FACIAL DYSMORPHISM IS INFLUENCED BY ETHNIC BACKGROUND OF THE PATIENT AND OF THE EVALUATOR

Aimé LUMAKA[1,2,3,4], Nele COSEMANS[1], Aimée LULEBO MAMPASI[5], Gerrye MUBUNGU[2,3], Nono MVUAMA[5], Toni LUBALA[6], Sebastien MBUYI-MUSANZAYI[6], Jeroen BRECKPOT[1], Maureen HOLVOET[1], Thomy de RAVEL[1], Griet Van BUGGENHOUT[1], Hilde PEETERS[1], Dian DONNAI[7], Leon MUTESA[8], Alain VERLOES[9], Prosper LUKUSA TSHILOBO [1,2,3,4], Koenraad DEVRIENDT[1]

Affiliations:

[1]Center for Human Genetics, University Hospitals Leuven, KU Leuven, Leuven Belgium

[2]Center for Human Genetics, Faculty of Medicine, University of Kinshasa, DR Congo

[3]Department of Paediatrics, Faculty of Medicine, University of Kinshasa, DR Congo

[4]Institut National de Recherche Biomédicale, DR Congo

[5]School Public Health, Faculty of Medicine, University of Kinshasa, DR Congo.

[6]Sendwe University Hospitals, University of Lubumbashi, DR Congo

[7]Manchester Centre for Genomic Medicine, Saint Mary's Hospital, Manchester Academic Health Science Centre, University of Manchester, Manchester, M13 9PL, United Kingdom

[8]Center for Human Genetics, College of Medicine and Health Sciences, University of Rwanda, Rwanda

[9]Département de Génétique, CHU Paris - Hôpital Robert Debré, Paris, France

Correspondence to:

Professor Koenraad DEVRIENDT, MD, PhD,

Centre for Human Genetics, University Hospitals Leuven, KU Leuven; Herestraat 49 BUS 602, 3000 Leuven, Belgium, E-mail: koenraad.devriendt@uzleuven.be

Tel secretary: + 32 16 34 59 03; Fax secretary: + 32 16 34 60 60

Conflicts of Interest

D.D. and K.D collaborated with FDNA® as member of the scientific advisory board. A.L., N.C., A.L.M., N.M., G.M., T.L., S.M.M., J.B., M.H., T.d.R., G.v.B., D.D, H.P., L.M., A.V., P.L.T. declare no conflict of interest.

Acknowledgements

AL received 2 travel grants from the FWO (Ref: V405213N and K210115) for patient recruitment in Kinshasa and a full-time IRO KU LEUVEN Scholarship. We acknowledge the collaboration of FDNA® team in the evaluation of Face2Gene solution.

ABSTRACT

The evaluation of facial dysmorphism is a critical step toward reaching a diagnostic. The aim of the present study was to evaluate the ability to interpret facial morphology in African children with intellectual disability (ID). First, 10 experienced clinicians (5 from Africa and 5 from Europe) rated gestalt in 127 African non-Down Syndrome (non-DS) patients using either the score 2 for "clearly dysmorphic", 0 for "clearly non dysmorphic" or 1 for "uncertain". The inter-rater agreement was determined using *kappa* coefficient. There was only fair agreement between African and European raters (*kappa*-coefficient = 0.29). Second, we applied the FDNA Face2Gene solution to assess Down Syndrome (DS) faces. Initially, Face2Gene showed a better recognition rate for DS in Caucasian (80 %) compared to African (36.8 %). We trained the Face2Gene with a set of African DS and non-DS photographs. Interestingly, the recognition in African increased to 94.7 %. Thus, training improved the sensitivity of Face2Gene.

Our data suggest that human based evaluation is influenced by ethnic background of the evaluator. In addition, computer based evaluation indicates that the ethnic of the patient also influences the evaluation and that training may increase the detection specificity for a particular ethnic.

