American Medical Association, 252*, 1905–1907.

Farkas, L. G. (1994). Anthropometry of the head and face in clinical practice. In L. G. Farkas (Ed.), *Anthropometry of the head and face* (2nd ed., pp. 71–77). New York: Raven Press.

Farkas, L. G., & Posnick, J. C. (1992). Growth and development of regional units in the head and face based on anthropometricmeasurements. *The Cleft Palate-Craniofacial Journal, 29*(4), 301–302.

Fraley, C., & Raftery, A. E. (2000). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.

García López, F., García Torres, M., Meliá Batista, B., Moreno Pérez, J. A., & Moreno-Vega, J. M. (2006). Solving feature subset selection problem by a parallel scatter search. *European Journal of Operational Research, 169*(2), 477–489.

Gilani, S. Z., Shafait, F., & Mian, A. (2013). Biologically significant facial landmarks: How significant are they for gender classification? In *IEEE international conference on digital image computing: Techniques and Applications (DICTA)* (pp. 1–8).

Gutlein, M., Frank, E., Hall, M., & Karwath, A. (2009). Large-scale attribute selection using wrappers. *IEEE symposium on computational intelligence and data mining CIDM'09, 2009* (pp. 332–339).

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Intelligent Information Systems Journal, 17*(2–3), 107–145.

Hamza, A. B., & Krim, H. (2006, August). Geodesic matching of triangulated surfaces. *IEEE Transactions on Image Processing, 15*(8), 2249–2258.

Han, J. C. (1990). Shortest paths on a polyhedron. *Sixth annual symposium on Computational geometry, SCG'90*, (pp. 360–369).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 100–108.

Jain, A. K. (2010). Data clustering: 50 Years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666.

Jain, A. K., Jianchang, M., & Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *Computer, 29*(3), 31–44.

Jones, L. A., & Campbell, J. M. (2010). Clinical characteristics associated with language regression for children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(1), 54–62.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kohonen, T. (1998). The Self-organizing maps. *Neurocomputing, 21*(1), 1–6.

Kovács, F., Legány, C., & Babos, A. (2005). Cluster validity measurement techniques. *6th International symposium of hungarian researchers on computational intelligence.*

Lord, C., Petkova, E., Hus, V., Gan, W., Lu, F., Martin, D. M., et al. (2012). A multisite study of the clinical diagnosis of different autism spectrum disorders. *Archives of General Psychiatry, 69*(3), 306–313.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C, et al. (2000, June). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., et al. (1989). Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal Autism Development Disorder, 19*, 185–212.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism Development Disorder, 24*, 659–685.

Miles, J. H. (2011). Autism subgroups from a medical genetics perspective. In *Autism spectrum disorders* (pp. 705–721). Oxford: Oxford University Press.

Miles, J. H., Hadden, L. L., Takahashi, T. N., & Hillman, R. E. (2000). Head circumference is an independent clinical finding associated with autism. *American Journal of Medical Genetics, 95*(4), 339–350.

Miles, J. H., Takahashi, T. N., Haber, A., & Hadden, L. (2003). Autism families with a high incidence of alcoholism. *Journal of Autism and Developmental Disorders, 33*(4), 403–415.

Miles, J. H., Takahashi, T. N., Hong, J., Munden, N., Flournoy, N., Braddock, S. R., et al. (2008). Development and validation of a measure of dysmorphology: Useful for autism subgroup classification. *American Journal of Medical Genetics Part A, 146A*, 1101–1116.

Mitchell, J. S., Mount, D. M., & Papadimitriou, C. H. (1987). The discrete geodesic problem. *SIAM Journal Computing, 16*(4), 647–668.

Quynh, D., He, Y., Xin, S.-Q., & Chen, Z. (2012). An intrinsic algorithm to compute geodesic distance fields on triangle meshes with holes, graphical models. *Proceedings of Geometric Modeling and Processing GMP'12, 74*(4), 209–220.

Rousseeuw, P. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics, 20*, 53–65.

Sasson, N. J., Lam, K. S., Parlier, M., Daniels, J. L., & Piven, J. (2013). Autism and the broad autism phenotype: Familial patterns and intergenerational transmission. *Journal of Neurodevelopmental Disorders, 5*(11).

Sparrow, S., Balla, D., & Cicchetti, D. (1984). *Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Service.

Stefanatos, G. A. (2008). Regression in autistic spectrum disorders. *Neuropsychology Review, 18*(4), 305–319.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics, 14*(3), 511–528.

Wade, J. L., Cox, N. B., Reeve, R. E., & Hull, M. (2014). Brief report: Impact of child problem behaviors and parental broad autism phenotype traits on substance use among parents of children with ASD. *Journal of Autism and Developmental Disorders*, 1–7.

Wang, K., Wang, B., & Peng, L. (2009). CVAP: Validation for cluster analyses. *Data Science Journal, 8*, 88–93.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Los Altos, CA: Morgan Kaufmann.

Xin, S.-Q., Quynh, D., Ying, X., & He, Y. (2012). A global algorithm to compute defect-tolerant geodesic distance. In *ACM SIGGRAPH ASIA 2012 Technical Briefs*, pp. 1–23.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678.

Xu, L., Yan, P., & Chang, T. (1988). Best first strategy for feature selection. In *Proceedings of ninth international conference on pattern recognition* (pp. 706–708).