AIDY : application of Artificial Intelligence to Dysmorphology

Quentin Hennocq

Université Paris Cité

École doctorale : Frontières de l'Innovation en Recherche et Éducation (ED 474)

Laboratoire : UMR1163/Institut Imagine, Plateforme Data science / Laboratoire « Forme et croissance du crâne »

AIDY: Application of Artificial Intelligence to Dysmorphology

AIDY

Par **Quentin HENNOCQ**

Thèse de doctorat en TECHNOLOGIES D'APPLICATION MEDICALE, DIAGNOSTIQUES, THERAPIES ET SANTE PUBLIQUE

Dirigée par **Roman Hossein KHONSARI**

Présentée et soutenue publiquement le 13/12/2023

Devant un jury composé de :

Roman Hossein KHONSARI, Professeur des Universités Praticien Hospitalier, Université Paris Cité, directeur
Peter CLAES, Professeur des Universités, KU Leuven, rapporteur

Charles Patrick EDERY, Professeur des Universités Praticien Hospitalier, Université Lyon 1, rapporteur
Michael SUTTIE, Docteur en science, Oxford, examinateur
Chloé BERTOLUS, Professeur des Universités Praticienne Hospitalière, Université Paris Sorbonne, examinatrice
Stanislas LYONNET, Professeur des Universités Praticien Hospitalier, Université Paris Cité, examinateur
Arnaud PICARD, Professeur des Universités Praticien Hospitalier, Université Paris Cité, examinateur
Marjolaine WILLEMS, Praticienne Hospitalière, Université Montpellier, examinatrice
Nicolas GARCELON, Docteur en science, Université Paris Cité, membre invité

“Who sees the human face correctly: the photographer, the mirror or the painter?”

Pablo Picasso

Acknowledgements

First of all, I would like to thank my thesis supervisor, Prof. Roman Hossein Khonsari, for his trust and constant encouragement after 5 years of collaboration, as well as my co-supervisor, Dr. Nicolas Garcelon, who enabled me to join this fantastic project at the Imagine Institute.

I would like to thank Dr. Marlène Rio for initiating and directing the AIDY project, and for her help throughout these three years.

My warmest thanks to Prof. Peter Claes and Patrick Edery for agreeing to act as rapporteurs for this document. Thank you for your interest in my work.

I would like to thank Prof. Stanislas Lyonnet, Prof. Arnaud Picard, Prof. Chloé Bertolus, Dr. Michael Suttie and Dr. Marjolaine Willems for agreeing to take part in the defence jury.

I would like to thank the team at the Data Science platform at the Institut Imagine, and in particular Thomas Bongibault, Hassan Faour and Maxime Douillet for their help throughout the project.

I would also like to express my gratitude to Hélène Chautard, Pauline Seiter, Marion Pilorge and Antoine Ferry for their interest in my work, and for their considerable contribution to the valuation of the AIDY project.

I would also like to thank all the geneticists with whom I was able to have discussions to identify clinical problems in dysmorphology, in particular Prof. Jeanne Amiel, Prof. Tania Attié-Bitach, Prof. Valérie Cormier Daire, Dr. Sandrine Marlin and Prof. David Geneviève.

I would like to thank the various head and neck surgeons who have shown an interest in the AIDY project and who have shared with me databases of excellent quality, in particular Dr. Eric Arnaud, Dr. Giovanna Paternoster, Dr. Françoise Firmin and the entire team at the Maxillofacial and Plastic Surgery Department at Necker Hospital.

I would also like to thank the entire plastic surgery team at Great Ormond Street Hospital for the warm welcome they extended to me and the trust they placed in me.

I would like to thank all those with whom I interacted during the 3 years of my thesis, each of whom helped me to progress in their own way.

I would like to thank my friends and family for their constant support and encouragement. Thank you to Albane for the everyday motivation you give me.

Publications

Hennocq Q, Khonsari RH, Benoît V, Rio M, Garcelon N. Computational diagnostic methods on 2D photographs: A review of the literature. J Stomatol Oral Maxillofac Surg. 2021 Sep;122(4):e71-e75. doi: 10.1016/j.jormas.2021.04.003. Epub 2021 Apr 20. PMID: 33848665.

Hennocq Q, Bongibault T, Bizière M, Delassus O, Douillet M, Cormier-Daire V, Amiel J, Lyonnet S, Marlin S, Rio M, Picard A, Khonsari RH, Garcelon N. An automatic facial landmarking for children with rare diseases. Am J Med Genet A. 2023 May;191(5):1210-1221. doi: 10.1002/ajmg.a.63126. Epub 2023 Jan 30. PMID: 36714960.

Hennocq Q, Bongibault T, Marlin S, Amiel J, Attie-Bitach T, Baujat G, Boutaud L, Carpentier G, Corre P, Denoyelle F, Djate Delbrah F, Douillet M, Galliani E, Kamolvisit W, Lyonnet S, Milea D, Pingault V, Porntaveetus T, Touzet-Roumazeille S, Willems M, Picard A, Rio M, Garcelon N, Khonsari RH. AI-based diagnosis in mandibulofacial dysostosis with microcephaly using external ear shapes. Front Pediatr. 2023 Aug 17;11:1171277. doi: 10.3389/fped.2023.1171277. PMID: 37664547; PMCID: PMC10469912.

Bayard C, Segna E, Taverne M, Fraissenon A, **Hennocq Q**, Periou B, Zerbib L, Ladraa S, Chapelle C, Hoguin C, Kaltenbach S, Villarese P, Asnafi V, Broissand C, Nemazanyy I, Autret G, Goudin N, Legendre C, Authier FJ, Viel T, Tavitian B, Gitiaux C, Fraitag S, Duong JP, Delcros C, Sergent B, Picard A, Dussiot M, Guibaud L, Khonsari R, Canaud G. Hemifacial myohyperplasia is due to somatic muscular PIK3CA gain-of-function mutations and responds to pharmacological inhibition. J Exp Med. 2023 Nov 6;220(11):e20230926. doi: 10.1084/jem.20230926. Epub 2023 Sep 15. PMID: 37712948; PMCID: PMC10503430.

Hennocq Q, Garcelon N, Bongibault T, Bessières B, Bouygues T, Marlin S, Amiel J, Boutaud L, Douillet M, Loeuillet L, Lyonnet S, Perrot A, Pingault V, Picard A, Quélin C, Rio M, Willems M, Attie-Bitach T, Khonsari RH, Roux N. AI-based diagnosis in fetal pathology using external ear shapes. Submitted in October 2023 in *Ultrasound in Obstetrics and Gynecology*.

Hennocq Q, Paternoster G, Collet C, Amiel J, Bongibault T, Bouygues T, Cormier-Daire V, Douillet M, Dunaway DJ, Jeelani NO, van de Lande LS, Lyonnet S, Ong J, Picard A, Rickart AJ, Rio M, Schievano S, Arnaud E, Garcelon N, Khonsari RH. AI-based diagnosis and phenotype – genotype correlations in syndromic craniosynostoses. Submitted in October 2023 in *The American Journal of Human Genetics*.

Hennocq Q, Willems M, Amiel J, Arpin S, Attie-Bitach T, Bongibault T, Bouygues T, Cormier-Daire V, Corre P, Dieterich K, Douillet M, Feydy J, Galliani E, Giuliano F, Kamolvisit W, Lyonnet S, Picard A, Porntaveetus T, Rio M, Rouxel F, Toutain A, Yauy K, Geneviève D, Khonsari RH, Garcelon N. Next Generation Phenotyping for diagnosis and phenotype – genotype correlations in Kabuki syndrome. Submitted in October 2023 in *Genetics in Medicine*.

National scientific communications

Hennocq Q. Diagnostic tools for craniofacial malformations based on artificial intelligence: are morphometric and cephalometric approaches still relevant? Jean Delaire Congress, June 12th 2021, Paris, France.

Hennocq Q. Application of AI to dysmorphology (AIDY). Fetal pathology days (SOFFOET), September 24th, 2021, Paris, France.

Hennocq Q. Image-based deep learning for surgeons. 56th National Congress of the French Society of Maxillofacial Surgery, Stomatology and Oral Surgery, September 29-30th 2023, Besançon, France.

Hennocq Q. Application of AI to dysmorphology (AIDY). RARE Meeting, October 19th 2021, Paris, France.

Hennocq Q. AI-based diagnosis and understanding of craniofacial microsomia (CFM) phenotype from 2D photographs. Jean Delaire Congress, June 8-10th 2023, Nantes, France.

Hennocq Q. Introducing an AI-based phenotyping model on 2D photographs for syndromic craniosynostoses. 58th National Congress of the French Society of Maxillofacial Surgery, Stomatology and Oral Surgery, October 4-7th 2023, Toulouse, France.

Hennocq Q. AI-based diagnosis in fetal pathology using external ear shapes. 58th National Congress of the French Society of Maxillofacial Surgery, Stomatology and Oral Surgery, October 4-7th 2023, Toulouse, France.

Hennocq Q. Genetic diagnosis of craniofacial malformations using a smartphone - why not? 58th National Congress of the French Society of Maxillofacial Surgery, Stomatology and Oral Surgery, October 4-7th 2023, Toulouse, France.

International scientific communications

Hennocq Q. External ear malformation in Guion Almeida syndrome: insights from geometric morphometrics. 26th Congress of the European Association for Cranio Maxillo Facial Surgery, September 26 – 30th 2022, Madrid, Spain.

Awards

Jury award. "My thesis in 180 seconds", 57th National Congress of the French Society of Maxillofacial Surgery, Stomatology and Oral Surgery, October 12 - 15th 2022, Reims, France.

Best investor pitch award. Application of AI to dysmorphology (AIDY), Alcimed, March 17th 2023, Paris, France.

Award of Excellence. Application of AI to dysmorphology (AIDY), AAIHP (Association of former residents at Paris hospitals) "Surgery Section", March 20th 2023, Paris, France.

Award for the best scientific communication. AI-based diagnosis and understanding of craniofacial microsomia (CFM) phenotype from 2D photographs. Jean Delaire Congress, June 8-10th 2023, Nantes, France.

Award for the best scientific communication (Madmax price). Genetic diagnosis of craniofacial malformations using a smartphone - why not? 58th National Congress of the French Society of Maxillofacial Surgery, Stomatology and Oral Surgery, October 4-7th 2023, Toulouse, France.

Education and student supervision

Invited teacher at University Diploma "AI in Health" (Dr. Laurent Marchand), Université Paris Cité, practical work (3h), 2021-2022 and 2022-2023, Paris, France.

3D design in surgery, practical work (3h), Medicine School, University of Paris, 2021-2022 and 2022-2023

Supervision of medical and engineering students:

- Céleste DE LA FOUCHARDIERE, Master 1 degree, Medicine school, University of Poitiers, May – August 2020.
- Thomas DAI PRA, Master 1 degree, Medicine school, University of Paris, May – August 2020.
- Ombline DELASSUS, Master 2 degree, EPF, May – August 2021.
- Matthieu BIZIERE, Master 2 degree, EPF, February – July 2021.
- Thomas BONGIBAULT, Master 2 degree, ESME Sudria, February – July 2021.
- François-Gaël DJATE DELBRAH, Master 2 degree, ESME Sudria, February – July 2022.
- Thomas BOUYGUES, Master 2 degree, EPITA, February – July 2023.

Abstract

Background. Thirty to 40% of the 7000 rare diseases present with craniofacial anomalies. Identifying these facial features requires the expert eye of dysmorphologist, and diagnosis in this field is based on experience. Hence, there have been a recent increase in the number of publications dedicated to the automatic diagnosis of rare conditions using facial photographs, an approach termed Next Generation Phenotyping (NGP).

We aimed to develop and evaluate the performances of a new NGP method on 2D photographs on an unprecedentedly large database, with a wide array of genetically proven syndromes, of various ages, genders, and ethnicities.

Methods. We included pictures from the photographic database of the maxillofacial surgery and plastic surgery department and from the medical genetics department of *Hôpital Necker – Enfants Malades* (AP-HP), Paris, France. This database contains 1,042,468 photographs from 22,000 patients followed in the department since 1981. The writing of this work is based on the different stages in the construction of this new tool. We first described the first stage in analyzing photographs: (1) automatically detecting regions of interest, i.e., frontal, lateral and external ear pictures; and (2) automatically placing a series of landmarks on these regions. We then used a combination of shape analysis methods based on geometric morphometrics and texture analysis on key areas of the face. Finally, these geometric and textural parameters were used to train machine learning models based on a XGboost classifier. These models were validated on independent data, from other national (Nantes, Lille, Montpellier) and international (London, Bangkok) hospitals.

Results. Object recognition was optimized with the Faster R-Convolutional Neural Network (CNN) based detector. The best landmarking model was the patch-based Active Appearance Model (AAM) and was able to significantly distinguish patients with Treacher Collins (TC) syndrome from control non-syndromic patients ($p <$

0.001). We were then able to train a detection model for Guion Almeida syndrome (Mandibulofacial Dysostosis with Microcephaly, MFDM) based on geometric morphometrics of the external ear, with an accuracy of 0.969 [0.838 - 0.999] ($p < 0.001$) among non-syndromic controls, and 0.813 [0.544 - 0.960] ($p = 0.003$) among 3 differential diagnoses of this condition. Then, the incorporation of frontal and lateral facial analysis, as well as texture analysis, enabled the diagnosis of Apert, Crouzon and Pfeiffer syndromes with respective accuracies of 0.879 [0.718 - 0.966] ($p < 0.001$), 0.932 [0.813 - 0.986] ($p < 0.001$) and 1.000 [0.815 - 1.000] ($p < 0.001$).

Conclusion. We were able to build a robust NGP tool, allowing automated analysis of the facial phenotype on 2D photographs of children with various genetic syndromes. In addition, three main types of deliverables were obtained: (1) diagnostic performances for one or more genetic syndromes, (2) phenotype-genotype correlations for certain syndromes with multiple genetic variants, and (3) an analysis of the effects of surgery or drugs on facial morphology. We will now extend this algorithm, whose methodology has been completed and validated, to the analysis of the total number of syndromes in our database.

Keywords: computer vision; machine learning; Artificial Intelligence; dysmorphology; genetic diseases; craniofacial syndromes; automatic diagnosis; Next Generation Phenotyping.

Résumé

Contexte. Trente à 40 % des 7 000 maladies rares présentent des anomalies craniofaciales. L'identification de ces caractéristiques faciales nécessite l'œil expert d'un dysmorphologiste, et le diagnostic dans ce domaine repose sur l'expérience. C'est pourquoi le nombre de publications consacrées au diagnostic automatique des maladies rares à partir de photographies faciales a récemment augmenté, une approche appelée "Next Generation Phenotyping" (NGP).

Nous avons cherché à développer et à évaluer les performances d'une nouvelle méthode NGP sur des photographies 2D dans une base de données d'une ampleur sans précédent, avec un large éventail de syndromes génétiquement prouvés, d'âges, de genres et d'ethnies différents.

Méthodes. Nous avons inclus des images provenant de la base de données photographiques du service de chirurgie maxillo-faciale et de chirurgie plastique et du service de génétique médicale de l'Hôpital Necker - Enfants Malades (AP-HP), Paris, France. Cette base de données contient 1 042 468 photographies de 22 000 patients suivis dans le service depuis 1981. La rédaction de ce travail s'appuie sur les différentes étapes de la construction de ce nouvel outil. Nous avons tout d'abord décrit la première étape de l'analyse des photographies : (1) la détection automatique des régions d'intérêt, c'est-à-dire les photos frontales, latérales et de l'oreille externe ; et (2) le placement automatique d'une série de points de repère sur ces régions. Nous avons ensuite utilisé une combinaison de méthodes d'analyse de la forme basées sur la morphométrie géométrique et l'analyse de la texture sur des zones clés du visage. Enfin, ces paramètres géométriques et texturaux ont été utilisés pour entraîner des modèles d'apprentissage automatique basés sur un classificateur XGboost. Ces modèles ont été validés sur des données indépendantes provenant d'autres hôpitaux nationaux (Nantes, Lille, Montpellier) et internationaux (Londres, Bangkok).

Résultats. La reconnaissance des objets a été optimisée à l'aide d'un détecteur basé sur un réseau neuronal CNN (Faster R-Convolutional Neural Network). Le meilleur modèle de repérage était le modèle d'apparence active (AAM) basé sur les patchs et a permis de distinguer de manière significative les patients atteints du syndrome de Treacher Collins (TC) des patients témoins non syndromiques ($p < 0,001$). Nous avons ensuite pu entraîner un modèle de détection du syndrome de Guion Almeida (Dysostose mandibulofaciale avec microcéphalie, MFDM) basé sur la morphométrie géométrique de l'oreille externe, avec une précision de 0,969 [0,838 - 0,999] ($p < 0,001$) parmi les témoins non syndromiques, et de 0,813 [0,544 - 0,960] ($p = 0,003$) parmi les 3 diagnostics différentiels de cette pathologie. Ensuite, l'incorporation de l'analyse frontale et latérale du visage, ainsi que de l'analyse de texture, a permis de diagnostiquer les syndromes d'Apert, de Crouzon et de Pfeiffer avec des précisions respectives de 0,879 [0,718 - 0,966] ($p < 0,001$), 0,932 [0,813 - 0,986] ($p < 0,001$) et 1,000 [0,815 - 1,000] ($p < 0,001$).

Conclusion. Nous avons pu construire un outil NGP robuste, permettant l'analyse automatisée du phénotype facial sur des photographies 2D d'enfants présentant divers syndromes génétiques. De plus, trois principaux types de livrables ont été obtenus : (1) des performances diagnostiques pour un ou plusieurs syndromes génétiques, (2) des corrélations phénotype-génotype pour certains syndromes avec plusieurs variantes génétiques, et (3) une analyse des effets de la chirurgie ou des médicaments sur la morphologie faciale. Nous allons maintenant étendre cet algorithme, dont la méthodologie a été complétée et validée, à l'analyse de l'ensemble des syndromes de notre base de données.

Mots-clés : apprentissage automatique ; intelligence artificielle ; dysmorphologie ; maladies génétiques ; syndromes craniofaciaux ; diagnostic automatique ; Next Generation Phenotyping.

Résumé substantiel

Introduction. La dysmorphologie est une discipline de la génétique clinique qui étudie et tente d'interpréter les variations de structure et de croissance chez l'homme. La reconnaissance de traits phénotypiques parfois subtils nécessite une expertise clinique. Moeschler et al admettent que le diagnostic est retardé ou reste inconnu dans 38% des évaluations de dysmorphologie. Selon ces auteurs, trois raisons expliquent cette incertitude : la rareté de nombreux syndromes génétiques, le manque d'accès des patients aux services spécialisés en dysmorphologie et l'expérience variable des praticiens. De plus, la plupart des ouvrages académiques et des cas publiés dans la littérature ne concernent que des populations caucasiennes. Certaines anomalies mineures observées dans une population caucasienne, telles que l'épaisseur des lèvres ou une large implantation nasale, peuvent être considérées comme un trait phénotypique normal dans une population africaine. L'anomalie nasale observée dans le syndrome de délétion 22q11 chez les Caucasiens se retrouve dans 89 % des populations asiatiques, contre 15 à 40 % des populations africaines. Ces inégalités sont encore aggravées par le fait que les centres spécialisés en génétique clinique et en dysmorphologie sont plus concentrés dans les pays développés que dans la plupart des pays à faible revenu dont la population n'est pas caucasienne. Les photographies permettent de stocker des informations identifiées par le praticien en vue d'une utilisation ultérieure (analogie avec de nouveaux patients ou suivi de patients), ou à des fins pédagogiques. Les bases de données photographiques des départements de génétique clinique ou d'autres spécialités en contact avec les syndromes génétiques sont donc souvent d'une grande valeur pour la recherche. La base de données complète contient 1 042 468 photographies d'environ 22 000 patients.

Les maladies congénitales sont la cinquième cause de mortalité néonatale, représentant 270 000 décès dans le monde en 2010. On estime à 7000 le nombre de maladies rares, dont 80% sont d'origine génétique. 30 à 40 % de ces 7 000 maladies rares présentent des anomalies

craniofaciales. Ces anomalies peuvent être majeures, comme une fente labiale et une fente palatine, et facilement détectables, mais elles peuvent aussi être mineures et n'être détectées que par un praticien qualifié. Certains de ces syndromes entraînent une mort prématurée, le plus souvent en raison de malformations des organes associés, de difficultés respiratoires ou de problèmes d'alimentation. On peut donc supposer que 2100 à 2800 affections rares pourraient être dépistées par une analyse détaillée du visage.

Dans les approches d'apprentissage automatique, les règles prédéfinies sont remplacées par la capacité du modèle à apprendre à partir d'exemples. Les applications de l'apprentissage automatique se multiplient dans le domaine de la santé, dans divers domaines tels que la radiologie, la dermatologie et la chirurgie. Les données d'entrée pertinentes de l'apprentissage automatique pour l'analyse d'images dans le domaine de la santé sont des évaluations de radiographies, des comptes-rendus, des coupes de pathologie et des photographies cliniques. Parmi ces sources de données, les photographies cliniques sont, comme nous l'avons vu précédemment, particulièrement importantes pour le diagnostic des malformations craniofaciales.

Les deux objectifs principaux de ce projet étaient :

- diagnostiquer un syndrome, dans une classification binaire puis multi-syndromes ;
- mettre en évidence des différences au sein d'un même syndrome, c'est-à-dire établir une corrélation phénotype-génotype.

Différents syndromes seront testés au cours du manuscrit afin de répondre à des questions diagnostiques pratiques posées par les généticiens. Nous conclurons par une analyse en situation multi-syndromique, basée sur l'ensemble des affections que nous avons pu inclure au cours de la thèse, à savoir : Crouzon - Pfeiffer, Apert, Saethre Chotzen, Muenke, Kabuki, Treacher Collins, Nager, Guion-Almeida, CHARGE et Silver Russell.

Méthodes. Nous présentons ici le pipeline de fonctionnement de notre modèle : prétraitement des images, placement automatique des *landmarks*, extraction des caractéristiques géométriques et texturales, stratification à l'aide de métadonnées et classification à l'aide d'un modèle d'apprentissage automatique.

Dans le set d'entraînement, nous avons inclus des images provenant de la base de données photographiques des services de chirurgie maxillo-faciale et plastique et de génétique médicale de l'Hôpital Necker - Enfants Malades (Assistance Publique - Hôpitaux de Paris). Cette base de données contient 1 042 468 photographies de 22 000 patients suivis dans le service depuis 1976. Toutes les photographies ont été prises par un photographe médical professionnel. Nous avons inclus rétrospectivement et prospectivement, de 1976 à 2023, toutes les photographies frontales et latérales de patients diagnostiqués avec 9 syndromes. Tous les patients ont eu une confirmation génétique de leur syndrome. Nous avons exclu les patients ayant des antécédents de chirurgie faciale. Plusieurs photographies par patient correspondaient à des âges différents. Les enfants non syndromiques ont été sélectionnés parmi les patients admis pour des plaies, des traumatismes, des infections et diverses lésions cutanées, sans aucun antécédent de maladie chronique. Plus précisément, le suivi de tout type de maladie chronique a été considéré comme un critère d'exclusion. Les comptes-rendus médicaux ont été récupérés à l'aide de Dr Warehouse (68). Pour chaque patient, la meilleure photographie de profil a été sélectionnée en termes de symétrie.

Dans le set de validation, nous avons récupéré des photographies frontales et latérales des syndromes d'intérêt auprès de plusieurs centres spécialisés, à Londres, Bangkok, Lausanne, Lille, Montpellier, Nantes, Tours et Grenoble. Cette diversité des données de validation ajoute une robustesse à l'étude, d'autant plus que certains centres comme Londres et Bangkok ont permis de tester les modèles sur des populations ethniquement diverses. Nous avons utilisé les mêmes critères d'inclusion et d'exclusion que pour le set d'entraînement.

Nous avons utilisé trois modèles différents basés sur 105 *landmarks* pour les vues frontales, 73 pour les vues latérales et 41 pour les images de l'oreille externe. Nous avons développé un modèle d'annotation automatique pour chaque modèle en suivant un pipeline comprenant : (1) la détection de la région d'intérêt (ROI) et (2) le placement automatique des *landmarks*.

Pour la détection des zones d'intérêt, un modèle RCNN (réseau neuronal convolutionnel plus rapide basé sur les régions) a été entraîné après l'augmentation des données. Pour le placement automatique des points de repère, nous avons utilisé un modèle AAM (Active Appearance Model) basé sur les patchs en utilisant la bibliothèque menpo sur Python 3.7. Chaque photographie annotée automatiquement a été vérifiée par deux auteurs en aveugle du diagnostic, et les *landmarks* ont été repositionnés manuellement si nécessaire, à l'aide de landmarker.io. L'ICC (coefficient de corrélation intraclasse) a été calculé entre les évaluateurs.

Nous avons effectué une analyse Procrustes généralisée (GPA) sur tous les nuages de *landmarks* en utilisant le paquetage geomorph sur R. Les coordonnées Procrustes ont ensuite été traitées à l'aide de l'analyse en composantes principales (ACP) pour la réduction des dimensions. Nous avons retenu les composantes principales expliquant 99% de la variance totale de la somme cumulée. Le dernier 1% a été considéré comme une information négligeable.

Nous avons divisé les photos de face et de profil en zones clés et appliqué des méthodes d'extraction de caractéristiques texturales à chacune de ces zones, ce qui permet de vérifier les résultats et de déterminer quelle zone a le plus contribué au diagnostic. Nous avons défini 14 zones clés, 11 à partir de la photographie frontale (œil droit, œil gauche, sourcil droit, sourcil gauche, glabelle, front, pointe nasale, philtrum, joue droite, joue gauche, menton) et 3 à partir de la photographie latérale (région pré-auriculaire, œil, relief malaire). Ces zones peuvent potentiellement contribuer au diagnostic et à la caractérisation du phénotype facial. Chaque

zone a été extraite automatiquement à l'aide des *landmarks* placés précédemment. Nous avons utilisé l'algorithme CLAHE (Contrast Limited Adaptative Histogram Equalization) pour l'égalisation de l'histogramme. Les méthodes de matrice de cooccurrence en niveaux de gris (GLCM), proposées par Haralick, sont basées sur l'estimation des fonctions de densité de probabilité conditionnelles conjointes du deuxième ordre.

Les caractéristiques géométriques et les composantes principales géométriques ont été combinées pour la suite de l'analyse. Pour tenir compte des métadonnées associées (âge et genre) et du fait que nous avons inclus plus d'une photographie par patient (c'est-à-dire la non-indépendance des données), un modèle mixte a été conçu pour chaque caractéristique. Les variables à expliquer étaient les caractéristiques (géométriques et texturales), l'âge, le genre et l'origine ethnique étant considérés comme des variables explicatives. L'âge, le genre et l'origine ethnique sont des facteurs importants en dysmorphologie car ils influencent le diagnostic et doivent donc être pris en compte. Un effet aléatoire sur l'âge et les individus a été introduit. Les résidus de chaque caractéristique ont été calculés pour prendre en compte les biais potentiels liés aux métadonnées.

Les entrées du modèle étaient les résidus des modèles linéaires décrits ci-dessus, pour chaque caractéristique géométrique ou texturale. Nous avons utilisé XGBoost (eXtreme Gradient Boosting), un classificateur d'apprentissage automatique supervisé, pour toutes les analyses. Nous avons choisi un booster basé sur les arbres, et la fonction de perte à minimiser était une régression logistique dans le cas d'une classification binaire, ou une fonction softmax pour une classification multi-classes. Le modèle présentant le taux d'erreur le plus faible dans le cas d'une classification binaire ou le taux d'erreur multi-classe dans le cas d'une classification multi-classe a été choisi pour l'analyse. Nous avons séparé l'ensemble de données en un ensemble d'apprentissage et un ensemble de test, et une validation croisée a été utilisée pour définir le nombre idéal d'itérations afin d'éviter le surajustement. Le modèle choisi avec le

nombre idéal d'itérations a ensuite été utilisé sur l'ensemble de validation indépendant pour tester les performances, en traçant la précision et l'AUC.

Résultats. Le set d'entraînement contenait 3330 photographies, correspondant à 1086 patients ; 47 % des patients étaient des témoins et 53 % présentaient un syndrome. Dans le groupe de contrôle, 54 % des patients étaient des femmes et l'âge moyen était de 7,0 +/- 4,6 ans. Dans le groupe de patients, les syndromes étaient les suivants : Crouzon-Pfeiffer (N = 348, 60%), Kabuki (N = 167, 29%), Treacher Collins (N = 149, 26%), Silver Russell (N = 131, 23%), Apert (N = 88, 15%), Muenke (N = 72, 13%), CHARGE (N = 69, 12%), MFDGA (N = 60, 10%), Saethre Chotzen (N = 47, 8%) et NAFD (N = 14, 2%). Le set de validation contenait 216 photographies, correspondant à 108 patients ; 18 % des patients étaient des témoins et 82 % présentaient un syndrome. Dans le groupe de contrôle, 58 % des patients étaient de genre féminin et l'âge moyen était de 6,1 +/- 4,5 ans. Dans le groupe de patients, les syndromes étaient les suivants : Crouzon-Pfeiffer (N = 27, 30%), Apert (N = 13, 15%), Kabuki (N = 11, 12%), Muenke (N = 7, 8%), Saethre Chotzen (N = 7, 8%), CHARGE (N = 6, 7%), Silver Russell (N = 6, 7%), MFDGA (N = 5, 6%), Treacher Collins (N = 5, 6%) et NAFD (N = 2, 2%).

Les représentations de Procrustes ont montré les caractéristiques morphologiques classiquement décrites dans la littérature pour ces 10 syndromes. Outre les traits du visage sur les vues frontales, les vues de profil et les oreilles externes étaient importantes pour la description phénotypique du syndrome, en particulier la brachycéphalie trouvée dans les syndromes d'Apert, de Crouzon-Pfeiffer, de Muenke et de Saethre Chotzen, les grandes oreilles proéminentes du syndrome de Kabuki, la conque triangulaire et l'hypoplasie du lobe de l'oreille du syndrome CHARGE, la rétrognathie des syndromes de Treacher Collins, NAFD et MFDGA, et la petite mandibule du syndrome de Silver Russell. Notre modèle a permis de classer correctement 71,3 % (61,8 - 79,6) des patients (précision top-1). La

précision du top-3 était de 93,5 %. Les performances varient en fonction des syndromes, avec une précision de 96,1 % pour les témoins (AUC = 1 000) ou de 90,7 % pour les groupes Apert (AUC = 0,992) et CHARGE (AUC = 0,912). Cependant, la précision était de 50 % dans le groupe NAFD (AUC = 0,450), correspondant au plus petit nombre de patients. Les courbes ROC et les AUC pour les ensembles de formation et de validation étaient comparables.

Dans le syndrome d'Apert, notre modèle n'a pas permis de classer correctement les patients entre les deux génotypes (*FGFR2* p.Pro253Arg vs. FGFR2 p.Ser252Trp) (AUC = 0.506 [0.215 - 0.797], p = 0.874). Dans le syndrome de Crouzon - Pfeiffer, les génotypes liés aux domaines IgI, IgII, IgIIIc et TK de FGFR2 et les variations de *FGFR3* ont été associés à des phénotypes faciaux plus sévères que les variations d'un site d'épissage de *FGFR2*. Dans ce dernier groupe, la brachycéphalie, le phénotype oculaire et la hauteur réduite du tiers moyen du visage semblaient moins sévères.

Le modèle a également permis de distinguer le syndrome de Kabuki 1 (KS1, lié à *KMT2D*) du syndrome de Kabuki 2 (KS2, lié à *KDM6A*) avec une AUC empirique de 0,805 (0,729 - 0,880, p < 0,001). Les patients atteints du syndrome KS2 avaient un visage plus rond, un nez plus court, une lèvre supérieure plus épaisse, des narines antéversées et un tiers moyen du visage plus court. Il n'y avait pas de différence évidente au niveau des sourcils et des yeux. Les oreilles externes étaient plus allongées verticalement chez KS2, avec un lobe hypoplasique et une rotation dans le sens inverse des aiguilles d'une montre. La conque semblait plus verticale en KS1.

Conclusion. Nous présentons ici un nouveau modèle de phénotypage de nouvelle génération, qui peut être utilisé pour détecter un syndrome génétique sur des photographies 2D du visage et du profil, afin d'aider les praticiens dans leur processus de diagnostic. Ce travail a été rendu possible grâce à l'incroyable quantité et qualité de la base de données photographiques de l'hôpital Necker - Enfants Malades. La valeur ajoutée par rapport aux outils existants est

l'analyse des vues de profil, la prise en compte de la forme de la voûte crânienne et la possibilité de détecter des patients non syndromiques. Ce modèle ne remplacera jamais l'expertise clinique, mais permettra de reconsidérer les approches phénotype-first, le but ultime étant de réduire la durée moyenne de l'errance diagnostique dans les maladies rares. Cette thèse n'est que l'introduction d'un grand projet, avec le recrutement de plusieurs ingénieurs et chercheurs, afin d'intégrer plusieurs centaines de syndromes génétiques dans l'algorithme, et d'être utilisé en pratique quotidienne par les médecins.

Table of contents

1 Introduction

Variations in the shape and proportions of the human face have long fascinated both scientists and artists. Giovanni Francesco Caroto painted a picture of a smiling child holding a drawing, which inspired Dr. Harry Angelman to describe the syndrome of the same name in 1965. The dysmorphologist admitted having made the connection between three children seen in consultation and this painting seen during his visit to the Castelvecchio museum in Verona (1). Similarly, the portrait by the Flemish painter Quinten Massy, "The Ugly Dutchess", painted in 1513 and exhibited at the National Gallery (London, UK), is the first phenotypic description of the *facies leonina*, associated with large nostrils, prominent cheekbones and pronounced superciliary arches. In 1989, Dequeker diagnosed Paget's disease in this woman. (2). Finally, the painting by an anonymous pupil of the Flemish painter Jan Joest van Kalkar, "Adoration of the Christ Child" (1515), depicts a shepherd and an angel with a typical Down's syndrome facies, with a narrowing of the middle third of the face, upslanted palpebral fissures and a small upturned nasal tip (Figure 1).

Pablo Picasso said, "Should you paint what's on a face? What's inside a face? Or what is hidden behind a face". With this in mind, the dysmorphologist identifies phenotypic facial features (among others), combines them and uses her or his experience and training to propose a diagnosis. Earlier diagnosis will reduce the morbidity and mortality associated with this syndrome (3).

Figure 1. Artistic representations of characters with facial dysmorphia. A. Painting by Giovanni Francesco Caroto, at the Castelvecchio Museum in Verona, depicting a child probably suffering from Angelman syndrome. **B.** « The Ugly Duchess », painting by Quinten Massy, exhibited at the National Gallery in London, UK (1513). This is the first description of the *facies leonina*, a term used by clinicians. **C.** "Adoration of the Christ Child", a painting by one of the anonymous pupils of the Flemish painter Jan Joest van Kalkar, exhibited at the Metropolitan Museum of Art in New York, USA (1515). The painting probably features a shepherd and an angel affected by Down's syndrome.

1.1 Dysmorphology and experience

Dysmorphology is a discipline of clinical genetics that studies and attempts to interpret human structural and growth variations (4). The recognition of sometimes subtle phenotypic traits requires expertise (5). In 1988, Diliberti described the 6 main tasks in this discipline: 1) identify a clinical sign and select a list of candidate syndromes from a book or through experience, 2) compare the patient's associated clinical presentation with the potential signs of the selected syndromes selected, in order to restrict the list, 3) screen for publications in the

scientific literature to support a diagnosis, 4) seek the opinion of a geneticist who is an expert in the syndrome, 5) if there is no specialist, seek the opinion of colleagues and possibly identify a new syndrome, and 6) carry out additional tests, particularly genetic tests, to confirm a diagnosis (6). Moeschler et al admit that the diagnosis is delayed or unknown in 38% of dysmorphology assessments. According to these authors, there are three reasons for this uncertainty: the rarity of many genetic syndromes, patients' lack of access to specialist dysmorphology services and the variable experience of practitioners (7). Furthermore, most of the academic books and cases published in the literature focus on Caucasian populations only (8). Certain minor anomalies found in a Caucasian population, such as lip thickness or broad nasal implantation, may be considered a normal phenotypic trait in an African population (9). The nasal anomaly found in Caucasian 22q11 deletion syndrome is found in 89% of an Asian population, compared with 15-40% of an African population (10,11). These inequalities are further exacerbated by the fact that centers specializing in clinical genetics and dysmorphology are more concentrated in developed countries than in most low-income countries with non-Caucasian populations (12).

1.2 Photographic data

Photographs allowed to store information identified by the practitioner for further use (analogy with new patients or patient follow-up), or for teaching purposes. The photographic databases of clinical genetics departments or other specialties in contact with genetic syndromes are therefore often incredibly valuable for research.

The photographic database we exploited in this work results from the fusion of three sources.

- The database of digitized photographs from the Maxillofacial surgery and plastic surgery department at Necker – Enfants Malades Hospital; each photograph was taken

by a professional medical photographer in standardized positions. The oldest photograph dates from 1995.

- The database of non-digitized photographs from the same department; each photograph was taken by the practitioner himself using different cameras, in standardized positions. The oldest photograph dates from 1976.
- The database of digitized photographs from the Clinical genetics department of the Necker – Enfants Malades Hospital and the Imagine Institute; each photograph was taken by the practitioner himself using different cameras, usually in standardized positions. The oldest photograph dates from 1998.

The complete database contains 1,042,468 photographs from about 22,000 patients. Necker-Enfants Malades is a hospital which accounts for a third of the pediatric activity of the Assistance Publique des Hôpitaux de Paris (APHP), the trust of the 38 Greater Paris academic hospital, and almost half of pediatric surgery cases yealry. Since the 1950s, the hospital has specialized in the management of genetic diseases. Necker – Enfants Malades is home to 32 local rare disease competence centers, and coordinates 15 national rare disease reference centers. As a specialist center, although it receives mostly patients from the Greater Paris region, more than 20% of the patients treated at Necker – Enfants Malades Hospital travel from other regions of France or from abroad (1). The Imagine Institute was created in 2007, and the building was inaugurated in 2014 on the Necker – Enfants Malades campus. Imagine brings together more than 1,000 researchers, doctors and healthcare staff with the aim of accelerating synergies, promoting the transfer of knowledge, and thus finding new treatments and diagnoses faster. Imagine Institute is a translational research center at the interface between care and research, where patients are monitored by Necker – Enfants Malades doctors and diagnosed by the Institute research teams. Imagine is Europe's leading center for research, care and teaching in the field of genetic diseases.

1.3 The burden of rare diseases

Congenital diseases are the 5th leading cause of neonatal death, accounting for 270,000 deaths worldwide in 2010 (14). There are an estimated 7000 rare diseases, 80% of which are genetic in origin (15). 30-40% of the 7000 rare diseases have craniofacial anomalies (15). These anomalies may be major, such as a cleft lip and cleft palate, and easily detected, but they may also be minor and only detectable by a trained practitioner. Some of these syndromes lead to premature death, most often due to associated organ malformations, breathing difficulties or feeding problems (2). We can therefore assume that 2100 to 2800 rare conditions could be screened by a detailed gestalt analysis.

1.4 Dysmorphology and machine learning

In machine learning approaches, pre-defined rules are replaced by the ability of the model to learn from examples (16). The applications of machine learning are increasing in healthcare (16), in diverse fields such as radiology (17), dermatology (18) and surgery (19). Relevant inputs of machine learning approached to shape analysis in healthcare are assessments of radiographs, texts, pathology sections, and clinical photographs. Among these sources of data, clinical photographs are, as we have seen previously, particularly important in the diagnosis of craniofacial malformations.

1.5 Objectives and prospects

The two main objectives of this project were:

- to diagnose a syndrome, in a binary then multi-syndromes classification;

- to highlight differences within the same syndrome, i.e., to establish a phenotype-genotype correlation.

Various syndromes will be tested in the course of the manuscript in order to answer practical diagnostic questions raised by geneticists. We will conclude with an analysis in a multi-syndrome situation, based on all the conditions that we were able to include during the thesis, namely: Crouzon - Pfeiffer, Apert, Saethre Chotzen, Muenke, Kabuki, Treacher Collins, Nager, Guion-Almeida, CHARGE and Silver Russell syndromes.

A secondary objective was to evaluate the effect of a treatment on facial morphology. In this prospect, we have evaluated: (1) the effect of a medical treatment, *alpelisib*, on a specific type of PIK3CA-related hypergrowth syndrome (hemifacial myohyperplasia), (2) the morphology of the nose after two different surgical corrections of cleft lip and palate, and (3) two surgical methods for treating non-syndromic scaphocephaly.

2 State-of-the-art

There has been a recent increase in the number of scientific publications dedicated to the diagnosis of rare conditions using facial photographs (20). We will divide this literature review into two parts: (1) assessment of algorithms designed for research purposes and (2) analysis of commercial tools for healthcare professionals.

2.1 Non-marketed algorithms

We first aimed at reporting all articles describing the computerized identification of a disease and/or a syndrome based on 2D facial photographs. PubMed was used to screen the MEDLINE database using MeSH (Medical Subject Heading) terms along with keywords from titles and abstracts.