Key words: facial dysmorphism, Down syndrome, DR Congo, Dysmorphology, Face2Gene, gestalt

INTRODUCTION

Since many syndromes have a characteristic craniofacial appearance, the recognition of facial dysmorphism makes an important contribution to reaching an etiological diagnosis. Facial dysmorphism results from the presence of multiple minor morphologic anomalies; several studies suggest that the presence of three or more minor anomalies in a patient can be used as an operational definition of 'dysmorphism' (reviewed by (1)). However, the evaluation of facial dysmorphism is hampered by several factors. For instance, the expected, "*normal*" facial features depend on the age and sex and may be influenced by the position during examination (static as in clinical pictures or dynamic when examining the individual live), by confounding factors such as strabismus and neurological manifestations or by the ethnic background. Considering that the evaluation of dysmorphism remains largely a subjective process requiring significant experience, one may question how much the ethnic background could influence its evaluation.

Studying dysmorphism in Central Africa is particularly challenging. First, certain features regarded as minor anomalies such as postaxial polydactyly, thick vermilion borders of the lips and a broad nasal tip, are common in Africans (2, 3). Second, the facial features of a known syndrome in a patient of African origin may differ from a Caucasian with the same syndrome, as reported for the Velocardiofacial syndrome (4, 5), Fragile-X syndrome (6) and fetal alcohol syndrome (7). These variations are thought to hinder the timely diagnosis of manageable diseases and may explain the apparently lower frequency of some disorders in non-Caucasian populations (5). Unfortunately, existing reference values for most quantitative traits as well as subjective description of many minor anomalies are largely based on Caucasians and are not available for Africans. This ethnic bias is further illustrated

by the fact that in a series of articles describing the terminology for minor anomalies, pictures from Africans are very underrepresented (8).

Of interest, efforts are ongoing to facilitate a more objective clinical evaluation of dysmorphism. Recent advances in morphometric analysis have explored the possibility to perform an objective evaluation of the facial *gestalt* from 2D or 3D facial images and reach a reliable syndrome diagnosis. This requires matching the face of a patient with similar patients in a database of individuals with known syndromes (9, 10). One such tool, Face2Gene (http://www.fdna.com/face2gene/) is available online, both as a web-based tool and as an application for smart-phones and tablets. The primary goal of these computed phenotyping tools is to narrow down the search space and to facilitate further confirmatory testing by excluding inconsistent diagnoses with great certainty. Such tools, being user-friendly, hold great promise of reaching a more rapid diagnosis for common genetic syndromes, and guide further confirmatory genetic testing more precisely (Gripp et al., 2016). Especially in low resource countries, where access to laboratory testing is limited, the potential of such tools is great.

The aim of the present study was to evaluate the ability to interpret facial morphology in children from Central Africa with intellectual disability (ID). First, we evaluated differences in the interpretation of facial *gestalt* in African children between experienced European and African dysmorphologists. Next, we assessed the performance of the existing computed phenotyping tool Face2Gene, at the current stage of its development, to recognize DS in Congolese versus Caucasian patients.

MATERIALS AND METHODS

This study was part of an etiological diagnostic study in index patients with ID, recruited in 6 specialized clinics and schools in Kinshasa in the DR Congo. A total of 127 patients were included, 33 females and 94 males aged between 1.24 and 24.65 years with a mean of 10.03 ± 4.68 years. We collected clinical information including photographs and obtained DNA from all of them. Of these, 19 had a genetically confirmed diagnosis of DS.

Comparison of scoring of facial (dys)morphology by African and European clinicians