(diagnosis[Title] OR detection[Title] OR dysmorphic[Title/Abstract] OR dysmorphology[Title/Abstract] OR disease[Title/Abstract] OR syndrome[Title/Abstract])

AND *(face[Title/Abstract] OR faces[Title/Abstract] OR facial[Title/Abstract])*

AND *("face-classification"[Title/Abstract] OR "facial analysis technology"[Title/Abstract] OR "artificial intelligence"[MeSH Terms] OR "biometric identification"[Mesh] OR "Medical informatics"[MeSH Terms] OR "Image Processing, Computer-Assisted"[MeSH Terms])*

All human studies involving 2D facial photographs used to diagnose one or several conditions in healthy or patient populations were included. Only studies with diagnosis confirmation by

clinical experts and/or molecular investigations were retained. We excluded studies based on the diagnosis of patients using a commercial tool.

We reported the pre-processing of the pictures, *i.e.*, every treatment the authors applied to the photograph before extracting features or using classifiers. Concerning feature extraction, we reported the use of: (1) landmarks, with manual and/or automatic placement, (2) geometric features, *i.e.*, distances, angles, or other morphometric data, and (3) texture features, based on the spatial arrangement of intensities on the image. We listed the classifiers and the validation methods of machine learning designs. We considered that authors had used machine learning methods when predictions were made based on a test set with a classifier, or on a model trained on a training set. With the pre-defined request, we listed 1515 results on PubMed. After applying all the exclusion criteria on reading title, abstract and full text when necessary, we included 27 articles. We defined two types of studies depending on the number of diseases to be diagnosed.

The first and most common study design – in 18/27 (67%) publications – intended to diagnose one condition in the general population versus healthy controls using facial photographs (11,21–35). The second most prevalent study design – in 9/27 (33%) publications – consisted in using multi-syndrome classification, *i.e.,* in differentiating syndromes from each other and from the general population (36–44). The photographic data had diverse origins: (1) pictures taken during clinics (45,46), (2) pictures from articles or books (36,44), (3) pictures from pre-existing cohorts (47) and (4) pictures publicly available on the internet (22). There was one photograph per patient in all publications. Authors used age-, sex-, and ethnicity-matched controls in 22% studies.

All publications applied two exclusive image processing approaches: (1) use of geometric and/or textural features to describe phenotypes or (2) deep learning process in order to detect a syndrome without describing phenotypes. Machine learning approaches used to diagnose

clinical conditions on 2D photographs generally followed similar frameworks: features were extracted from photographs to diagnose a disease or the syndrome was detected directly without feature extraction by deep learning. In case of features extraction, the steps were: image pre-processing, image processing with or without landmark placement, extraction of geometrical and/or textural features, reduction of the dimensions of the input, and classification and validation using training and test sets.

2.1.1 Features extraction

Landmarks are defined by reproductible points of interest on an image. They were one of the most common tools used to extract phenotypic data, as in 22/27 (81%) publications. Other authors had applied global shape and texture analysis approaches without landmark placement.

Geometric features were extracted from landmarks and used for disease detection. One straightforward approach was to measure Euclidean distances and angles from landmarks (29,30,41,42,48). Another approach to geometrical information extraction, used in 4/27 (15%) publications, was Procrustes superimposition (21,25,36,49). Procrustes distance was defined as the deformation necessary to fit a dysmorphic face on a control face via an iterative least-squares process. Each face was centered and scaled using a centroid (50) to minimize the sum of squared differences between the landmarks. Douglas and Mutsvangwa (51) concluded that Procrustes-based approaches were the best methods for shape analysis in terms of prediction accuracy.

A texture extraction system was used in 18/27 (67%) publications, in association with geometric features extraction or independently. The most common methods were Gabor Wavelet Transformation (GWT) used in 12/27 (44%) (52,53) and/or a Local Binary Patterns (LBP) in 7/27 (26%) publications (54). These methods provided information about the surface

located between the landmarks. For example, zygomatic hypoplasia in Treacher Collins syndrome could be better detected by analyzing the shadows on a frontal facial photograph than by measuring angles between landmarks.

Machine learning was defined by using a training set to make predictions on a validation or test set. Based on this, only 3/27 (11%) publications were not using machine learning methods. Diverse classifiers were applied, such as Generalized Linear Models (GLM), Support Vector Machines (SVM), k-nearest neighbors (kNN), or random forests (RF). Deep learning – Convolutional Neural Networks (CNN) or Artificial Neural Networks (AAN) – was used in 3/27 (11%) publications. Kong et al. (28) also used CNN with convolutional layers, pooling layers and fully-connected layers.

2.1.2 Performances

Saraydemir et al. (55) found an overall accuracy of 97% with 15 patients for diagnosing Down syndrome within a healthy population. This performance was higher than clinical experts. Kruszka et al (11) found an accuracy of 95% for detecting 22q11.2 deletion syndrome in 156 patients vs. the general population. Kong et al. (28) used the F1-score to evaluate the performances of their classifier in detecting acromegaly in 527 patients. They obtained a F1-score of 91% without landmarks and with deep learning methods.

2.1.3 Published paper

ARTICLE IN PRESS
JID: JORMAS [m5G;April 30, 2021;11:40]
J Stomatol Oral Maxillofac Surg 000 (2021) 1–5
Available online at ScienceDirect www.sciencedirect.com
Elsevier Masson France EM|consulte www.em-consulte.com

Review

Computational diagnostic methods on 2D photographs: A review of the literature☆,☆☆

Quentin Hennocq[a,b,*], Roman Hossein Khonsari[a,b], Vincent Benoît[a], Marlène Rio[a,c], Nicolas Garcelon[a]

[a] Institut Imagine, Paris Descartes University-Sorbonne Paris Cité, Paris, France
[b] Department of Maxillo-Facial Surgery and Plastic Surgery, Hôpital Universitaire Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris, Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou, Faculté de Médecine, Université de Paris, Paris, France
[c] Department of Genetics, IHU Necker-Enfants Malades, University Paris Descartes, Paris, France

ARTICLE INFO

Article History:
Received 11 September 2020
Accepted 7 April 2021
Available online xxx

Keywords:
Literature review
Machine learning
Deep learning
Photograph
Dysmorphology
Diagnosis

ABSTRACT

Here we provide a literature review of all the methods reported to date for analyzing 2D pictures for diagnostic purposes. Pubmed was used to screen the MEDLINE database using MeSH (Medical Subject Heading) terms and keyworks. The different recognition steps and the main results were reported. All human studies involving 2D facial photographs used to diagnose one or several conditions in healthy populations or in patients were included. We included 1515 articles and 27 publications were finally retained. 67% of the articles aimed at diagnosing one particular syndrome versus healthy controls and 33% aimed at performing multi-class syndrome recognition. Data volume varied from 15 to 17,106 patient pictures. Manual or automatic landmarks were one of the most commonly used tools in order to extract morphological information from images, in 22/27 (81%) publications. Geometrical features were extracted from landmarks based on Procrustes superimposition in 4/27 (15%). Textural features were extracted in 19/27 (70%) publications. Features were then classified using machine learning methods in 89% of publications, while deep learning methods were used in 11%. Facial recognition tools were generally successful in identifying rare conditions in dysmorphic patients, with comparable or higher recognition accuracy than clinical experts.

Introduction

Shape analysis on 2D pictures can involve computations based on pre-defined rules; in machine learning approaches, these rules are replaced by the ability of the model to learn from examples [1]. Applications of machine learning are increasing in healthcare [1], in a variety of fields such as radiology [2], dermatology [3] and surgery [4]. Inputs or features are converted into outputs or labels. Deep learning is defined as a subtype of machine learning in which artificial neural networks are used. These neural networks can encode relationships between features and labels that are not accessible to the human brain, because of their complexity or because they do not match the way our brain processes data [1].

Examples of relevant inputs in healthcare are X-rays, texts, pathology slices and clinical photographs. Among these inputs, clinical photographs are particularly important in the diagnosis of craniofacial malformations. In fact, 30–40% of the 7000 rare diseases have craniofacial anomalies [5]. Many among these 2–3000 syndromes have minor facial features or are very uncommon: their screening requires the eye of an expert in dysmorphology. Due to this specific clinical issue, there have been a recent increase in the number of scientific publications dedicated to the diagnosis of rare conditions using computational methods on facial photographs [6].

Here we provide a literature review of all the methods used for computational diagnosis on facial photographs. This overview will be of use for physicians and scientists interested in syndrome recognition, and more generally, in facial recognition.

Methods

We performed a review of the literature according to the framework described by Arksey and O'Malley [7]. We aimed at reporting all articles describing the computerized identification of a disease and/or a syndrome based on 2D facial photographs. PubMed was used to screen the MEDLINE database using MeSH (Medical Subject

☆ No financial assistance was received in support of the study.
☆☆ **Declarations of interest:** none.
* Corresponding author at: Institut Imagine; 24, boulevard du Montparnasse, 75015 Paris.
E-mail address: quentin.hennocq@aphp.fr (Q. Hennocq).

https://doi.org/10.1016/j.jormas.2021.04.003
2468-7855/

Please cite this article as: Q. Hennocq, R.H. Khonsari, V. Benoit et al., Computational diagnostic methods on 2D photographs: A review of the literature, Journal of Stomatology oral and Maxillofacial Surgery (2021), https://doi.org/10.1016/j.jormas.2021.04.003

Heading) terms along with keywords from titles and abstracts (see request Table 1).

All human studies involving 2D facial photographs used to diagnose one or several conditions in healthy or patient populations were included. Only studies with diagnosis confirmation by clinical experts and/or molecular investigations were retained. Articles on prenatal diagnosis, extra-facial diseases and dermatological diseases of the face (e.g. facial acnea or naevi) were excluded. Studies assessing treatment results, such as craniofacial surgery results, were excluded, as well as studies dedicated to the detection of pain or other emotions. We also excluded studies based on the diagnosis of patients using a pre-existing algorithm, such as *Face2Gene* [8].

One reviewer (QH) screened all the titles and/or abstracts. When exclusion criteria were not clear after reading title and abstract, the full text was assessed. The relevant publications were then charted with the following items: year of publication, first author, country of affiliation, study design or objective, disease of interest, origin of data and number of patients. We reported the pre-processing of the pictures, i.e. every treatment the authors applied to the photograph before extracting features or using classifiers. Concerning feature extraction, we reported the use of: [1] landmarks, with manual and/or automatic placement, [2] geometric features, i.e. distances, angles or other morphometric data and [3] texture features, based on the spatial arrangement of intensities on the image. We listed the classifiers and the validation methods of machine learning designs. We considered that authors had used machine learning methods when predictions were made on a test set with a classifier or a model trained on a training set.

Results

With the pre-defined request, we listed 1515 results on PubMed. After applying all the exclusion criteria on reading title, abstract and full text when necessary, we included 27 articles. These publications were fully read and charted (Table 2). The region of affiliation of the first author was North America – USA only in 37%, Europe – Germany, UK and Croatia in 33%, and Asia – China, India and Turkey in 30%.

1. Study designs and conditions

We defined two types of studies depending of the number of diseases to be diagnosed.

The first and most common study design – in 18/27 (67%) publications – intended to diagnose one condition in the general population versus healthy controls using facial photographs. Facial pictures of the following conditions were tested versus healthy controls: Down syndrome [9–12], Cornelia de Lange syndrome [13], Rubinstein-Taybi syndrome [13], 22q11.2 deletion syndrome [14], Laron syndrome [15], Williams-Beuren syndrome [16], acromegaly [17–21], Cushing disease [22], Turner syndrome [23], Chronic Fatigue Syndrome [24] and schizophrenia [25].

The second most prevalent study design – in 9/27 (33%) publications – consisted in using multi-syndrome classification, i.e. in differentiating syndromes from each other and from the general population [8,26–35]. For example, Ferry et al. [26] aimed at diagnosing 8 syndromes – Angelman, Apert, Cornelia de Lange, Down, fragile X, progeria, Treacher-Collins and Williams-Beuren syndromes – in the general population. Gurovich et al. [8] aimed at diagnosing 216 different syndromes in a population of affected patients.

The conditions reported in the included studies were:

- genetic syndromes: Down [26], Angelman [8], Prader-Willi [31], Cornelia de Lange [8], X fragile [31], Apert [26], Progeria [26], Treacher-Collins [26], Williams-Beuren [29], Sotos [27], Cri-du-chat [31], Smith-Lemli-Opitz [31], 22q11.2 deletion [31], Turner [8], Noonan [8], Mucopolysaccharidosis [31], Rubinstein-Taybi [8], Hurler [33], Wolf-Hirschhorn [33] and 183 others in [8],
- endocrine conditions with a facial phenotype (e.g. acromegaly and Cushing syndrome) [20]
- psychiatric conditions: schizophrenia [25] and chronic fatigue syndrome [24].

Table 1
Pubmed request.

	(diagnosis[Title] OR detection[Title] OR dysmorphic[Title/Abstract] OR dysmorphology[Title/Abstract] OR disease[Title/Abstract] OR syndrome[*Title/Abstract*])
AND	(face[Title/Abstract] OR faces[Title/Abstract] OR facial[*Title/Abstract*])
AND	("face-classification"[Title/Abstract] OR "facial analysis technology"[-Title/Abstract] OR "artificial intelligence"[MeSH Terms] OR "biometric identification"[Mesh] OR "Medical informatics"[MeSH Terms] OR "Image Processing, Computer-Assisted"[MeSH *Terms*])

2. Data description

The photographic data had diverse origins: [1] pictures taken during clinics [9,11,14–22,24,27,29–33,36,37,38–40, 2] pictures from articles or books [13,14,26,35,41, 3] pictures from pre-existing cohorts [8,34,42,43] and [4] pictures publicly available on the internet [10,26]. The data volume varied from 15 pictures of affected patients [36] to 17,106 for Gurovich et al. [8]. The average number of affected (non-control) patients in the studies was 844 (+/- 3272); more precisely, 114 (+/- 138) affected individuals were included in average in order to detect one syndrome versus the general population and 2085 (+/- 5309) affected individuals in order perform multi-syndrome classification. There was one photograph per patient in all publications. Authors used age-, sex-, [16,19,22,37,39] and ethnicity-matched controls [11] in 6/27 (22%) studies.

3. Image pre-processing

Pre-processing corresponded to the preparation of pictures before extracting data and applying classifiers. It is a useful step in machine learning: images are grouped (face, profile), resized and transformed (rotation, cropping, lighting) to optimize analysis.

An automated facial detection system was used in several studies to sort frontal views from lateral views and then crop the pictures. The most common algorithm in this field was the *Haar cascade* tool developed by Viola and Jones [1]. One author used manual cropping, scaling and rotation to maximize the classifier performances [36]. This approach was not adapted to large datasets. Manual or automated cropping was useful for little standardized databases, for instance when several individuals were present on pictures (e.g. the child and his mother) [6].

4. Image processing

All authors applied two exclusive image processing approaches: [1] use of geometric and/or textural features in order to describe phenotypes or [2] deep learning process in order to detect a syndrome without describing phenotypes.

Landmarks positioning

Landmarks are defined by reproductible points of interest on an image. They are one of the most common tools used to extract phenotypic data, as in 22/27 (81%) publications. In 7/27 (26%)

Table 2
General characteristics for each of the studies, and method of classification. Publications not using machine learning (n = 3) were excluded. SVM = Support Vector Machine. kNN = k-Nearest Neighbors. RF = Random forests. CNN = Convolutional Neural Network. AAN = Artificial Neural Networks.

Multisyndrome	Data volume	Number of syndromes	Linear functions	SVM	kNN	RF	CNN / ANN
Boehringer, 2006	147	10		✓	✓		
Vollmar, 2007	200	14		✓	✓		
Boehringer, 2011	202	14		✓	✓		
Ferry, 2014	1363	8			✓		
Kuru, 2014	92	15			✓		
Shu, 2015	13	100		✓	✓		
Özdemir, 2018	124	5			✓		✓
Gurovich, 2019	17,106	216					✓
Monosyndrome		Condition / syndrome					
Schaefer, 1994	49	Laron	✓				
Learned-Miller, 2006	24	Acromegaly		✓			
Miller, 2011	24	Acromegaly		✓			
Saraydemir, 2011	15	Down		✓	✓		
Schneider, 2011	57	Acromegaly	✓				
Buretic-Tomljanovic, 2012	58	Schizophrenia	✓				
Kosilek, 2013	20	Cushing	✓				
Zhao, 2014	100	Down		✓	✓	✓	
Chen, 2015	294	Chronic fatigue syndrome			✓		
Kong, 2017	527	Acromegaly		✓	✓	✓	✓
Kruszka, 2017	156	22q11.2		✓			
Kruszka, 2017	129	Down		✓			
Kruszka, 2017	161	Cornelia de Lange		✓			
Chen, 2018	32	Turner		✓			
Kruszka, 2018	286	Williams-Beuren		✓			
Dowsett, 2019	246	Cornelia de Lange		✓			

publications, the authors placed landmarks manually and automatic landmarking was used in 16/27 (59%) publications.

In 4/27 (15%) publications [10,21,36,40], the authors did not use landmarks to analyze faces and phenotypes and preferred assessing the overall shape and textures.

Frontalization and facial alignment

Frontalization is a process modifying a frontal image in order to improve the orientation of the frontal view. Kong et al. [18] used alignment, based on Procrustes superimposition [44,45], in order to compute frontal pictures from photos with capture angles varying from −45° to +45° around the standard coronal plane. The final frontal face was produced using the corresponding symmetric sides, which raised numerous issues in conditions with facial asymmetry.

Geometric features extraction

Geometric features were extracted from landmarks and used for disease detection. The main advantage of using landmarks was to be more flexible with pose changes and different angles of view. One straightforward approach was to measure Euclidean distances and angles from landmarks [19,20,22,23,32,33].

Another approach to geometrical information extraction, used in 4/27 (15%) publications, was Procrustes analysis [9,13,26,46,47]. Procrustes distance was defined as the deformation necessary to fit a dysmorphic face on a control face via an iterative least-squares process. Each of the faces were centered and scaled using a centroid [48] to minimize sum of squared differences between the landmarks. Douglas and Mutsvangwa [49] concluded that Procrustes-based approaches were the best methods for shape analysis in terms of prediction accuracy.

Texture features extraction

A texture extraction system was used in 18/27 (67%) publications, in association with geometric features extraction [8,10,11,19,20,22–24,29,34,40,41,45] or independently [10,36,40]. The most common methods were Gabor Wavelet Transformation (GWT) used in 12/27 (44%) [50,51] and/or a Local Binary Patterns (LBP) in 7/27 (26%) publications [52]. These methods provide information about the surface located between the landmarks. For example, zygomatic hypoplasia in Treacher Collins syndrome would be better detected by analyzing the shadows on a photo than by measuring angles between landmarks.

The information on geometric / texture features created extensive data matrices, with the need to reduce dimensionality before incorporation into classifiers. For this purpose, Principal Component Analysis (PCA) was used in 17/27 (70%) publications.

Gurovich et al. [8] did not use any geometric or textural features extraction step for prediction making after facial alignment based on different landmarks. Their approach is referred to as deep learning because it corresponds to a *black box* approach diagnosing dysmorphic features without going through a phenotype description process. When using deep learning, authors cannot easily explain what contributed to the diagnosis of a particular syndrome.

5. Classification

Machine learning was defined by using a training set to make predictions on a validation or test set. Based on this, 3/27 (11%) publications were not using machine learning methods.

In these studies, diverse classifiers were used. The simplest classifiers were Generalized Linear Models (GLM) such as logistic regression and linear discriminant analysis, as used in 4/27 (15%) publications [13,18]. Their output was a linear combination of input facial data.

SVM – used in 13/27 (48%) publications – was a classification method aiming at separating two data sets with distance maximization based on an optimal separating hyperplane [53]. Bounds between datasets and the optimal separating hyperplane were referred to as 'support vectors' (68).

The k-nearest neighbors (kNN) rule – used in 11/27 (41%) publications – classified each unlabeled example by the majority label of its k-nearest neighbors in the training set [54]. It stored instances of previously earned images. Classification was obtained from a

majority vote of the nearest neighbors [18]. The kNN classifier was particularly adapted for multiclass detection.
Random forests (RT) – used in 2/27 (7%) publications – were another technique [18] consisting in growing many classification trees. The forest chose the classification tree having the best prediction. In addition, a subset of features was randomly selected from the optional features in order to grow the tree at each node. Random forests were particularly adapted for large number of weak classifiers.
Deep learning – Convolutional Neural Networks (CNN) or Artificial Neural Networks (AAN) – was used in 3/27 (11%) publications. Kong et al. [18] also used CNN with convolutional layers, pooling layers and fully-connected layers. Gurovich et al. [8] used Deep Convolutional Neural Networks (DCNN) [55] for each of the cropped facial areas, which were then combined by averaging to form a model for a multiclass problem.
Kong et al. [18] compared the performances of these machine learning techniques (linear functions, SVM, kNN, RF and CNN): better classification abilities were found in terms of accuracy for SVM with detected features and for CNN without detected features.

6. Clinical results

Saraydemir et al. [36] and Zhao et al. [9] found an overall accuracy of respectively 97% and 97.9% with 15 and 24 patients for diagnosing Down syndrome in a healthy population. Both these performances were higher than clinical experts. Kruszka and his team found an accuracy of 95% for detecting 22q11.2 deletion syndrome in 156 patients vs. the general population [14], 94% for Down syndrome [11] in 129 patients, 90% for Williams-Beuren syndrome in 286 patients [16], 94% for Cornelia de Lange in 246 patients [37] and 89% for Noonan syndrome in 161 patients [39]. Kong et al. [18] used the F1-score in order to evaluate the performances of their classifier in detecting acromegaly in 527 patients. They obtained a F1-score of 92% with landmarks and SVM and 91% without landmarks and with deep learning methods. Accuracies of 86% and 82% were obtained to detect acromegaly among a healthy population in 24 [17] and 57 patients [19].
Ferry et al. [26] used the *Clustering Improvement Factor* (CIF) after implementing kNN on 8 syndromes and 1363 patients. Treacher-Collins syndrome was then clustered 23.5-fold better than by chance alone, Down syndrome 11.6-fold better, and Williams-Beuren 9.4-fold-better. The overall accuracy for detecting one syndrome vs. the 7 others was 76% and the overall accuracy for detecting syndromes vs. a healthy population was 95%. Another method for displaying results in a multi-syndrome classification design was to provide a Top-5 or Top-10 accuracy, i.e. to evaluate the probability of a condition to be ranked as the 5 or 10 most probable syndromes by the classifier. Gurovich et al. [8], in 17,106 patients with 216 syndromes, found a Top-10 accuracy of 91%, a Top-5 accuracy of 85% and a Top-1 accuracy of 61%. Kuru et al. [35] found 53% accuracy on a multi-class syndrome detection of 15 conditions with 92 patients. Boehringer et al. [29] had an overall accuracy of 21% for multi-syndrome detection on 14 conditions in 202 patients, which is very low compared to the performances of the clinical experts.

Discussion

Machine learning approaches used to diagnose clinical conditions on 2D photographs generally followed similar frameworks: either features were extracted from photographs to diagnose a disease or the syndrome was detected directly without feature extraction by deep learning. In case of features extraction, the steps were: image pre-processing, image processing with or without landmark placement, extraction of geometrical and/or textural features, reduction of the dimensions of the input, and classification and validation using training and test sets. Procrustes analysis was a reliable technique for the extraction of morphometric parameters [49]. Texture extraction techniques seemed of interest in adding information to the geometric features [24].

Deep learning methods are a promising tool with much space for performance improvement [8]. The number of publications using deep learning was however limited because this approach requires large data volumes – 17,106 patients in Gurovich et al. [8] – in order to train the model. The challenge in the future years will be to build large photographic databases and train neural networks in order to improve predictions.

Comparing publications and diagnosis performances of machine learning tools was difficult due to the variety of assessment parameters: Boehringer [29] used accuracy, Zhao [9] precision and recall, Dudding-Byth [34] top-10, top-5 and top-1 accuracies, and Ferry [26] the Clustering Improvement Factor (CIF). Evidence-based comparisons of machine learning in facial recognition would greatly benefit from more homogenous result assessment approaches.

Conclusion

Facial recognition approaches are efficient in identifying rare conditions using 2D photographs. Machine-learning algorithms improve recognition accuracy. Deep learning techniques such as CNNs are promising, but larger datasets are needed. Based on the current results of available algorithms, we can predict that collaborative studies with massive shared datasets could shortly lead to the development of algorithms with almost perfect diagnostic abilities.

References

[1] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019 Apr 4;380(14):1347–58.
[2] Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. Radiology 2018 Aug;288(2):318–28.
[3] Novoa RA, Gevaert O, Ko JM. Marking the path toward artificial intelligence –based image classification in dermatology. JAMA Dermatol 2019 Oct 1;155 (10):1105–6.
[4] Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. JAMA Surg [Internet] 2019 Dec 11 [cited 2019 Dec 14]Available from: https://jamanetwork.com/journals/jamasurgery/fullarticle/2756311.
[5] Hart TC, Hart PS. Genetic studies of craniofacial anomalies: clinical implications and applications. Orthod Craniofac Res 2009 Aug;12(3):212–20.
[6] Thevenot J, López MB, Hadid A. A survey on computer vision for assistive medical diagnosis from faces. IEEE J Biomed Health Inform 2018 Sep;22(5):1497–511.
[7] Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb 1;8(1):19–32.
[8] Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med 2019 Jan;25 (1):60–4.
[9] Zhao Q, Rosenbaum K, Okada K, Zand DJ, Sze R, Summar M, et al. Automated down syndrome detection using facial photographs. 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2013. p. 3670–3.
[10] Burçin K, Vasif NV. Down syndrome recognition using local binary patterns and statistical evaluation of the system. Expert Syst Appl 2011 Jul 1;38(7):8690–5.
[11] Kruszka P, Porras AR, Sobering AK, Ikolo FA, Qua SL, Shotelersuk V, et al. Down syndrome in diverse populations. Am J Med Genet A 2017;173(1):42–53.
[12] Basel-Vanagaite L, Wolf L, Orin M, Larizza L, Gervasini C, Krantz ID, et al. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. Clin Genet 2016;89(5):557–63.
[13] Kruszka P, Addissie YA, McGinn DE, Porras AR, Biggs E, Share M, et al. 22q11.2 deletion syndrome in diverse populations. Am J Med Genet A 2017 Apr;173 (4):879–88.
[14] Schaefer GB, Rosenbloom AL, Guevara-Aguirre J, Campbell EA, Ullrich F, Patil K, et al. Facial morphometry of Ecuadorian patients with growth hormone receptor deficiency/Laron syndrome. J Med Genet 1994 Aug;31(8):635–9.
[15] Kruszka P, Porras AR, de Souza DH, Moresco A, Huckstadt V, Gill AD, et al. Williams-Beuren syndrome in diverse populations. Am J Med Genet A 2018 May;176 (5):1128–36.

[16] Miller RE, Learned-Miller EG, Trainer P, Paisley A, Blanz V. Early diagnosis of acromegaly: computers vs clinicians. Clin Endocrinol (Oxf) 2011;75(2):226–31.
[17] Kong X, Gong S, Su L, Howard N, Kong Y. Automatic detection of acromegaly from facial photographs using machine learning methods. EBioMedicine 2017 Dec 15;27:94–102.
[18] Schneider HJ, Kosilek RP, Günther M, Roemmler J, Stalla GK, Sievers C, et al. A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification. J Clin Endocrinol Metab 2011 Jul 1;96(7):2074–80.
[19] Kosilek RP, Frohner R, Würtz RP, Berr CM, Schopohl J, Reincke M, et al. Diagnostic use of facial image analysis software in endocrine and genetic disorders: review, current results and future perspectives. Eur J Endocrinol 2015 Oct 1;173(4):M39–44.
[20] Learned-Miller E, Lu Q, Paisley A, Trainer P, Blanz V, Dedden K, et al. Detecting acromegaly: screening for disease with a morphable model. Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv 2006;9(Pt 2):495–503.
[21] Kosilek RP, Schopohl J, Grunke M, Reincke M, Dimopoulou C, Stalla GK, et al. Automatic face classification of Cushing's syndrome in women – a novel screening approach. Exp Clin Endocrinol Diab 2013 Oct;121(9):561–4.
[22] Chen S, Pan Z, Zhu H, Wang Q, Yang J-J, Lei Y, et al. Development of a computer-aided tool for the pattern recognition of facial features in diagnosing Turner syndrome: comparison of diagnostic accuracy with clinical workers. Sci Rep [Internet] 2018:8. [cited 2019 Dec 9]Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6006259/.
[23] Chen Y, Liu W, Zhang L, Yan M, Zeng Y. Hybrid facial image feature extraction and recognition for non-invasive chronic fatigue syndrome diagnosis. Comput Biol Med 2015 Sep 1;64:30–9.
[24] Buretic-Tomljanovic A, Petaros A, Suc E, Nincoletti S, Rubeša G, Tomljanović D, et al. Craniofacial morphologic and anthropometric features of Croatian schizophrenia patients and non-psychiatric controls - a pilot study. Anthropol Anz Ber Uber Biol-Anthropol Lit 2012 Nov 1;69:379–97.
[25] Ferry Q, Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, et al. Diagnostically relevant facial gestalt information from ordinary photos. eLife [Internet] 2014:3. [cited 2019 Dec 9]Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4067075/.
[26] Boehringer S, Vollmar T, Tasse C, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, et al. Syndrome identification based on 2D analysis software. Eur J Hum Genet 2006 Oct;14(10):1082–9.
[27] Shukla P, Gupta T, Saini A, Singh P, Balasubramanian R. A deep learning framework for recognizing developmental disorders. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV; 2017. p. 705–14.
[28] Boehringer S, Guenther M, Sinigerova S, Wurtz RP, Horsthemke B, Wieczorek D. Automated syndrome detection in a set of clinical facial photographs. Am J Med Genet A 2011;155(9):2161–9.
[29] Loos HS, Wieczorek D, Würtz RP, von der Malsburg C, Horsthemke B. Computer-based recognition of dysmorphic faces. Eur J Hum Genet 2003 Aug;11(8):555–60.
[30] Vollmar T, Maus B, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, Wieczorek D, et al. Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces. Eur J Med Genet 2008 Jan 1;51(1):44–53.
[31] Cerrolaza JJ, Porras AR, Mansoor A, Zhao Q, Summar M, Linguraru MG. Identification of dysmorphic syndromes using landmark-specific local texture descriptors. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016. p. 1080–3.
[32] Ozdemir ME, Telatar Z, Eroğul O, Tunca Y. Classifying dysmorphic syndromes by using artificial neural network based hierarchical decision tree. Australas Phys Eng Sci Med 2018 Jun 1;41(2):451–61.
[33] Dudding-Byth T, Baxter A, Holliday EG, Hackett A, O'Donnell S, White SM, et al. Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. BMC Biotechnol 2017;17(1):90. 19.
[34] Kuru K, Niranjan M, Tunca Y, Osvank E, Azim T. Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. Artif Intell Med 2014 Oct 1;62(2):105–18.
[35] Knaus A, Pantel JT, Pendziwiat M, Hajjir N, Zhao M, Hsieh T-C, et al. Characterization of glycosylphosphatidylinositol biosynthesis defects by clinical features, flow cytometry, and automated image analysis. Genome Med 2018 09;10(1):3.
[36] Saraydemir S, Taşpınar N, Eroğul O, Kayserili H, Dinçkan N. Down syndrome diagnosis based on gabor wavelet Transform. J Med Syst 2012 Oct;36(5):3205–13.
[37] Dalal AB, Phadke SR. Morphometric analysis of face in dysmorphology. Comput Methods Programs Biomed 2007 Feb 1;85(2):165–72.
[38] Kruszka P, Porras AR, Addissie YA, Moresco A, Medrano S, Mok GT, et al. Noonan syndrome in diverse populations. Am J Med Genet A 2017 Sep;173(9):2323–34.
[39] Shu T, Zhang B. Non-invasive health status detection system using Gabor filters based on facial block texture features. J Med Syst 2015 Feb 27;39(4):41.
[40] Valentine M, Bihm DCJ, Wolf L, Hoyme HE, May PA, Buckley D, et al. Computer-aided recognition of facial attributes for fetal alcohol spectrum disorders. Pediatrics [Internet] 2017 Dec 1;140(6) [cited 2019 Dec 9]Available from: https://pediatrics.aappublications.org/content/140/6/e20162028.
[41] Dowsett L, Porras AR, Kruszka P, Davis B, Hu T, Honey E, et al. Cornelia de Lange syndrome in diverse populations. Am J Med Genet A 2019;179(2):150–8.
[42] Kuru K, Niranjan M, Tunca Y, Osvank E, Azim T. Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. Artif Intell Med 2014 Oct 1;62(2):105–18.
[43] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition CVPR 2001 [Internet]. IEEE Comput. Soc, Kauai, HI, USA; 2001. [cited 2019 Dec 9]. p. I-511-I–518. Available from: http://ieeexplore.ieee.org/document/990517/.
[44] Rohlf FJ. Bias and error in estimates of mean shape in geometric morphometrics. J Hum Evol 2003 Jun 1;44(6):665–83.
[45] Peters JR, Campbell RM, Balasubramanian S. Characterization of the age-dependent shape of the pediatric thoracic spine and vertebrae using generalized procrustes analysis. J Biomech 2017 Oct 3;63:32–40.
[46] Douglas TS, Mutsvangwa TEM. A review of facial image analysis for delineation of the facial phenotype associated with fetal alcohol syndrome. Am J Med Genet A 2010;152A(2):528–36.
[47] Liehr T, Acquarola N, Pyle K, St-Pierre S, Rinholm M, Bar O, et al. Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. Clin Genet 2018;93(2):378–81.
[48] Shen L, Bai L. A review on Gabor wavelets for face recognition. Pattern Anal Appl 2006 Oct 1;9(2):273–92.
[49] Wiskott L, von der Malsburg C. Recognizing faces by dynamic link matching. Neuroimage 1996 Dec 1;4(3):S14–8.
[50] Gumus E, Kilic N, Sertbas A, Ucan ON. Evaluation of face recognition techniques using PCA, wavelets and SVM. Expert Syst Appl 2010 Sep 1;37(9):6404–8.
[51] Vapnik VN. Statistical learning theory. New York: Wiley; 1998. p. 736 p..
[52] Weinberger K.Q., Saul L.K. Distance metric learning for large margin nearest neighbor classification.:38.
[53] Li H, Lin Z, Shen X, Brandt J, Hua G. A convolutional neural network cascade for face detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. Boston, MA, USA. IEEE; 2015. p. 5325–34http://ieeexplore.ieee.org/document/7299170/.
[54] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models editors.. In: Burkhardt H, Neumann B, editors. Computer vision — ECCV'98. Berlin, Heidelberg: Springer; 1998. p. 484–98.
[55] Ashraf AB, Lucey S, Cohn JF, Chen T, Ambadar Z, Prkachin KM, et al. The painful face – pain expression recognition using active appearance models. Image Vis Comput 2009 Oct;27(12):1788–96.

2.2 Marketed algorithms

Two models based on deep learning methods were commercialized and are currently used by physicians to help them with diagnosis.

Face2Gene (FDNA, Inc.), using their DeepGestalt algorithm, operates according to two processes: (1) CLINIC, where the user uploads a facial photograph and obtains a list of 30 most-likely syndromes, with a *Gestalt* score per syndrome, grading the similarities between the tested face and the average face of the syndrome; and a *Features* score, which is obtained after manually selecting HPO terms allowing to refine the diagnosis on photographs; (2) RESEARCH, which allows users to compare up to 6 cohorts uploaded from personal databases by providing classification performances based on AUC and accuracies between these different groups. GestaltMatcher, the other commercially available algorithm, now integrated into the Face2Gene technology through the ULTRA-RARE tab in CLINIC, creates a 'Clinical Face Phenotype Space' which establishes distances between photographs and similarities between patients, and may suggest a molecular diagnosis.

PubMed was used to screen the MEDLINE database using MeSH (Medical Subject Heading) terms along with keywords from titles and abstracts:

("next generation phenotyping"[Text Word] AND "photos"[Text Word]

OR *("Face2Gene"[Text Word] OR "DeepGestalt"[Text Word] OR "FDNA"[Text Word])*

The results were initially screened with the abstract only, to exclude unrelated articles. The articles finally included were analyzed by a single author (QH), according to the following template: lead author, year of publication, syndrome of interest, number of patients included, country of origin of the first author, use of Face2Gene, GestaltMatcher or other tool, use of

the CLINIC or RESEARCH version of Face2Gene, control population (for Face2Gene RESEARCH), comparisons with clinical experts, ethnicity considerations, multiclass or binary comparison, performance, and presence of conflict of interest (defined by one of the authors belonging to FDNA, Inc. or GestaltMatcher).

The request resulted in 155 Pubmed results. After reading the abstracts, 87 were selected: 83 evaluated the DeepGestalt tool (95%), 3 the GestaltMatcher tool (3%), and 1 another tool (1%). 16/87 (18%) were case reports, reporting either the top 10 syndrome ranking proposed by Face2Gene CLINIC or the Gestalt score. 42/87 (48%) used the RESEARCH version of Face2Gene, with comparison to healthy controls or to a differential diagnosis. Controls were systematically matched on age and gender to cases. 8 (9%) papers compared performances to clinical experts. 23/87 (26%) considered the effects of ethnic differences on performance. The design was a multiclass classification in 33/87 (38%), and binary in 54/87 (62%). Performance when using Face2Gene CLINIC was expressed in terms of top 1, top 3, top 10 or top 30 sensitivities, rank position, or Gestalt scores. Performance when using Face2Gene RESEARCH was expressed in terms of Area Under the Curve (AUC), mean accuracy, and/or sensitivity / specificity. Finally, 39/87 (45%) papers declared conflicts of interest with FDNA, Inc. (Supplemental Table 1).

This literature review provided an update on advances in Next-Generation Phenotyping on 2D photographs. The DeepGestalt method from Face2Gene (FDNA, Inc.) offered an algorithm based on deep learning methods following automatic landmark detection. The lead article of Gurovich et al (56) described an algorithm trained on more than 200 syndromes with more than 17,000 photographs and obtained a top 10 sensitivity of 91%. The results of the Face2Gene CLINIC tool were variable across studies, and depended on the performance metrics. Zarate et al (57) found a top 10 sensitivity of 27% in the detection of SATB2-associated disorders, while Martinez-Monseny et al (58) found a top 10 sensitivity of 100% in

the diagnosis of PMM2-CDG. Pascolini et al (59) reports a top 1 sensitivity of only 9% in a multiclass design on 16 syndromes. The results in using the Face2Gene RESEARCH tool were also variable in the studies: Stamberger et al (60) found a non-significant AUC of 61.7% in the NEXMIF encephalopathy classification versus age-, gender-, ethnicity-matched healthy controls.

The study of Porras et al (61) described a new tool, not owned by FDNA Inc., also based on deep learning methods with landmark detection, using geometric morphometrics parameters, referred to a Statistical Shape Model (SSM). These authors trained their model on 2800 photographs comprising 128 syndromes using data augmentation methods. Porras et al (3) described an accuracy of 88% for detecting the presence of a syndrome (90% sensitivity and 86% specificity) and explained that the DeepGestalt tool would identify a syndrome with an accuracy of only 61 to 69% and only perform well on frequent genetic syndromes. We can however note some limitations in this study. Firstly, not all patients had genetic confirmation of their syndrome. Furthermore, Mensah et al (62), a team including members of FDNA Inc., explained in a response to Porras et al (3) that 5 syndromes out of 129 were largely over-represented, accounting for 77% of the photos.

2.3 Prospects for a new Next Generation Phenotyping model

In the Porras et al (61) tool and in DeepGestalt / GestaltMatcher (56), the algorithms were based on a deep learning analysis of the face. Phenotypic elements in profiles, skull outline, hairline, and external ear position and shape were not taken into account. The profile contains essential information for diagnosis in dysmorphology. Todd et al (63) described as major features the facial tilt, nasolabial angle, and glabellar projection in Congenital Central Hypoventilation Syndrome due to the *PHOX2B* variant; Abell et al (64) explained that retrognathia is a major highlight of the diagnosis of *EFTUD2*-related Mandibulo-Facial

Dysostosis with Microcephaly; finally, Tüysüz et al (65) defined prognathism as a common sign in Beckwith-Widemann syndrome. Legendre et al (66) described a characteristic asymmetric external ear with a triangular concha and an absent lobe in a cohort of 119 patients with CHARGE syndrome; Gordon et al (67) described a characteristic ear in *PLCB4*-related auriculo-condylar syndrome, including the presence of a question-mark ear with post-auricular tag. The existing tools did not mention whether facial surgery had been performed on the patients at the time the photograph was taken. Some photographs in the GestaltMatcher database contained images of Treacher Collins syndrome patients with mandibular distractors in place. Bone surgeries (distraction, orthognathic surgery, zygomatic osteotomies, craniofacial surgery) or soft tissue surgeries (canthopexy, lipofilling) may be performed as part of the patient follow-up. It seems questionable to define an average face of a syndrome if the subset of included patients was benefited from mandibular advancement for instance.
Finally, neither of the two tools can be used to detect controls. Even with low reliability, they will suggest the diagnosis of a genetic syndrome for all subjects.

We therefore aimed to develop a new Next Generation Phenotyping tool, trained on full frontal and lateral facial photographs of non-syndromic control children with genetically confirmed genetic syndromes who have not undergone surgery.

3 Material and methods

Here we report the operating pipeline of our model: image pre-processing, automatic landmark placement, extraction of geometric and textural features, stratification using metadata, and classification using a machine learning model **(Figure 2)**.

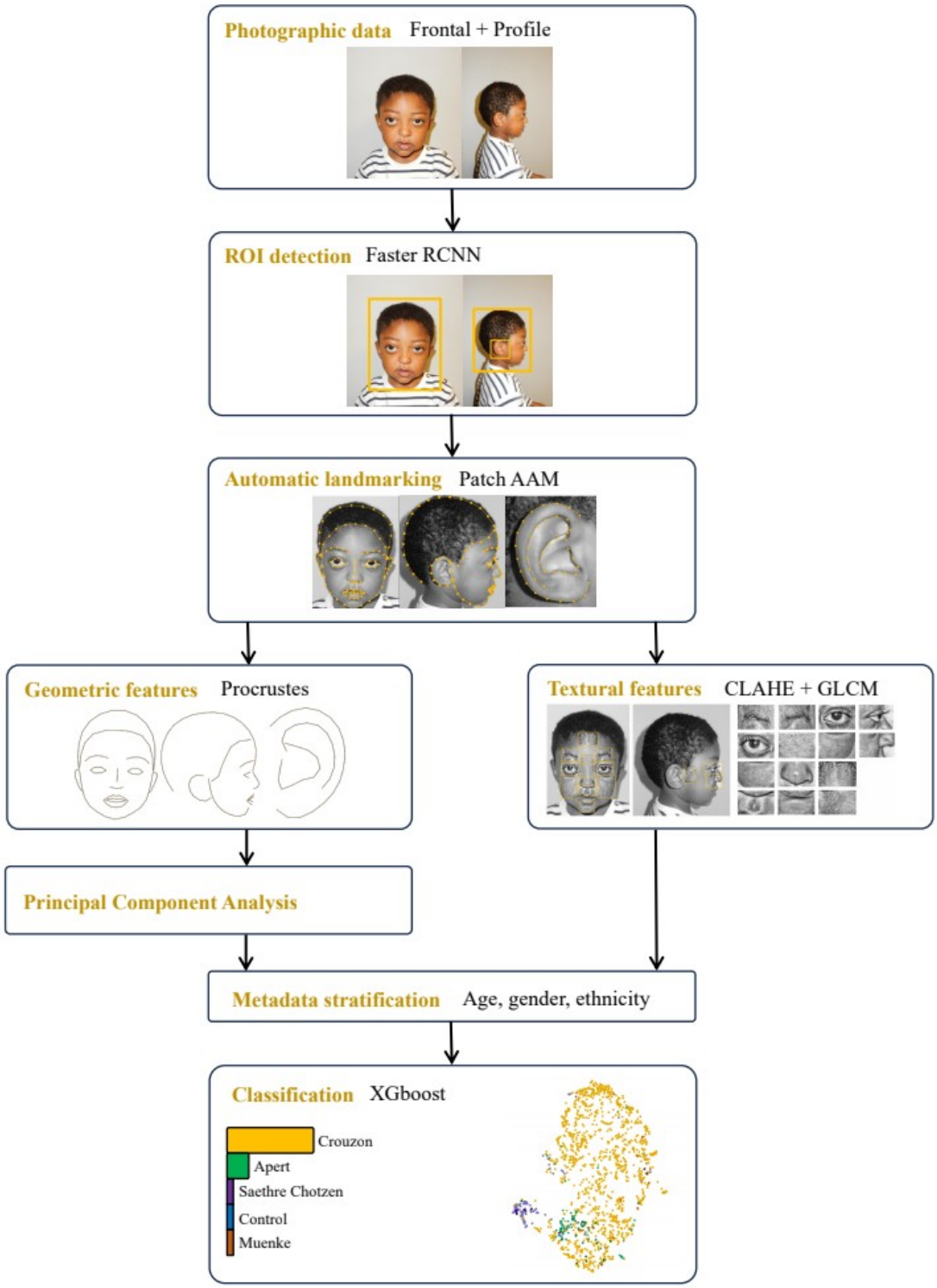


Figure 2. From the initial photograph to diagnostic probability. Faster RCNN = Faster Region-based Convolutional Neural Network**;** CLAHE = Contrast Limited Adaptative Histogram Equalization; GLCM = Gray-level Co-occurrence Matrix; XGboost = eXtreme Gradient Boosting.

3.1 Data description

3.1.1 Training set

As explained in the introduction, we included pictures from the photographic database of the Maxillofacial surgery and plastic surgery and the Medical genetics departments of *Hôpital Necker – Enfants Malades* (Assistance Publique – Hôpitaux de Paris), Paris, France. This database contained 1,042,468 photographs from 22,000 patients followed in the department since 1976. All photographs were taken by a professional medical photographer using a Nikon D7000 device in standardized positions. The non-digitized slide data (before 1985) was scanned using an Epson Perfection V850 Pro scanner.

We included retrospectively and prospectively, from 1976 to 2023, all frontal and lateral pictures of patients diagnosed with 9 syndromes. The photographs were not calibrated. All patients had genetic confirmation of their syndrome. We excluded patients with a history of facial surgery. Multiple photographs per patient corresponded to different ages. Duplicate photographs were excluded.

Non-syndromic children were selected among patients admitted for wounds, trauma, infection, and various skin lesions, without any record of chronic conditions. More precisely, follow-up for any type of chronic disease was considered as an exclusion criterion. The reports were retrieved using Dr Warehouse (68). For each patient, the best profile picture was selected in terms of symmetry.

3.1.2 Validation set

We retrieved frontal and lateral photographs of the syndromes of interest from several specialized centers:

- the Craniofacial unit of Great Ormond Street Hospital (GOSH) in London (UK): Crouzon, Pfeiffer, Apert, Muenke, Saethre Chotzen and Treacher Collins syndromes (Pr. David Dunaway);
- the Center of excellence in genomics and precision dentistry, Chulalongkorn University, Bangkok (Thailand): Treacher Collins, CHARGE, Kabuki syndromes (Dr. Thantrira Porntaveetus);
- the Clinical genetic department of Lausanne University Hospital, Lausanne, (Switzerland): Kabuki syndrome (Dr. Fabienne Giuliano);
- the Maxillofacial surgery department of Lille University Hospital (France): Nager syndrome (Pr. Joël Ferri);
- the Clinical genetics department of Montpellier University Hospital (France): Kabuki, CHARGE, Treacher Collins syndromes (Dr. Marjolaine Willems, Pr. David Geneviève);
- the Maxillofacial department of Nantes University Hospital (France): Guion Almeida, Nager, Treacher Collins syndromes (Pr. Pierre Corre);
- the Clinical genetics department of Tours University Hospital (France): Kabuki syndrome (Pr. Annick Toutain);
- the Clinical genetics department of Grenoble University Hospital (France): Kabuki syndrome (Pr. Klaus Dieterich);

This diversity of validation data adds real strength to the study, especially as certain centers such as London and Bangkok allowed to test the models on ethnically diverse populations. All patients had genetic confirmation of their syndrome. None of the patients in the validation set were present twice, and none were issued from the training set. We excluded patients with a history of facial surgery.

For the control group, we selected photographs from our local database, without any redundancy with the training set, using similar inclusion criteria. We extracted data on age at the time of the photograph and gender. Informed and written consents were obtained from the legal representatives of each child, or from the patient himself if he was of age. All photographs in the validation group were manually annotated by two independent raters, blinded for the diagnosis. The ICC (Intraclass Correlation Coefficient) was computed. ICC values greater than 0.9 corresponded to excellent reliability of the manual annotation (69).

3.2 Model for Automated Landmark Annotation

We used three different templates based on 105 landmarks for the frontal view, 73 for the lateral view and 41 for the external ear. We developed an automatic annotation model for each template following a pipeline including: (1) detection of the region of interest and (2) automatic placement of the landmarks.

3.2.1 Model selection

Step 1: Region Of Interest (ROI) detection

An initial step was the detection of the ROI, *i.e.,* a frontal, a lateral view or an external ear, using bounding boxes. We assessed two bounding box detection methods.

- (1) *Haar* cascades (70), which is the historical method, trained from scratch with the *OpenCV* library (71); a *Haar* feature was a calculation performed on adjacent rectangular regions in a detection window. This calculation was based on pixel sum differences between the regions, on which cascade classifiers were used.
- (2) *Faster R-CNN* (Faster Region-based Convolutional Neural Network) (72), more recent, which were incremented with *detectron2* (73) from a pre-trained model. *Faster*

R-CNN was based on a Convolution Neural Network (CNN) preceded by a region proposal algorithm to generate 'bounding boxes' or locations of possible objects in the image. The Fast RCNN model was trained after data augmentation (images and their +10° and -10° rotations), with a learning rate of 0.001, a batch size of 4, and a gamma of 0.05 and 2000 iterations.

The manual annotations of the bounding boxes were performed with *labelme* (74). We manually annotated a total of 100 facial photographs, 100 profiles, and 100 ears. Data augmentation techniques, i.e., +/– 10 degrees of rotation and flipping, were used to increase the size of the training set (75). In case of multiple detections, the bounding box with the highest probability of detection was selected. The highest weight of the *Haar* cascade and the best confidence score of the Faster R-CNN were selected. Intersection over Union (IoU) was used to define true positives, false positives, and false negatives for the predicted bounding boxes. Then, all methods were compared using precision, recall and F1-score at IoU of 0.5 and 0.75 for models which had been trained with and without data augmentation (76). An IoU of 0.5 would limit false negatives and therefore sensitivity, while an IoU of 0.75 would limit false positives and therefore specificity. The objects of interest were then reduced to 600 × 600 pixels and converted to gray levels.

Step 2: Automatic landmark detection

The objective of this second step was (1) to evaluate the open-source algorithms on their ability to automatically place landmarks and (2) in case of failure, to determine the best method using our data.

We manually annotated a total of 1709 facial photographs, 2186 profiles, and 1443 ears.

The manual landmarking of the frontal, profile, and external ear photographs was performed according to the templates proposed by the *IBUG* group (Intelligent Behavior Understanding

Group, Department of Computing, Imperial College London, United Kingdom) which defined a consensual list of facial landmarks, that we later extended.

The manual annotation of the true landmarks was performed using the *landmarker.io* tool (77).

For each of the three sets, a 5% sample was randomly selected to define a test set. Patients in the test set were removed from the training set to obtain two completely independent sets. Since a patient could have several photographs in the database, all photographs of patients selected for the test set were removed from the training set. The final fitting error was defined as the distance between the final automatic annotation on the validation set and the manual annotation, considered as the gold standard. Random samples were created to induce variations in the size of the training set (n = 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, … , 1500) and each sample was repeated 10 times. The final errors per patient, per run, and per sample size were reported. A linear hierarchical model was produced to account for inter- and intra-individual variability, with the introduction of a polynomial term of degree 3.

Three automatic annotation models were tested using to the *Menpo* platform (78).

1) The **holistic Active Appearance Model** (holistic AAM), a statistical deformable model of the shape and appearance of a deformable object class (79). AMM was a generative model which aims to recover a parametric description of a certain object through optimization during fitting. AAM contained: (1) a shape model, based on a Generalized Procrustes Analysis (GPA) and dimension reduction by Principal Component Analysis (PCA), (2) a motion model, based on a warp function, and (3) an appearance model, after applying the warp function on the shapes (80). The holistic AAM used a holistic appearance representation obtained by warping the texture into the reference frame with a non-linear function.

2) The **patch-based AAM** (patch AAM), representing the appearance in a patch-based fashion: rectangular patches were extracted around the landmark points. Fitting an AAM on a test image involved the optimization of the following cost function with respect to the shape and appearance parameters. The Lucas-Kanade optimization belonged to the family of gradient-descent algorithms (81,82).
3) The **Constrained Local Model** (CLM) was a class of methods for locating sets of points (constrained by a statistical shape model) on a target image. The general approach was to sample a region from the image around the current estimate, project it into a reference frame and search for a combination of points which optimized the total cost, by manipulating the shape model parameters (83,84).

As described on their homepage, the Menpo Project provided: a web-based tool for annotation of bulk data for model training, a command line tool for landmark localization with state-of-the-art pre-trained models, and a generic object detection in terms of a bounding box.
The fitting error per iteration was then calculated. The model with the best performance in the previous step was pre-trained on public databases to determine whether the contribution of a pre-trained model increased performance (Figure 3). Annotated and publicly available online databases of photographs are available (85), i.e., *Labeled Face Parts in the Wild* (LFPW) (86), *Annotated faces-on-the-wild* (AFW) (87), Helen (88) and XM2VTS (89) for frontal facial pictures with a total of 3402 images; *Annotated Facial Landmarks in the Wild* (AFLW) (90) for profiles with a total of 1526 images; Ears in-the-Wild (91) for ears with a total of 2858 images.

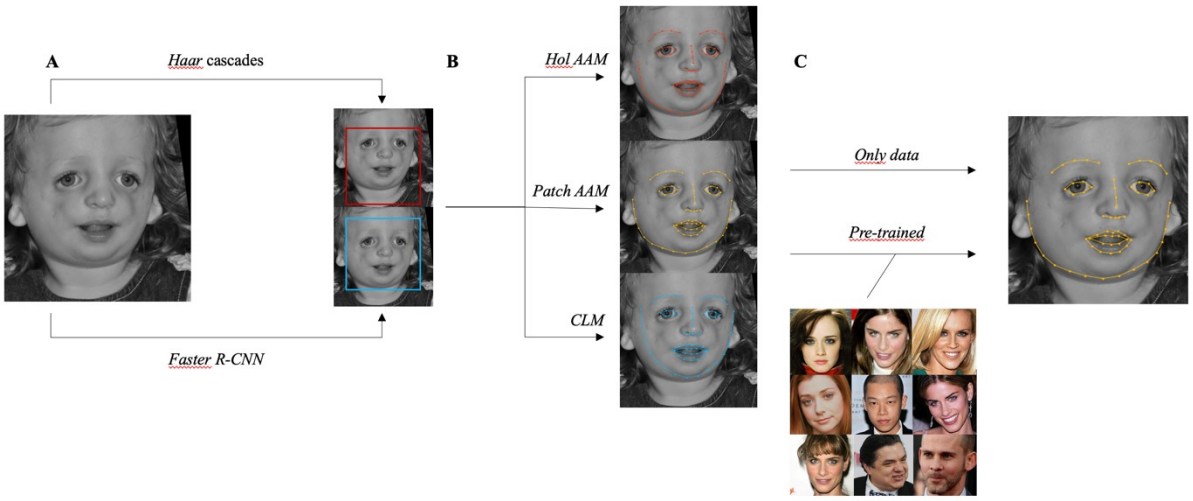


Figure 3. Selection of the best automatic landmarking model. A. ROI detection. B. Testing for three automatic annotation models (Hol AAM, Patch AAM, CLM). C. Comparison of the model only trained on the data versus the model pre-trained on public data, on frontal face photographs. The public photographs in step C were extracted from the LPFW database. Hol = Holistic.

3.2.2 Model optimization

We have attempted to optimize the chosen automatic annotation model using three methods.

- A raw annotation on the base image.
- Annotation by groups: after a raw annotation, a second model for each group of landmarks (eye, nose, mouth, mandibular contour, etc.) was reapplied to refine the placement.
- Double-scale annotation: a raw annotation is made on the image, then a second annotation is performed after resizing the image around the landmarks initially placed.

The best model was selected based on the minimization of the final fitting error.

3.3 Geometric features extraction

As underlined before, Douglas and Mutsvangwa (51) concluded that Procrustes-based approaches were the best methods for shape analysis in terms of prediction accuracy.

We performed Generalized Procrustean Analysis (GPA) (49) on all landmark clouds using the *geomorph* package on R (Figure 4). Since the data were uncalibrated photographs, ROI sizes were not available: shape parameters only were assessed and not centroid sizes.

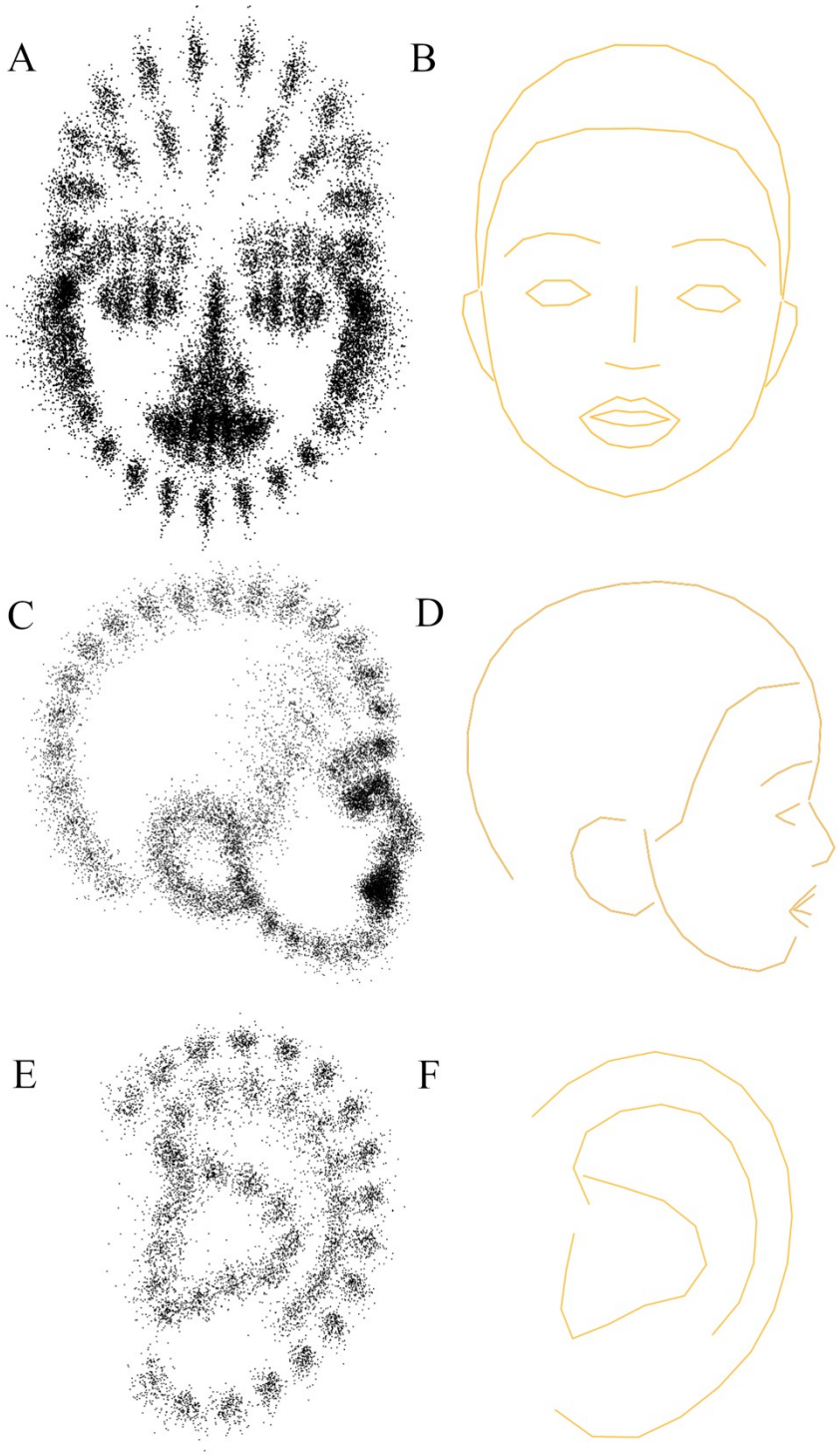


Figure 4. Superposition (A/C/E) and average (B/D/F) of Procrustes coordinates for frontal views (A/B), profiles (C/D), and external ears (E/F). This example includes all patients with Crouzon syndrome.

3.4 Textural features extraction

3.4.1 Image partitioning

We decided to separate the frontal and profile views of the face into key areas and apply textural feature extraction methods to each of these zones. This allowed us to check the results and determine which zone had contributed most to the diagnosis.

We defined 14 key areas, 11 on the frontal (right / left eye, right / left eyebrow, glabella, forehead, nasal tip, philtrum, right / left cheek, chin) and 3 from the profile (pre-auricular region, eye, malar relief) views. Each of these areas could potentially guide towards the diagnosis of facial dysmorphia. Each zone was extracted automatically using the previously placed landmarks (Figure 5).

3.4.2 Histogram equalization

We used the CLAHE (Contrast Limited Adaptative Histogram Equalization) algorithm for histogram equalization. This method has been used used by several authors before the use of feature extractors (92,93). CLAHE enhances contrast by evenly dispersing gray values (94). The aim was to reduce the influence of illumination, particularly at the time of the photograph was taken, or of skin color. Kiflie et al recommend CLAHE as a first-choice equalization method (95).

3.4.3 Gray-level co-occurrence matrix

Gray-level co-occurrence matrix (GLCM) methods, as proposed by Haralick (96), are based on the estimation of the second-order joint conditional probability density functions. These matrices characterize the spatial relationships between pixels. GLCM is one of the most

popular and actual methods for texture analysis (97,98), recently used extensively in radiomics, for instance to analyze CT-scan or MRI images (99–101). GLCM has also proven its effectiveness in analyzing skin texture (102). The co-occurrence matrix contains information on:

1) entropy: local variations in the GLCM;
2) homogeneity: the closeness of the distribution of elements in the GLCM to the GLCM diagonal;
3) contrast, in the GLCM;
4) energy, or angular second moment: sum of squared elements in the GLCM;
5) correlation: the joint probability occurrence of the specified pixel pairs (103).

There are 28 GLCM features, taking into account the average and range for each item of information, for each zone, so 28 x 14 = 394 textural features for each patient (Figure 5).

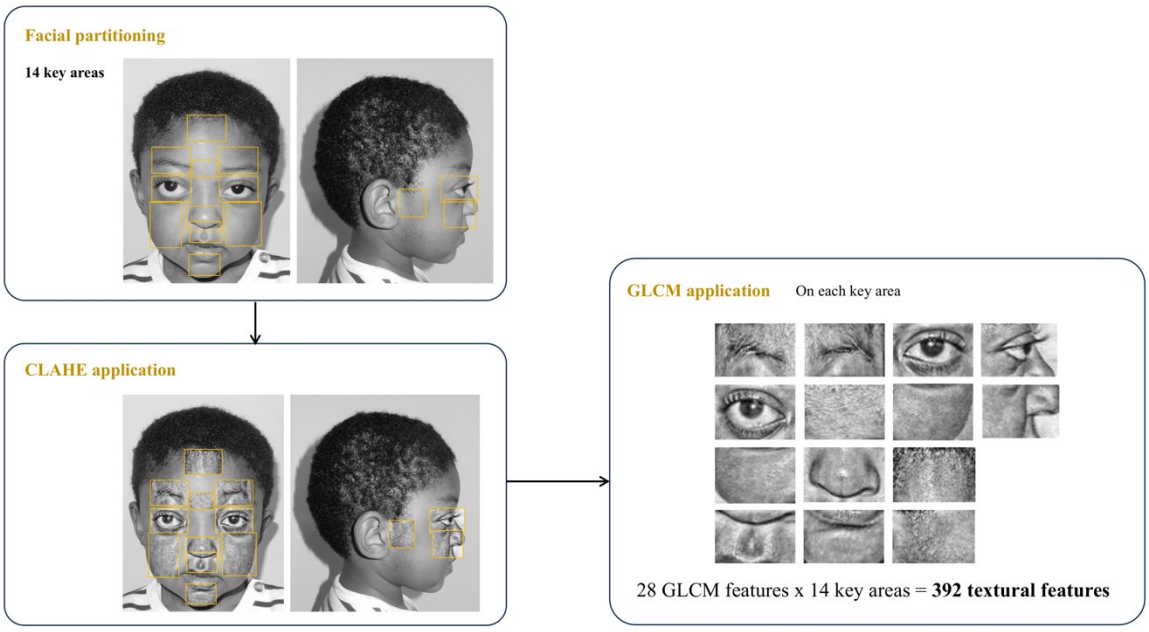


Figure 5. Extraction procedure for textural features. CLAHE = Contrast Limited Adaptative Histogram Equalization; GLCM = Gray-Level Co-occurrence Matrix.

3.5 Dimension reduction and stratification using metadata

Procrustean coordinates were processed using Principal Component Analysis (PCA) for dimension reduction (104,105). We retained the principal components explaining in cumulative sum 99% of the total variance. The last 1% was considered as negligible information.

The geometric features described in the previous paragraph, and the geometric principal components, were combined for further analysis. To consider associated metadata (age and gender) and the fact that we had included more than one photograph per patient (that is the non-independence of the data), a mixed model was designed for each feature. The variables to be explained were the features (geometric and textural), with age, gender, and ethnicity considered as explanatory variables. A random effect on age and individuals was introduced. The equation of the mixed model was:

$$Features_{i,j} \quad \alpha + age.\beta_1 + gender.\beta_2 + ethnicity.\beta_3 + age.\beta_{1,i} + \varepsilon_{i,j}$$

where $age.\beta_{1,i}$ corresponded to a random slope for age per individual and $\varepsilon_{i,j}$ was a random error term. We did not use an interaction term between age and gender and age and ethnicity as it did not increase the likelihood of the model. Age, gender and ethnicity are significant factors in dysmorphology because they influence the diagnosis, and must therefore be taken into account (8,106).

The residuals of each feature were calculated to take into account potential biases linked to the metadata:

$$\varepsilon_{i,j} = Features_{i,j} - \alpha + age.\beta_1 + gender.\beta_2 + ethnicity.\beta_3 + age.\beta_{1,i}$$

3.6 Training the classification model

The inputs to the model were the residuals from the linear models described above, for each geometric or textural feature.

We used XGBoost (eXtreme Gradient Boosting), a supervised machine learning classifier, for all the analyses (107). We chose a tree-based booster, and the loss function to be minimized was a logistic regression in the case of binary classification, or a softmax function in case of multi-class classification. We set a number of hyperparameters to improve the performance and effect of the machine learning model: learning rate = 0.3, gamma = 0, maximum tree depth = 6. The model with the lowest error rate in case of binary classification, or multiclass error rate in case of multi-class classification, was chosen for analysis. We separated the dataset into a training set and a testing set, and a 5-fold cross-validation was used to define the ideal number of iterations to avoid overfitting.

The chosen model with the ideal number of iterations was then used on the independent validation set to test performances by plotting accuracy, sensitivity, specificity, F1-score, precision and recall, and AUC. The ROC (Receiver Operating Characteristics) curves were plotted in R using the *plotROC* package (108).

3.7 UMAP (Uniform Manifold Approximation and Projection) representations

The residuals $\varepsilon_{i,j}$ were represented using UMAP for visual clustering, a nonlinear dimension reduction technique (109). We retained the residuals associated with features with a classification gain (in their cumulative sum) > 0.75 in the importance matrix associated with the XGboost model. A k (local neighborhood size) value of 15 was used. A cosine metric was

introduced to compute distances in high dimensional spaces: the effective minimal distance between embedded points was 10^{-6}. The three conditions of UMAP, namely uniform distribution, local constancy of the Riemannian metric, and local connectivity were verified. UMAP analyses were performed using the package *umap* on R (110).

4 Choice of automatic annotation model

4.1 ROI detection

For all types of ROI (frontal pictures, profiles, and external ears), and for an IoU of 0.5, the best performances based on the F1-score were obtained for the *Faster R-CNN*, in comparison to *Haar cascades*. F1-score were 1.000, 0.999, and 0.969 respectively for frontal pictures, profiles, external ears (111). As stated before, in (111), the Faster RCNN model was trained after data augmentation (images and their +10° and -10° rotations), with a learning rate of 0.001, a batch size of 4, a gamma of 0.05 and 2000 iterations. The Faster RCNN model has since been optimized and split into two stages:

- 1) *ROI detection:* Faster RNN trained on 15633 images after data augmentation (images and their +10° and -10° rotations). There were 6186 frontal images (2062 x 3) and 9447 right and left profile images (3159 x 3). The batch size was set to 2, the learning rate was set to 0.0025, and the maximum number of iterations was 2800.
- 2) *Determination of profile laterality:* pre-trained ResNet50 network (112) using the Pytorch library (113). The training images included 1,570 left profiles and 1,579 right profiles. The batch size was set at 16, an Adam optimizer (114) was used with a learning rate of 0.001, a step of 7, a gamma of 0.1, trained over 25 epochs.

4.1.1 Automatic landmarking

Based on the final fitting error, for all type of annotation, the best model was the patch-based AAM ($p < 0.001$). The pretrained patch-based AAM led to a final error higher than the patch-

based AAM trained from scratch ($p < 0.001$) (111). Since the publication of (111), the patch AAM has been improved:

- 1) *Modification of the templates* proposed by the Menpo platform (85) (Figure 6): 105 frontal landmarks, to take into account the shape of the cranial vault, the hairline and the position of the ears; 73 profile landmarks, to take into account the shape of the cranial vault, the hairline and the position of the ear; 45 ear landmarks, after removing the 'anthelix' group, which was considered to be not reproducible, according to the publication below (4).

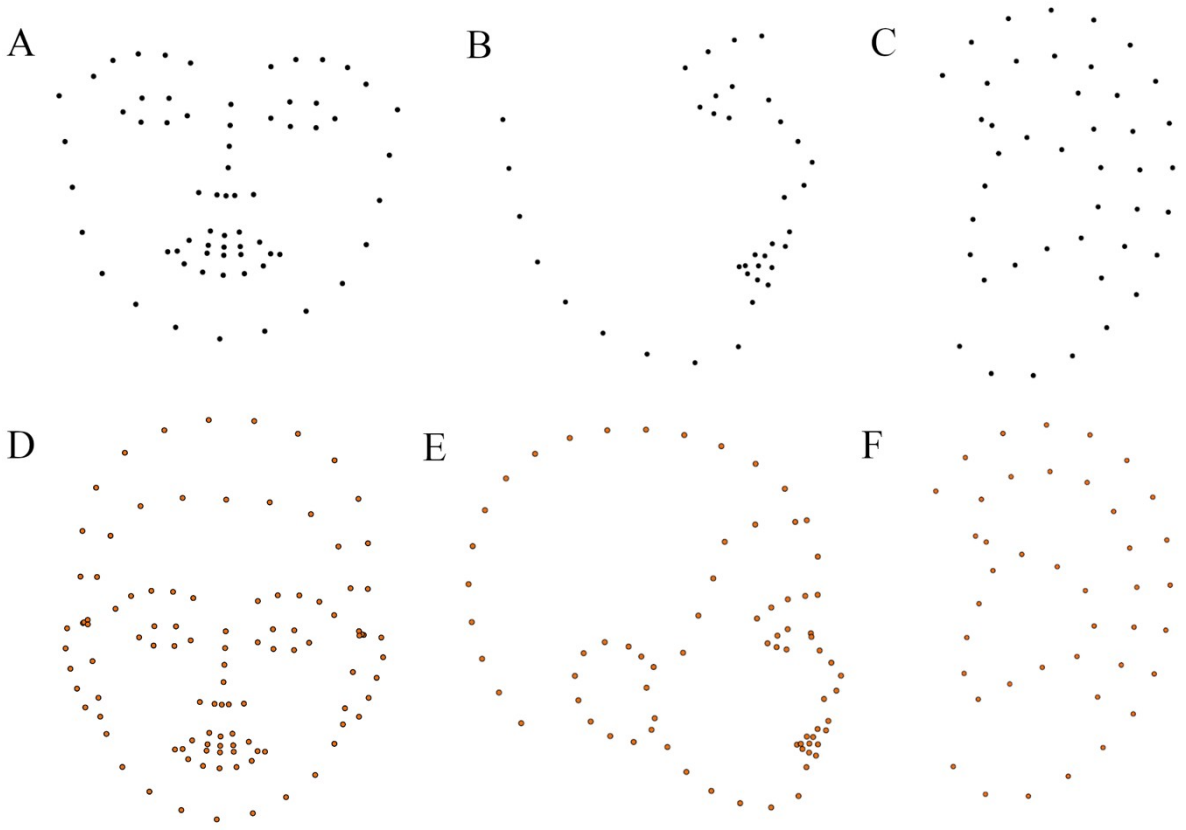


Figure 6. Initial (A/B/C) landmark templates proposed by the Menpo platform versus the updated templates, for the frontal pictures (A/D), profile views (B/E), and external ears (C/F). Note the addition of the shape of the cranial vault, the position of the hairline and

the position and size of the ears for the frontal and profile faces, and the deletion of the anthelix landmarks for the ear.

- 2) *Two-scales landmarking:* the model for frontal pictures was trained on 904 manually annotated photographs, with a first stage of dimensioning (diagonal = 150), a patch shape of [(15, 15), (23, 23)], and 50 iterations, then a second stage without resizing, with a patch shape of [(20, 20), (30, 30)] and 10 new iterations. The final fitting error was then the lowest in this design, with an average of 7.68 +/- 8.14. The model for profile views was trained on 1,439 manually annotated photographs, with a first stage of dimensioning (diagonal = 150), a patch shape of [(15, 15), (23, 23)] and 25 iterations, then a second stage without resizing, with a patch shape of [(15, 15), (23, 23)], and 5 new iterations. The final fitting error was then the lowest in this design, with an average of 11.65 +/- 14.30. The model for ears was trained on 1221 manually annotated photographs, with a first stage of dimensioning (diagonal = 100), a patch shape of [(15, 15), (23, 23)], and 50 iterations, then a second stage without resizing, with a patch shape of [(20, 20), (30, 30)] and 20 new iterations. The final fitting error was then the lowest in this design, with an average of 29.87 +/- 14.50. All three models used the Lucas Kanade optimizer (115).

4.1.2 Manual landmarking reproducibility

The comparison between two raters was excellent with an ICC of 0.999 [0.999-0.999] for frontal pictures, 0.999 [0.999-0.999] for profile views, and 0.992 [0.991-0.993] for external ears. The landmarks defining the antihelix were the less reproducible, because this anatomical region can be missing in some individuals.

4.1.3 Published paper

Received: 2 November 2022 | Revised: 10 January 2023 | Accepted: 12 January 2023

DOI: 10.1002/ajmg.a.63126

ORIGINAL ARTICLE

AMERICAN JOURNAL OF medical genetics PART A WILEY

An automatic facial landmarking for children with rare diseases

Quentin Hennocq[1,2] | Thomas Bongibault[1] | Matthieu Bizière[1] | Ombline Delassus[1] | Maxime Douillet[1] | Valérie Cormier-Daire[3] | Jeanne Amiel[3] | Stanislas Lyonnet[3] | Sandrine Marlin[3] | Marlène Rio[3] | Arnaud Picard[2] | Roman Hossein Khonsari[2] | Nicolas Garcelon[1]

[1]Imagine Institute, INSERM UMR 1163, Paris, France

[2]Département de chirurgie maxillo-faciale et chirurgie plastique pédiatrique, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris, Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou, Faculté de Médecine, Université de Paris Cité, Paris, France

[3]Fédération de médecine génomique, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris, Faculté de Médecine, Université de Paris Cité, Paris, France

Correspondence
Quentin Hennocq, Institut Imagine, 24, boulevard du Montparnasse, 75015 Paris, France.
Email: quentin.hennocq@aphp.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-10-IAHU-01

Abstract

Two to three thousand syndromes modify facial features: their screening requires the eye of an expert in dysmorphology. A widely used tool in shape characterization is geometric morphometrics based on landmarks, which are precise and reproducible anatomical points. Landmark positioning is user dependent and time consuming. Many automatic landmarking tools are currently available but do not work for children, because they have mainly been trained using photographic databases of healthy adults. Here, we developed a method for building an automatic landmarking pipeline for frontal and lateral facial photographs as well as photographs of external ears. We evaluated the algorithm on patients diagnosed with Treacher Collins (TC) syndrome as it is the most frequent mandibulofacial dysostosis in humans and is clinically recognizable although highly variable in severity. We extracted photographs from the photographic database of the maxillofacial surgery and plastic surgery department of Hôpital Necker–Enfants Malades in Paris, France with the diagnosis of TC syndrome. The control group was built from children admitted for craniofacial trauma or skin lesions. After testing two methods of object detection by bounding boxes, a *Haar* Cascade-based tool and a Faster Region-based Convolutional Neural Network (Faster R-CNN)-based tool, we evaluated three different automatic annotation algorithms: the patch-based active appearance model (AAM), the holistic AAM, and the constrained local model (CLM). The final error corresponding to the distance between the points placed by automatic annotation and those placed by manual annotation was reported. We included, respectively, 1664, 2044, and 1375 manually annotated frontal, profile, and ear photographs. Object recognition was optimized with the Faster R-CNN-based detector. The best annotation model was the patch-based AAM ($p < 0.001$ for frontal faces, $p = 0.082$ for profile faces and $p < 0.001$ for ears). This automatic annotation model resulted in the same classification performance as manually annotated data. Pretraining on public photographs did not improve the performance of the model. We defined a pipeline to create automatic

annotation models adapted to faces with congenital anomalies, an essential prerequisite for research in dysmorphology.

KEYWORDS
dysmorphology, geometric morphometrics, landmarks, machine learning, Treacher Collins syndrome

1 | INTRODUCTION

Shape analysis on two-dimensional (2D) pictures can involve computations based on predefined rules, such as in multivariate models used in geometric morphometrics. In machine learning approaches, these predefined rules are replaced by the ability of the model to learn from examples (Rajkomar et al., 2019). The applications of machine learning are increasing in healthcare (Rajkomar et al., 2019), in diverse fields such as radiology (Choy et al., 2018), dermatology (Novoa et al., 2019), and surgery (Loftus et al., 2020).

Relevant inputs of machine learning approached to shape analysis in healthcare are assessments of radiographs, texts, pathology slices, and clinical photographs (Gurovich et al., 2019). Among these sources of data, clinical photographs are particularly important in the diagnosis of craniofacial malformations. In fact, 30%–40% of the 7000 rare diseases have craniofacial anomalies (Hart & Hart, 2009). Many of these syndromes have facial features that are subtle to describe or not known to nonspecialist physicians: their screening requires the eye of an expert in dysmorphology. Due to this specific clinical issue, there have been a recent increase in the number of scientific publications dedicated to the diagnosis of rare conditions using facial photographs (Thevenot et al., 2018).

A tool developed in 2011 by the facial dysmorphology novel analysis (FDNA) group and used by geneticists is *Face2gene* (Gurovich et al., 2019). This algorithm is based on deep learning techniques to make a diagnosis using a photograph as input. The performance of deep learning is robust due to a large amount of training data (Latorre-Pellicer et al., 2020), but a shortcoming raised is that some classification steps are not controlled, leading to the so-called black box effect (Hennocq et al., 2021). The alternative is to decompose the classification phenomenon and add certain amount of patient metadata, essential for diagnosis, such as age, ethnicity, or gender (Lumaka et al., 2017).

A widely used tool in this context is landmark-based geometric morphometrics, as this was used in 81% of the studies, according to a literature review by our team (Hennocq et al., 2021) on phenotyping methods on 2D frontal photographs. Landmarks are precise and reproducible anatomical points. Positioning landmarks is user-dependent and time-consuming. To tackle this issue, automatic landmarking tools have been developed, such as the active appearance model (AAM) and the constrained local model (CLM) for frontal and profile faces but are mainly trained on photographs of healthy adults. The growth of the human face is an allometric, that is, nonlinear phenomenon (Gayon, 2000). A child is not simply a small adult; there is a change in facial ratios at different ages and a tool for automatic annotation of children's faces would probably need to be trained with children (Larson et al., 2018; Nyemb et al., 2014).

Here, we described the pipeline for automatic annotation of frontal, profile, and ear photographs in normal children and in children with a specific craniofacial malformation—Treacher Collins (TC) syndrome—and we evaluated the performance of this tool. This annotation tool will allow to improve the diagnostic performance of the practitioners, to teach the practitioners the features that allowed the diagnosis, and finally to compare patients within the same syndrome. The first step was to detect the object of interest, that is, a face, a profile or an ear, and the second step to train an automatic annotation model using manually annotated images. A supplementary step was to evaluate the quality of the manual annotation and in particular the interrater variability.

2 | MATERIALS AND METHODS

2.1 | Editorial policies and ethical considerations

The work has been carried out in accordance with The Code of Ethics of the World Medical Association. The study was approved by the CESREES (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé) ethics committee (reference 4570023 Bis) and by the CNIL (Commission Nationale Informatique et Libertés, reference MLD/MFI/AR221900). Informed and written consent was obtained from the legal representatives of each child.