Pictures of all 127 index cases (frontal view, and in most cases also a profile), taken during the clinical examination, were evaluated and rated by 10 clinicians, 5 from Africa (DR Congo and South Africa) and 5 from Europe (Belgium and United Kingdom). Raters were asked to score the face of each patient based on the facial *gestalt*, meaning the face as a whole without formal and detailed analysis. They could score the face as either normal (score 0), clearly dysmorphic (score 2) or uncertain (score 1). We hypothesized that African and European physicians would score in the same way (null Hypothesis) and used SPSS to calculate the Cronbach-**α** coefficient for intra-rater agreement and the *kappa coefficient* for inter-rater agreement (11). A Cronbach-**α** value higher than 0.7 reflects a satisfactory internal consistency. The interpretation of *kappa* coefficient was made as follows: less than chance agreement when < 0, slight agreement when 0.01–0.20, fair agreement if 0.21– 0.40, moderate agreement if 0.41–0.60, substantial agreement when 0.61–0.80 and almost perfect agreement when 0.81–0.99 (12). We were also interested in the correlation between the group-score that a patient received and the number of minor facial anomalies as observed during the clinical examination (conducted by K.D. and A.L). We used the Pearson correlation coefficient (r) to calculate the correlation and used the t-test to measure the significance. Interpretation of the Pearson correlation coefficient in terms of magnitude of

effect sizes was done according to Cohen J (13) as follows: “small” if r 0.10-0.30, “medium” when 0.30-0.50 and “large” for r of 0.50 and above.

Facial Dysmorphology Novel Analysis (FDNA) technology

The FDNA technology enables automatic detection and evaluation of subtle cranio-facial dysmorphism, as well as identification of recognizable facial patterns, associated with multiple rare diseases. Face2Gene is a phenotyping tool powered by the FDNA technology. The input consists of regular 2D facial frontal photos that are uploaded onto the tool. The recognition algorithm and the training process were described elsewhere (14). The predictive power to provide possibly matching syndromes increases as the tool is populated with more data extracted from diagnosed cases. This study was conducted between 20 March and 20 August 2015. On its public version, the output from Face2Gene is a list of 10 syndromes, referred to as “matches”, with the most matching syndrome ranked at top and the least matching at the bottom. For each match, Face2Gene also returns a heat map scaled such as the most distinguishing local markers are in red whereas the least ones are in blue (supplemental Figure 1). Face2Gene combines several face recognition methodologies, mainly the local appearance-based recognition and the *gestalt*-based recognition. With the local-based recognition, the system directly evaluates the probability of a match for every part of the face, based on visual cues. For the *gestalt*-based recognition, the system considers the entire face pattern, without focusing on specific regions.

Computed Evaluation of Congolese DS patients and training of Face2Gene

Facial photographs were obtained from 75 genetically confirmed DS patients with African ethnicity recruited from Kinshasa (DR Congo, n=24), Lubumbashi (DR Congo, n=11), Kigali (Rwanda, n=27) and Paris (France, n=13). A cohort of 109 non-DS patients, with African

ethnicity (26 females and 83 males), aged between 1.24 and 24.65 years (average 10.09 ± 4.76 years) were from the study on the etiology of ID mentioned above.

An African DS Test Set of 19 cases (12 boys and 7 girls, mean age of 10.06 years and range from 1.86 to 17.08 years) was used to evaluate the performance of Face2Gene before and after training.

The results were compared with a 54 African non-DS test set and a cohort of 20 Flemish DS test set, sex and age matched to the 19 African DS patients. The training set, comprising 55 African DS patients and 55 African non-DS, served to train the system as previously reported (14). During the training, the system depicted average appearances and constructed typical templates (masks) for each group namely African DS and African non-DS (Supplemental figure 1).

We recorded both the mean rank and the metrics mean(1/rank) results for each patient in the different test sets. In contrast to the public version, we had access to up to 100 matches.

Ethics

The participants were informed about the structure and aims of the study. For each participant, parents or legal representatives provided written consent for study participation. We applied an anonymous and non-personal coding system to protect participants' privacy. Our research protocol was approved under the number ESP/CE/008/2015 by the National Ethical Committee of the Public Health School of the University of Kinshasa, Kinshasa, DR Congo.

RESULTS

Comparison of the scoring by African and European clinicians

African physicians attributed more score "2" (clearly dysmorphic) (290 times) than Europeans (222 times) and conversely scored "0" (normal) less often (185 times) than their European colleagues (296 times) (Table 1). We defined the rater's cumulative score as the sum of the 127 individual scores given by each rater. Mean cumulative score from the African raters was higher, 148 ± 27.78 (range 110 to 176) compared to Europeans 112.20 ± 27.48 (range 84 to 142).