2.2 | Dataset

We included pictures from the photographic database of the maxillofacial surgery and plastic surgery department of *Hôpital Necker Enfants Malades* in Paris, France. This database contains 594,000 photographs of 22,000 patients followed in the department since 1981. The photographs were taken by a professional medical photographer using Nikon D7000 in standardized positions, that is, frontal facial pictures with neutral expression and right and left profiles along the Frankfort horizontal plane (Hexsel et al., 2017).

We included all pictures of patients diagnosed with TC syndrome during a medical genetics consultation. Non syndromic children were selected among patients admitted for wounds, trauma, infection and various skin lesions not being part of a chronic disease. Follow-up for any type of chronic disease was considered as an exclusion criterion, as this

chronic disease could be related to a genetic syndrome with a potential described or undescribed facial phenotype (Khonsari et al., 2013). The reports were retrieved using Dr Warehouse (Garcelon et al., 2018).

All frontal and profile pictures of control and TC children were included. Only photographs of control ears were included, as the ears of children with TC were too abnormal ranging from dysplastic ears to microtia or anotia (McElrath & Winters, 2022). For profiles and ears, for each patient, pictures of right and left sides were included.

2.3 | Step 1: Preprocessing by object detection

An initial step was the detection of the object of interest, that is, a frontal face, a profile face or an ear, using bounding boxes. We assessed two bounding box detection methods: (1) the *Haar* cascades (Viola & Jones, 2001) and (2) Faster Region-based Convolutional Neural Network (*Faster R-CNN*) (Ren et al., 2017). The *Haar* cascade model, which is the historical method, was trained from scratch with the *OpenCV* library (Bradski, 2000); the *Faster R-CNN*, more recent, was increment with *detectron2* (Wu et al., 2019) from a pretrained model.

A *Haar* feature is a calculation performed on adjacent rectangular regions in a detection window. This calculation is based on pixel sum differences between the regions, on which cascade classifiers are then used (Viola & Jones, 2001). *Faster R-CNN* is based on a convolution neural network (CNN) preceded by a region proposal algorithm to generate "bounding boxes" or locations of possible objects in the image (Ren et al., 2017).

The manual annotations of the bounding boxes were performed with *labelme* (Labelme, 2022). We manually annotated a total of 100 facial photographs, 100 profiles and 100 ears. Data augmentation techniques, that is, ±10 degrees of rotation and flipping, were used to increase the size of the training set (Pawara et al., 2017).

In case of multiple detections, the bounding box with the highest probability of detection was selected.

The highest weight of the *Haar* cascade and the best confidence score of the Faster R-CNN were selected. Intersection over union (IoU) was used to define true positives, false positives, and false negatives for the predicted bounding boxes. Then, methods were compared with precision, recall and F1-score at IoU of 0.5 and 0.75 for models, which had been trained with and without data augmentation (Padilla et al., 2020). An IoU of 0.5 would limit false negatives and therefore sensitivity, while an IoU of 0.75 would limit false positives and therefore specificity. The objects of interest were then reduced to 600 × 600 pixels and converted to gray levels.

2.4 | Step 2: Training of an automatic landmark detection algorithm

The objective of this second step was (1) to evaluate the open-source algorithms on their ability to automatically place landmarks, and (2) in case of failure, to determine the best method using our data.

We (author QH) manually annotated a total of 1709 facial photographs, 2186 profiles, and 1443 ears.

The manual landmarking of the frontal (68 points), profile (39 points), and ear (55 points) photographs was performed according to the templates proposed by the *IBUG* group (Intelligent Behavi Understanding Group, Department of Computing, Imperial College London, UK) which define a consensual list of facial landmarks (Figure S1).

The manual annotation of the true landmarks was performed using the *landmarker.io* tool (landmarker.io, n.d.).

For each of these three sets, a 5% sample was randomly selected to define a validation set. Patients in the validation set were removed from the training set to obtain two completely independent sets. Since a patient could have several photographs taken, all photographs of patients selected for the validation set were removed from the training set.

The final fitting error was defined as the distance between the final automatic annotation on the validation set and the manual annotation, considered as the gold standard.

Random samples were created to vary the size of the training set (n = 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, ..., 1500) and each sample was repeated 10 times. The final error per patient, per run and per sample size was reported. A linear hierarchical model was produced to account for interindividual and intraindividual variability, with the introduction of a polynomial term of degree 3.

Three automatic annotation models were tested, such as (i) the AAM with its holistic, (ii) the AAM patch-based variants, and (iii) the CLM, using to the *Menpo* platform (Alabort-i-Medina et al., 2014). As described on their homepage, the Menpo Project provides a web-based tool for annotation of bulk data for model training, a command line tool for landmark localization with state-of-the-art pretrained models and a generic object detection in terms of a bounding box.

(i) AAM is a statistical deformable model of the shape and appearance of a deformable object class (Matthews & Baker, 2004). It is a generative model, which aims to recover a parametric description of a certain object through optimization during fitting. AAM contained: (1) a shape model, based on a generalized procrustes analysis (GPA) and a reduction of dimensions by principal component analysis (PCA), (2) a motion model, based on a warp function, (3) an appearance model, after applying the warp function on the shapes (Active Appearance Model, n.d.). The holistic AAM uses a holistic appearance representation obtained by warping the texture into the reference frame with a nonlinear warp function.

(ii) A patched AAM model represents the appearance in a patch-based fashion: rectangular patches are extracted around the landmark points. Fitting an AAM on a test image involves the optimization of the following cost function with respect to the shape and appearance parameters. The Lucas–Kanade optimization belongs to the family of gradient-descent algorithms (Antonakos et al., 2015; Xiong & De la Torre, 2013). The fitting error per iteration was then calculated.

(iii) A CLM is a class of methods for locating sets of points (constrained by a statistical shape model) on a target image. The general approach is to sample a region from the image around the current

estimate, project it into a reference frame and search for a combination of points, which optimized the total cost, by manipulating the shape model parameters (Cristinacce & Cootes, 2006, 2008).

The model with the best performance in the previous step was then pretrained on public databases to find out whether the contribution of a pretrained model increases performance. Indeed, already annotated and publicly available online databases of photographs are available (Deng et al., 2019), that is, *Labeled Face Parts in the Wild* (LFPW) (Belhumeur et al., 2011), *Annotated faces-on-the-wild* (AFW) (Zhu & Ramanan, 2012), Helen (Le et al., 2012), and XM2VTS (Messer et al., 2003) for frontal facial pictures with a total of 3402 images; *Annotated Facial Landmarks in the Wild* (AFLW) (Kostinger et al., 2011) for profiles with a total of 1526 images; Ears in-the-Wild (Zhou & Zaferiou, 2017) for ears with a total of 2858 images.

These steps were summarized in Figure 1. The code is available to readers on the website https://framagit.org/imagine-plateforme-bdd/aidy.

2.5 | Supplementary step: Analysis of manual landmarking reproducibility

The objective of this step was to determine whether manual landmarking between two raters was reproducible, and which landmarks were the most difficult to place that is, those with the most inter-rater variability. A sample of 55 patients was double annotated by two independent reviewers to assess reproducibility. Results were in the form of Kendall's tau (Kendall, 1947), Pearson's r^2 (Gaddis & Gaddis, 1990), and intraclass correlation coefficient (ICC) (Bartko, 1966) on both axes and on centroid size (Lam et al., 2015; Naji et al., 2014). A Kendall's tau greater than 0.7 and an r^2 close to 1 corresponded to a good agreement between the two annotations (Kendall, 1947). An ICC greater than 0.9 corresponded to excellent reliability of the manual annotation (Bartko, 1966).

A landmark reproducibility map based on Pearson's correlation coefficient was produced for the facial, profile, and ear pictures to determine which points were reproducible and therefore the most difficult to place between two different raters.

2.6 | Statistical assessment

A linear hierarchical model, used to represent the final fitting error as a function of sample size, was then designed to account for interindividual and intraindividual variability, as the data were repeated in the same individual and therefore not independent, with the introduction of a polynomial term of degree 3. This polynomial term was introduced as it maximized the likelihood of the model. Secondary models were then designed to compare the different algorithms (patch-based AAM, holistic AAM, and CLM) and to assess the effect of group (controls/TC syndrome) and age category (<2 years old (y.o.), 2–8 y.o., and >8 y.o) variables on the final error. The significance threshold was defined as $p < 0.05$; a significant parameter had an effect on the relevant variables for each model. Assumptions of normality and homoscedasticity of errors were tested. The statistical analyses were performed on *R 3.6.2* using the *nlme* (Pinheiro et al., 2012) and *ggplot* packages (Wickham, 2016).

2.7 | Clinical application: Classification

A classification between TC patients versus controls in the validation set was performed, using either ground-truth landmarks (placed manually) or automatically placed landmarks with the best model from the previous step. For nonanatomically defined landmarks—that is, landmarks 2–16 for frontal faces; 2–11 for profiles; 2–19, 22–35, and 37–50 for ears—we used the sliding semi-

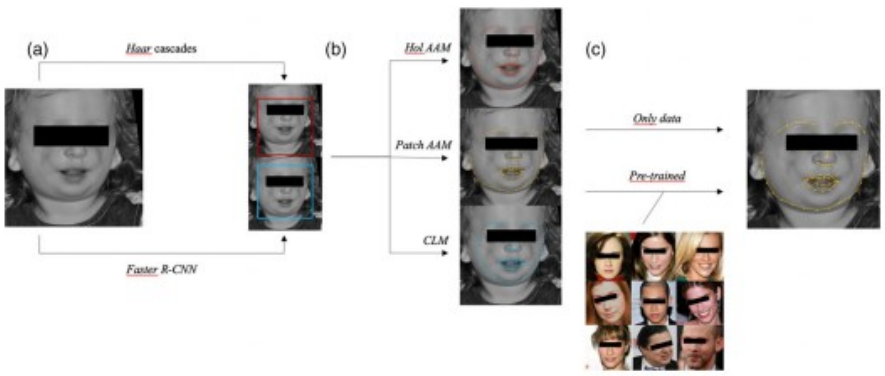


FIGURE 1 Steps for creating an automatic landmarking tool. (a). Bounding box detection. (b). Testing of three automatic annotation models. (c). Comparison of the model only trained on the data versus the model pretrained on public data. Here the examples chosen are frontal face photographs. The public photographs in Step C were extracted from the LPFW database.

landmarks method (Bookstein, 1997) to remove the artifacts from equidistant sampling. We then performed a GPA (Rohlf & Slice, 1990), followed by a PCA. We kept the first components that explained more than 95% of the total variance. We used a MANOVA to determine whether these components were sufficient to classify a patient into the two groups. The significance threshold was defined as $p < 0.05$. These statistical analyses were performed on *R 3.6.2* using the *geomorph* package (Baken et al., 2021).

3 | RESULTS

3.1 | Dataset

In the TC syndrome group, 56% of included patients had a genetic confirmation (*TCOF1* or *POLR1D* mutations) (Table 1).

For frontal facial pictures ($N = 1664$), we included 618 photographs of patients with TC corresponding to 60 patients and 1046 photographs of control patients corresponding to 842 patients. Mean age was 8 years (±5.5) and there were 46% of girls in the dataset of photographs.

For the profile pictures ($N = 2044$), we included 451 photographs of patients with TC corresponding to 60 patients and 1593 photographs of control patients corresponding to 576 patients. Mean age was 7.8 years (±5.2) and there were 53% of girls in the dataset of photographs.

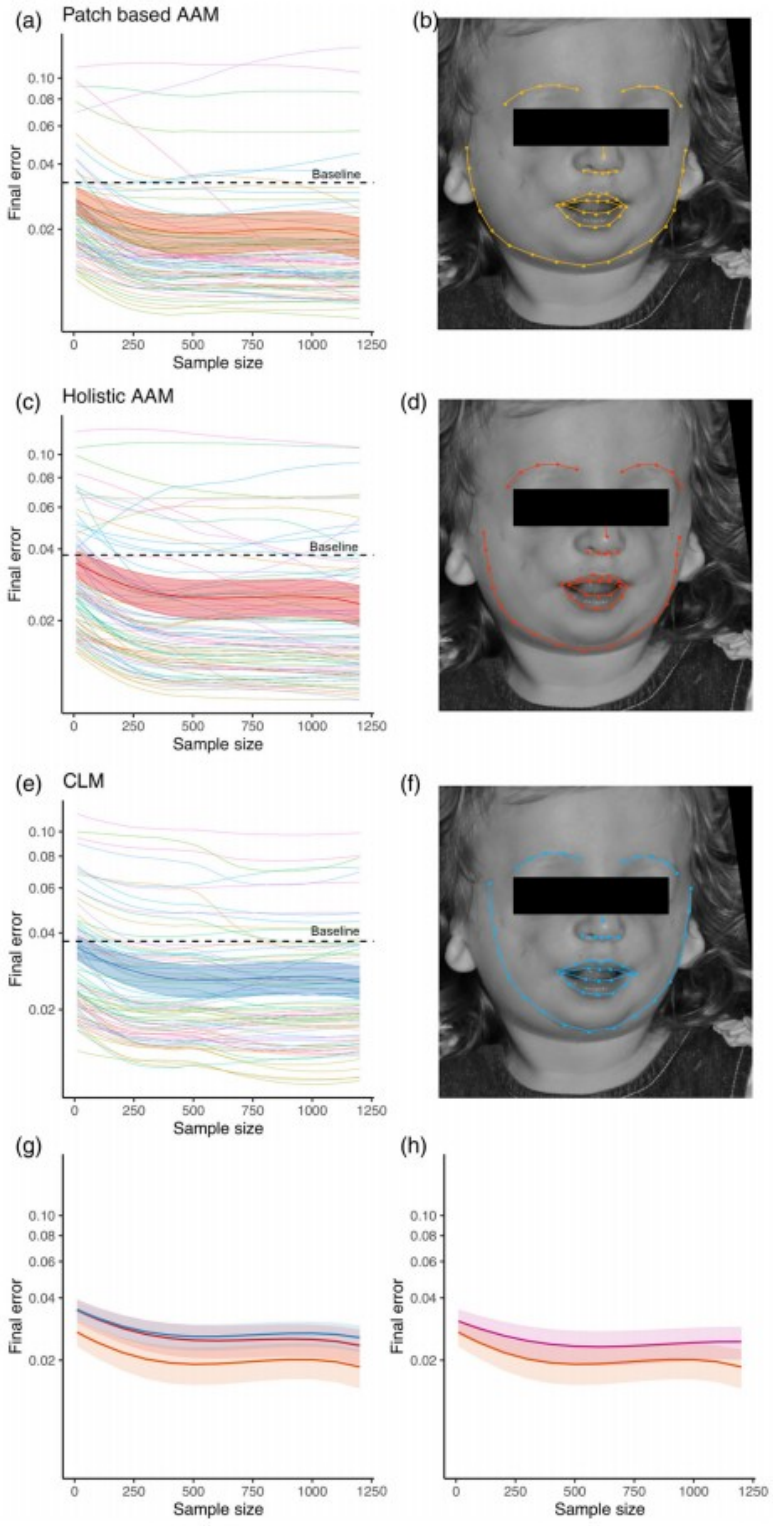
For the ears ($N = 1375$), we included 100% photographs of control patients corresponding to 503 patients. Mean age was 7.5 years (±5.1) and there were 53% of girls.

3.2 | Step 1 evaluation: Preprocessing by object detection

For frontal facial pictures and for an IoU of 0.5, the best performances based on the F1-score were obtained for the *Faster R-CNN* tool with or without data augmentation (F1-score = 1) (Table S1). For an IoU of 0.75, the best performances were obtained for the *Faster R-CNN* tool without data augmentation (F1-score = 1).

TABLE 1 (A) Dataset of photographs and (B) population description.

(A)		Frontal	Profile	Ears
N		1664	2044	1375
Gender	Girls	770 (46%)	1093 (53%)	733 (53%)
Group				
	Treacher Collins	618 (37%)	451 (22%)	0 (0%)
	Controls	1046 (63%)	1593 (78%)	1375 (100%)
Age (years)				
	Mean ± SD	8.0 ± 5.5	7.8 ± 5.2	7.5 ± 5.1
	Median	7.4	7.6	7.7
	Min	0	0	0
	Max	27	27	22
	<2 y.o	257 (16%)	326 (16%)	276 (20%)
	2–8 y.o	655 (40%)	753 (37%)	450 (67%)
	>8 y.o	739 (45%)	949 (47%)	649 (47%)
(B)		**Frontal**	**Profile**	**Ears**
N		902	576	503
Gender	Girls	463 (51%)	303 (53%)	268 (53%)
Group				
	Treacher Collins	60 (7%)	60 (10%)	0 (0%)
	Controls	842 (93%)	516 (90%)	503 (53%)
Age (years)				
	Mean ± SD	6.6 ± 4.8	7.5 ± 5.9	7.2 ± 5.3
	Median	6.2	7.2	7.2
	Min	0	0	0
	Max	27	27	22
	<2 y.o	204 (23%)	103 (18%)	102 (20%)
	2–8 y.o	377 (42%)	225 (39%)	183 (36%)
	>8 y.o	321 (36%)	248 (43%)	218 (43%)



FIUGRE 2 Automatic annotation models for frontal facial pictures. (a) Patch-based AAM. Prediction of the final error as a function of the training sample size, with its credibility interval. Dotted line: final error obtained only with the public photographic base. (b). Automatic annotation with the patch-based AAM trained on 1200 photographs, on a child with Treacher Collins syndrome. (**c**). Holistic AAM: prediction of the final error as a function of the training sample size, with its credibility interval. (d) Automatic annotation with the holistic AAM trained on 1200 photographs. (**e**) CLM: prediction of the final error as a function of the training sample size, with its credibility interval. (f). Automatic annotation with the CLM trained on 1200 photographs. (g). Comparison of the patch-based AAM (orange), holistic AAM (red) and CLM (blue) models, with their credibility intervals. (**h**) Comparison of the patch-based AAM trained from scratch (orange), and the pretrained patch-based AAM (purple).

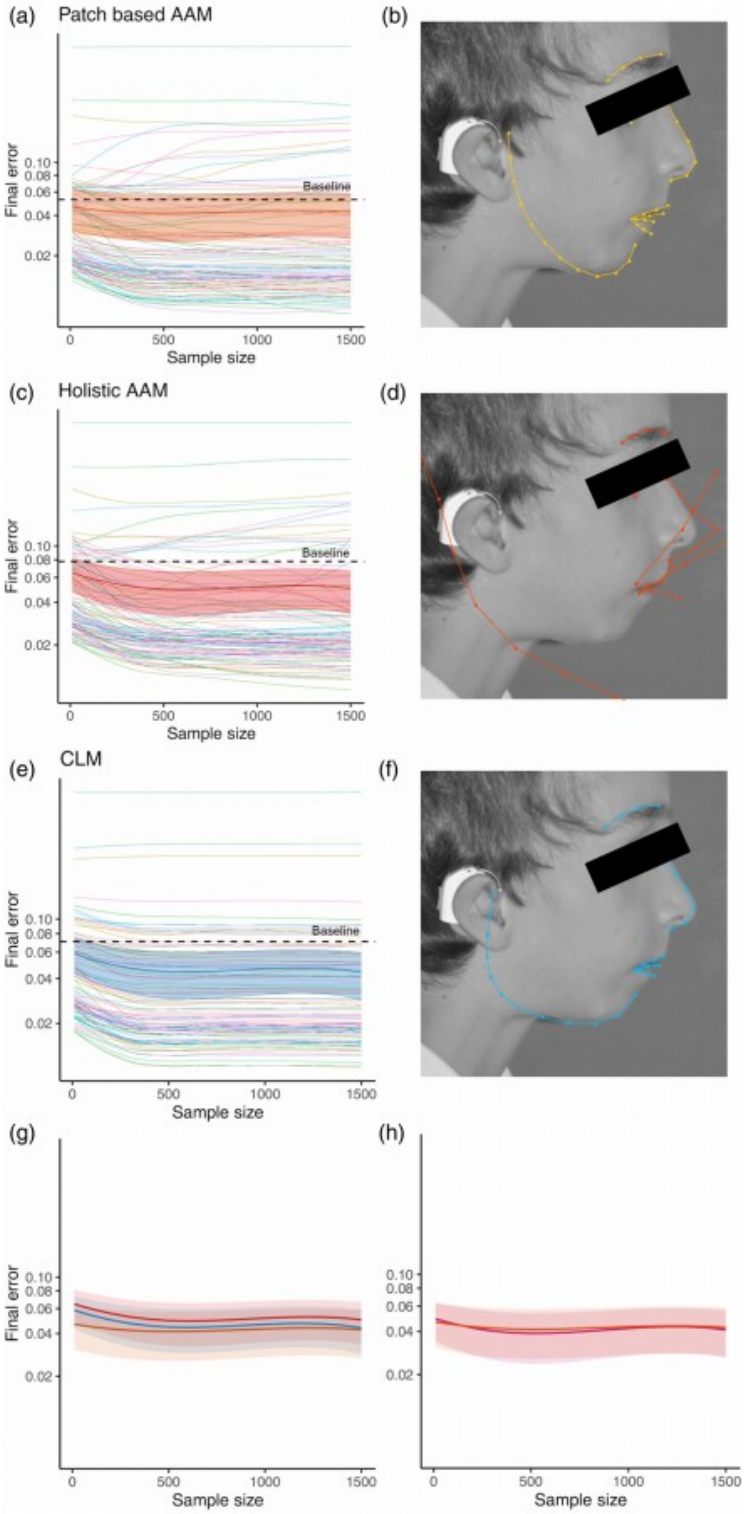


FIGURE 3 Legend on next page.

For profile pictures and for an IoU of 0.5, the best performances based on the F1-score were obtained for the *Faster R-CNN* tool without data augmentation (F1-score = 0.999). For an IoU of 0.75, the best performances were also obtained for the *Faster R-CNN* tool without data augmentation (F1-score = 0.969).

For ears and for an IoU of 0.5, the best performances were obtained for the *Faster R-CNN* tool with or without data augmentation (F1-score = 0.984). For an IoU of 0.75, the best performances were also obtained for the *Faster R-CNN* tool without data augmentation (F1-score = 0.944).

3.3 | Step 2 evaluation: Training of an automatic landmark detection algorithm

First, each photo was automatically annotated only with publicly available data corresponding to the baseline (Table S3 and S4). For front faces, the final errors were 0.033 (±0.021), 0.038 (±0.024), and 0.037 (±0.037) for the AAM patch, holistic AAM, and CLM models, respectively. For profile faces, the final errors were 0.053 (±0.076), 0.078 ± 0.091, and 0.071 ± 0.079. For the ears, the results were 0.450 (±0.045), 0.485 (±0.055), and 0.420 (±0.043). These final errors, represented by the dotted lines in Figures 2–4, were all larger than those obtained by algorithms trained on our data.

Using our annotated data, a final fitting error rate was calculated for training sets containing from 10 to 1200 images for frontal facial pictures, 1500 for profiles, and 1100 for ears.

For frontal facial pictures, the best model was the patch-based AAM ($p < 0.001$) (Figure 2). The pretrained patch-based AAM led to a final error higher than the patch-based AAM trained from scratch ($p < 0.001$). Patients with TC had a higher final error than controls ($p < 0.001$). There was no significant difference for age categories, but a trend toward less error for older children (>8 y.o vs. <2 y.o; $p = 0.077$) (Figure S3).

3.4 | Supplementary step evaluation: Analysis of manual landmarking reproducibility

For frontal facial pictures, the comparison between two raters was excellent with a Kendall's tau of 0.972, an r^2 (Pearson coefficient) of 0.999, an ICC of 0.999 [0.999–0.999] on the axes and 1 (Rajkomar et al., 2019) on the centroid size (Figure S2).

For profiles, the comparison between two raters was excellent with a Kendall's tau of 0.963, an r^2 (Pearson coefficient) of 0.998, an ICC of 0.999 [0.999–0.999] on the axes and 1 (Rajkomar et al., 2019) on the centroid size. The mandibular angle did not show satisfactory reproducibility, as this anatomical region is less pronounced in patients with TC syndrome.

For the ears, the comparison between two evaluators was also excellent with a Kendall's tau of 0.936, an r^2 of 0.992, an ICC of 0.992 [0.991–0.993] on the axes and 0.999 [0.999–0.999] on the centroid size. The landmarks defining the antihelix were the less reproducible, because this anatomical region can be missing in some individuals.

3.5 | Clinical application: Classification

GPA followed by PCA was then performed for the frontal and profile images, for (1) the manually annotated photos ("ground-truth") and (2) the automatically annotated photos by our best model (the patch-based AAM). The first 8 principal components, explaining more than 95% of the total variance were tested using MANOVA to predict if the patient had TC or was a control. Thus, face landmarks were predictive of TC syndrome in the ground-truth data and in the automatically annotated data, whereas profile landmarks were not in both groups. The results of the manually annotated data were reproducible with the automatically annotated data (Table 2).

4 | DISCUSSION

Based on our results, it appears that the best tools to create an automatic annotation tool are (1) a detection algorithm, based on a Faster R-CNN to generate a reliable bounding box around the object without data augmentation, then (2) a patch-based AAM not pretrained on public data. AAM seems to be a robust algorithm in medical image analysis (Cootes et al., 2001; Wilms et al., 2017). The use of open-source algorithms trained on healthy adults all performed worse than the algorithms trained on our data.

Most papers used *Haar* cascades for object detection, but it turns out that this detection tool was described by some authors as not very effective in uncontrolled environments with varying exposure angles and lighting (Hennocq et al., 2021; Zhang & Zhang, 2010). A

FIGURE 3 Automatic annotation models for profile faces. (a) Patch-based AAM: prediction of the final error as a function of the training sample size, with its credibility interval. Dotted line: final error obtained only with the public photographic base. (b) Automatic annotation with the patch-based AAM trained on 1500 photographs, on a child with Treacher Collins syndrome. (c) Holistic AAM: prediction of the final error as a function of the training sample size, with its credibility interval. (d) Automatic annotation with the holistic AAM trained on 1500 photographs. (e) CLM: prediction of the final error as a function of the training sample size, with its credibility interval. (f) Automatic annotation with the CLM trained on 1500 photographs. (g) Comparison of the patch-based AAM (orange), holistic AAM (red), and CLM (blue) models, with their credibility intervals. (h) Comparison of the patch-based AAM trained from scratch (orange), and the pretrained patch-based AAM (purple).

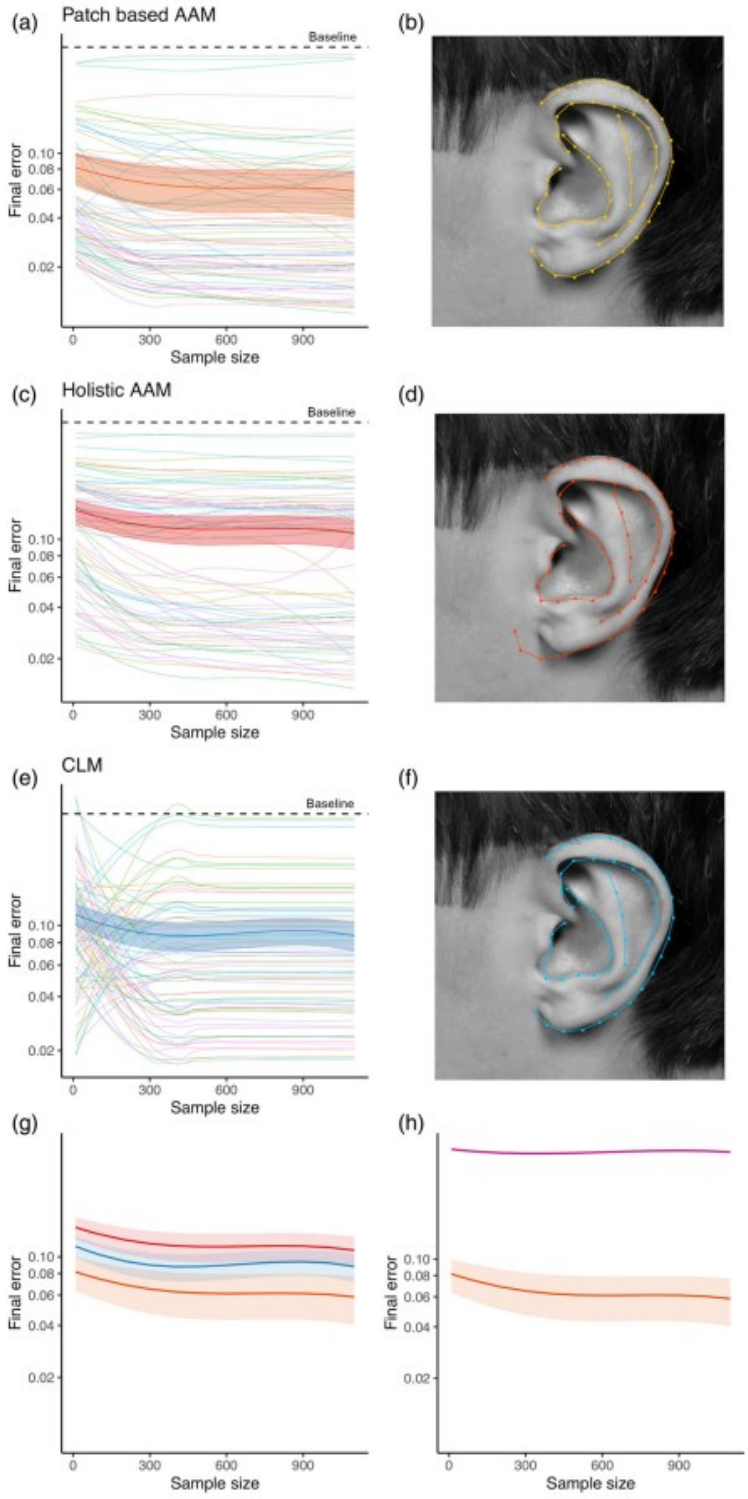


FIGURE 4 Legend on next page.

few authors used a R-CNN-based tool for frontal and profile facial detection (Gurovich et al., 2019; Li et al., 2015).

Among the studies using landmarks (81%), 27% used manual landmarking and 73% used automatic landmarking. Some authors used the AAM as an automatic annotation tool (Yi et al., 2014). None has studied the external ear.

Some authors (Gurovich et al., 2019) did not include any landmarks or geometric features extraction step in their diagnostic tool for facial malformations. In their approach, the face was automatically cropped into several regions such as the eyes or the mouth and CNN were applied to classify patients straight on the pixels. Their approach was therefore a strict deep learning technique, using a *black box* diagnosis principle without a phenotype description process. The system propose allows, through the placement of landmarks and the extraction and transformation of geometric features, to carry out a clinically relevant phenotypic description for each syndrome, while taking into account all the data useful for diagnosis, such as age, sex, ethnicity, or genotype of the patient. This method will allow to compare patients with no diagnosis among them and also with patients considered similar and diagnosed with a syndromic condition, in order to guide genetic investigations.

As previously described in the results section, issues with manual landmarking were reported on facial pictures, profiles, and ears. For profile pictures, the mandibular angle did not show satisfactory reproducibility, as this anatomical region is less pronounced in patients with TC syndrome. For ears, the landmarks defining the antihelix were the less reproducible, because this anatomical region can be missing in some individuals (Figure S2). These results were nevertheless excellent, with, respectively, a general Kendall's tau of 0.972, 0.963, and 0.936 for frontal facial pictures, profiles, and ears, which can be considered as a proof of the excellent reliability of manual annotation.

For all annotated structures, the final error seemed to stabilize for a training sample size of 500. By observing the results of automatic annotation beyond this threshold of 500 images with a patch-based AAM, which was significantly the best annotation model for our type of data, the results were satisfactory. The pretrained models based on public photographs did not improve the annotation performance, probably because the public data were too different from the photographs included into the study (several very young children and children with craniofacial malformations).

In the process of developing a clinically relevant diagnostic tool based on a machine learning approach of shape analysis for facial 2D picture, we performed a GPA and then a dimension reduction using PCA to compare our two groups. The results were reproducible between the manually annotated photos and the automatically annotated photos by patch-based AAM, a point that adds strength to our tool. The final step of the process will consist in the application of machine learning methods to enable the classification of children according to a syndrome, or within the same syndrome for different genotypes. For instance, our approach will allow to objectively determine whether TC includes different phenotypes, with potential phenotype/genotype correlations based on the presence of *TCOF1* or *POLR1D* mutations. These methods can thus be applied to other craniofacial syndromes that are sometimes difficult to diagnose. The clinical features of a syndrome can thus be objectively described, and contribute to deeper phenotyping. Genotype/phenotype correlations within the same syndrome can furthermore be established by including metadata.

The use of open-source algorithms trained on healthy adults all performed worse than the algorithms trained on our data. Here, we provide a pipeline to develop an automatic annotator for assessing craniofacial anomalies of frontal, profiles, and ears photographs for children of various age, sometimes with significant facial dysmorphia. This annotation tool could also be used for the assessment of other part of the face such as the oculo-palpebral region, the front, the lips, or the extremities, using the same pipeline. Such an automatic annotation tool is a prerequisite in the field of artificial intelligence, where large volumes of data are often required, and for which manual annotation is difficult to implement.

TABLE 2 Comparison of the predictability of TC syndrome for manually placed (ground-truth) and automatically placed (patch-based AAM) landmarks.

	Ground-truth	Patch-based AAM
Frontal	0.001*	0.040*
Profile	0.719	0.621

Note: A significant result (*$p < 0.05$) by MANOVA means that the eight principal components are statistically different between the TC and control groups.

AUTHOR CONTRIBUTIONS

Quentin **Hennocq**: Conceptualization, methodology, writing; Thomas **Bongibault**, Matthieu **Bizière**, Ombline **Delassus** and Maxime **Douillet**: Conceptualization, methodology; Arnaud **Picard**, Valérie **Cormier-Daire**, Jeanne **Amiel**, Stanislas **Lyonnet** and Sandrine **Marlin**: Reviewing and Editing; Marlène **Rio**, Roman Hossein **Khonsari** and Nicolas **Garcelon**: Conceptualization, methodology, supervision, reviewing and editing.

FIGURE 4 Automatic annotation models for ears. (a) Patch-based AAM: prediction of the final error as a function of the training sample size, with its credibility interval. Dotted line: final error obtained only with the public photographic base. (b) Automatic annotation with the patch-based AAM trained on 1100 photographs, on an example of ear. (c) Holistic AAM: prediction of the final error as a function of the training sample size, with its credibility interval. (d) Automatic annotation with the holistic AAM trained on 1100 photographs. (e) CLM: prediction of the final error as a function of the training sample size, with its credibility interval. (f) Automatic annotation with the CLM trained on 1100 photographs. (g) Comparison of the patch-based AAM (orange), holistic AAM (red) and CLM (blue) models, with their credibility intervals. (h) Comparison of the patch-based AAM trained from scratch (orange), and the pretrained patch-based AAM (purple).

FUNDING INFORMATION

This work was funded by the "Agence Nationale de la Recherche," "Investissements d'Avenir" program (ANR-10-IAHU-01), and by France 2030 grant "Face4Kids" (ANR-21-PMRB-0004).

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Quentin Hennocq https://orcid.org/0000-0001-7882-8287

REFERENCES

Active Appearance Model. (n.d.). The Menpo Project. https://www.menpo.org/menpofit/aam.html

Alabort-i-Medina, J., Antonakos, E., Booth, J., Snape, P., & Zafeiriou, S. (2014). Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 679–682). ACM. https://doi.org/10.1145/2647868.2654890

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., & Zafeiriou, S. P. (2015). Feature-based Lucas–Kanade and active appearance models. *IEEE Transactions on Image Processing*, *24*, 2617–2632. https://doi.org/10.1109/TIP.2015.2431445

Baken, E. K., Collyer, M. L., Kaliontzopoulou, A., & Adams, D. C. (2021). Geomorph v4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution*, *12*, 2355–2363. https://doi.org/10.1111/2041-210X.13723

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*, 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2011). Localizing parts of faces using a consensus of exemplars. In *CVPR 2011* (Vol. 8). IEEE.

Bookstein, F. L. (1997). Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Medical Image Analysis*, *1*, 225–243. https://doi.org/10.1016/S1361-8415(97)85012-8

Bradski, G. (2000). The OpenCV library. *Dr Dobb's Journal: Software Tools for the Professional Programmer*, *120*, 25.

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., & Dreyer, K. J. (2018). Current applications and future impact of machine learning in radiology. *Radiology*, *288*, 318–328. https://doi.org/10.1148/radiol.2018171820

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 681–685. https://doi.org/10.1109/34.927467

Cristinacce, D., & Cootes, T. (2008). Automatic feature localisation with constrained local models. *Pattern Recognition*, *41*, 3054–3067. https://doi.org/10.1016/j.patcog.2008.01.024

Cristinacce, D., & Cootes, T. F. (2006). Feature detection and tracking with constrained local models. In *Proceedings of the British machine vision conference 2006* (pp. 95.1–95.10). British Machine Vision Association. https://doi.org/10.5244/C.20.95

Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., & Zafeiriou, S. (2019). The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. *International Journal of Computer Vision*, *127*, 599–624. https://doi.org/10.1007/s11263-018-1134-y

Gaddis, M. L., & Gaddis, G. M. (1990). Introduction to biostatistics: Part 6, correlation and regression. *Annals of Emergency Medicine*, *19*, 1462–1468. https://doi.org/10.1016/s0196-0644(05)82622-8

Garcelon, N., Neuraz, A., Salomon, R., Faour, H., Benoit, V., Delapalme, A., Munnich, A., Burgun, A., & Rance, B. (2018). A clinician friendly data warehouse oriented toward narrative reports: Dr Warehouse. *Journal of Biomedical Informatics*, *80*, 52–63. https://doi.org/10.1016/j.jbi.2018.02.019

Gayon, J. (2000). De la Croissance relative à l'allométrie (1918-1936)/from relative growth to allometry (1918-1936). *Revue d'Histoire Des Sciences*, *53*, 475–498. https://doi.org/10.3406/rhs.2000.2095

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P. M., Kamphausen, S. B., Zenker, M., Bird, L. M., & Gripp, K. W. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, *25*, 60–64. https://doi.org/10.1038/s41591-018-0279-0

Hart, T. C., & Hart, P. S. (2009). Genetic studies of craniofacial anomalies: Clinical implications and applications. *Orthodontics & Craniofacial Research*, *12*, 212–220. https://doi.org/10.1111/j.1601-6343.2009.01455.x

Hennocq, Q., Khonsari, R. H., Benoît, V., Rio, M., & Garcelon, N. (2021). Computational diagnostic methods on 2D photographs: A review of the literature. *Journal of Stomatology, Oral and Maxillofacial Surgery*, *122*, e71–e75. https://doi.org/10.1016/j.jormas.2021.04.003

Hexsel, D., Hexsel, C. L., Dal'Forno, T., Schilling de Souza, J., Silva, A. F., & Siega, C. (2017). Standardized methods for photography in procedural dermatology using simple equipment. *International Journal of Dermatology*, *56*, 444–451. https://doi.org/10.1111/ijd.13500

Kendall, M. G. (1947). The variance of tau when both rankings contain ties. *Biometrika*, *34*, 297.

Khonsari, R. H., Ohazama, A., Raouf, R., Kawasaki, M., Kawasaki, K., Porntaveetus, T., Ghafoor, S., Hammond, P., Suttie, M., Odri, G. A., Sandford, R. N., Wood, J. N., & Sharpe, P. T. (2013). Multiple postnatal craniofacial anomalies are characterized by conditional loss of polycystic kidney disease 2 (Pkd2). *Human Molecular Genetics*, *22*, 1873–1885. https://doi.org/10.1093/hmg/ddt041

Kostinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 2144–2151). IEEE. https://doi.org/10.1109/ICCVW.2011.6130513

Labelme, W. K. (2022). Image polygonal annotation with Python. https://doi.org/10.5281/zenodo.5711226

Lam, S. H. F., Bailitz, J., Blehar, D., Becker, B. A., Hoffmann, B., Liteplo, A. S., Rajan, K. B., & Lambert, M. (2015). Multi-institution validation of an emergency ultrasound image rating scale-A pilot study. *The Journal of Emergency Medicine*, *49*, 32–39.e1. https://doi.org/10.1016/j.jemermed.2015.01.010

landmarker.io. (n.d.). The Menpo Project. https://www.menpo.org/landmarkerio/

Larson, J. R., Manyama, M. F., Cole, J. B., Gonzalez, P. N., Percival, C. J., Liberton, D. K., Ferrara, T. M., Riccardi, S. L., Kimwaga, E. A., Mathayo, J., Spitzmacher, J. A., Rolian, C., Jamniczky, H. A., Weinberg, S. M., Roseman, C. C., Klein, O., Lukowiak, K., Spritz, R. A., & Hallgrimsson, B. (2018). Body size and allometric variation in facial shape in children. *American Journal of Physical Anthropology*, *165*, 327–342. https://doi.org/10.1002/ajpa.23356

Latorre-Pellicer, A., Ascaso, Á., Trujillano, L., Gil-Salvador, M., Arnedo, M., Lucia-Campos, C., Antoñanzas-Pérez, R., Marcos-Alcalde, I., Parenti, I., Bueno-Lozano, G., Musio, A., Puisac, B., Kaiser, F. J., Ramos, F. J., Gómez-Puertas, P., & Pié, J. (2020). Evaluating Face2Gene as a tool to identify Cornelia de Lange syndrome by facial phenotypes.