To assess whether the two groups of physicians had the same perception of dysmorphism in African patients, we calculated the inter-rater coefficient of agreement. The *kappa*-coefficient was 0.29, corresponding to a fair agreement between the 2 groups.

Correlation between score and number of minor anomalies on the face

We wished to evaluate whether the *gestalt*-based scoring for a patient correlated with number of minor facial anomalies. During the detailed dysmorphic examination (A.L, K.D.), all minor facial anomalies were recorded and counted. The Cronbach-**α** coefficient for intra-rater agreement was 0.76 within African raters and 0.85 within European raters, suggesting a satisfactory internal consistency. To derive the correlation between groups, we tailed the scores from the 5 raters for the number of minor anomalies and derived the mean and standard deviation (Figure 1). The Pearson correlation coefficient was 0.409 (p=0.000) for African raters and 0.417 (p=0.000) for Europeans, consistent with medium correlation effects in both groups.

Computed Evaluation of Congolese DS patients and training of Face2gene

Photographs of the three test sets were evaluated using Face2Gene, which reported DS within the first 10 matches in respectively 36.8 % (7/19) of African DS, 0% (0/54) of African non-DS and 80% (16/20) of Flemish DS cases (Table 2). The mean rank of the DS match was 23 in African DS, 65.7 in African non-DS and 7.2 in Flemish DS. After training the system using 55 African DS patients and 55 African non-DS, DS ranked among the first 10 matches in 94.7 % of African DS (mean rank 5.2), in 85.2 % in African non-DS (mean rank 9.4) and in 70 % of Flemish DS (mean rank 8.8) (Table 2).

Training of the system thus led to a significant increase in mean ranking of DS cases (table 2, figure 2). The metric of *mean(1/rank)* of the Congolese DS group improved from 0.23 before re-training to 0.88 after re-training, while the Flemish group score did not change significantly (0.52 before and 0.53 after re-training). However, the scores of the non-DS group also increased, from 0.02 before re-training to 0.42 after re-training.

DISCUSSION

In a first part of this study, we have compared the way African and European clinicians with experience in dysmorphology evaluate facial dysmorphism. In a series of 127 individuals with ID, there was a good correlation between the number of minor facial anomalies, assessed independently during the clinical examination (by K.D. and A.L.) and a "*gestalt*" evaluation of the facial dysmorphism by experienced clinicians. This indicates that a "*gestalt*" evaluation by an experienced clinician is a valuable alternative for a detailed evaluation and counting the number of minor anomalies. The higher the number of minor facial anomalies, the more likely a person was regarded as dysmorphic, which is not unexpected. Of interest, on average, European clinicians were less likely to score an individual as dysmorphic compared to the African clinicians, regardless of the number of minor facial anomalies. This means that

the ability to recognize dysmorphic features in African individuals was lower in European clinicians compared to African clinicians. We have no clear explanation for this observation, but this variable may need to be taken into account in dysmorphology studies in different ethnicities. It would be of interest to perform the inverse experiment, i.e. scoring faces of Caucasian children with ID by both groups, to see whether the same differences in scoring are present. It should be noted that there was significant consistency within each group of clinicians.

These results thus indicate that there is an important subjective component in the evaluation of (facial) dysmorphism. For this reason, there is interest in tools to obtain an objective evaluation or even a possible syndrome diagnosis. Online applications that can establish a syndrome diagnosis based on a facial picture of a person are becoming a reality. Such systems offer the potential for a universal, rapid and cheap syndrome diagnosis. The performance of such systems is obviously a critical factor. We tested the Face2Gene tool, because we wished to compare its performance in Europeans and Africans.

Our data indicate that the tool has a high precision in both groups. However, the accuracy in the Caucasian cohort was much higher compared to the African cohort. This is interesting, since it confirms that there are differences in facial appearance of Caucasian versus African DS patients. The most likely explanation why Face2Gene is underperforming in Congolese DS is that the tool is trained mostly with Caucasian cases. We therefore hypothesized that the performance will improve after training the system with more DS cases from Central Africa. This is exactly what we observed.

Our findings clearly show that increasing the number of training images of a specific syndrome from a specific ethnic group significantly improves the algorithm's ability to

discriminate the syndrome without degradation in the performance of the same syndrome for other ethnic groups. However, also the DS scores of the non-Down group increased significantly. We suspect that the reason behind was lack of training images from a comparable ethnic group affected with non-DS, as well as unaffected individuals. Therefore, training the algorithm with a relatively significant number of training images from African ethnicity without representing other syndromes as well, created a bias towards DS for all patients from this ethnic group. Thus we expect that the performance of Face2Gene will improve if an increasing number of cases from African ethnicity with a known diagnosis are uploaded.

The current experiment suffered from a number of limitations. First, the number of cases included was limited. Also the Face2Gene application is an emerging tool and the number of training cases included might still be insufficient to permit an optimal performance. Most importantly, from a clinical point of view, the cases we tested were young children and adolescents, but no neonates. An early, neonatal diagnosis of DS is important for correct counseling and adequate follow-up. Given the changing facial phenotype with ageing, more studies are needed to evaluate the performance of Face2Gene with neonates with DS.

REFERENCES

1. Hennekam RC. A newborn with unusual morphology: some practical aspects. Semin Fetal Neonatal Med 2011: 16: 109-113.
2. Talbert L, Kau CH, Christou T et al. A 3D analysis of Caucasian and African American facial morphologies in a US population. J Orthod 2014: 41: 19-29.
3. Ofodile FA, Bokhari FJ, Ellis C. The black American nose. Ann Plast Surg 1993: 31: 209-218; discussion 218-209.
4. McDonald-McGinn DM, Minugh-Purvis N, Kirschner RE et al. The 22q11.2 deletion in African-American patients: an underdiagnosed population? Am J Med Genet A 2005: 134: 242-246.
5. Veerapandiyan A, Abdul-Rahman OA, Adam MP et al. Chromosome 22q11.2 deletion syndrome in African-American patients: a diagnostic challenge. American journal of medical genetics Part A 2011: 155A: 2186-2195.
6. Schwartz CE, Phelan MC, Pulliam LH et al. Fragile X syndrome: incidence, clinical and cytogenetic findings in the black and white populations of South Carolina. Am J Med Genet 1988: 30: 641-654.
7. Moore ES, Ward RE, Wetherill LF et al. Unique facial features distinguish fetal alcohol syndrome patients and controls in diverse ethnic populations. Alcohol Clin Exp Res 2007: 31: 1707-1713.
8. Allanson JE, Cunniff C, Hoyme HE et al. Elements of morphology: standard terminology for the head and face. American journal of medical genetics Part A 2009: 149A: 6-28.
9. Hammond P, Hutton TJ, Allanson JE et al. Discriminating power of localized three-dimensional facial morphology. Am J Hum Genet 2005: 77: 999-1010.
10. Ferry Q, Steinberg J, Webber C et al. Diagnostically relevant facial gestalt information from ordinary photos. Elife 2014: 3: e02020.
11. Tang W, Hu J, Zhang H et al. Kappa coefficient: a popular measure of rater agreement. Shanghai Arch Psychiatry 2015: 27: 62-67.
12. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005: 37: 360-363.
13. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hilsdale, New Jersey: Lawrence Erlbaum Asociates, 1988.
14. Basel-Vanagaite L, Wolf L, Orin M et al. Recognition of the Cornelia de Lange Syndrome Phenotype with Facial Dysmorphology Novel Analysis. Clin Genet 2015.

Legends

Figure 1.