International Journal of Molecular Sciences, *21*, E1042. https://doi.org/10.3390/ijms21031042

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer vision – ECCV 2012. ECCV 2012* (Vol. 7574, pp. 679–692). Springer. https://doi.org/10.1007/978-3-642-33712-3_49

Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5325–5334). IEEE. https://doi.org/10.1109/CVPR.2015.7299170

Loftus, T. J., Tighe, P. J., Filiberto, A. C., Efron, P. A., Brakenridge, S. C., Mohr, A. M., Rashidi, P., Upchurch, G. R., Jr., & Bihorac, A. (2020). Artificial intelligence and surgical decision-making. *JAMA Surgery*, *155*, 148–158. https://doi.org/10.1001/jamasurg.2019.4917

Lumaka, A., Cosemans, N., Lulebo Mampasi, A., Mubungu, G., Mvuama, N., Lubala, T., Mbuyi-Musanzayi, S., Breckpot, J., Holvoet, M., de Ravel, T., van Buggenhout, G., Peeters, H., Donnai, D., Mutesa, L., Verloes, A., Lukusa Tshilobo, P., & Devriendt, K. (2017). Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clinical Genetics*, *92*, 166–171. https://doi.org/10.1111/cge.12948

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*, 135–164. https://doi.org/10.1023/B:VISI.0000029666.37597.d3

McElrath, A. D., & Winters, R. (2022). *Mandibulofacial dysostosis*. StatPearls Publishing.

Messer, K., Kittler, J., Sadeghi, M., Marcel, S., Marcel, C., Bengio, S., Cardinaux, F., Sanderson, C., Czyz, J., Vandendorpe, L., Srisuk, S., Petrou, M., Kurutach, W., Kadyrov, A., Paredes, R., Kepenekci, B., Tek, F. B., Akar, G. B., Deravi, F., & Mavity, N. (2003). Face verification competition on the XM2VTS. *Database*, *2688*, 964–974. https://doi.org/10.1007/3-540-44887-X_112

Naji, P., Alsufyani, N. A., & Lagravère, M. O. (2014). Reliability of anatomic structures as landmarks in three-dimensional cephalometric analysis using CBCT. *The Angle Orthodontist*, *84*, 762–772. https://doi.org/10.2319/090413-652.1

Novoa, R. A., Gevaert, O., & Ko, J. M. (2019). Marking the path toward artificial intelligence-based image classification in dermatology. *JAMA Dermatology*, *155*, 1105–1106. https://doi.org/10.1001/jamadermatol.2019.1633

Nyemb, P. M. M., Sankale, A.-A., Ndiaye, L., Ndiaye, A., & Gaye, M. (2014). Etude morphométrique de l'oreille externe chez les jeunes adultes. *The Pan African Medical Journal*, *19*, 355. https://doi.org/10.11604/pamj.2014.19.355.4632

Padilla, R., Netto, S. L., da Silva, E. A. B., & A Survey on Performance Metrics for Object-Detection Algorithms. (2020). International conference on systems. *Signals and Image Processing (IWSSIP)*, *2020*, 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

Pawara, P., Okafor, E., Schomaker, L., & Wiering, M. (2017). Data augmentation for plant classification. https://doi.org/10.1007/978-3-319-70353-4_52

Pinheiro, J. C., Bates, D. J., DebRoy, S., & Sakar, D. (2012). The Nlme package: Linear and nonlinear mixed effects models. *R package version 3.1-161*, *6*. https://CRAN.R-project.org/package=nlme

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *The New England Journal of Medicine*, *380*, 1347–1358. https://doi.org/10.1056/NEJMra1814259

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rohlf, F. J., & Slice, D. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology*, *39*, 40–59. https://doi.org/10.2307/2992207

Thevenot, J., Lopez, M. B., & Hadid, A. (2018). A survey on computer vision for assistive medical diagnosis from faces. *IEEE Journal of Biomedical and Health Informatics*, *22*, 1497–1511. https://doi.org/10.1109/JBHI.2017.2754861

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, p. I-511-I–518). IEEE Comput. Soc.

Wickham, H. (2016). ggplot2.

Wilms, M., Handels, H., & Ehrhardt, J. (2017). Representative patch-based active appearance models generated from small training populations. In M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, & S. Duchesne (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2017* (pp. 152–160). Springer International Publishing. https://doi.org/10.1007/978-3-319-66182-7_18

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, *2013*, 532–539. https://doi.org/10.1109/CVPR.2013.75

Yi, J., Mao, X., Chen, L., Xue, Y., & Compare, A. (2014). Facial expression recognition considering individual differences in facial structure and texture. *Computer Vision, IET*, *8*, 429–440. https://doi.org/10.1049/iet-cvi.2013.0171

Zhang, C., & Zhang, Z. (2010). A survey of recent advances in face detection.

Zhou, Y., & Zaferiou, S. (2017). Deformable models of ears in-the-wild for alignment and recognition. In *2017 12th IEEE international conference on automatic face gesture recognition (FG 2017)* (pp. 626–633). IEEE. https://doi.org/10.1109/FG.2017.79

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hennocq, Q., Bongibault, T., Bizière, M., Delassus, O., Douillet, M., Cormier-Daire, V., Amiel, J., Lyonnet, S., Marlin, S., Rio, M., Picard, A., Khonsari, R. H., & Garcelon, N. (2023). An automatic facial landmarking for children with rare diseases. *American Journal of Medical Genetics Part A*, 1–12. https://doi.org/10.1002/ajmg.a.63126

5 Results for the first main objective: diagnosis support

Here we report the results corresponding to the first objective, i.e., diagnostic support. We will first describe a diagnostic task based on the external ear for a group of 4 syndromes validated on children and fetuses. We will then present the results of our classification into 10 groups (9 syndromes and controls).

5.1 Diagnosis in Mandibulofacial Dysostosis with Microcephaly (MFDM) using external ear shapes

Mandibulo-Facial Dysostosis with Microcephaly (MFDM) is a rare disease with a broad spectrum of symptoms, characterized by zygomatic and mandibular hypoplasia, microcephaly, and ear abnormalities. Here, we aimed at describing the external ear phenotype of MFDM patients and train a model to differentiate MFDM ears from non-syndromic control ears (binary classification) and from ears of the main differential diagnoses of this condition (multi-class classification): Treacher Collins (TC), Nager (NAFD), and CHARGE syndromes.

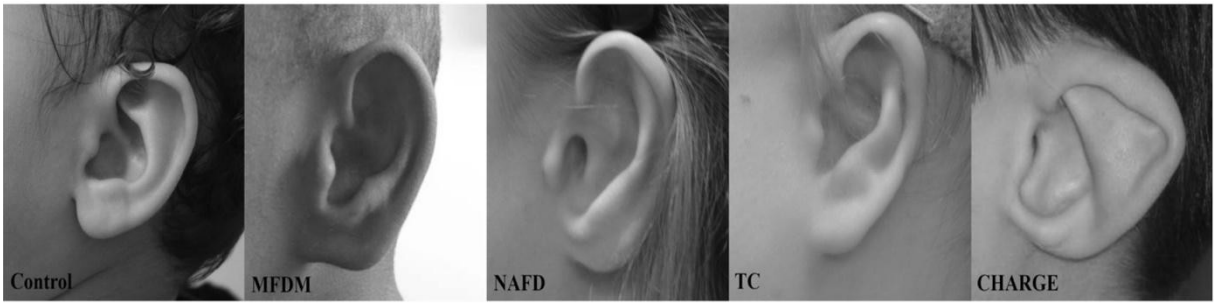


Figure 7. External ear photographs for each patient group: controls, Mandibulo-Facial Dysostosis with Microcephaly (MFDM), Nager type Acro-Facial Dysostosis (NAFD), Treacher Collins (TC), and CHARGE syndromes.

The training set contained 1592 ear photographs, corresponding to 550 patients; 52% of patients were female and the mean age was 7.2 +/- 5.9 years, ranging from 0 to 60.7 years. We included 1296 photographs of control ears, corresponding to 471 patients; 53% of controls were female, with a mean age of 7.2 +/- 5.4 years. The MFDM group included 105 photographs from 31 patients, all genetically confirmed (*EFTUD2* heterozygous pathogenic variations), the NAFD group included 33 pictures from 9 patients, all genetically confirmed (*SF3B4*). We included 70 photographs corresponding to 15 patients in the TC group. All had genetic confirmation (*TCOF1* or *POLR1D*). The CHARGE group included 88 photos from 24 patients. All were genetically confirmed (*CHD7*).

Design №1

The best performances were obtained after 114 iterations. Patients could be classified into MFDM or control groups in the validation set with a balanced accuracy of 0.969 [0.838 – 0.999] ($p < 0.001$) and an AUC of 0.975. Only one patient was misclassified.

Design №2.1

The classification into MFDM, TC, CHARGE and control groups in the validation set was optimized after 76 iterations. On the validation data, the overall balanced accuracy was 0.811 [0.648 – 0.920] ($p = 0.002$). The balanced accuracy was 0.769 for the classification into MFDM, 0.721 for TC, 0.752 for CHARGE, and 0.938 for controls. AUC in the validation set was 0.837 for MFDM, 1.000 for controls, 0.857 for CHARGE, and 0.500 for TC.

Design №2.2

The classification into MFDM, TC and CHARGE groups in the validation set was optimized after 91 iterations. On the validation data, the overall balanced accuracy was 0.813 [0.544 –

0.960] (p = 0.003). With this classifier, the balanced accuracy was 0.944 for the classification into MFDM, 0.873 for CHARGE, and 0.500 for TC. AUC in the validation set was 1.000 for MFDM, 0.969 for CHARGE, and 0.500 for TC.

We also illustrated the potential for clinical use of automatic ear-based diagnosis on a preliminary case study. A non-premature female child aged 9 days was admitted in fetal pathology with bilateral choanal atresia, inner ear malformations, agenesis of the acoustic-facial bundle and cerebello-pontine hypoplasia. She had died within a few days after birth. CHARGE syndrome was confirmed post-mortem by a heterozygous *de novo* pathogenic variation in the *CHD7* gene (c. 4353+1G>A). The patient also carried a heterozygous *de novo* variation of unknown significance in the *EFTUD2* gene (c. 1954G>A, p.Asp652Asn). Our ear-based model on the ears of this patient (with a XGBoost classifier) proposed: CHARGE syndrome 84%, control patient 11%, MFDM 3%, NAFD 2% or TC 1% (Figure 8), supporting the diagnosis of CHARGE syndrome, and showing little tendency towards an MFDM ear. As systematic *EFTUD2* heterozygous pathogenic variation screening is currently recommended in unusual CHARGE cases (116): our model, with further clinical validation, could be used as a clinical support for directing genetic investigations (117).

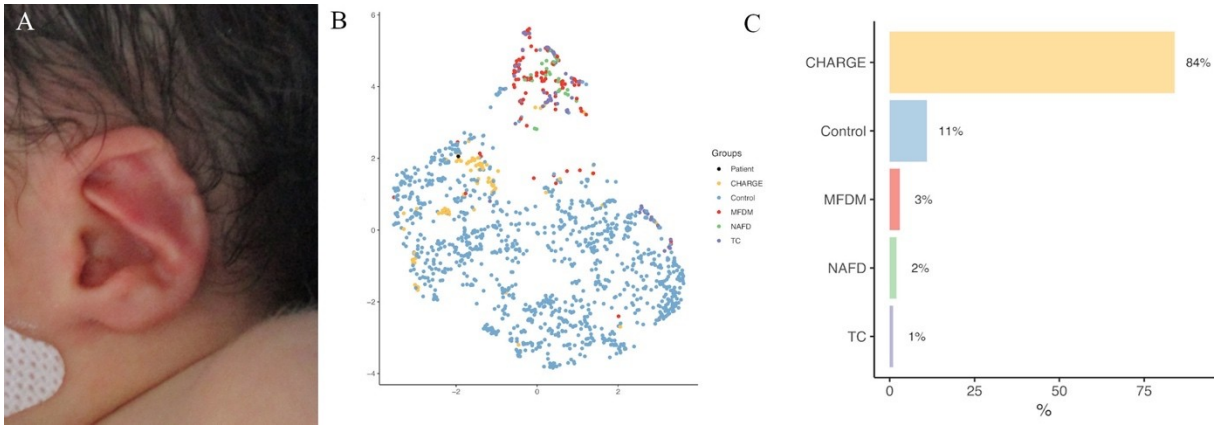


Figure 8. Case study for automatic ear-based diagnosis in CHARGE syndrome (A). (B) UMAP clustering of design №2.1; black dot: patient. (C) probability histogram with a XGBoost classifier.

frontiers | Frontiers in Pediatrics

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fped.2023.1171277

OPEN ACCESS

EDITED BY
Stephen Aronoff,
Temple University, United States

REVIEWED BY
Livia Garavelli,
AUSL IRCCS Reggio Emilia, Italy
Ewelina Bukowska-Olech,
Poznan University of Medical Sciences, Poland

*CORRESPONDENCE
Quentin Hennocq
quentin.hennocq@aphp.fr

RECEIVED 27 February 2023
ACCEPTED 26 July 2023
PUBLISHED 17 August 2023

CITATION
Hennocq Q, Bongibault T, Marlin S, Amiel J, Attie-Bitach T, Baujat G, Boutaud L, Carpentier G, Corre P, Denoyelle F, Djate Delbrah F, Douillet M, Galliani E, Kamolvisit W, Lyonnet S, Milea D, Pingault V, Porntaveetus T, Touzet-Roumazeille S, Willems M, Picard A, Rio M, Garcelon N and Khonsari RH (2023) AI-based diagnosis in mandibulofacial dysostosis with microcephaly using external ear shapes. Front. Pediatr. 11:1171277. doi: 10.3389/fped.2023.1171277

AI-based diagnosis in mandibulofacial dysostosis with microcephaly using external ear shapes

Quentin Hennocq[1,2,3]*, Thomas Bongibault[1,3], Sandrine Marlin[1,4], Jeanne Amiel[1,4], Tania Attie-Bitach[1,4], Geneviève Baujat[1,4], Lucile Boutaud[4], Georges Carpentier[5], Pierre Corre[6,7], Françoise Denoyelle[8], François Djate Delbrah[1], Maxime Douillet[1], Eva Galliani[2], Wuttichart Kamolvisit[9,10], Stanislas Lyonnet[1,4], Dan Milea[11], Véronique Pingault[1,4], Thantrira Porntaveetus[9,10], Sandrine Touzet-Roumazeille[5], Marjolaine Willems[12], Arnaud Picard[2], Marlène Rio[1,4], Nicolas Garcelon[1] and Roman H. Khonsari[1,2,3]

[1]Imagine Institute, INSERM UMR1163, Paris, France, [2]Service de Chirurgie Maxillo-Faciale et Chirurgie Plastique, Hôpital Necker—Enfants Malades, Assistance Publique—Hôpitaux de Paris, Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou, Faculté de Médecine, Université de Paris Cité, Paris, France, [3]Laboratoire 'Forme et Croissance du Crâne', Faculté de Médecine, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris, Université Paris Cité, Paris, France, [4]Service de Médecine Génomique des Maladies Rares, Hôpital Necker—Enfants Malades, Assistance Publique—Hôpitaux de Paris, Faculté de Médecine, Université de Paris Cité, Paris, France, [5]CHU Lille, Inserm, Service de Chirurgie Maxillo-Faciale et Stomatologie, U1008-Controlled Drug Delivery Systems and Biomaterial, Université de Lille, Lille, France, [6]Department of Oral and Maxillofacial Surgery, INSERM U1229—Regenerative Medicine and Skeleton RMeS, Nantes, France, [7]Department of Oral and Maxillofacial Surgery, Nantes University, CHU Nantes, Nantes, France, [8]Department of Paediatric Otolaryngology, AP-HP, Hôpital Necker-Enfants Malades, Paris, France, [9]Center of Excellence for Medical Genomics, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand, [10]Center of Excellence in Genomics and Precision Dentistry, Department of Physiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand, [11]Duke-NUS Medical School Singapore, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore, [12]Département de Génétique Clinique, CHRU de Montpellier, Hôpital Arnaud de Villeneuve, Institute for Neurosciences of Montpellier, INSERM, Univ Montpellier, Montpellier, France

Introduction: Mandibulo-Facial Dysostosis with Microcephaly (MFDM) is a rare disease with a broad spectrum of symptoms, characterized by zygomatic and mandibular hypoplasia, microcephaly, and ear abnormalities. Here, we aimed at describing the external ear phenotype of MFDM patients, and train an Artificial Intelligence (AI)-based model to differentiate MFDM ears from non-syndromic control ears (binary classification), and from ears of the main differential diagnoses of this condition (multi-class classification): Treacher Collins (TC), Nager (NAFD) and CHARGE syndromes.

Methods: The training set contained 1,592 ear photographs, corresponding to 550 patients. We extracted 48 patients completely independent of the training set, with only one photograph per ear per patient. After a CNN-(Convolutional Neural Network) based ear detection, the images were automatically landmarked. Generalized Procrustes Analysis was then performed, along with a dimension reduction using PCA (Principal Component Analysis). The principal components were used as inputs in an eXtreme Gradient Boosting (XGBoost) model, optimized using a 5-fold cross-validation. Finally, the model was tested on an independent validation set.

Results: We trained the model on 1,592 ear photographs, corresponding to 1,296 control ears, 105 MFDM, 33 NAFD, 70 TC and 88 CHARGE syndrome ears. The

model detected MFDM with an accuracy of 0.969 [0.838–0.999] ($p < 0.001$) and an AUC (Area Under the Curve) of 0.975 within controls (binary classification). Balanced accuracies were 0.811 [0.648–0.920] ($p = 0.002$) in a first multiclass design (MFDM vs. controls and differential diagnoses) and 0.813 [0.544–0.960] ($p = 0.003$) in a second multiclass design (MFDM vs. differential diagnoses).

Conclusion: This is the first AI-based syndrome detection model in dysmorphology based on the external ear, opening promising clinical applications both for local care and referral, and for expert centers.

KEYWORDS

AI, machine learning, dysmorphology, craniofacial malformation, MFDM

1. Introduction

Mandibulo-Facial Dysostosis with Microcephaly (MFDM), formerly named Mandibulo-Facial Dysostosis Guion Almeida type (MFDGA) (1, 2), is a rare disease with a broad spectrum of symptoms, characterized by zygomatic (92%) and mandibular (93%) hypoplasia, microcephaly (88%, 64% congenital or 36% postnatal), cognitive impairment (97%–100%), small or dysplastic external ear (97%) and deafness (83%), most often conductive (3). MFDM may also include choanal atresia (30%–33%), cleft palate (43%–47%), facial asymmetry (53%–58%), and extra-facial abnormalities, such as heart malformations (30%–35%), thumb abnormalities (31%), esophageal involvement (atresia/fistulae, 27%–33%), short stature (30%), vertebral abnormalities (28%) and epilepsy (27%) (4). Facial dysostoses are subdivided into two groups: Mandibulo-Facial Dysostoses (MFD) and Acro-Facial Dysostoses (AFD), the latter including limb abnormalities (5). Because there may be associated with spine abnormalities, some authors have listed MFDM as a pre-axial acrofacial dysostosis, Guion Almeida type (AFDGA) (5–7).

Since 2012, the diagnosis of MFDM is established based on clinical features and the screening for a heterozygous pathogenic variant of the *EFTUD2* gene (17q21.31) coding for the nuclear ribonucleoprotein component of 116 KDA U5 protein (8). This variant occurs frequently *de novo* (80%) (4, 9). The main mechanism of disease is haploinsufficiency (10), caused in 18% of cases by a missense substitution, in 38% by a stop-gain *EFTUD2* heterozygous pathogenic variation and in 43% by a splice site variation (4, 11). No genotype-phenotype correlations in patients with *EFTUD2* heterozygous pathogenic variations have been identified (8, 12).

Regarding deformities of the external ear in MFDM, Lines et al. (3, 8) described microtia (grades I-III), abnormalities of the superior helix and antihelix, preauricular tags and auditory canal atresia/stenosis. The posterior-inferior margin of the lobule can have a right-angle ("squared-off") configuration (3, 8, 13).

The main differential diagnoses of MFDM are other mandibulofacial dysostoses — i.e., Nager type Acro-Facial Dysostosis (NAFD), Postaxial acrofacial dysostosis Miller type, and Treacher Collins (TC) syndromes — and CHARGE syndrome (14, 15). MFDM patients are often misdiagnosed within this spectrum. Distinguishing MFDM ears from CHARGE ears can sometimes be tricky, and *EFTUD2* heterozygous pathogenic variation screening is recommended in patients with unusual forms of CHARGE syndrome (14).

Based on these clinical questions, the three objectives of this study were: (1) objectively determine the phenotype of pinna malformations in MFDM using geometric morphometrics and machine learning techniques vs. controls (design № 1), (2) compare the ears of MFDM patients with ears from the main differential diagnoses, with or without controls (respectively design № 2.1 and № 2.2) and (3) compare phenotypes from the different genotypes causing MFDM (design № 3).

2. Material and methods

2.1. Training set

We included pictures from the photographic database of the Maxillofacial surgery and Plastic Surgery department and from the Medical genetics department of *Hôpital Necker—Enfants Malades* (Assistance Publique—Hôpitaux de Paris), Paris, France. This database contains 594,000 photographs from 22,000 patients followed in the department since 1981. All photographs were taken by a professional medical photographer using a Nikon D7000 device in standardized positions.

We included retrospectively and prospectively, from 1981 to 2023, all profile pictures of patients diagnosed with MFDM, TC, NAFD and CHARGE syndromes, with a visible pinna (**Figure 1**). The photographs were not calibrated. All patients had genetic confirmation of their syndrome. We excluded patients with ear reconstruction surgery. Multiple photographs per patient corresponded to different ages. Duplicate photographs were excluded.

Non-syndromic children were selected among patients admitted for wounds, trauma, infection and various skin lesions, without any record of chronic conditions. More precisely, follow-up for any type of chronic disease was considered as an exclusion criterion. The reports were retrieved using Dr Warehouse (16). For each patient, right and left sides were included.

The study was approved by the CESREES (Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé, № 4570023bis) and by the CNIL (Commission Nationale Informatique et Libertés, № MLD/MFI/

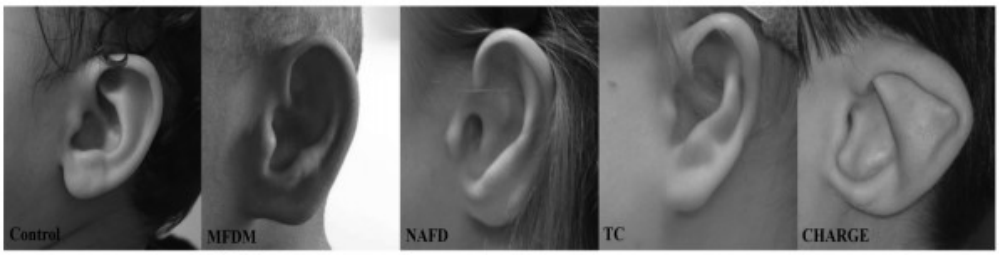


FIGURE 1
Examples of external ear photographs for each patient group: controls, mandibulo-facial dysostosis with microcephaly (MFDM), Nager type acro-facial dysostosis (NAFD), Treacher Collins (TC), and CHARGE syndromes.

AR221900). Informed and written consents were obtained from the legal representatives of each child, or from the patient himself if he was of age.

2.2. Validation set

A fully independent validation set was designed using publicly available data published in the literature. We included patients with MFDM (6, 14, 17), NAFD (18–20), CHARGE syndrome (21–24) and TC syndrome (25, 26); all had genetic confirmation of their syndromes.

We also retrieved ear photographs of these syndromes of interest from the databases of the Maxillofacial surgery and/or Genetics departments of the University Hospitals of Lille (France), Montpellier (France), Nantes (France) and the King Chulalongkorn Memorial Hospital in Bangkok (Thailand). None of the patients in the validation set were present twice, and none were from the training set. For the control group, we selected a group of photographs from our local database, without any redundancy with the training set, using similar inclusion criteria.

We extracted data on age at the time of the photograph and gender. We excluded patients with no information on the contralateral ear to take into account asymmetry or severity.

All photographs in the validation group were manually annotated by two independent raters (QH and MD), blinded for the diagnosis. The ICC (Intraclass Correlation Coefficient) was computed. ICC values greater than 0.9 corresponded to excellent reliability of the manual annotation (27).

2.3. Landmarking

We used an available template (28) based on 55 landmarks placed on the outer helix, the antihelix, the lobe, the tragus, the antitragus, the helix, the crus helicis, and the concha. We developed an automatic annotation model trained on 1,592 manually annotated ear photographs following a pipeline including: (1) a Faster R-CNN (Convolution Neural Network) to detect ears on the pixels of lateral face photographs and (2) a patch-AAM (Active Appearance Model), to automatically place landmarks.

The Fast RCNN model (29) was trained on 5,154 ear photographs after data augmentation (1,718 images and their +10° and −10° rotations), with a learning rate of 0.001, a batch size of 4, a gamma of 0.05 and 2,000 iterations. The patch-AAM was trained on 1,221 ear photos, after 50 iterations, with a Lucas-Kanade optimization (30). The Faster R-CNN was developed in Pytorch on Python 3.7 (31). The patch-AAM was developed using the *menpo* library on Python 3.7 (32). These two methods and the choice of hyperparameters have been described in a previous report by our team (33).

Each automatically annotated photograph was checked by the first author (QH) and landmarks were manually re-positioned when necessary, using *landmarker.io* (34).

To ensure a uniform distribution of landmarks along the curves of the ear (outer helix, inner helix, antihelix, concha), anatomical landmarks were transformed into sliding semi-landmarks using the *geomorph* package on R (35). Landmarks corresponding to the antihelix were removed because Hennocq et al. (33) showed that they were not reproducible between two annotators.

Ears were finally annotated based on 41 anatomical landmarks and semi-landmarks, placed automatically and double-checked manually.

2.4. Geometric morphometrics

We performed Generalized Procrustes Analysis (GPA) (36) on all landmark clouds using the *geomorph* package on R. Since the data were uncalibrated photographs, ear sizes were not available: shape parameters only were assessed and not centroid sizes.

Procrustes coordinates were processed using Principal Component Analysis (PCA) for dimension reduction (37): 8 principal components (PC) accounting for more than 90% of the global variance were retained.

To take into account associated metadata (age and gender) and the fact that we had included more than one photograph per patient (that is the non-independence of the data), a mixed model was designed for each principal component. The variable

to be explained was PC, with age and gender considered as explanatory variables. A random effect on age and individuals was introduced. The equation of the mixed model was:

$$PC_{i,j} \sim \alpha + age.\beta_1 + gender.\beta_2 + age.\beta_{1,i} + \varepsilon_{i,j}$$

where $age.\beta_{1,i}$ corresponded to a random slope for age per individual, and $\varepsilon_{i,j}$ was a random error term. We did not use an interaction term between age and gender as it did not increase the likelihood of the model. Age, gender and ethnicity are significant factors in dysmorphology because they influence the diagnosis, and must therefore be taken into account (38).

2.5. Asymmetry and severity of microtia

Accounting for the heterogeneity of external ear anomalies was difficult. We graded microtia in stages I-IV according to the Marx classification (39). Only grade I ears could be annotated, as the main anatomical structures were missing in grades II, III et IV. However, the frequency of ears >grade I had to be considered for each disease group as it was a potential diagnostic feature. Information on the left/right asymmetry was also included as it could have been variable according to syndromes.

The overall severity for each patient was defined as the sum of microtia grades on each ear. Asymmetry was quantified using a mixed scale ranging from 0 to 3, corresponding to the subtraction of the left and right microtia grades. A high score corresponded to high left/right asymmetry. For bilateral grade I ears, we computed an asymmetry index based on fluctuating asymmetry (40, 41), normalized between 0 and 1. A patient with two grade II ears had a symmetry score of 0. A patient with one grade III ear and one grade I ear had a symmetry score of 2. A patient with two grade I ears had an asymmetry score corresponding to his normalized asymmetry index, ranging between 0 and 1.

The severity and asymmetry scores were compared between different groups using mixed linear models to take into account repeated data per patient. The model coefficients for each group were compared to 0 by Student's t tests. The significance level was set at $p < 0.05$.

2.6. Uniform manifold approximation and projection (UMAP) representations

The residuals $\varepsilon_{i,j}$ were represented using UMAP (42), a nonlinear dimension reduction technique for data visualization. Each design was plotted with and without the severity and asymmetry scores. A k (local neighborhood size) value of 15 was used. A cosine metric was introduced to compute distances in high dimensional spaces: the effective minimal distance between embedded points was 10^{-6}. The three conditions of UMAP, namely uniform distribution, local constancy of the Riemannian metric and local connectivity were verified. UMAP analyses were performed using the package *umap* on R (43).

2.7. Machine learning models and metrics

The landmark clouds were superimposed with the previous generalized Procrustes analysis and PCA. With the metadata (age and gender), the residuals $\varepsilon_{i,j}$ were reported for each PC and each ear of the validation group. The inputs to the model were the residuals from the linear models described above.

We used XGBoost (eXtreme Gradient Boosting), a supervised machine learning classifier, for all the analyses (44). We set a number of hyperparameters to improve the performance and effect of the machine learning model: learning rate = 0.3, gamma = 0, maximum tree depth = 6. We separated the dataset into a training set and a testing set, and a 5-fold cross-validation was used to define the ideal number of iterations to avoid overfitting. The model with the lowest logloss-score was chosen for analysis. The chosen model was then used on the independent validation set to test performances, by plotting accuracy, sensitivity, specificity, F1-score, precision and recall, AUC (in a one vs. all design). The ROC (Receiver Operating Characteristics) curves were plotted in R using the *plotROC* package (45).

3. Results

3.1. Training set

The training set contained 1,592 ear photographs, corresponding to 550 patients; 52% of patients were female and the mean age was 7.2 ± 5.9 years, ranging from 0 to 60.7 years.

We included 1,296 photographs of control ears, corresponding to 471 patients; 53% of controls were female, with a mean age of 7.2 ± 5.4 years.

The MFDM group included 105 photographs from 31 patients, all genetically confirmed (*EFTUD2* heterozygous pathogenic variations); 52% were female and the mean age was 9.2 ± 9.8 years. Regarding ear aplasia, 92% of the ears were normal or grade I, 3% were grade 22, 5% were grade III, and 0% was grade IV.

The NAFD group included 33 pictures from 9 patients, all genetically confirmed (SF3B4), with 56% females, and a mean age of 11.8 ± 8.8 years. All ears were normal or grade I.

We included 70 photographs corresponding to 15 patients in the TC group. The mean age was 5.5 ± 4.2 years and 40% were female. All had genetic confirmation (*TCOF1* or *POLR1D*). Eighty percent of the ears were normal or grade I, 17% grade II, 3% grade III, and 0% grade IV.

The CHARGE group included 88 photos from 24 patients; 42% were female and mean age was 5.1 ± 5.9 years. All were genetically confirmed (*CHD7*). All ears were normal or grade I (**Table 1**).

In the MFDM group, 11 out of 31 patients (35%) had a heterozygous pathogenic variation in a splice site of *EFTUD2*. One of these patients had a Lys620Asn variant (1860G > C) which could be considered as a splice site variation and not as missense (35). Nine out of 31 patients (29%) had a frameshift *EFTUD2* heterozygous pathogenic variation, 7/31 (23%) a nonsense variation, and 4/31 (13%) an intragenic deletion. No patient had a missense variation (**Supplementary Table S1**).

TABLE 1 Description of the training set population.

	Total	Controls	MFDM	NAFD	TC	CHARGE
N (ears)	1,592	1,296	105	33	70	88
N (patients)	550	471	31	9	15	24
Gender						
Female	288/550 (52%)	251/471 (53%)	16/31 (52%)	5/9 (56%)	6/15 (40%)	10/24 (42%)
Age						
Mean ± SD	7.2 ± 5.9	7.2 ± 5.4	9.2 ± 9.8	11.8 ± 8.8	5.5 ± 4.2	5.1 ± 5.9
Median	6.9	7.3	5.0	9.8	5.3	6.5
Min	0.0	0.1	0.0	0.0	0.0	0.0
Max	60.7	60.7	39.6	33.5	17.2	21.6
Identified pathogenic genetic variation		NA	31/31 (100%)	9/9 (100%)	15/15 (100%)	24/24 (100%)
Grade of aplasia						
0–1		1,296/1,296 (100%)	97/105 (92%)	33/33 (100%)	56/70 (80%)	88/88 (100%)
2		0/1,296 (0%)	3/105 (3%)	0/33 (0%)	12/70 (17%)	0/88 (0%)
3		0/1,296 (0%)	5/105 (5%)	0/33 (0%)	2/70 (3%)	0/88 (0%)
4		0/1,296 (0%)	0/105 (0%)	0/33 (0%)	0/70 (0%)	0/88 (0%)

MFDM, mandibulo-facial dysostosis with microcephaly; NAFD, nager type acro-facial dysostosis; TC, treacher collins; CHARGE, coloboma, heart defect, atresia choanae, retarded growth and development, genital hypoplasia, ear anomalies/deafness; SD, standard deviation.

Average models per group were designed after Procrustes transformation, and compared (Figures 2, 3). Ears in the MFDM group had a clockwise rotation and a vertical shift of the concha (Figure 2) when compared to controls. Previously described features—thickened helix, enlarged and square lobe—were also reported.

3.2. Validation set

We extracted a total of 48 patients completely independent of the training set, with only one photograph per ear per patient. Severity and asymmetry scores were computed and only one side was then randomly selected. The validation set included 11

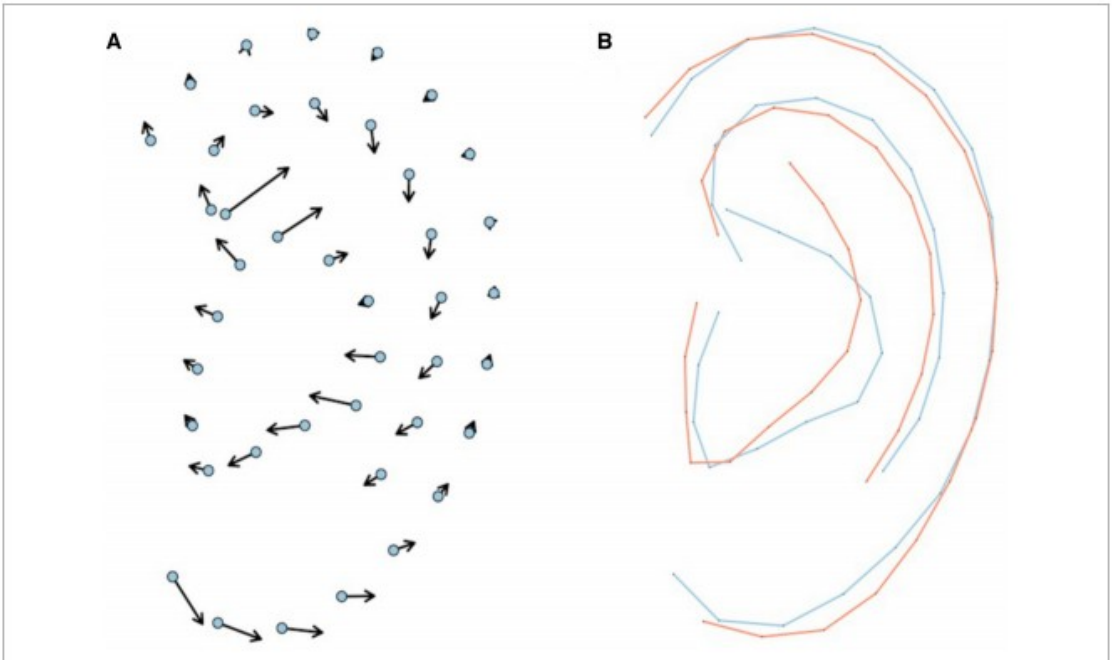


FIGURE 2
Vectors represent distances between MFDM mean landmarks and the control mean landmarks (A). Comparison of average MFDM (red) and control (blue) ear models after Procrustes transformation (B).

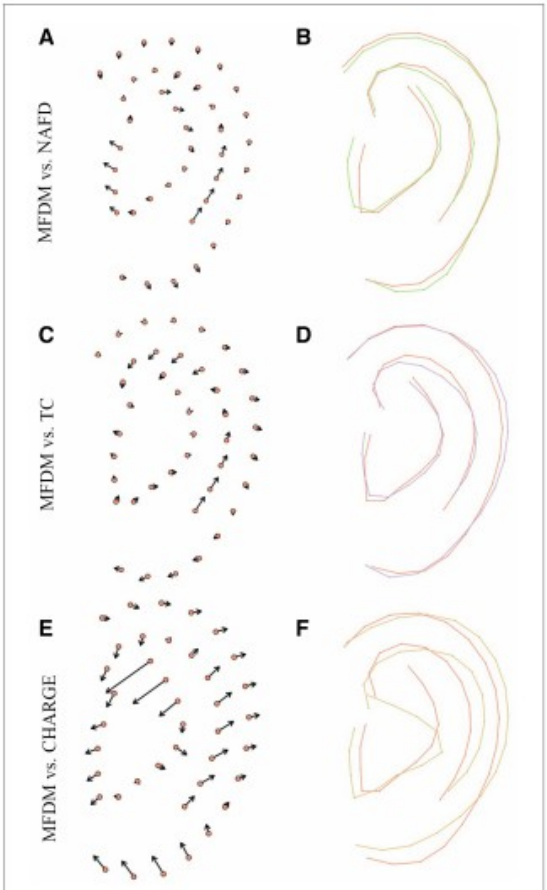


FIGURE 3
Comparison of average MFDM (red) and the main differential diagnoses: NAFD (green) (A, B), TC (purple) (C, D) and CHARGE (yellow) (E, F), after Procrustes transformation. Vectors (A, C, E) represent distances between MFDM mean landmarks and other groups mean landmarks.

MFDM patients (23%), 2 NAFD (4%), 6 TC (13%), 8 CHARGE (17%) and 21 controls (44%) (Supplementary Table S3). We did not have access to the other ear for NAFD patients in the validation set and therefore the asymmetry and severity scores were not obtained.

ICC was 0.991 between the two annotators and the reliability of the annotation was therefore considered as excellent (27).

3.3. Severity and asymmetry

Severity and asymmetry scores were compared between groups. In design № 1, TC ears were statistically more severely affected ($p < 0.001$). CHARGE and control groups had lower severity grades ($p = 0.027$ and $p < 0.001$, respectively), compared to MFDM. Control ears were less asymmetric ($p < 0.001$) than MFDM ears. CHARGE ears were less asymmetric than MFDM ears in design № 2.2 (Supplementary Table S2).

3.4. UMAP representations

Patients were clustered using UMAP (Figure 4). MFDM patients were distinct from controls (design № 1, Figure 4A), and CHARGE patients, but not from NAFD and TC patients (designs № 2.1 and № 2.2, Figures 4B,C).

3.5. Machine learning models and metrics

3.5.1. Design № 1

The best performances were obtained without integrating the asymmetry and severity parameters, after 114 iterations. The AUC was 0.985 in the training set (Figure 5A). Patients could be classified into MFDM or control groups in the validation set with a balanced accuracy of 0.969 [0.838–0.999] ($p < 0.001$) and

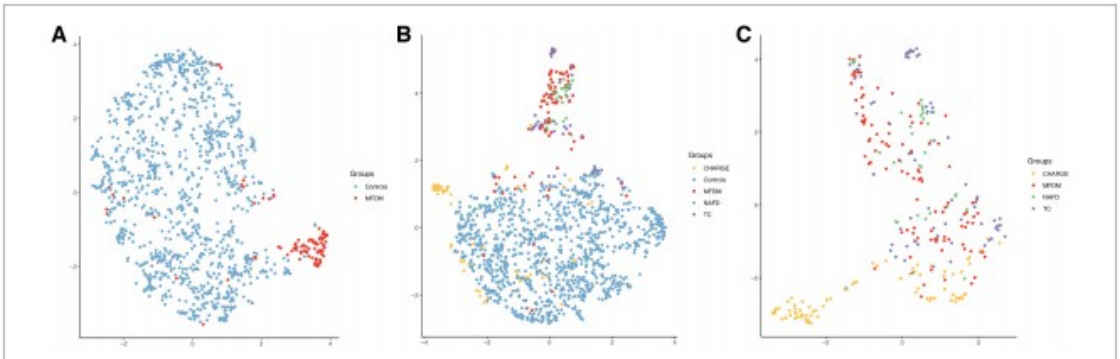


FIGURE 4
UMAP representations for designs №1 (A), № 2.1 (B) and № 2.2 (C), including severity and asymmetry parameters. Each color corresponds to a patient group. MFDM, Mandibulo-Facial Dysostosis with Microcephaly; NAFD, Nager type Acro-Facial Dysostosis; TC, Treacher Collins; CHARGE, Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness.