This graph shows the relation between the number of dysmorphic features (X axis, number patients with the corresponding number of minor facial dysmorphic features between brackets) and the average score (Y axis) obtained by all individuals with that number of dysmorphism, as scored by African raters (Blue dashed line) and Europeans (Orange solid line). The figure is read as follows: people with the number of minor dysmorphic features on the X axis, obtained on average the score on the Y axis. The mean score shows an increase as the number of minor anomalies increases.

Figure 2.

A. Facial dysmorphology novel analysis (FDNA) technology detection score curves of the three populations [Congolese DS (purple), Congolese control (red) and Flemish DS (green). The X-axis represents the DS score and the Y-axis represents probability. The three populations were best fit to a Gaussian distribution to extrapolate the score distribution of each group. The curves for each population were distinct. Separation was evident between the Congolese and African DS and the African non-DS curves
B. Retraining the system with African DS patients had a significant effect on the scores in the Congolese DS (purple) group but not the Flemish DS group (green). However, also the score of the Congolese Control Group (red) changed.

Legend Table 1. Columns 1 to 5 correspond to African raters and A to E to Europeans. Rows contain numbers of times the score in the first column has been used by the rater. For instance the African rater 1 attributed the score 0 to 58 pictures and all 5 African physicians used the score 0 a total of 185 times out of 635.

Figures

Figure 1. **Correlation between number of minor anomalies and average score**

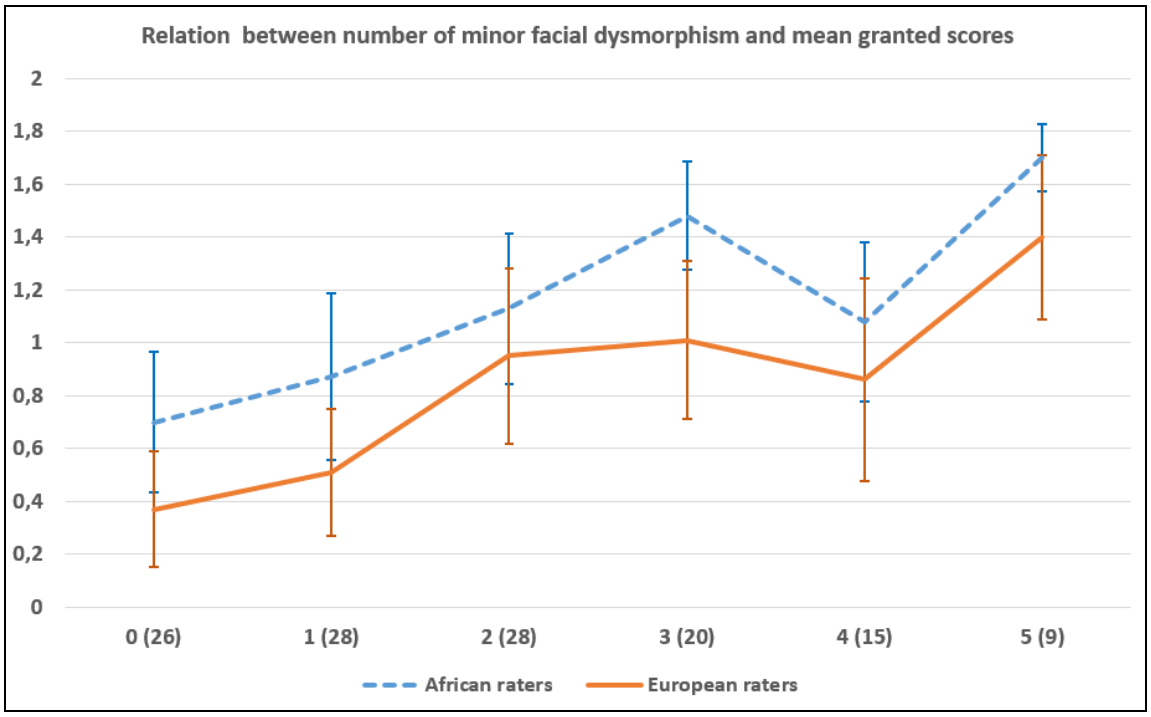
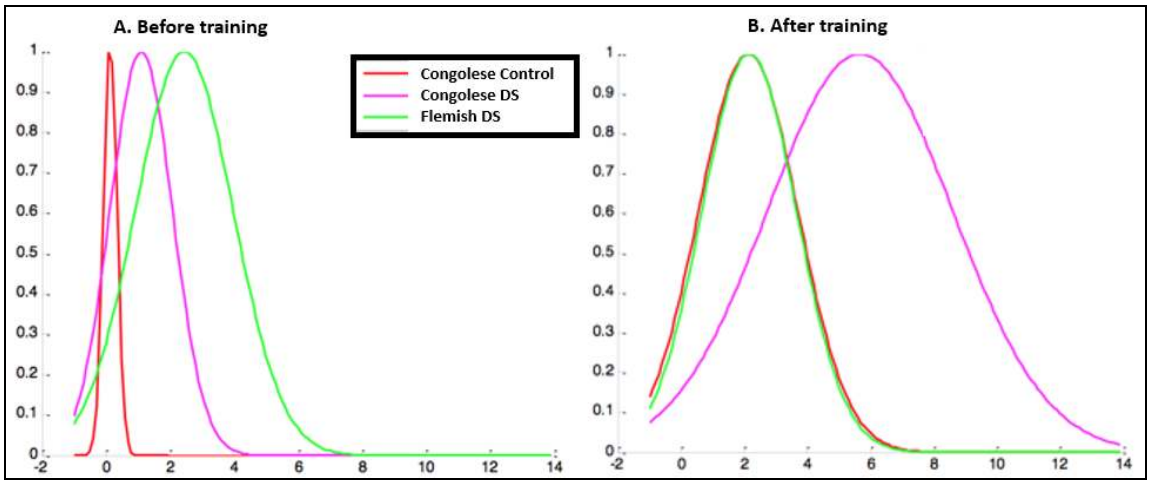