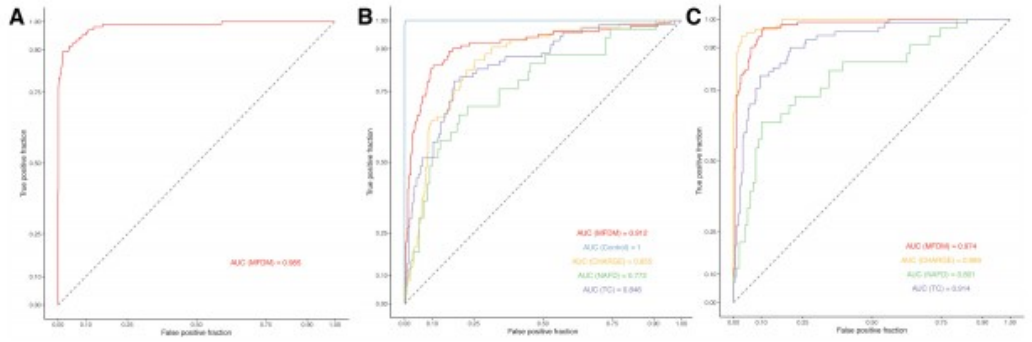


FIGURE 5
Empirical ROC curves for designs № 1 (A), № 2.1 (B) and № 2.2 (C). MFDM, Mandibulo-Facial Dysostosis with Microcephaly; NAFD, Nager type Acro-Facial Dysostosis; TC, Treacher Collins; CHARGE, Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness.

an AUC of 0.975 (Table 2). Only one patient was misclassified (Table 3).

3.5.2. Design № 2.1

The best performances were obtained by integrating the asymmetry and severity parameters. The classification into MFDM, TC, CHARGE and control groups in the validation set was optimized after 76 iterations. The AUC was 0.912 for MFDM, 1.000 for controls, 0.855 for CHARGE, 0.772 for NAFD and 0.846 for TC in the training set (Figure 5B). On the validation data, the overall balanced accuracy was 0.811 [0.648–0.920] ($p = 0.002$). The balanced accuracy was 0.769 for the classification into MFDM, 0.721 for TC, 0.752 for CHARGE and 0.938 for controls. AUC in the validation set was 0.837 for MFDM, 1.000 for controls, 0.857 for CHARGE and 0.500 for TC (Tables 4, 5).

TABLE 2 Classification results on the validation set for design № 1.

Design № 1	
Accuracy	0.969 [0.838–0.999] $p < 0.001$*
Sensitivity (Se)	1.000
Specificity (Sp)	0.909
Balanced Accuracy	0.954
AUC	0.975

Se, Sensitivity; Sp, Specificity.
*Statistically significant test result ($p \leq 0.05$).

TABLE 3 Confusion matrix on the validation set for design № 1.

		Reference	
		MFDM	Control
Prediction	MFDM	**10**	0
	Control	1	**21**

MFDM, mandibulo-facial dysostosis with microcephaly.
Bolded values denote True Positives (TP).

TABLE 4 Classification results on the validation set for design № 2.1.

Design № 2.1		
Overall (multiclass design)		
Accuracy		0.811 [0.648–0.920] $p = 0.002$*
Binary (one-vs.-all design)		
Sensitivity (Se)	MFDM	0.571
	Control	1.000
	CHARGE	0.571
	TC	0.500
Specificity (Sp)	MFDM	0.967
	Control	0.875
	CHARGE	0.933
	TC	0.943
Balanced Accuracy	MFDM	0.769
	Control	0.938
	CHARGE	0.752
	TC	0.721
AUC	MFDM	0.837
	Control	1.000
	CHARGE	0.857
	TC	0.500

MFDM, mandibulo-facial dysostosis with microcephaly; TC, treacher collins; CHARGE, coloboma, heart defect, atresia choanae, retarded growth and development, genital hypoplasia, ear anomalies/deafness.
*Statistically significant test result ($p \leq 0.05$).

TABLE 5 Confusion matrix on the validation set for design № 2.1.

		Reference			
		MFDM	Control	CHARGE	TC
Prediction	MFDM	**4**	0	1	0
	Control	0	**21**	2	0
	CHARGE	1	0	**4**	1
	TC	2	0	0	**1**

MFDM, mandibulo-facial dysostosis with microcephaly; TC, treacher collins; CHARGE, coloboma, heart defect, atresia choanae, retarded growth and development, genital hypoplasia, ear anomalies/deafness.
Bolded values denote True Positives (TP).

3.5.3. Design № 2.2

The best performances were obtained by integrating the asymmetry and severity parameters. The classification into MFDM, TC and CHARGE groups in the validation set was optimized after 91 iterations. The AUC was 0.974 for MFDM, 0.889 for CHARGE, 0.801 for NAFD and 0.914 for TC in the training set (**Figure 5C**). On the validation data, the overall balanced accuracy was 0.813 [0.544–0.960] ($p = 0.003$). With this classifier, the balanced accuracy was 0.944 for the classification into MFDM, 0.873 for CHARGE and 0.500 for TC. AUC in the validation set was 1.000 for MFDM, 0.969 for CHARGE and 0.500 for TC (**Tables 6, 7**).

3.5.4. Design № 3

AUC was 0.602 [0.483–0.734] ($p = 0.370$) on the training set. This classification was not statistically significant and was therefore not tested on the validation set. The UMAP representation did not find any clusters based on *EFTUD2* heterozygous pathogenic variation type and site (**Supplementary Figure S1**).

4. Discussion

Applications of machine learning are increasing in healthcare (46–49). The field of dysmorphology has been transformed by the framework for genetic syndrome classification called DeepGestalt (50), produced by the Face2Gene group. Publications comparing human performances to DeepGestalt performances are flourishing (51–54), and some authors state that digital tools provide better results than human experts in terms of diagnosis. We do not believe that Artificial Intelligence (AI) algorithms can fully replace the experience of an expert practitioner, but AI-based tools can considerably increase diagnostic performances, and also contribute to the diffusion of specialized expertise. However, as in all deep learning approaches, DeepGestalt predictions are tricky to explain (50): the phenotypic traits leading to diagnosis cannot be traced. Moreover, only the frontal facial pictures are considered within this framework, that does not take into account the profile pictures and external ears. To our knowledge, we report the first machine learning classifier based on external ear shape. Even though the diagnosis of a given syndrome is never fully based on ear anomalies, this anatomical region is a major source of distinctive phenotypic features in a large array of syndromes (42–44).

TABLE 6 Classification results on the validation set for design № 2.2.

Design № 2.2		
Overall (multiclass design)		
Accuracy		0.813 [0.544–0.960] $p = 0.003$*
Binary (one-vs.-all design)		
Sensitivity (Se)	MFDM	1.000
	CHARGE	0.857
	TC	0.000
Specificity (Sp)	MFDM	0.889
	CHARGE	0.889
	TC	0.929
Balanced accuracy	MFDM	0.944
	CHARGE	0.873
	TC	0.464
AUC	MFDM	1.000
	CHARGE	0.969
	TC	0.500

MFDM, Mandibulo-Facial Dysostosis with Microcephaly; TC, Treacher Collins; CHARGE, Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness.
*Statistically significant test result ($p \leq 0.05$).

TABLE 7 Confusion matrix on the validation set for design № 2.2.

		Reference		
		MFDM	CHARGE	TC
Prediction	MFDM	**7**	0	1
	CHARGE	0	**6**	1
	TC	0	1	**0**

MFDM, mandibulo-facial dysostosis with microcephaly; TC, treacher collins; CHARGE, coloboma, heart defect, atresia choanae, retarded growth and development, genital hypoplasia, ear anomalies/deafness.
Bolded values denote True Positives (TP).

Ear phenotype in MFDM has been previously reported. Guion-Almeida et al. described 4 Brazilian children with small ears, a large lobe, and preauricular skin tags in years 2000 (55) and 2006 (1). In 2009 (2), the same team described small and cup-shaped ears with atretic external auditory canal in two other cases. Smigiel et al. (56) reported three MFDM cases with asymmetric microtia, a thickened helix, and protruding ear lobes. Lehalle et al. (17) described abnormalities of the external ear in 100% out of 34 MFDM cases, with minor abnormalities in 29/34 cases (squared, flattened and externally deviated ear lobe), asymmetric ears in 24% of cases and preauricular tags in 33% of cases. Voigt et al. (6), Huang et al. (4), Lines et al. (8) et Yu et al. (57) described similar abnormal pinnae. We could not find any information in the literature on the frequency of grade >I ear involvement in MFDM, or on the asymmetry of microtia.

In TC, Katsanis & Jabs (58) reported absent or small, malformed, sometimes rotated ears. Abdollahi Fakhim et al. (59) compared NAFD and TC without mentioning ears. Bernier et al. (18) described pinnae malformations in NAFD without providing further details. We did not find detailed phenotypic descriptions of the external ear in TC and NAFD in the literature.

In contrast, Davenport et al. (60) described the ear phenotype of CHARGE ears in greater details. CHARGE ears were small, wide and 'looked as if they were stretched or bent' (60). The most distinctive feature according to these authors was the triangular shape of the concha and a discontinuity between the antihelix and the antitragus. Davenport et al. (60) also explained that many patients had small or absent lobes, with significant left/right asymmetry.

We thus report new features for MFDM ears: clockwise rotation and vertical shift of the concha (**Figure 2**). We confirm previously described features such as helix thickening, and enlarged and squared lobes. MDFM ears were also more asymmetric than controls. These overall features were shared with the NAFD and TC groups. Microtia grades were nevertheless higher in TC.

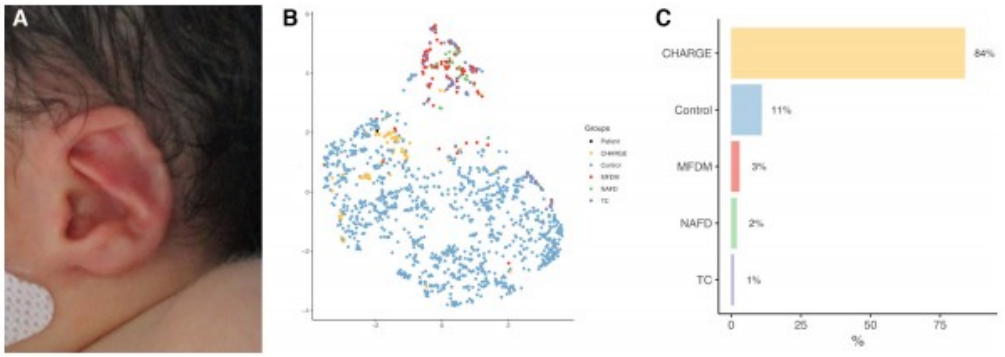


FIGURE 6
Case study of automatic ear-based CHARGE syndrome diagnosis (A). (B) UMAP clustering of design № 2.1; black dot: patient. (C) probability histogram with a XGboost classifier.

CHARGE ears had a specific shape, with a triangular concha, a smaller but wider overall size with a thinner helix and a smaller lobe. In brief, the shape of the pinna can be considered as a relevant feature to differentiate MFDM from CHARGE.

The classification algorithm from design № 1 provides an accuracy of 96.9% for distinguishing MDFM from controls, with only 1 patient misclassified in the validation set. with poorer results when using multi-class classification, which provides an overall balanced accuracy of 81.1% in design № 2.1 (MFDM and its differential diagnoses + controls) and 81.3% in design № 2.2 (MFDM and its differential diagnoses). These results account for the difficulty to diagnose MFDM from NAFD and TC. On the other hand, our results were satisfactory for detecting CHARGE ears, with an AUC reaching 85.7% in design № 2.1, and 96.9% in design № 2.2. We could not detect any genotype-phenotype correlations (design № 3).

The clinical use of automatic ear-based diagnosis can be highlighted based on a preliminary case study. A non-premature female child aged 9 days was admitted in fetal pathology with bilateral choanal atresia, inner ear malformations, agenesis of the acoustic-facial bundle and cerebellopontine hypoplasia. She had died within a few days after birth. CHARGE syndrome was confirmed post-mortem by a heterozygous *de novo* pathogenic variation in the *CHD7* gene (c. 4,353 + 1G > A). The patient also carried a heterozygous *de novo* variation of unknown significance in the *EFTUD2* gene (c. 1954G > A, p.Asp652Asn). Our ear-based model on the ears of this patient (with a XGBoost classifier) proposed: CHARGE syndrome 84%, control patient 11%, MFDM 3%, NAFD 2% or TC 1% (Figure 6), supporting the diagnosis of CHARGE syndrome, and showing little tendency towards MFDM ear. As systematic *EFTUD2* heterozygous pathogenic variation screening being currently recommended in unusual CHARGE cases [9], our model, with further clinical validation, could be used as a clinical support for directing genetic investigations.

Here we report the first attempt of automatic ear-based diagnosis in craniofacial dysmorphology. The algorithms we propose have been tested on independent and international validation sets involving rare disease centers in Europe and Asia. Validation data was nevertheless limited for NAFD, highlighting the need for data sharing when designing machine learning-based clinical tools. AI-based automatic facial diagnostic algorithms, including profile and ear analysis, are powerful approaches in supporting practitioners in diagnostic processes.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

Ethics statement

The studies involving humans were approved by CESREES (Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé, № 4570023bis). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

QH: Conceptualization, Methodology, Investigation, Software, Data curation, Formal analysis, Validation, Writing—Original Draft. TB: Data curation, Software. SM: Data curation, Conceptualization. JA: Data curation. TA-B: Data curation. GB: Data curation. LB: Data curation. GC: Validation. PC: Validation. FD: Data curation. FD: Data curation, Software. MD: Data curation. EG: Data curation. WK: Validation. SL: Project

administration. DM: Validation. VP: Data curation, Validation. TP: Validation. ST-R: Validation. MW: Validation. AP: Data curation, Project administration. MR: Project administration, Funding acquisition. NG: Conceptualization, Methodology, Investigation, Software, Project administration, Funding acquisition, Writing—Original Draft. RK: Conceptualization, Methodology, Investigation, Project administration, Funding acquisition, Writing—Original Draft. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the "Agence Nationale de la Recherche", "Investissements d'Avenir" program (ANR-10-IAHU-01), by France 2030 grant "Face4Kids" (ANR-21-PMRB-0004) and by Université Paris Cité with National University of Singapore (2021-05-R/UP-NUS).

Acknowledgments

We thank Françoise Firmin and Joël Ferri for sharing patient data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fped.2023.1171277/full#supplementary-material

SUPPLEMENTARY FIGURE S1
UMAP representations for design №3. (A). Type of variation. (B). Site of variation on EFTUD2: first half ("beginning") or second half ("end").

Supplementary Table S1
EFTUD2 heterozygous pathogenic variations in patients with MFDM MFDM, Mandibulo-Facial Dysostosis with Microcephaly.

Supplementary Table S2
Comparisons of severity and asymmetry scores by study design. MFDM, Mandibulo-Facial Dysostosis with Microcephaly; NAFD, Nager type Acro-Facial Dysostosis; TC, Treacher Collins; CHARGE, Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness.

Supplementary Table S3
Description of the validation set population. MFDM, Mandibulo-Facial Dysostosis Guion Almeida type; NAFD, Nager type Acro-Facial Dysostosis; TC, Treacher Collins; CHARGE, Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; SD, Standard Deviation.

References

1. Guion-Almeida ML, Zechi-Ceide RM, Vendramini S, Ju Nior AT. A new syndrome with growth and mental retardation, mandibulofacial dysostosis, microcephaly, and cleft palate. *Clin Dysmorphol.* (2006) 15(3):171–4. doi: 10.1097/01.mcd.0000220603.09661.7e

2. Guion-Almeida ML, Vendramini-Pittoli S, Passos-Bueno MRS, Zechi-Ceide RM. Mandibulofacial syndrome with growth and mental retardation, microcephaly, ear anomalies with skin tags, and cleft palate in a mother and her son: autosomal dominant or X-linked syndrome? *Am J Med Genet A.* (2009) 149A(12):2762–4. doi: 10.1002/ajmg.a.32816

3. Lines M, Hartley T, MacDonald SK, Boycott KM. Mandibulofacial dysostosis with microcephaly. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, editors. *Genereviews®*. Seattle (WA): University of Washington (1993). (Cited 2022 May 25). Available at: http://www.ncbi.nlm.nih.gov/books/NBK214367/

4. Huang L, Vanstone MR, Hartley T, Osmond M, Barrowman N, Allanson J, et al. Mandibulofacial dysostosis with microcephaly: mutation and database update. *Hum Mutat.* (2016) 37(2):148–54. doi: 10.1002/humu.22924

5. Wieczorek D. Human facial dysostoses. *Clin Genet.* (2013) 83(6):499–510. doi: 10.1111/cge.12123

6. Voigt C, Mégarbané A, Neveling K, Czeschik JC, Albrecht B, Callewaert B, et al. Oto-facial syndrome and esophageal atresia, intellectual disability and zygomatic anomalies—expanding the phenotypes associated with EFTUD2 mutations. *Orphanet J Rare Dis.* (2013) 8:110. doi: 10.1186/1750-1172-8-110

7. Bukowska-Olech E, Materna-Kiryluk A, Walczak-Sztulpa J, Popiel D, Badura-Stronka M, Koczyk G, et al. Targeted next-generation sequencing in the diagnosis of facial dysostoses. *Front Genet.* (2020) 11:580477. doi: 10.3389/fgene.2020.580477

8. Lines MA, Huang L, Schwartzentruber J, Douglas SL, Lynch DC, Beaulieu C, et al. Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. *Am J Hum Genet.* (2012) 90(2):369–77. doi: 10.1016/j.ajhg.2011.12.023

9. RESERVES IUTD. Orphanet: mandibulofacial dysostosis guion almeida type (Cited 2022 May 13). Available at: https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=FR&data_id=11150&Disease_Disease_Search_diseaseType=ORPHA&Disease_Disease_Search_diseaseGroup=79113&Disease(s)/group%20of%20diseases=Mandibulofacial-dysostosis-Guion-Almeida-type&title=Mandibulofacial-dysostosis-Guion-Almeida-type&search=Disease_Search_Simple

10. Ryu JH, Kim HY, Ko JM, Kim MJ, Seong MW, Choi BY, et al. Clinical and molecular delineation of mandibulofacial dysostosis with microcephaly in six Korean patients: when to consider EFTUD2 analysis? *Eur J Med Genet.* (2022) 65(5):104478. doi: 10.1016/j.ejmg.2022.104478

11. Kim SY, Lee DH, Han JH, Choi BY. Novel splice site pathogenic variant of EFTUD2 is associated with mandibulofacial dysostosis with microcephaly and extracranial symptoms in Korea. *Diagnostics (Basel).* (2020) 10(5):296. doi: 10.3390/diagnostics10050296

12. Lines M, Hartley T, MacDonald SK, Boycott KM, et al. Mandibulofacial dysostosis with microcephaly. In: Adam MP, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, Gripp KW, editors. *Genereviews®*. Seattle (WA): University of Washington (1993). (cited 2023 May 13). Available at: http://www.ncbi.nlm.nih.gov/books/NBK214367/

13. Silva JB, Soares D, Leão M, Santos H. Mandibulofacial dysostosis with microcephaly: a syndrome to remember. *BMJ Case Rep.* (2019) 12(8):e229831. doi: 10.1136/bcr-2019-229831

14. Luquetti DV, Hing AV, Rieder MJ, Nickerson DA, Turner EH, Smith J, et al. "Mandibulofacial dysostosis with microcephaly" caused by EFTUD2 mutations: expanding the phenotype. *Am J Med Genet A.* (2013) 161A(1):108–13. doi: 10.1002/ajmg.a.35696

15. Lacour JC, McBride L, St Hilaire H, Mundinger GS, Moses M, Koon J, et al. Novel *de novo* EFTUD2 mutations in 2 cases with MFDM, initially suspected to have alternative craniofacial diagnoses. *Cleft Palate Craniofac J.* (2019) 56(5):674–8. doi: 10.1177/1055665618806379

16. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: dr. Warehouse. *J Biomed Inform.* (2018) 80:52–63. doi: 10.1016/j.jbi.2018.02.019

17. Lehalle D, Gordon CT, Oufadem M, Goudefroye G, Boutaud L, Alessandri JL, et al. Delineation of EFTUD2 haploinsufficiency-related phenotypes through a series of 36 patients. *Hum Mutat.* (2014) 35(4):478–85. doi: 10.1002/humu.22517

18. Bernier FP, Caluseriu O, Ng S, Schwartzentruber J, Buckingham KJ, Innes AM, et al. Haploinsufficiency of SF3B4, a component of the pre-mRNA spliceosomal complex, causes Nager syndrome. *Am J Hum Genet.* (2012) 90(5):925–33. doi: 10.1016/j.ajhg.2012.04.004

19. Gorlin RJ, Cohen MM Jr, Hennekam RCM. *Syndromes of the head and neck*. Oxford, UK: Oxford University Press (2001). 1332 p.

20. Zhao J, Yang L. Broad-spectrum next-generation sequencing-based diagnosis of a case of Nager syndrome. *J Clin Lab Anal.* (2020) 34(9):e23426. doi: 10.1002/jcla.23426

21. Chang JH, Park DH, Shin JP, Kim IT. Two cases of CHARGE syndrome with multiple congenital anomalies. *Int Ophthalmol.* (2014) 34(3):623–7. doi: 10.1007/s10792-013-9817-4

22. Husu E, Hove HD, Farholt S, Bille M, Tranebjærg L, Vogel I, et al. Phenotype in 18 Danish subjects with genetically verified CHARGE syndrome. *Clin Genet.* (2013) 83(2):125–34. doi: 10.1111/j.1399-0004.2012.01884.x

23. Blake KD, Prasad C. CHARGE Syndrome. *Orphanet J Rare Dis.* (2006) 1:34. doi: 10.1186/1750-1172-1-34

24. Lalani SR, Safiullah AM, Fernbach SD, Harutyunyan KG, Thaller C, Peterson LE, et al. Spectrum of CHD7 mutations in 110 individuals with CHARGE syndrome and genotype-phenotype correlation. *Am J Hum Genet.* (2006) 78(2):303–14. doi: 10.1086/500273

25. Marszałek-Kruk BA, Wójcicki P, Dowgierd K, Śmigiel R. Treacher collins syndrome: genetics, clinical features and management. *Genes (Basel).* (2021) 12(9):1392. doi: 10.3390/genes12091392

26. Liu J, Dong J, Li P, Duan W. *De novo* TCOF1 mutation in treacher collins syndrome. *Int J Pediatr Otorhinolaryngol.* (2021) 147:110765. doi: 10.1016/j.ijporl.2021.110765

27. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep.* (1966) 19(1):3–11. doi: 10.2466/pr0.1966.19.1.3

28. Zhou Y, Zaferiou S. Deformable models of ears in-the-wild for alignment and recognition. In: *2017 12th IEEE international conference on automatic face gesture recognition (FG 2017)*, London, UK. (2017). p. 626–33.

29. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* (2017) 39(6):1137–49. doi: 10.1109/TPAMI.2016.2577031

30. Lucas B, Kanade T. An iterative image registration technique with an application to stereo vision (IJCAI). In: *IJCAI'81: 7th international joint conference on artificial intelligence*, Vol. 2. Pittsburgh, Pennsylvania: ACM Digital Library (1981). p. 674–9.

31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. arXiv. (2019). (Cited 2023 Jul 17). Available at: http://arxiv.org/abs/1912.01703

32. Alabort-i-Medina J, Antonakos E, Booth J, Snape P, Zafeiriou S. Menpo: a comprehensive platform for parametric image alignment and visual deformable models. In: *Proceedings of the 22nd ACM international conference on multimedia*. Orlando Florida USA: ACM Digital Library (2014) p. 679–82. (Cited 2022 Feb 24). Available at: https://dl.acm.org/doi/10.1145/2647868.2654890

33. Hennocq Q, Bongibault T, Bizière M, Delassus O, Douillet M, Cormier-Daire V, et al. An automatic facial landmarking for children with rare diseases. *Am J Med Genet A.* (2023) 191(5):1210–21. doi: 10.1002/ajmg.a.63126

34. Landmarker.io. The menpo project. (Cited 2022 Mar 20) Available at: https://www.menpo.org/landmarkerio/

35. Baken EK, Collyer ML, Kaliontzopoulou A, Adams DC. Geomorph v4.0 and gmShiny: enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods Ecol Evol.* (2021) 12(12):2355–63. doi: 10.1111/2041-210X.13723

36. Rohlf FJ, Slice D. Extensions of the procrustes method for the optimal superimposition of landmarks. *Syst Zool.* (1990) 39(1):40–59. doi: 10.2307/2992207

37. Pearson K. LIII On lines and planes of closest fit to systems of points in space. *Lond Edinb,Dublin Philos Mag J Sci.* (1901) 2(11):559–72. doi: 10.1080/14786440109462720

38. Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, et al. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med.* (2003) 348(12):1170–5. doi: 10.1056/NEJMsb025007

39. Marx H. Die missblindungen des ohreds. *Handb Spez Pathol Anat Histol.* (1926) 12:620–5. ISBN: 978-3-642-66019-1.

40. Klingenberg CP, Barluenga M, Meyer A. Shape analysis of symmetric structures: quantifying variation among individuals and asymmetry. *Evolution.* (2002) 56(10):1909–20. doi: 10.1111/j.0014-3820.2002.tb00117.x

41. Palmer AR. Fluctuating asymmetry analyses: a primer. In: Markow TA, editor. *Developmental instability: its origins and evolutionary implications: proceedings of the international conference on developmental instability: its origins and evolutionary implications, tempe, Arizona, 14–15 June 1993*. Dordrecht: Springer Netherlands (1994). p. 335–64. (Contemporary Issues in Genetics and Evolution). Available at: doi: 10.1007/978-94-011-0830-0_26 (Cited 2022 Aug 30).

42. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. (2020). (Cited 2022 Aug 30). Available at: http://arxiv.org/abs/1802.03426

43. R Core Team. European environment agency. (2020). (Cited 2023 Jan 18). Available at: https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006

44. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. New York, NY, USA: Association for Computing Machinery (2016). p. 785–94. (KDD '16). Available at: https://dl.acm.org/doi/10.1145/2939672.2939785 (Cited 2023 Jul 4).

45. Sachs MC. plotROC: a tool for plotting ROC curves. *J Stat Softw.* (2017) 79:2. doi: 10.18637/jss.v079.c02

46. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* (2019) 380(14):1347–58. doi: 10.1056/NEJMra1814259

47. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. *Radiol.* (2018) 288(2):318–28. doi: 10.1148/radiol.2018171820

48. Novoa RA, Gevaert O, Ko JM. Marking the path toward artificial intelligence-based image classification in dermatology. *JAMA Dermatol.* (2019) 155(10):1105–6. doi: 10.1001/jamadermatol.2019.1633

49. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. *JAMA Surg.* (2020) 155(2):148–58. doi: 10.1001/jamasurg.2019.4917

50. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med.* (2019) 25(1):60–4. doi: 10.1038/s41591-018-0279-0

51. Zhang Q, Ding Y, Feng B, Tang Y, Chen Y, Wang Y, et al. Molecular and phenotypic expansion of alström syndrome in Chinese patients. *Front Genet.* (2022) 13:808919. doi: 10.3389/fgene.2022.808919

52. Javitt MJ, Vanner EA, Grajewski AL, Chang TC. Evaluation of a computer-based facial dysmorphology analysis algorithm (Face2Gene) using standardized textbook photos. *Eye.* (2022) 36(4):859–61. doi: 10.1038/s41433-021-01563-5

53. Latorre-Pellicer A, Ascaso Á, Trujillano L, Gil-Salvador M, Arnedo M, Lucia-Campos C, et al. Evaluating Face2Gene as a tool to identify cornelia de lange syndrome by facial phenotypes. *Int J Mol Sci.* (2020) 21(3):E1042. doi: 10.3390/ijms21031042

54. Mishima H, Suzuki H, Doi M, Miyazaki M, Watanabe S, Matsumoto T, et al. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. *J Hum Genet.* (2019) 64(8):789–94. doi: 10.1038/s10038-019-0619-z

55. Guion-Almeida ML, Kokitsu-Nakata NM, Richieri-Costa A. Clinical variability in cerebro-oculo-nasal syndrome: report on two additional cases. *Clin Dysmorphol.* (2000) 9(4):253–7. doi: 10.1097/00019605-200009040-00004

56. Smigiel R, Bezniakow N, Jakubiak A, Bloch M, Patkowski D, Obersztyn E, et al. Phenotype analysis of Polish patients with mandibulofacial dysostosis type guion-almeida associated with esophageal atresia and choanal atresia caused by EFTUD2 gene mutations. *J Appl Genet.* (2015) 56(2):199–204. doi: 10.1007/s13353-014-0255-4

57. Yu KPT, Luk HM, Gordon CT, Fung G, Oufadem M, Garcia-Barcelo MM, et al. Mandibulofacial dysostosis guion-almeida type caused by novel EFTUD2 splice site variants in two Asian children. *Clin Dysmorphol.* (2018) 27(2):31–5. doi: 10.1097/MCD.0000000000000214

58. Katsanis SH, Jabs EW. Treacher collins syndrome. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, et al., editors. *Genereviews®*. Seattle (WA): University of Washington (1993). (Cited 2022 Sep 8). Available at: http://www.ncbi.nlm.nih.gov/books/NBK1532/

59. Fakhim S A, Shahidi N, Mousaviagdas M. A case report: nager acrofacial dysostosis. *Iran J Otorhinolaryngol.* (2012) 24(66):45–50. PMCID: PMC3846201.

60. Davenport SL, Hefner MA, Thelin JW. CHARGE Syndrome. Part I. External ear anomalies. *Int J Pediatr Otorhinolaryngol.* (1986) 12(2):137–43. doi: 10.1016/S0165-5876(86)80071-4

5.2 AI-based diagnosis in fetal pathology using external ear shapes

We applied the previous model to a new validation population, corresponding to fetuses. Facial analysis of fetuses is crucial for antenatal and post-mortem diagnosis (118). Prenatal abnormalities of the external ear are common in many syndromes (119,120). For instance, Nunez-Castruita et al (121) reported that the insertion position of the external ears supports the diagnostic process in genetic disorders, as fetuses with low set ears are at greater risk of congenital anomalies. Similarly, Sacchini et al (122) recommended the analysis of the length of the external ear in the antenatal diagnosis of Down syndrome.

The aim of this study was to test the model described in the previous section on photographs of fetal ears, with the aim of supporting the medical genetics diagnosis. The two genetic syndromes studied were Mandibulo-Facial Dysostosis with Microcephaly (MFDM) and CHARGE syndromes. These two syndromes present, as explained previously, characteristic external ear phenotypes (123,124), enabling post-mortem diagnosis.

We tested the trained model on photographs of fetuses from the Fetal pathology department of *Hôpital Necker - Enfants Malades*. Control came from a large data set of fetuses including intra uterine fetal death, termination of pregnancy and late miscarriage, autopsied between 2022 and 2023. From this data, 14 subjects were selected according to several criteria. We excluded: macerated fetuses, fetuses with polymalformative syndrome (except isolated malformation), fetuses with dysmorphic features, fetuses with chromosomal abnormalities, and fetuses with oligoamnios. Post-mortem examination was performed in the Fetal pathology department following a standard protocol including frontal and profile X-rays, photographs, external and internal examination, and histological assessment (125). Each fetus was photographed using a Nikon D7000 device. The associated metadata were reported, i.e., term and gender. The term was then converted to age (then negative) by subtracting the theoretical term (41 SA) from the term of the fetus for input into the linear models previously

described. We included 14 photographs of control ears, corresponding to 14 fetuses; 71% of controls were female, with a mean term of 27 +/- 7.2 WA. Controls were one late miscarriage, two non-macerated intra-uterine fetal death without malformation in post-mortem examination, and eleven terminations of pregnancy (ten for isolated cardiopathy and one for low urinary tract obstruction without oligoamnios). The MFDM group included 12 photographs from 12 fetuses, all genetically confirmed (*EFTUD2* heterozygous pathogenic variations); 50% were female and the mean term was 31 +/- 4.5 WA. The CHARGE group included 25 photos from 25 fetuses; 60% were female and mean term was 31 +/- 4.5 WA. All were genetically confirmed (*CHD7* variation).

37/51 (72.5%) of the fetuses were correctly classified (Table 1). The overall accuracy was therefore 72.6% (58.3 - 84.1%, $p < 0.001$), and the balanced accuracies were 76.4%, 86.2%, and 74.9% respectively for CHARGE, MFDM and control fetuses (Table 2). The AUC was 86.8%, 90.3%, and 87.5% respectively for CHARGE, MFDM and control fetuses.

		Reference		
		CHARGE	**MFDM**	**Control**
Prediction	**CHARGE**	**18**	1	4
	MFDM	1	**9**	0
	Control	6	2	**10**

Table 3. Confusion matrix for the model predictions on the validation set. The green boxes corresponded to true positives (correctly classified fetuses). MFDM = Mandibulo-Facial Dysostosis with Microcephaly; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness.

Overall (multiclass design)		
Accuracy		**72.6%** [58.3 - 84.1%] p < 0.001 *
Binary (one-versus-all design)		
Sensitivity (Se)	CHARGE	72.0%
	Control	71.4%
	MFDM	75.0%
Specificity (Sp)	CHARGE	80.8%
	Control	78.3%
	MFDM	97.4%
Balanced Accuracy	CHARGE	76.4%
	Control	74.9%
	MFDM	86.2%
AUC	CHARGE	86.8%
	Control	87.5%
	MFDM	90.3%

Table 2. Final model performances on the validation set. AUC = Area Under the Curve.

* Statistical significance compared to the *No Information Rate* (NIR) = 0.490.

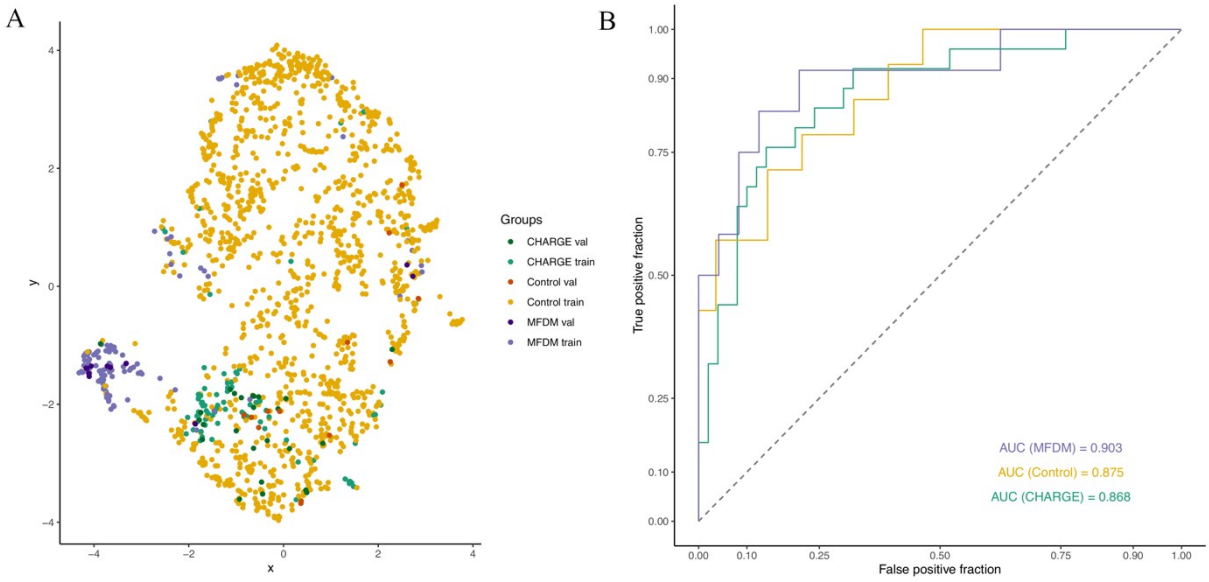


Figure 9. A. UMAP visualization for the training set (children) and validation set (fetuses). Note that the MFDM cluster is more distinct than the two other ones. **B. ROC curves for the validation data for each syndrome, with the associated AUC (in a one-versus-all design).** AUC = Area Under the Curve.

We were able to prove the efficiency of an automatic detection model for external ear anomalies of two genetic syndromes versus controls, trained on children and tested on fetuses. External ears alone are not sufficient for antenatal diagnosis. Prospects based on our current results are (1) to integrate information on the face (frontal and profile views) in order to offer a comprehensive diagnostic support model for fetal pathologists and (2) to validate the performances of the model on 3D ultrasound data to include our approach in antenatal diagnosis. This publication is currently under review.

AI-based diagnosis in fetal pathology using external ear shapes

Submitted to Ultrasound in Obstetrics, october 2023

Quentin **Hennocq**[1,2,3*], Nicolas **Garcelon**[1], Thomas **Bongibault**[1,3], Bettina **Bessières**[1,4], Thomas **Bouygues**[1,3], Sandrine **Marlin**[1,4], Jeanne **Amiel**[1,4], Lucile **Boutaud**[4], Maxime **Douillet**[1], Laurence **Loeuillet**[1,4], Stanislas **Lyonnet**[1,4], Adélie **Perrot**[5], Véronique **Pingault**[1,4], Arnaud **Picard**[2], Chloé **Quélin**[5], Marlène **Rio**[1,4], Marjolaine **Willems**[6], Tania **Attie-Bitach**[1,4], **Roman** H. **Khonsari** [1,2,3#], Nathalie **Roux**[4,7#]

[1]Imagine Institute, Data Science Platform, Université Paris Cité, INSERM UMR 1163, Paris, France.

[2]Département de chirurgie maxillo-faciale et chirurgie plastique, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou ; Faculté de Médecine, Université Paris Cité ; Paris, France

[3]Laboratoire 'Forme et Croissance du Crâne', Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris, Paris, France.

[4]Service de médecine génomique des maladies rares, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Faculté de Médecine, Université Paris Cité ; 75015 Paris, France

[5]Département de génétique clinique, CHU Hôpital Sud, Rennes, France.

[6]Département de Génétique Clinique, CHRU de Montpellier, Hôpital Arnaud de Villeneuve, Institute for Neurosciences of Montpellier, Univ Montpellier, INSERM, 34000 Montpellier, France.

[7]EA fetus 7328 and LUMIERE Platform, Université de Paris, Paris, France

# These authors contributed equally to this work.

Abstract. External ear analysis in fetuses provides key clinical elements for antenatal and post-mortem diagnoses. Here we trained an automatic phenotype assessment tool to recognize syndromic ears in two syndromes, CHARGE and Mandibulo-Facial Dysostosis Guion Almeida type (MFDGA), versus controls. We trained an automatic model on all profile pictures of children diagnosed with genetically confirmed MFDGA and CHARGE syndromes, and a cohort of control patients, collected from 1981 to 2023 in Necker-Enfants Malades Hospital (Paris) with a visible external ear. The model consisted in extracting landmarks from photographs of external ears, in applying geometric morphometry methods (Procrustes transformation) and in classification using an eXtreme Gradient Boosting (XGboost) model. The approach was then tested on photographs of two groups of fetuses: controls or with CHARGE and MFDGA syndromes. The training set contained a total of 1592 ear photographs, corresponding to 550 children. The validation set contained a total of 51 ear photographs, corresponding to 51 fetuses. The overall accuracy was 72.6% (58.3 - 84.1%, $p < 0.001$), and 76.4%, 74.9% and 86.2% respectively for CHARGE, control and MFDGA fetuses. The AUC were 86.8%, 87.5% and 90.3% respectively for CHARGE, controls and MFDGA fetuses. We report the first automatic fetal phenotyping model, with satisfactory classification performances. Further validations are required before using this approach as a diagnostic tool.

Keywords: AI; machine learning; dysmorphology: fetal pathology; external ear; MFDGA; CHARGE

Introduction. Applications of machine learning are increasing in healthcare [1–4]. The field of dysmorphology has been challenged and transformed by the framework for genetic syndrome classification called DeepGestalt [5], produced by the Face2Gene group. Publications comparing human performances to DeepGestalt performances are flourishing [6–9], and some of them suggest that digital tools do it better than human experts in terms of diagnosis. However, these tools do not currently use any information about the morphology of the outer ear; andnone of these tools has yet been tested on fetuses. Even though the diagnosis of a given syndrome is never fully based on ear anomalies, this anatomical region is a major source of distinctive phenotypic features in a large array of syndromes [10–12]

Facial analysis of fetuses is crucial for antenatal and post-mortem diagnosis [13]. Prenatal abnormalities of the external ear are common in many syndromes [14,15]. For instance, Nunez-Castruita et al [16] report that the insertion position of the external ears supports the diagnostic process in genetic disorders, as fetuses with low set ears are at greater risk of congenital anomalies. Similarly, Sacchini et al [17] recommend the analysis of the length of the external ear in the antenatal diagnosis of Down syndrome.

The aim of this study was (1) to train a tool using artificial intelligence (AI) methods on photographs of the external ears of control children and children with 2 genetic syndromes, and (2) to test this tool on photographs of fetal ears, with the aim of supporting the medical genetics diagnosis. The genetic syndromes studied were Mandibulo-Facial Dysostosis Guion Almeida type (MFDGA) and CHARGE syndromes. These two syndromes present characteristic external ear phenotypes [18,19], enabling post-mortem diagnosis.