Figure 2. **The effect of re-training Face2Gene algorithms on the scores of DS**



Tables

Table 1. Comparison of the scoring by African and European clinicians

	African physicians						European physicians					
Raters	1	2	3	4	5	Total	A	B	C	D	E	Total
Score 0	58	26	20	42	39	**185**	80	54	75	48	39	**296**
Score 1	28	26	49	9	48	**160**	10	22	17	34	34	**117**
Score 2	41	75	58	76	40	**290**	37	51	35	45	54	**222**
Cumulative score	110	176	165	161	128	**740**	84	124	87	124	142	**561**

Table 2. Effect of the training on performance of Face2Gene

	Before training			After training		
	Af. DS	Af. Non-DS	Flemish DS	Af. DS	Af. Non-DS	Flemish DS
Ranks	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
0 to 10	7 (36.8)	0 (0.0)	16 (80.00)	18 (94.7)	46 (85.2)	14 (70.00)
11 to 20	3 (15.8)	3 (5.6)	2 (10.00)	0 (0.0)	4 (7.4)	4 (20.00)
21 to 30	4 (21.1)	2 (3.7)	1 (5.00)	0 (0.0)	1 (1.9)	0 (0.00)
31 to 40	2 (10.5)	7 (13.0)	1 (5.00)	0 (0.0)	0 (0.0)	1 (5.00)
41 to 50	1 (5.3)	4 (7.4)	0 (0.0)	0 (0.0)	0 (0.0)	1 (5.00)
51 to 60	0 (0.0)	4 (7.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.00)
61 to 70	0 (0.0)	4 (7.4)	0 (0.0)	0 (0.0)	1 (1.9)	0 (0.00)
71 to 80	1 (5.3)	9 (16.7)	0 (0.0)	1 (5.3)	2 (3.7)	0 (0.00)
81 to 90	1 (5.3)	15 (27.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.00)
91 to 100	0 (0.0)	6 (11.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.00)
Total	19	54	20	19	54	20 (100)