Material and Methods

Training set

We screened the photographic databases of the Maxillofacial surgery & Plastic surgery and Medical genetics departments of *Hôpital Necker – Enfants Malades* (Assistance Publique – Hôpitaux de Paris), Paris, France. This database contained 594,000 photographs from 22,000 patients followed in these departments since 1981. All photographs were taken by a professional medical photographer using a Nikon D7000 device in standardized positions.

We included retrospectively and prospectively, from 1981 to 2023, all profile pictures of patients diagnosed with MFDGA and CHARGE syndromes, with a visible external ear (Figure 1). All patients had genetic confirmation of their syndrome. We excluded patients with ear reconstruction surgery.

Control children were selected among patients admitted for lacerations, trauma, infection and various skin lesions, without any records of a chronic disease. More precisely, follow-up for any type of chronic disease was considered as an exclusion criterion. The reports were retrieved using Dr Warehouse [20], the local hospital data warehouse. For each patient, right and left sides were included.

The study was approved by the CESREES (Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé, №4570023bis) and by the CNIL (Commission Nationale Informatique et Libertés, №MLD/MFI/AR221900). Informed and written consents were obtained from the legal representatives of each child, or from the patient himself when over 18 years of age.

Landmarking

We used an available template [21] based on 55 landmarks placed on the outer helix, the antihelix, the lobe, the tragus, the antitragus, the helix, the crus helicis, and the concha. We developed an automatic annotation model trained on 1592 manually annotated ear photographs following a pipeline including: (1) a Faster R-CNN (Convolution Neural Network) to detect ears on lateral face photographs and (2) a patch-AAM (Active Appearance Model), to automatically place landmarks. Each automatically annotated photograph was checked by the first author (QH) and landmarks were manually re-positioned when necessary, using *landmarker.io* [22].

To ensure a uniform distribution of landmarks along the curves of the ear (outer helix, inner helix, antihelix, concha), anatomical landmarks were transformed into sliding semi-landmarks using the *geomorph* package on R [23]. Landmarks corresponding to the antihelix were removed because our team [24] has previously showed that they were not reproducible between two annotators.

Ears were finally modeled based on 41 anatomical landmarks and semi-landmarks, placed automatically and double-checked manually.

Geometric morphometrics

We performed Generalized Procrustes Analysis (GPA) [25] on all landmark clouds using the *geomorph* package on R [23]. Since the data were uncalibrated photographs, ear sizes were not available: shape parameters only were assessed and not centroid sizes.

Procrustean coordinates were processed using Principal Component Analysis (PCA) for dimension reduction: 8 principal components (PC) accounting for more than 90% of the global variance were retained.

To take into account associated metadata (age and gender) and the fact we had included more than one photograph per patient (non-independence), a mixed model was designed for each principal component. A random effect on age and individuals was introduced. The equation of the mixed model was:

$$PC_{i,j} \sim \alpha + age.\beta_1 + gender.\beta_2 + age.\beta_{1,i} + \varepsilon_{i,j}$$

where $age.\beta_{1,i}$ corresponded to a random slope for age per individual, and $\varepsilon_{i,j}$ was a random error term. We did not use an interaction term between age and gender as it did not increase the likelihood of the model. A model residual could thus be plotted for each PC and for each ear,

Uniform Manifold Approximation and Projection (UMAP) representations

The residuals $\varepsilon_{i,j}$ were represented using UMAP for visual clustering, a nonlinear dimension reduction technique [26]. A k (local neighborhood size) value of 15 was used. A cosine metric was introduced to compute distances in high dimensional spaces: the effective minimal distance between embedded points was 10^{-6}. The three conditions of UMAP, namely uniform distribution, local constancy of the Riemannian metric and local connectivity were verified. UMAP analyses were performed using the package *umap* on R [27].

Machine learning models and metrics

The landmark clouds were superimposed with the previous generalized Procrustes analysis and PCA. With the metadata (age and gender), the residuals $\varepsilon_{i,j}$ were reported for each PC and each ear of the validation group.

We used XGBoost (eXtreme Gradient Boosting), a supervised machine learning classifier, for all the analyses [28]. We set a number of hyperparameters to improve the performance and effect of the machine learning model: learning rate = 0.3, gamma = 0, maximum tree depth = 6. We separated the dataset into a training set and a testing set, and a 5-fold cross-validation was used to define the ideal number of iterations to avoid overfitting. The model with the lowest logloss-score was chosen for analysis. The chosen model was then used on the independent validation set to test performances, by plotting accuracy, sensitivity, specificity, F1-score, precision and recall, AUC (in a one versus all design). The ROC (Receiver Operating Characteristics) curves were plotted in R using the *plotROC* package [29].

Validation set

We tested the trained model on photographs of fetuses from the Fetal pathology department of *Hôpital Necker - Enfants Malades*. Control came from a large data set of fetuses including intra uterine fetal death, termination of pregnancy and late miscarriage, autopsied between 2022 and 2023. From this data, 14 subjects were selected according to several criteria. We excluded: macerated fetuses, fetuses with polymalformative syndrome (except isolated malformation), fetuses with dysmorphic features, fetuses with chromosomal abnormalities, fetuses with oligoamnios. Post mortem examination was performed in the Fetal pathology department following a standard protocol including face and profile radiographies, photographies, external and internal examination and histological [30]. Each fetal profile was photographed using a Nikon D7000 device. The associated metadata were reported, i.e. term and gender. The term was then converted to age (then negative), by subtracting the theoretical term (41 SA) from the term of the fetus, for input into the linear models previously described.

Each affected fetus had genetic confirmation of the syndrome. Manual landmarking was performed by the same author (QH), blind to the diagnosis. None of the fetus in the validation set were present twice (Figure 1).

Ears were finally modeled based on 41 anatomical landmarks and semi-landmarks, placed automatically and double-checked manually.

Geometric morphometrics

We performed Generalized Procrustes Analysis (GPA) [25] on all landmark clouds using the *geomorph* package on R [23]. Since the data were uncalibrated photographs, ear sizes were not available: shape parameters only were assessed and not centroid sizes.

Procrustean coordinates were processed using Principal Component Analysis (PCA) for dimension reduction: 8 principal components (PC) accounting for more than 90% of the global variance were retained.

To take into account associated metadata (age and gender) and the fact we had included more than one photograph per patient (non-independence), a mixed model was designed for each principal component. A random effect on age and individuals was introduced. The equation of the mixed model was:

$$PC_{i,j} \sim \alpha + age.\beta_1 + gender.\beta_2 + age.\beta_{1,i} + \varepsilon_{i,j}$$

where $age.\beta_{1,i}$ corresponded to a random slope for age per individual, and $\varepsilon_{i,j}$ was a random error term. We did not use an interaction term between age and gender as it did not increase the likelihood of the model. A model residual could thus be plotted for each PC and for each ear,

Uniform Manifold Approximation and Projection (UMAP) representations

The residuals $\varepsilon_{i,j}$ were represented using UMAP for visual clustering, a nonlinear dimension reduction technique [26]. A k (local neighborhood size) value of 15 was used. A cosine metric was introduced to compute distances in high dimensional spaces: the effective minimal distance between embedded points was 10^{-6}. The three conditions of UMAP, namely uniform distribution, local constancy of the Riemannian metric and local connectivity were verified. UMAP analyses were performed using the package *umap* on R [27].

Machine learning models and metrics

The landmark clouds were superimposed with the previous generalized Procrustes analysis and PCA. With the metadata (age and gender), the residuals $\varepsilon_{i,j}$ were reported for each PC and each ear of the validation group.

We used XGBoost (eXtreme Gradient Boosting), a supervised machine learning classifier, for all the analyses [28]. We set a number of hyperparameters to improve the performance and effect of the machine learning model: learning rate = 0.3, gamma = 0, maximum tree depth = 6. We separated the dataset into a training set and a testing set, and a 5-fold cross-validation was used to define the ideal number of iterations to avoid overfitting. The model with the lowest logloss-score was chosen for analysis. The chosen model was then used on the independent validation set to test performances, by plotting accuracy, sensitivity, specificity, F1-score, precision and recall, AUC (in a one versus all design). The ROC (Receiver Operating Characteristics) curves were plotted in R using the *plotROC* package [29].

Validation set

We tested the trained model on photographs of fetuses from the Fetal pathology department of *Hôpital Necker - Enfants Malades*. Control came from a large data set of fetuses including intra uterine fetal death, termination of pregnancy and late miscarriage, autopsied between 2022 and 2023. From this data, 14 subjects were selected according to several criteria. We excluded: macerated fetuses, fetuses with polymalformative syndrome (except isolated malformation), fetuses with dysmorphic features, fetuses with chromosomal abnormalities, fetuses with oligoamnios. Post mortem examination was performed in the Fetal pathology department following a standard protocol including face and profile radiographies, photographies, external and internal examination and histological [30]. Each fetal profile was photographed using a Nikon D7000 device. The associated metadata were reported, i.e. term and gender. The term was then converted to age (then negative), by subtracting the theoretical term (41 SA) from the term of the fetus, for input into the linear models previously described.

Each affected fetus had genetic confirmation of the syndrome. Manual landmarking was performed by the same author (QH), blind to the diagnosis. None of the fetus in the validation set were present twice (Figure 1).

Results

Training set

The training set contained a total of 1592 ear photographs, corresponding to 550 patients (Figure 2); 52% of patients were female and the mean age was 7.2 +/- 5.9 years, ranging from 0 to 60.7 years.

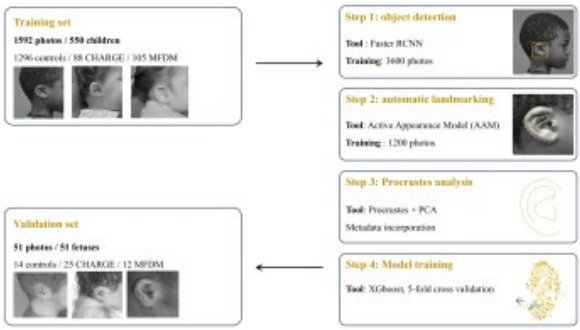


Figure 1. Description of the model steps, and the training and validation sets. MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; RCNN = Convolutional Neural Network; PCA = Principal Component Analysis; AAM = Active Appearance Model; XGboost = eXtreme Gradient Boosting.

Figure 2. Typical examples of external ear photographs for each patient group in the training set (children): controls, CHARGE syndrome and Mandibulo-Facial Dysostosis Guion Almeida type (MFDGA). Phenotypes are sometimes not as pronounced.

We included 1296 photographs of control ears, corresponding to 471 patients; 53% of controls were female, with a mean age of 7.2 +/- 5.4 years. The MFDGA group included 105 photographs from 31 patients, all genetically confirmed (*EFTUD2* heterozygous pathogenic variations); 52% were female and the mean age was 9.2 +/- 9.8 years. The CHARGE group included 88 photos from 24 patients; 42% were female and mean age was 5.1 +/- 5.9 years. All were genetically confirmed (*CHD7* variation) (Table 1).

Average models per group were then designed after Procrustes transformation (Figures 2). The MFDGA group had a clockwise rotation and a vertical shift of the concha when compared to controls. Previously described features - thickened helix, enlarged and square lobe - were also reported. In the CHARGE group, we reported badly hemmed, squared and cup-shaped external ear, as previously described in the literature [31].

		Total	Controls	MFDGA	CHARGE
N (ears)		1592	1296	105	88
N (patients)		550	471	31	24
Gender	Females	288/550 (52%)	251/471 (53%)	16/31 (52%)	10/24 (42%)
Age					
	Mean +/- SD	7.2 +/- 5.9	7.2 +/- 5.4	9.2 +/- 9.8	5.1 +/- 5.9
	Median	6.9	7.3	5.0	6.5
	Min	0.0	0.1	0.0	0.0
	Max	60.7	60.7	39.6	21.6
Identified pathogenic genetic variation			NA	31/31 (100%)	24/24 (100%)

Table 1. Description of the training set population. MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; SD = Standard Deviation.

Validation set

The validation set contained a total of 51 ear photographs, corresponding to 51 fetuses (Figure 4); 61% of were female and the mean term was 29 +/- 5.4 WA, ranging from 15 to 39 WA.

We included 14 photographs of control ears, corresponding to 14 fetuses; 71% of controls were female, with a mean term of 27 +/- 7.2 WA. Controls were one late miscarriage and two non-macerated intra uterine fetal death without malformation in post mortem examination and eleven terminations of pregnancy (ten for isolated cardiopathy and one low urinary tract obstruction without oligoamnios). The MFDGA group included 12 photographs from 12 fetuses, all genetically confirmed (*EFTUD2* heterozygous pathogenic variations); 50% were female and the mean term was 31 +/- 4.5 WA. The CHARGE group included 25 photos from 25 fetuses; 60% were female and mean term was 31 +/- 4.5 WA. All were genetically confirmed (*CHD7* variation) (Table 2).

Classification performances

The optimal performance of the model without overfitting was obtained after 19 iterations. 37/51 (72.5%) of the fetuses were correctly classified (Table 3).

The overall accuracy was therefore 72.6% (58.3 - 84.1%, p < 0.001), and the balanced accuracies were 76.4%, 86.2% and 74.9% respectively for CHARGE, MFDGA and control fetuses (Table 4). The AUC was 86.8%, 90.3% and 87.5% respectively for CHARGE, MFDGA and control fetuses (Figure 5).

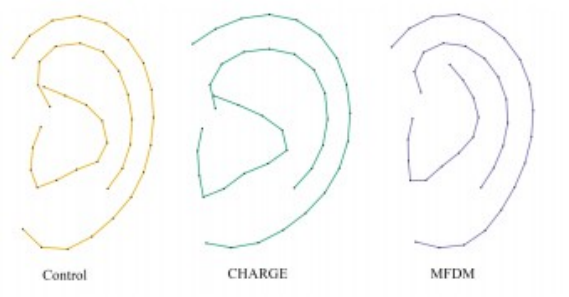


Figure 3. Mean shapes after Procrustes superimposition for controls, CHARGE, and MFDGA.

Figure 4. Examples of external ear photographs for each group in the validation set (fetuses): control (23 WA), CHARGE syndrome (34 WA) and Mandibulo-Facial Dysostosis Guion Almeida type (MFDGA) (27 WA).

		Total	Controls	MFDGA	CHARGE
N (ears)		51	14	12	25
N (patients)		51	14	12	25
Gender	Females	31/51 (61%)	10/14 (71%)	6/12 (50%)	15/25 (60%)
Term (WA)					
	Mean +/- SD	29 +/- 5.4	27 +/- 7.2	29 +/- 4.0	31 +/- 4.5
	Median	30	27	29	32
	Min	15	15	22	18
	Max	39	39	35	37
Identified pathogenic genetic variation			NA	12/12 (100%)	25/25 (100%)

Table 2. Description of the validation set population. MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; WA = Weeks of Amenorrhea; SD = Standard Deviation.

		Reference		
		CHARGE	**MFDGA**	**Control**
Prediction	**CHARGE**	**18**	1	4
	MFDGA	1	**9**	0
	Control	6	2	**10**

Table 3. Confusion matrix for our model predictions on the validation set. The green boxes correspond to true positives (correctly classified fetuses). MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness.

Overall (multiclass design)		
Accuracy		**72.6%** [58.3 - 84.1%] p < 0.001 *
Binary (one-versus-all design)		
Sensitivity (Se)	CHARGE	72.0%
	Control	71.4%
	MFDGA	75.0%
Specificity (Sp)	CHARGE	80.8%
	Control	78.3%
	MFDGA	97.4%
Balanced Accuracy	CHARGE	76.4%
	Control	74.9%
	MFDGA	86.2%
AUC	CHARGE	86.8%
	Control	87.5%
	MFDGA	90.3%

Table 4. Final model performances on the validation set. AUC = Area Under the Curve.
* Statistical significance compared to the *No Information Rate* (NIR) = 0.490.

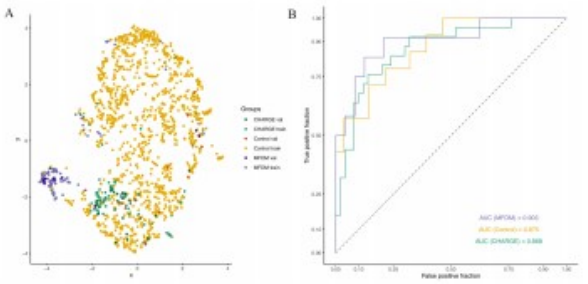


Figure 5. A. UMAP visualization for the training set (children) and validation set (fetuses). Note that the MFDGA cluster is more distinct than the two other ones. **B. ROC curves for the validation data for each syndrome, with the associated AUC (in a one-versus-all design).** AUC = Area Under the Curve.

Discussion. We were able to prove the efficiency of an automatic detection model for external ear anomalies of two genetic syndromes versus controls, trained on children and tested on fetuses. Performance was satisfactory, with a significant overall accuracy of 74.9%, and 76.4% and 86.2% respectively for the specific detection of control, CHARGE and MFDGA fetuses.

Facial analysis of fetuses is crucial for antenatal and post-mortem diagnosis [13]. Abnormalities of the external ear are common in many syndromes [14,15]. Sondern et al [32] proposes charts for the size/width ratio of the prenatal external ear using 3D ultrasound, and proposes a pathological score in a fetus with proven Noonan syndrome. These authors suggest that the external ear is crucial for the antenatal diagnosis of craniofacial anomalies.

A study conducted by Biard et al [33] showed that neonatal clinical examination of children with CHARGE syndrome revealed abnormal external ears in 100% of cases, with low set, posteriorly rotated, symmetrical and badly hemmed, squared or cup-shaped ears. Other studies [19,34,35] also report a 100% prevalence of external ear anomalies in this condition. Biard et al. [33] therefore stated that external ear analysis should be considered as the keystone of antenatal diagnosis in CHARGE syndrome. The detection rate for these anomalies is 48% by ultrasound and 100% by antenatal MRI, according to the study conducted by Millischer et al [36]. Regarding MFDGA, Lehalle et al [18] described abnormalities of the external ear in 100% out of 34 MFDGA cases, with minor abnormalities in 29/34 cases (squared, flattened and externally deviated ear lobe), asymmetric ears in 24% of cases and preauricular tags in 33% of cases.

External ear development begins in the sixth embryonic week with the formation of six mesenchymal hillocks and is achieved by the tenth embryonic week. However, the external ear continues to grow throughout the gestation and it is not until 32 weeks that the ear pinna completes its migration to the side of the head [37]. The characteristic surface configuration is thus detectable from the 4th month of gestation onwards, when the auricular hillocks merge and form the outer ear [14,32]. The examination of the fetal ear with 2D ultrasound is most informative between 20 and 24 weeks of gestation and is currently part of the standard ultrasound screening [32].

Previous studies on the fetal ear development consistently showed a positive linear correlation of both ear length and width with gestational age and established reference values [38–40]. In addition to 2D ultrasound, 3D ultrasound has proven useful in the differentiation between normal and abnormal fetal anatomy of the external ear [41,42].

Here we report the first automatic fetal phenotyping model, focused on the external ear. External ears alone are not sufficient for antenatal diagnosis. Future prospects based on our current results are (1) to integrate information on the face (frontal and profile views) in order to offer a comprehensive diagnostic support model for fetal pathologists and (2) to validate the performances of the model on 3D ultrasound data to include our approach in antenatal diagnosis.

Acknowledgements

We thank Dr Françoise Firmin for sharing patient data. This work was supported by the 'Agence Nationale de la Recherche', 'Investissements d'Avenir' program (ANR-10-IAHU-01), by France 2030 grant "Face4Kids" (ANR-21-PMRB-0004) and by Université Paris Cité with National University of Singapore (2021-05-R/UP-NUS).

References

1. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259

2. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, Geis JR, Pandharipande PV, Brink JA, Dreyer KJ. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*. 2018;288(2):318-328. doi:10.1148/radiol.2018171820

3. Novoa RA, Gevaert O, Ko JM. Marking the Path Toward Artificial Intelligence-Based Image Classification in Dermatology. *JAMA Dermatol*. 2019;155(10):1105-1106. doi:10.1001/jamadermatol.2019.1633

4. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, Rashidi P, Upchurch GR, Bihorac A. Artificial Intelligence and Surgical Decision-making. *JAMA Surg*. 2020;155(2):148-158. doi:10.1001/jamasurg.2019.4917

5. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, Basel-Salmon L, Krawitz PM, Kamphausen SB, Zenker M, Bird LM, Gripp KW. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med*. 2019;25(1):60-64. doi:10.1038/s41591-018-0279-0

6. Zhang Q, Ding Y, Feng B, Tang Y, Chen Y, Wang Y, Chang G, Liu S, Wang J, Li Q, Fu L, Wang X. Molecular and Phenotypic Expansion of Alström Syndrome in Chinese Patients. *Front Genet*. 2022;13:808919. doi:10.3389/fgene.2022.808919

7. Javitt MJ, Vanner EA, Grajewski AL, Chang TC. Evaluation of a computer-based facial dysmorphology analysis algorithm (Face2Gene) using standardized textbook photos. *Eye*. 2022;36(4):859-861. doi:10.1038/s41433-021-01563-5

8. Latorre-Pellicer A, Ascaso Á, Trujillano L, Gil-Salvador M, Arnedo M, Lucia-Campos C, Antoñanzas-Pérez R, Marcos-Alcalde I, Parenti I, Bueno-Lozano G, Musio A, Puisac B, Kaiser FJ, Ramos FJ, Gómez-Puertas P, Pié J. Evaluating Face2Gene as a Tool to Identify Cornelia de Lange Syndrome by Facial Phenotypes. *Int J Mol Sci*. 2020;21(3):E1042. doi:10.3390/ijms21031042

9. Mishima H, Suzuki H, Doi M, Miyazaki M, Watanabe S, Matsumoto T, Morifuji K, Moriuchi H, Yoshiura KI, Kondoh T, Kosaki K. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. *J Hum Genet*. 2019;64(8):789-794. doi:10.1038/s10038-019-0619-z

10. Rooijers W, Tio P a. E, van der Schroeff MP, Padwa BL, Dunaway DJ, Forrest CR, Koudstaal MJ, Caron CJJM. Hearing impairment and ear anomalies in craniofacial microsomia: a systematic review. *Int J Oral Maxillofac Surg*. Published online February 3, 2022:S0901-5027(22)00007-8. doi:10.1016/j.ijom.2022.01.005

11. Rymer K, Shiang R, Hsiung A, Pandya A, Bigdeli T, Webb BT, Rhodes J. Expanding the phenotype for the recurrent p.Ala391Glu variant in FGFR3: Beyond crouzon syndrome and acanthosis nigricans. *Mol Genet Genomic Med*. 2019;7(6):e656. doi:10.1002/mgg3.656

12. Peart LS, Gonzalez J, Bivona S, Latchman K, Torres L, Undiagnosed Diagnosis Network, Tekin M. Bilateral choanal stenosis in auriculocondylar syndrome caused by a PLCB4 variant. *Am J Med Genet A*. 2022;188(4):1307-1310. doi:10.1002/ajmg.a.62634

13. Andresen C, Matias A, Merz E. Fetal face: the whole picture. *Ultraschall Med*. 2012;33(5):431-440. doi:10.1055/s-0031-1299482

14. Bartel-Friedrich S, Wulke C. Classification and diagnosis of ear malformations. *GMS Curr Top Otorhinolaryngol Head Neck Surg*. 2008;6:Doc05.

15. Dudarewicz L, Kałuzewski B. Prenatal screening for fetal chromosomal abnormalities using ear length and shape as an ultrasound marker. *Med Sci Monit*. 2000;6(4):801-806.

16. Nuñez-Castruita A, López-Serna N. Low-set ears and associated anomalies in human foetuses. *Int J Pediatr Otorhinolaryngol*. 2018;104:126-133. doi:10.1016/j.ijporl.2017.11.008

17. Sacchini C, El-Sheikhah A, Cicero S, Rembouskos G, Nicolaides KH. Ear length in trisomy 21 fetuses at 11-14 weeks of gestation. *Ultrasound Obstet Gynecol*. 2003;22(5):460-463. doi:10.1002/uog.903

18. Lehalle D, Gordon CT, Oufadem M, Goudefroye G, Boutaud L, Alessandri JL, Baena N, Baujat G, Baumann C, Boute-Benejean O, Caumes R, Decaestecker C, Gaillard D, Goldenberg A, Gonzales M, Holder-Espinasse M, Jacquemont ML, Lacombe D, Manouvrier-Hanu S, Marlin S, Mathieu-Dramard M, Morin G, Pasquier L, Petit F, Rio M, Smigiel R, Thauvin-Robinet C, Vasiljevic A, Verloes A, Malan V, Munnich A, de Pontual L, Vekemans M, Lyonnet S, Attié-Bitach T, Amiel J. Delineation of EFTUD2 haploinsufficiency-related phenotypes through a series of 36 patients. *Hum Mutat*. 2014;35(4):478-485. doi:10.1002/humu.22517

19. Sanlaville D, Etchevers HC, Gonzales M, Martinovic J, Clément-Ziza M, Delezoide AL, Aubry MC, Pelet A, Chemouny S, Cruaud C, Audollent S, Esculpavit C, Goudefroye G, Ozilou C, Fredouille C, Joye N, Morichon-Delvallez N, Dumez Y, Weissenbach J, Munnich A, Amiel J, Encha-Razavi F, Lyonnet S, Vekemans M, Attié-Bitach T. Phenotypic spectrum of CHARGE syndrome in fetuses with CHD7 truncating mutations correlates with expression during human development. *J Med Genet*. 2006;43(3):211-217. doi:10.1136/jmg.2005.036160

Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, Munnich A, Burgun A, Rance B. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. *J Biomed Inform*. 2018;80:52-63. doi:10.1016/j.jbi.2018.02.019

21. Zhou Y, Zaferiou S. Deformable Models of Ears in-the-Wild for Alignment and Recognition. In: *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. ; 2017:626-633. doi:10.1109/FG.2017.79

22. landmarker.io · The Menpo Project. Accessed March 20, 2022. https://www.menpo.org/landmarkerio/

23. Baken EK, Collyer ML, Kaliontzopoulou A, Adams DC. geomorph v4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution*. 2021;12(12):2355-2363. doi:10.1111/2041-210X.13723

24. Hennocq Q, Bongibault T, Bizière M, Delassus O, Douillet M, Cormier-Daire V, Amiel J, Lyonnet S, Marlin S, Rio M, Picard A, Khonsari RH, Garcelon N. An automatic facial landmarking for children with rare diseases. *American Journal of Medical Genetics Part A*. n/a(n/a). doi:10.1002/ajmg.a.63126

25. Rohlf FJ, Slice D. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. *Systematic Zoology*. 1990;39(1):40-59. doi:10.2307/2992207

26. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Published online September 17, 2020. doi:10.48550/arXiv.1802.03426

27. R Core Team (2020). — European Environment Agency. Accessed January 18, 2023. https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006

28. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

29. Sachs MC. plotROC: A Tool for Plotting ROC Curves. *J Stat Softw*. 2017;79:2. doi:10.18637/jss.v079.c02

30. Protocole type d'examen autopsique foetal ou néonatal. Haute Autorité de Santé. Accessed August 11, 2023. https://www.has-sante.fr/jcms/c_1747470/fr/protocole-type-d-examen-autopsique-foetal-ou-neonatal

31. Blake KD, Prasad C. CHARGE syndrome. *Orphanet J Rare Dis*. 2006;1:34. doi:10.1186/1750-1172-1-34

32. Sondern K, Kreitz K, Hammer K, Möllers M, Oelmeier de Murcia K, Köster HA, Braun J, Klockenbusch W, Schmitz R. 3D Ultrasound Evaluation of the Fetal Outer Ear: Novel Biometry Ratio and Comparison of Different Surface Display Modes. *Fetal Diagn Ther*. 2019;46(3):200-206. doi:10.1159/000497145

33. Biard JM, Payrat S, Clapuyt P, Barrea C, Benoit V, Baldin P, Bernard P, Van Grambezen B, Sznajer Y. Antenatal diagnosis of CHARGE syndrome: Prenatal ultrasound findings and crucial role of fetal dysmorphic signs. About a series of 10 cases and review of literature. *Eur J Med Genet*. 2021;64(4):104189. doi:10.1016/j.ejmg.2021.104189

34. Legendre M, Gonzales M, Goudefroye G, Bilan F, Parisot P, Perez MJ, Bonnière M, Bessières B, Martinovic J, Delezoide AL, Jossic F, Fallet-Bianco C, Bucourt M, Tantau J, Loget P, Locuillet L, Laurent N, Leroy B, Salhi H, Bigi N, Rouleau C, Guimiot F, Quélin C, Bazin A, Alby C, Ichkou A, Gesny R, Kitzis A, Ville Y, Lyonnet S, Razavi F, Gilbert-Dussardier B, Vekemans M, Attié-Bitach T. Antenatal spectrum of CHARGE syndrome in 40 fetuses with CHD7 mutations. *J Med Genet*. 2012;49(11):698-707. doi:10.1136/jmedgenet-2012-100926

35. Busa T, Legendre M, Bauge M, Quarello E, Bretelle F, Bilan F, Sigaudy S, Gilbert-Dussardier B, Philip N. Prenatal findings in children with early postnatal diagnosis of CHARGE syndrome. *Prenat Diagn*. 2016;36(6):561-567. doi:10.1002/pd.4825

36. Millischer AE, Sonigo P, Attie T, Spaggiari E, O'Gorman N, Bessieres B, Kermorvant E, Boddaert N, Salomon LJ, Grevent D. Fetal MRI findings in a retrospective cohort of 26 cases of prenatally diagnosed CHARGE syndrome individuals. *Prenat Diagn*. 2019;39(9):781-791. doi:10.1002/pd.5429

37. Batool S. The Developing Human-Clinically Oriented Embryology by Keith L. Moore. Accessed June 24, 2023. https://www.academia.edu/44502808/The_Developing_Human_Clinically_Oriented_Embryology_by_Keith_L_Moore

38. Chang CH, Chang FM, Yu CH, Liang RI, Ko HC, Chen HY. Fetal ear assessment and prenatal detection of aneuploidy by the quantitative three-dimensional ultrasonography. *Ultrasound Med Biol*. 2000;26(5):743-749. doi:10.1016/s0301-5629(00)00206-4

39. Hatanaka AR, Rolo LC, Mattar R, Araujo Júnior E, Nardozza LMM, Moron AF. Reference intervals for fetal ear length between 19 and 24 weeks of pregnancy on 3-dimensional sonography. *J Ultrasound Med*. 2011;30(9):1185-1190. doi:10.7863/jum.2011.30.9.1185

40. Chitkara U, Lee L, Oehlert JW, Bloch DA, Holbrook RH, El-Sayed YY, Druzin ML. Fetal ear length measurement: a useful predictor of aneuploidy? *Ultrasound Obstet Gynecol*. 2002;19(2):131-135. doi:10.1046/j.0960-7692.2001.00558.x

41. Merz E, Welter C. 2D and 3D Ultrasound in the evaluation of normal and abnormal fetal anatomy in the second and third trimesters in a level III center. *Ultraschall Med*. 2005;26(1):9-16. doi:10.1055/s-2004-813947

42. Shih JC, Shyu MK, Lee CN, Wu CH, Lin GJ, Hsieh FJ. Antenatal depiction of the fetal ear with three-dimensional ultrasonography. *Obstet Gynecol*. 1998;91(4):500-505. doi:10.1016/s0029-7844(98)00021-0

5.3 Multi-syndromes classification

5.3.1 Training set

The training set contained 3330 photographs, corresponding to 1086 patients; 47 % of patients were controls and 53% presented with a syndrome. In the control group, 54% of patients were females and the mean age was 7.0 +/- 4.6 years. In the patient group, the syndromes were: Crouzon-Pfeiffer (N = 348, 60%), Kabuki (N = 167, 29%), Treacher Collins (N = 149, 26%), Silver Russell (N = 131, 23%), Apert (N = 88, 15%), Muenke (N = 72, 13%), CHARGE (N = 69, 12%), MFDGA (N = 60, 10%), Saethre Chotzen (N = 47, 8%) and NAFD (N = 14, 2%) (Table 3).

		Controls	Syndromic patients
N			
	Consultations	520 (31%)	1145 (69%)
	Photographs	1040 (31%)	2290 (69%)
	Patients	510 (47%)	576 (53%)
Gender			
	Female	279 (54%)	513 (45%)
	Male	241 (46%)	632 (55%)
Age (years)			
	Mean +/- SD	7.0 +/- 4.6	5.9 +/- 8.4
	Median	7.1	2.9
	Min	0.2	0
	Max	22.1	80.0
Ethnicity			
	African	28 (5%)	101 (9%)
	Asian	9 (2%)	28 (2%)
	Caucasian	483 (93%)	1016 (89%)
Syndrome			
	Apert		88 (15%)
	CHARGE		69 (12%)
	Crouzon Pfeiffer		348 (60%)
	Kabuki		167 (29%)

MFDGA	60 (10%)
Muenke	72 (13%)
NAFD	14 (2%)
Saethre Chotzen	47 (8%)
Silver Russell	131 (23%)
Treacher Collins	149 (26%)

Table 3. Description of the training set population. CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; NAFD = Nager type Acro-Facial Dysostosis; SD = Standard Deviation.

5.3.2 Validation set

The training set contained 216 photographs, corresponding to 108 patients; 18 % of patients were controls and 82% presented with a syndrome. In the control group, 58% of patients were females and the mean age was 6.1 +/- 4.5 years. In the patient group, the syndromes were: Crouzon-Pfeiffer (N = 27, 30%), Apert (N = 13, 15%), Kabuki (N = 11, 12%), Muenke (N = 7, 8%), Saethre Chotzen (N = 7, 8%), CHARGE (N = 6, 7%), Silver Russell (N = 6, 7%), MFDGA (N = 5, 6%), Treacher Collins (N = 5, 6%) and NAFD (N = 2, 2%) (Table 4).

		Controls	**Syndromic patients**
N			
	Consultations	19 (18%)	89 (82%)
	Photographs	38 (18%)	178 (82%)
	Patients	19 (18%)	89 (82%)
Gender			
	Female	11 (58%)	51 (57%)
	Male	8 (42%)	38 (43%)
Age (years)			
	Mean +/- SD	6.1 +/- 4.5	5.3 +/- 5.9

	Median	6.6	3.6
	Min	0.1	0.0
	Max	15.8	33.5
Ethnicity			
	African	1 (5%)	7 (8%)
	Asian	1 (5%)	0 (0%)
	Caucasian	17 (90%)	82 (92%)
Syndrome			
	Apert		13 (15%)
	CHARGE		6 (7%)
	Crouzon Pfeiffer		27 (30%)
	Kabuki		11 (12%)
	MFDGA		5 (6%)
	Muenke		7 (8%)
	NAFD		2 (2%)
	Saethre Chotzen		7 (8%)
	Silver Russell		6 (7%)
	Treacher Collins		5 (6%)

Table 4. Description of the validation set population. CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; NAFD = Nager type Acro-Facial Dysostosis; SD = Standard Deviation.

5.3.3 Results of the multi-syndrome classification model

The Procrustes representations showed the morphological characteristics classically described in the literature for these 10 syndromes (Figure 10). In addition to the facial features on frontal views, the profile views and external ears were important for the phenotypic description of the syndrome, in particular the brachycephaly found in the Apert, Crouzon-Pfeiffer, Muenke and Saethre Chotzen syndromes, the large, prominent ears of Kabuki syndrome, the triangular concha and earlobe hypoplasia of CHARGE syndrome, the

retrognathia of Treacher Collins, NAFD and MFDGA syndromes, and the small mandible of Silver Russell syndrome.

Our model was able to correctly classify 71.3% (61.8 - 79.6) of patients (top-1 accuracy). Top-3 accuracy was 93.5%. Performance varied according to syndromes, with 96.1% accuracy for the controls (AUC = 1,000) or 90.7% for Apert (AUC = 0.992) and CHARGE (AUC = 0.912) groups. However, the accuracy was 50% in the NAFD group (AUC = 0.450), corresponding to the smallest number of patients. The ROC curves and AUCs for the training and validation sets were comparable (Figure 11) (Tables 5, 6).

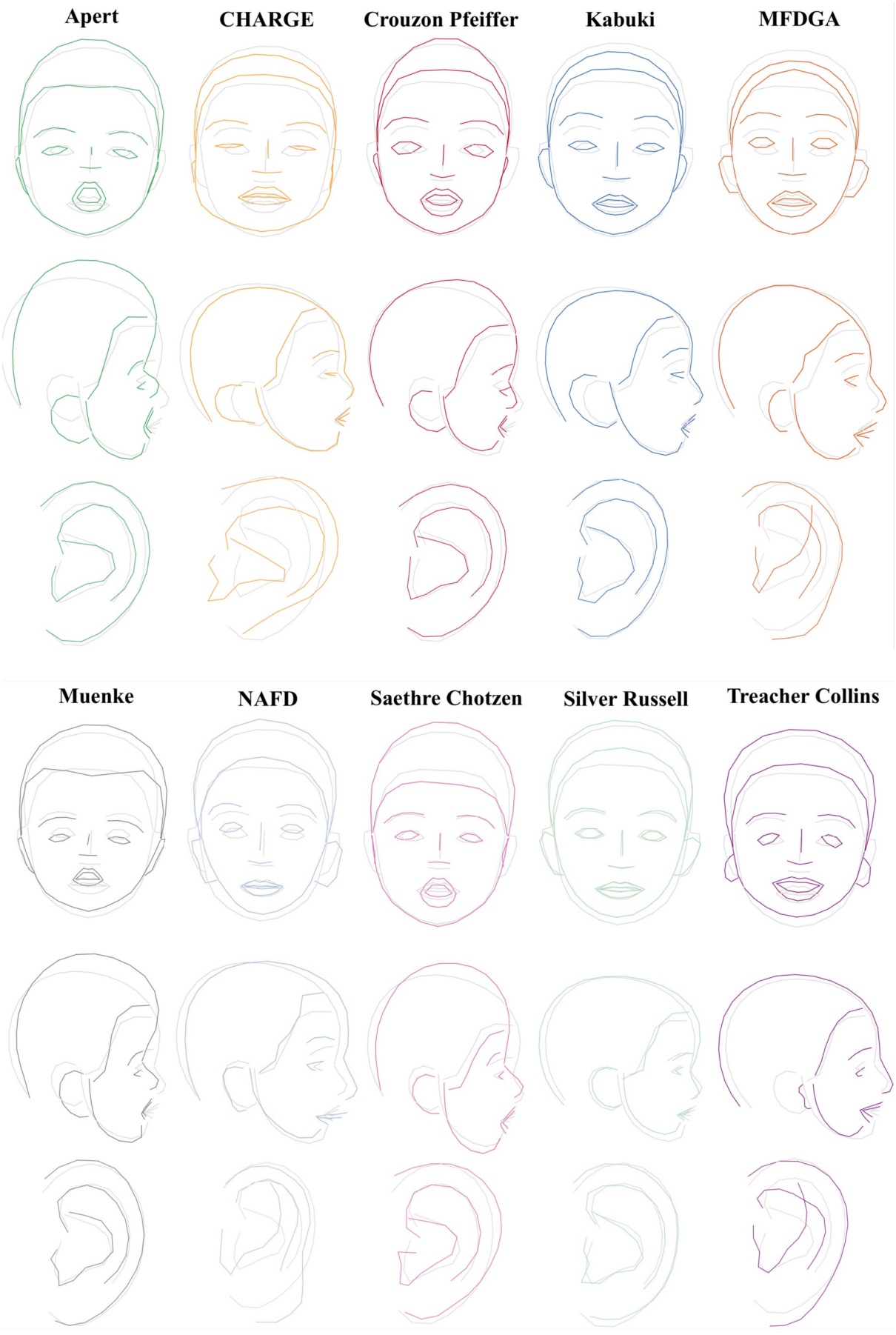

Apert
CHARGE
Crouzon Pfeiffer
Kabuki
MFDGA
Muenke
NAFD
Saethre Chotzen
Silver Russell
Treacher Collins

Figure 10. Average shapes for 10 syndromes vs controls after Procrustes superimposition of frontal views, lateral views, and external ears, before 5 years of age. Grey = controls. CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; NAFD = Nager type Acro-Facial Dysostosis.

Group	AUC (training)	AUC (validation)	Sensitivity	Specifity	Accuracy
Overall					0.713 (0.618 - 0.796) *
Control	1.000	1.000	1.000	0.921	0.961
Apert	0.921	0.992	0.846	0.968	0.907
CHARGE	0.912	0.986	0.833	0.980	0.907
Crouzon Pfeiffer	0.994	0.970	0.667	0.975	0.821
Kabuki	0.941	0.983	0.727	0.959	0.843
MFDGA	0.867	0.888	0.400	0.990	0.695
Muenke	0.985	0.804	0.286	0.950	0.618
NAFD	0.586	0.450	0.000	1.000	0.500
Saethre Chotzen	0.865	0.778	0.429	0.990	0.709
Silver Russell	0.951	0.986	0.833	0.971	0.902
Treacher Collins	0.974	0.972	0.800	0.971	0.885

Table 5. Classification performances in the validation set. AUC = Area Under the Curve; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; NAFD = Nager type Acro-Facial Dysostosis.

		Group									
		Controls	**AS**	**CHARGE**	**CPS**	**KS**	**MFDGA**	**MS**	**NAFD**	**SCS**	**SRS**	**TCS**
Prediction	**Controls**	**19**	0	0	3	2	0	2	0	0	0	0
	AS	0	**11**	0	0	0	0	2	0	1	0	0
	CHARGE	0	1	**5**	1	0	0	0	0	0	0	0
	CPS	0	0	0	**18**	1	0	1	0	0	0	0
	KS	0	0	0	1	**8**	1	0	0	0	1	1
	MFDGA	0	0	0	0	0	**2**	0	1	0	0	0
	MS	0	1	0	2	0	0	**2**	0	2	0	0
	NAFD	0	0	0	0	0	0	0	**0**	0	0	0

SCS	0	0	0	0	0	1	0	0	**3**	0	0
SRS	0	0	1	2	0	0	0	0	0	**5**	0
TCS	0	0	0	0	0	1	0	1	1	0	**4**

Table 6. Confusion matrix in the validation set. AS = Apert Syndrome; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; CPS = Crouzon Pfeiffer Syndrome; KS = Kabuki Syndrome; MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; MS = Muenke Syndrome; NAFD = Nager type Acro-Facial Dysostosis; SCS = Saethre Chotzen Syndrome; Silver Russell Syndrome; Treacher Collins Syndrome.

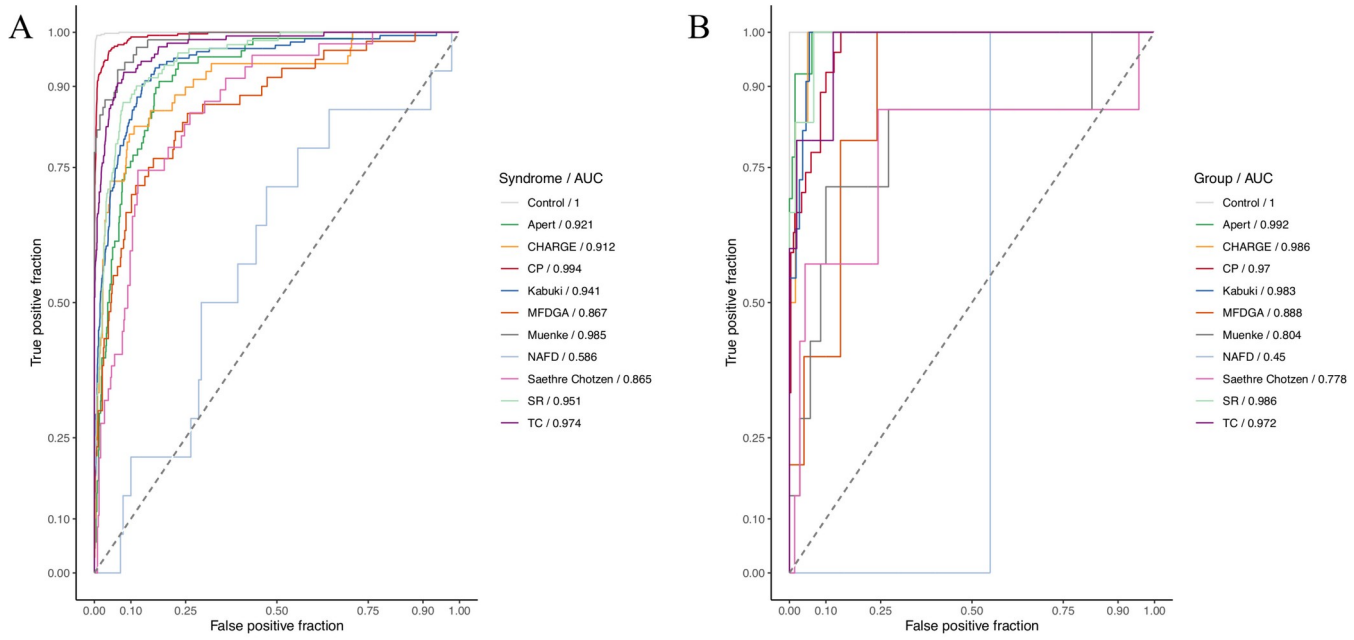


Figure 11. A. Empirical ROC curves (training set) for the different groups with AUC. B. ROC curves (validation set) for the different groups with AUC. AUC = Area Under the Curve; AS = Apert Syndrome; CHARGE = Coloboma, Heart defect, Atresia choanae, Retarded growth and development, Genital hypoplasia, Ear anomalies/deafness; CP =

Crouzon Pfeiffer Syndrome; MFDGA = Mandibulo-Facial Dysostosis Guion Almeida type; NAFD = Nager type Acro-Facial Dysostosis; Silver Russell Syndrome; Treacher Collins Syndrome.

On Figure 12, we have reported two examples of the use of our model on two patients from the validation set. The results are given in the form of a probability for each group. The first proband has a 99% probability of CHARGE syndrome and a variant in the *CHD7* gene should be investigated. The second proband has an 88% probability of Apert syndrome and a variant in the *FGFR2* gene should be investigated.

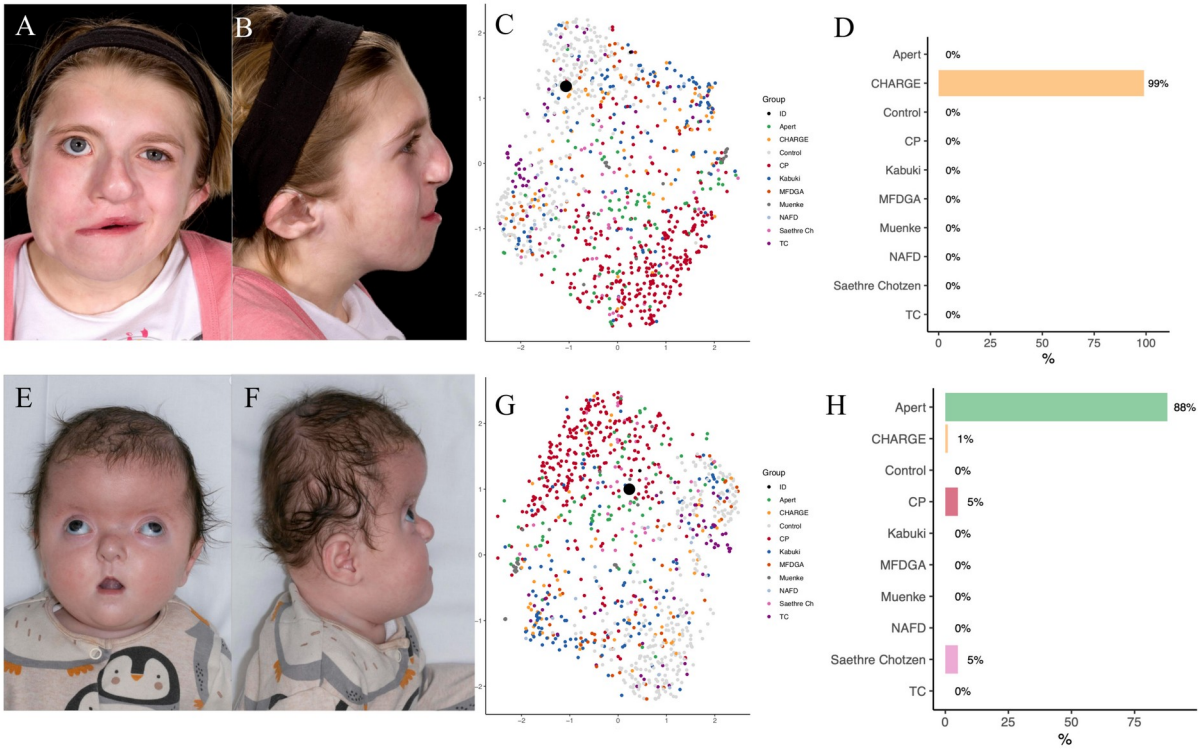


Figure 12. Example of classification using our model for two probands of the validation set. A and B. Frontal and profile faces of proband 1. C. UMAP representation of the training data according to the different groups, with positioning of proband 1. D. Histogram of

predictions by our model. E and F. Frontal and profile faces of proband 2. G. UMAP representation of the training data according to the different groups, with positioning of proband 2. H. Histogram of predictions by our model.

6 Results for the second main objective: phenotype – genotype correlations

We will present two use cases of phenotype-genotype correlations using our model:

- Diagnosis and phenotype-genotype correlations in the group of syndromic craniosynostoses
- Diagnosis and phenotype-genotype correlations in the Kabuki syndrome group

6.1 Facial AI-based diagnosis and phenotype – genotype correlations in syndromic craniosynostoses

Craniosynostosis is a group of diseases characterized by premature fusion of one or more cranial sutures (126). Around 8% of craniosynostoses are genetic or syndromic in origin (127). Many syndromes combine craniosynostosis, facial abnormalities and various extra-cranial malformations. Genetic variations in the *FGFR2* (Fibroblast Growth Factor Receptor 2) and *FGFR3* (Fibroblast Growth Factor Receptor 3) genes lead to gains in function and accelerated differentiation of osteoblasts (128,129), eventually causing the premature closing of craniofacial sutures. These genetic variations are the most frequent causes of syndromic craniosynostoses (130). Among less frequent genes implicated in syndromic forms, *TWIST1* mutations cause Saethre Chotzen syndrome (SCS).

The *FGFR2* and *FGFR3* genes code for a tyrosine kinase receptor and are composed of: (1) an extracellular portion containing Immunoglobulin-like domains (IgI, IgII, IgIIIa and IgIIIc), (2) a transmembrane region, and (3) an intracellular tyrosine kinase region (TK1 and TK2) (131–136). Phenotype/genotype correlations are not well understood in *FGFR*-related

craniosynostoses despite several reports focused on specific anatomical regions, such as Morice et al (137), who have reported that genetic variations on the different portions of the *FGFR2* and *FGFR3* receptor were associated with phenotypic variability in the mandible, and Bouaoud et al (138) who have associated the thickness of the supra-orbital bar with specific *FGFR* mutations.

Crouzon syndrome (CS) is an autosomal dominant condition caused by activating mutations in *FGFR2* or exceptionally *FGFR3* genes (139,140). The facial presentation of CS is characterized by ocular proptosis, hypertelorism, brachycephaly, and midface retrusion, but familial cases with high variability in penetrance are reported (141).

Pfeiffer syndrome (PS) is also autosomal dominant, caused by activating *FGFR1* or *FGFR2* mutations. In addition to a facial phenotype similar to CS, PS affects feet and hands, with large thumbs or toes, brachydactyly, and syndactyly (142).

Apert syndrome (AS) is another autosomal dominant *FGFR2*-related craniosynostosis mostly affecting the coronal suture (brachycephaly) and associated with midface hypoplasia, hypertelorism, a small mandible, and syndactyly of the hands and feet (126). (142).

Muenke syndrome (MS) is an autosomal dominant disorder affecting the *FGFR3* gene (p.Pro250Arg), with involvement of the coronal suture, hypoplasia of the midface, hypertelorism and macrocephaly (142).

Saethre-Chotzen syndrome (SCS) is caused by an autosomal dominant mutation in the *TWIST1* gene, resulting in fusions of the coronal, lambdoid and/or metopic sutures, facial asymmetry, a variable degree of syndactyly and hearing loss (143,144).

The aims of this study were (1) to train a tool using artificial intelligence (AI) methods on facial frontal, lateral and external ear pictures to support diagnosis for syndromic craniosynostoses vs controls and (2) to screen for genotype / phenotype correlations in AS, CS and PS.

Ranging between 1979 and 2023, we included 2228 frontal and lateral facial photographs, corresponding to 541 patients. The control group comprised 1122 photographs, corresponding to 304 patients. The AS, CS, MS, PS, and SCS groups comprised 146, 604, 52, 236, and 68 photographs respectively, corresponding to 44, 98, 16, 51, and 28 patients. The validation set was composed of 168 frontal and lateral facial photographs corresponding to 84 patients from Great Ormond Street Hospital (GOSH, London). The control, AS, CS, MS, PS, and SCS groups comprised 26, 40, 62, 12, 10, and 18 photographs respectively, corresponding to 13, 20, 31, 6, 5, and 9 patients.

We found the facial characteristics classically found in the literature for these syndromes. There were very few phenotypic differences between the CS and PS groups.

Each of the 5 syndromes was significantly distinguishable from control patients in the validation set, with high AUCs going from 0.983 [0.954 – 1.000] ($p < 0.001$) for CS to 1.000 [1.000 – 1.000] ($p < 0.001$) for MS and PS. AS was significantly distinguishable from CS (0.955 [0.893 – 1.000], $p < 0.001$) and from SCS (0.906 [0.799 – 1.000], $p < 0.001$), but not from MS (0.658 [0.394 – 0.923], $p = 0.546$) and PS (0.700 [0.470 – 0.930], $p = 0.348$). Interestingly, the model was not able to distinguish CS and PS, with an AUC of 0.548 [0.270 – 0.827] ($p = 0.672$). MS was not discernable from PS and SCS with AUCs of 0.567 [0.157 – 0.977] ($p = 0.841$) and 0.667 [0.370 – 0.963] ($p = 0.642$) (Figure 13). Finally, the model was able to distinguish PS from SCS with an AUC of 0.844 [0.606 – 1.000] ($p = 0.021$) (Table 7).

	Controls	AS	CS	MS	PS
Controls					
AS	0.992 [0.974 - 1.000] *				
CS	0.983 [0.954 - 1.000] *	0.955 [0.893 - 1.000] *			
MS	1.000 [1.000 - 1.000] *	0.658 [0.394 - 0.923]	0.785 [0.522 - 1.000] *		
PS	1.000 [1.000 - 1.000] *	0.700 [0.470 - 0.930]	0.548 [0.270 - 0.827]	0.567 [0.157 - 0.977]	
SCS	0.940 [0.834 - 1.000] *	0.906 [0.799 - 1.000] *	0.871 [0.704 - 1.000] *	0.667 [0.370 - 0.963]	0.844 [0.606 - 1.000] *

Table 7. AUC for each binary classification on the validation set. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre-Chotzen Syndrome. * = statistically significant ($p < 0.05$).

For further analyses, we grouped CS and PS patients together under the name Crouzon-Pfeiffer syndrome (CPS), given the absence of significant phenotypic differences in binary classification. The number of iterations retained for this multi-class model was 248 and 70.2% [0.593 - 0.797] of patients in the validation set were correctly diagnosed. The confusion matrix revealed 13 correctly predicted controls, 13 AS, 27 CPS, 3 MS, and 3 SCS correctly diagnosed (Table 8).

		Reference				
		Control	**AS**	**CPS**	**MS**	**SCS**
Prediction	**Control**	**13**	2	4	1	1
	AS	0	**13**	2	2	1
	CPS	0	3	**27**	0	0
	MS	0	2	3	**3**	4
	SCS	0	0	0	0	**3**

Table 8. Confusion matrix in a multi-syndrome design. AS = Apert Syndrome, CPS = Crouzon-Pfeiffer Syndrome, MS = Muenke Syndrome, SCS = Saethre-Chotzen Syndrome. Bold values: True Positives (TP).

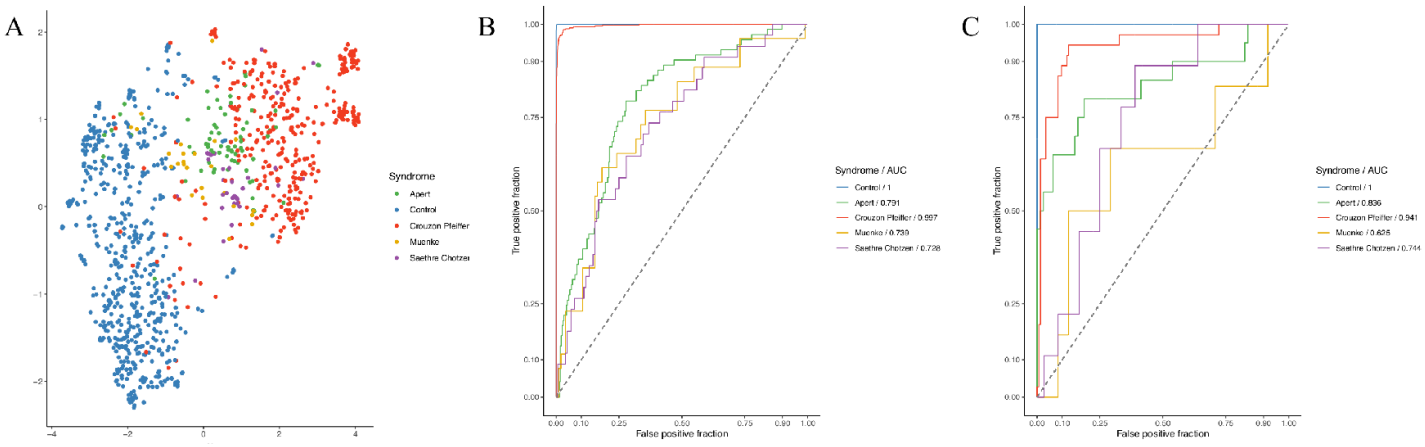


Figure 13. Multi-syndrome classification in syndromic craniosynostoses. A. UMAP representation of the training data according to the different groups. B. Empirical ROC curves (training set) for each group with AUC. C. ROC curves (validation set) for each group, with AUC. AUC = Area Under the Curve.

In AS, our model was unable to correctly classify patients between the two genotypes (*FGFR2* p.Pro253Arg vs. *FGFR2* p.Ser252Trp) (AUC = 0.506 [0.215 - 0.797], p = 0.874).

In CPS, genotypes linked to the IgI, IgII, IgIIIc and TK domains of *FGFR2* and variations in *FGFR3* were associated with more severe facial phenotypes than variations in a splice site of *FGFR2*. In the latter group, brachycephaly, ocular phenotype and reduced height of the mid-face seemed to be less severe.

A binary classification between the 'IgIIIc' and 'Splicing domain' groups was significant, with an AUC = 0.786 [0.554 - 1.000] ($p < 0.023$) on the validation set. In addition, the controls were indistinguishable from the 'Splicing domain' group (AUC = 0.577 [0.332 - 0.822], p = 0.657).

To conclude, we were able to detect more than 70% of patients, with excellent performance particularly in the CPS group (84.4% and AUC = 0.941). Finally, we were able to describe the significant phenotypic variability within the CPS group.

AI-based diagnosis and phenotype – genotype correlations in syndromic craniosynostoses

Submitted to The American Journal of Human Genetics, october 2023

Quentin **Hennocq**,[1,2,3*] Giovana **Paternoster**,[1,4] Corinne **Collet**,[5] Jeanne **Amiel**,[1,6] Thomas **Bongibault**,[1,3] Thomas **Bouygues**,[1,3] Valérie **Cormier-Daire**,[1,6] Maxime **Douillet**,[1] David J **Dunaway**,[7] Nu Owase **Jeelani**,[7] Lara S **van de Lande**,[7,8] Stanislas **Lyonnet**,[1,6] Juling **Ong**,[7] Arnaud **Picard**,[2] Alexander J **Rickart**,[7] Marlène **Rio**,[1,6] Silvia **Schievano**,[7] Eric **Arnaud**,[4,9] Nicolas **Garcelon**,[1**] Roman H. **Khonsari** [1,2,3,4**]

[1]Imagine Institute, INSERM UMR1163, 75015 Paris, France
[2]Département de chirurgie maxillo-faciale et chirurgie plastique, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou ; Faculté de Médecine, Université de Paris Cité ; 75015 Paris, France
[3]Laboratoire 'Forme et Croissance du Crâne', Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris ; Faculté de Médecine, Université Paris Cité, Paris, France.
[4]Département de neurochirurgie, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou ; Faculté de Médecine, Université de Paris Cité ; 75015 Paris, France
[5]Département de génétique moléculaire, Hôpital Robert Debré, Université de Paris Cité ; Paris, France.
[6]Service de médecine génomique des maladies rares, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Faculté de Médecine, Université de Paris Cité ; 75015 Paris, France [7]UCL Great Ormond Street Institute of Child Health and Craniofacial Unit, Great Ormond Street Hospital for Children, London, UK.
[8]Department of Oral and Maxillofacial Surgery, Erasmus Medical Center, Rotterdam, The Netherlands.
[9]Clinique Marcel Sembat (Ramsay), Boulogne, France.

** These authors contributed equally to this work.

Abstract. Apert (AS), Crouzon (CS), Muenke (MS), Pfeiffer (PS), and Saethre Chotzen (SCS) are among the most frequently diagnosed syndromic craniosynostoses. The aim of this study was (1) to train an innovative model using AI-based methods on 2D facial frontal, lateral, and external ear photographs to assist diagnosis for syndromic craniosynostoses vs controls, and (2) to screen for genotype / phenotype correlations in AS, CS, and PS. We included retrospectively and prospectively, from 1979 to 2023, all frontal and lateral pictures of patients genetically diagnosed with AS, CS, MS, PS and SCS syndromes. After a deep learning-based preprocessing, we extracted geometric and textural features and used XGboost (eXtreme Gradient Boosting) to classify patients. The model was tested on an independent international validation set of genetically confirmed AS, CS, PS, MS, SCS and non-syndromic controls. Between 1979 and 2023, we included 2228 frontal and lateral facial photographs corresponding to 541 patients. 70.2% [0.593 - 0.797] ($p < 0.001$) of patients in the validation set were correctly diagnosed. The best performances were obtained by the control group (balanced accuracy = 0.943 and AUC = 1.000), the Crouzon Pfeiffer syndrome group (balanced accuracy = 0.844 and AUC = 0.941) and the AS group (balanced accuracy = 0.786 and AUC = 0.836). We found no statistically significant facial phenotype difference between CS and PS (AUC = 0.548 [0.270 – 0.827], $p = 0.672$). Genotypes linked to a splice donor site of *FGFR2* in Crouzon Pfeiffer syndrome (CPS) caused a milder phenotype in CPS.

Keywords: artificial intelligence; machine learning; dysmorphology; syndromic craniosynostosis; Crouzon syndrome; Pfeiffer syndrome; Apert syndrome; *FGFR2*, Saethre Chotzen syndrome; Muenke syndrome

Introduction

Applications of AI (Artificial Intelligence) are increasing in healthcare [1–4]. The field of dysmorphology is currently being transformed by these new methods [5]. Studies comparing human performances to AI are flourishing [6–9], and some of them suggest that digital diagnostic tools outperform human experts [10].

Craniosynostosis is a group of diseases characterized by premature fusion of one or more cranial sutures [11]. Around 8% of craniosynostoses are genetic or syndromic in origin [12]. Many syndromes combine craniosynostosis, facial abnormalities and various extra-cranial malformations. Genetic variations in the *FGFR2* (Fibroblast Growth Factor Receptor 2) and *FGFR3* (Fibroblast Growth Factor Receptor 3) genes lead to gains in function and accelerated differentiation of osteoblasts [13,14], eventually causing the premature closing of craniofacial sutures. These genetic variations are the most frequent causes of syndromic craniosynostoses [15]. Among less frequent genes implicated in syndromic forms, *TWIST1* mutations cause Saethre Chotzen syndrome (SCS).

The *FGFR2* and *FGFR3* genes code for a tyrosine kinase receptor and are composed of: (1) an extracellular portion containing Immunoglobulin-like domains (IgI, IgII, IgIIIa and IgIIIc), (2) a transmembrane region, and (3) an intracellular tyrosine kinase region (TK1 and TK2) [16–21]. Phenotype/genotype correlations are not well understood in *FGFR*-related craniosynostoses despite several reports focused on specific anatomical regions, such as Morice et al [22], who have reported that genetic variations on the different portions of the *FGFR2* and *FGFR3* receptor were associated with phenotypic variability in the mandible, and Bouaoud et al [23] who have associated the thickness of the supra-orbital bar with specific *FGFR* mutations.

Crouzon syndrome (CS) is an autosomal dominant condition caused by activating mutations in *FGFR2* or exceptionally *FGFR3* genes [24,25]. The facial presentation of CS is characterized by ocular proptosis, hypertelorism, brachycephaly, and midface retrusion, but familial cases with high variability in penetrance are reported [26].

Pfeiffer syndrome (PS) is also autosomal dominant, caused by activating *FGFR1* or *FGFR2* mutations. In addition to a facial phenotype similar to CS, PS affects feet and hands, with large thumbs or toes, brachydactyly and syndactyly [27]. Apert syndrome (AS) is another autosomal dominant *FGFR2*-related craniosynostosis mostly affecting the coronal suture (brachycephaly) and associated with midface hypoplasia, hypertelorism, a small mandible and syndactyly of the hands and feet [11]. [27]. Muenke syndrome (MS) is an autosomal dominant disorder affecting the *FGFR3* gene (p.Pro250Arg), with involvement of the coronal suture, hypoplasia of the midface, hypertelorism and macrocephaly [27]. Finally, Saethre-Chotzen syndrome (SCS) is caused by an autosomal dominant mutation in the *TWIST1* gene, resulting in fusions of the coronal, lambdoid and/or metopic sutures, facial asymmetry, a variable degree of syndactyly and hearing loss [28,29].

The aims of this study were (1) to train a tool using artificial intelligence (AI) methods on facial frontal, lateral and external ear pictures to support diagnosis for syndromic craniosynostoses vs controls and (2) to screen for genotype / phenotype correlations in AS, CS and PS.

Material and Methods. The study was approved by the CESREES (Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé, №4570023bis) and by the CNIL (Commission Nationale Informatique et Libertés, №MLD/MFI/AR221900). Informed and written consents were obtained from the legal representatives of each child or from the patients themselves if they were of age.

Training set

We included pictures from the photographic database of the Maxillofacial surgeryand Plastic surgery and Neurosurgery departments of *Hôpital Necker – Enfants Malades* (Assistance Publique – Hôpitaux de Paris), Paris, France. This database contains 594,000 photographs from 22,000 patients followed in these departments since 1979. All photographs from 1995 were taken by a professional medical photographer using a Nikon D7000 device in standardized positions. The photographs between 1979 and 1995 were not digitized and were therefore scanned using an Epson Perfection V850 Pro scanner.

We included retrospectively from 1979 to 2019 and prospectively from 2019 to 2023, all frontal and lateral pictures of patients diagnosed with CS, AS, PS, MS and SCS syndromes (Figure 1). The photographs were not calibrated. All patients had genetic confirmation of their syndrome. PS was diagnosed, in addition to the presence of a genetic variation in *FGFR1* or *FGFR2*, by clinical involvement of the feet and/or hands. We excluded patients with any history of cranial or facial surgery. Multiple photographs per patient corresponded to different ages of follow-up. Duplicate photographs were excluded.

Non-syndromic children were selected among patients admitted for lacerations, trauma, infection and various skin lesions, without any record of chronic conditions. More precisely, follow-up for any type of chronic disease was considered as an exclusion criterion. The reports were retrieved using the local data warehouse Dr Warehouse [30].

Validation set

Similar inclusion and exclusion criteria as the training set were used to build the validation set, from the database of the Craniofacial Unit of Great Ormond Street Hospital, London (UK). All patients had genetic confirmation of their syndrome. None of the patients in the validation set was present twice, and none was from the training set. For the control group, we selected photographs from our local database, without any redundancy with the training set, using similar inclusion criteria. We extracted data on age at the time of the photograph and gender. Informed and written consents were obtained from the legal representatives of each child or from the patient himself if he was of age.

Landmarking

We used three different templates based on 105 landmarks for the frontal, 73 for the lateral views and 41 for the external ear pictures. We developed an automatic annotation model for each template following a pipeline including: (1) detection of the Region Of Interest (ROI) and (2) automatic placement of the landmarks.

For ROI detection, a Faster RCNN (Faster Region-based Convolutional Neural Network) model was trained after data augmentation (images and their +10° and -10° rotations), with a learning rate of 0.001, a batch size of 4, a gamma of 0.05 and 2000 iterations. Our Faster RCNN model was optimized and split into two stages.

(1) *ROI detection:* Faster RNN trained on 15633 images, after data augmentation (images and their +10° and -10° rotations). There were 6186 frontal images (2062 x 3) and 9447 right and left profile images (3159 x 3). The batch size was set to 2, the learning rate was set to 0.0025 and the maximum number of iterations was 2800.

(2) *Determination of profile laterality:* pre-trained ResNet50 network [31] using the Pytorch library [32]. The training images included 1,570 left profiles and 1,579 right profiles. The batch size was set at 16, an Adam optimizer [33] was used with a learning rate of 0.001, a step of 7, a gamma of 0.1, trained over 25 epochs.

For the automatic placement of landmarks, we used a patch-based AAM (Active Appearance Model) using the *menpo* library on Python 3.7 [34], as shown being the most accurate method in [35]. We used a two-scale landmarking: the model for frontal pictures was trained on 904 manually annotated photographs, with a first stage of dimensioning (diagonal = 150), a patch shape of [(15, 15), (23, 23)] and 50 iterations, then a second stage without resizing, with a patch shape of [(20, 20), (30, 30)] and 10 new iterations. The model for profile pictures was trained on 1,439 manually annotated photographs, with a first stage of dimensioning (diagonal = 150), a patch shape of [(15, 15), (23, 23)] and 25 iterations, then a second stage without resizing, with a patch shape of [(15, 15), (23, 23)] and 5 new iterations. The model for ears was trained on 1221 manually annotated photographs, with a first stage of dimensioning (diagonal = 100), a patch shape of [(15, 15), (23, 23)] and 50 iterations, then a second stage without resizing, with a patch shape of [(20, 20), (30, 30)] and 20 new iterations. All three models used the Lucas Kanade optimizer [36].

Each automatically annotated photograph was checked by two authors blinded for the diagnosis, QH and MD, and landmarks were manually re-positioned when necessary, using *landmarker.io* [37]. The ICC (Intraclass Correlation Coefficient) was computed between the raters. ICC values greater than 0.9 corresponded to excellent reliability of the manual annotation [38].

Geometric morphometrics

We performed Generalized Procrustes Analysis (GPA) [39] on all landmark clouds using the *geomorph* package on R [40].

Since the data were uncalibrated photographs, ROI sizes were not available: shape parameters only were assessed and not centroid sizes. Procrustes coordinates were then processed using Principal Component Analysis (PCA) for dimension reduction. We retained the principal components explaining 99% of the total variance in cumulative sum. The last 1% was considered as negligible information.

Texture extraction

We partitioned the frontal and profile pictures into key areas, and applied textural feature extraction methods to each of these zones, allowing to check the results and determine which zone has contributed most to the diagnosis.

We defined 14 key areas, 11 from the frontal (right eye, left eye, right eyebrow, left eyebrow, glabella, forehead, nasal tip, philtrum, right cheek, left cheek, chin) and 3 from the lateral (pre-auricular region, eye, malar relief) pictures. These areas could potentially contribute to the diagnosis and characterization of the facial phenotype. Each zone was extracted automatically using the previously placed landmarks.

We used the CLAHE (Contrast Limited Adaptative Histogram Equalization) algorithm for histogram equalization, a previously reported before the use of feature extractors [41,42]. CLAHE enhanced contrast by evenly dispersing gray values [43], by reducing the influence of illumination when the picture was taken, and of skin color. Kiflie et al recommended CLAHE as a first choice equalization method [44].

Gray-level co-occurrence matrix (GLCM) methods, as proposed by Haralick [45], are based on the estimation of the second-order joint conditional probability density functions. These matrices characterize the spatial relationships between pixels. GLCM constitutes one of the most commonly used methods in texture analysis [46,47], for instance in radiomics to analyze CT or MRI images [48–50], or for skin texture assessment [51]. In GLCM, the co-occurrence matrix contains information on entropy, homogeneity, contrast, energy and correlation between pixels. There are 28 GLCM features, taking into account the average and range for each item of information and for each zone, representing 28 x 14 = 394 textural features for each patient.

Stratification using metadata

The geometric features and the geometric principal components were combined for further analysis. To consider associated metadata (age and gender) and the fact that we included more than one photograph per patient (that is the non-independence of the data), a mixed model was designed for each feature. The variables to be explained were the features (geometric and textural), with age, gender and ethnicity considered as explanatory variables. Age, gender and ethnicity are significant factors in dysmorphology because they influence the diagnosis, and must therefore be taken into account [52,53].

A random effect on age and individuals was introduced. The equation of the mixed model was:

$$\boldsymbol{Features_{i,j}} \sim \alpha + age.\beta_1 + gender.\beta_2 + ethnicity.\beta_3 + age.\beta_{1,i} + \varepsilon_{i,j}$$

where $age.\beta_{1,i}$ corresponded to a random slope for age per individual, and $\varepsilon_{i,j}$ was a random error term. We did not use an interaction term between age and gender and age and ethnicity as it did not increase the likelihood of the model.

The residuals of each feature were computed to consider potential biases linked to the metadata:

$$\boldsymbol{\varepsilon_{i,j}} = \boldsymbol{Features_{i,j}} - \alpha + age.\beta_1 + gender.\beta_2 + ethnicity.\beta_3 + age.\beta_{1,i}$$

Classification model

The inputs to the model were the residuals from the linear models described above, for each geometric or textural feature. We used XGBoost (eXtreme Gradient Boosting), a supervised machine learning classifier, for all the analyses [54]. We chose a tree-based booster, and the loss function to be minimized was a logistic regression in the case of binary classification, or a softmax function for multi-class classification. We set several hyperparameters to improve the performance and effect of the machine learning model: learning rate = 0.3, gamma = 0, maximum tree depth = 6. The model with the lowest error rate in case of binary classification or multiclass error rate in case of multi-class classification was chosen for analysis. We separated the dataset into a training set and a testing set, and a 5-fold cross-validation was used to define the ideal number of iterations to avoid overfitting.

The chosen model with the ideal number of iterations was then used on the independent validation set to test performances, by plotting accuracy and AUC. The ROC (Receiver Operating Characteristics) curves were plotted in R using the *plotROC* package [55].

Uniform Manifold Approximation and Projection (UMAP) representations

The residuals $\varepsilon_{i,j}$ were represented using UMAP for visual clustering, a nonlinear dimension reduction technique [56]. We retained the residuals associated with features with a classification gain (in their cumulative sum) > 0.75 in the importance matrix associated with the XGboost model. A k (local neighborhood size) value of 15 was used. A cosine metric was introduced to compute distances in high dimensional spaces: the effective minimal distance between embedded points was 10^{-6}. The three conditions of UMAP, namely uniform distribution, local constancy of the Riemannian metric and local connectivity were verified. UMAP analyses were performed using the package *umap* on R [57].

Classification designs

We produced several models to meet different objectives (Figure 1).

1) Syndrome diagnosis support: each group (controls, AS, CS, MS, PS, SCS) was tested against all others in binary classifications; a multi-class model was then designed.
2) Genotype-phenotype correlations: the two genotypes associated with AS (*FGFR2* p.Pro253Arg vs. *FGFR2* p.Ser252Trp) were tested in binary classifications with the control group; the 5 groups of genotypes associated with CS and PS (*FGFR2* IgI-IIIa, IgIIIc, Splicing domain, TK, *FGFR3* and *FGFR1*) were tested in binary classifications.

Results

1. Training set

A total of 2228 frontal and lateral facial photographs, corresponding to 541 patients, were included. The mean age was 6.2 +/- 6.5 and ranged from 0 to 44 years; 44% were girls. Ethnicity was 92% Caucasian, 7% African or Caribbean, and 1% Asian.

The control group comprised 1122 photographs, corresponding to 304 patients with a mean age of 6.1 +/- 4.8 years.

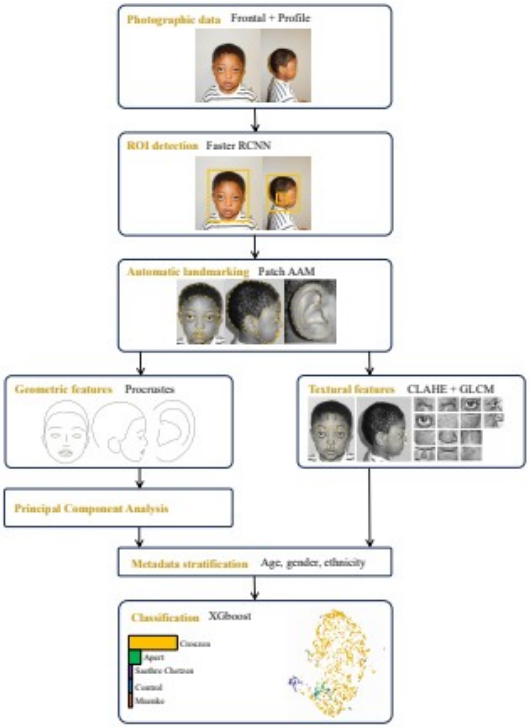


Figure 1. Analysis pipeline, from the initial photograph to diagnostic probability. Faster RCNN = Faster Region-based Convolutional Neural Network; CLAHE = Contrast Limited Adaptative Histogram Equalization; GLCM = Gray-level Co-occurrence Matrix; XGboost = eXtreme Gradient Boosting.

The AS, CS, MS, PS and SCS groups comprised 146, 604, 52, 236 and 68 photographs respectively, corresponding to 44, 98, 16, 51 and 28 patients. Examples from the database are shown Figure 2.

Figure 2. Examples from the training set for each group. A, G: Control, male (dental abscess). B, H: Apert syndrome, male. C, I: Crouzon syndrome, male. D, J: Muenke syndrome, male. E,K: Pfeiffer syndrome, female. F, L: Saethre Chotzen syndrome, female.

Within the CS group, 99% were affected by the *FGFR2* gene mutation and 1% *FGFR3*. For PS, this was 8% on the *FGFR1* gene and 92% on the *FGFR2* gene. 24% of the genetic variations were located on the extracellular portion of *FGFR2* (IgI, IgII, IgIIIa) with the most frequent variation being p.Cys278Phe (7%); 57% were located on the trans-membrane portion of *FGFR2* (IgIIIc) with the most frequent variation being p.Cys342Tyr (10%); 7% affected a splice donor site, with the most common variant being p.Ala344Ala (4%); 6% of the variations were located on the *FGFR2* intra-cellular domain (TK), with p.Asn549His as the most prevalent variation (1%); finally, 2% affected *FGFR3* (p.Ala391Glu) and 4% *FGFR1* (p.Pro252Arg).

In 100% of cases, genetic variation was identified in the *FGFR2* gene in the AS group (of which 57% p.Ser252Trp and 43% p.Pro253Arg), in the *FGFR3* gene in the MS group (p.Pro250Arg) and in the *TWIST1* gene in the SCS group (of which subtype p.Glu65* in 25% of the cases) (Table 1, Supp. Table 1).

The comparison between two raters was excellent with an ICC of 0.999 [0.999-0.999] for frontal pictures, 0.999 [0.999-0.999] for lateral pictures, and 0.992 [0.991-0.993] for external ears.

	Total	Controls	AS	CS	MS	PS	SCS
N							
Consultations	1114	561 (50%)	73 (7%)	302 (27%)	26 (2%)	118 (11%)	34 (3%)
Photographs	2228	1122 (50%)	146 (7%)	604 (27%)	52 (2%)	236 (11%)	68 (3%)
Patients	541	304 (56%)	44 (8%)	98 (18%)	16 (3%)	51 (9%)	28 (5%)
Gender							
Female	490 (44%)	274 (49%)	42 (58%)	92 (30%)	19 (73%)	43 (36%)	20 (61%)
Male	624 (56%)	287 (51%)	31 (42%)	210 (70%)	7 (27%)	75 (64%)	14 (39%)
Age (years)							
Mean +/- SD	6.2 +/- 6.5	6.1 +/- 4.8	2.1 +/- 2.8	6.0 +/- 8.1	5.4 +/- 12.6	3.4 +/- 7.4	6.2 +/- 10.8
Median	3.1	7.6	0.7	1.6	0.9	1.4	0.7
Minimum	0.0	0.2	0.1	0.1	0.2	0.1	0.0
Maximum	44.0	18.3	11.8	44.0	42.0	37.1	40.0
Ethnicity							
African / Caribbean	80 (7%)	29 (5%)	12 (16%)	35 (12%)	0 (0%)	3 (3%)	1 (3%)
Asian	10 (1%)	10 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Caucasian	1024 (92%)	522 (93%)	61 (84%)	267 (88%)	26 (100%)	115 (97%)	32 (97%)
Genetic variation							
FGFR1	11 (1%)		0 (0%)	0 (0%)	0 (0%)	9 (8%)	0 (0%)
FGFR2	480 (43%)		73 (100%)	298 (99%)	0 (0%)	109 (92%)	0 (0%)
FGFR3	30 (3%)		0 (0%)	4 (1%)	26 (100%)	0 (0%)	0 (0%)
TWIST1	33 (3%)		0 (0%)	0 (0%)	0 (0%)	0 (0%)	34 (100%)

Table 1. Clinical description of the training set. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre-Chotzen Syndrome, SD = Standard Deviation, *FGFR* = Fibroblast Growth Factor Receptor, *TWIST1* = Twist Family bHLH Transcription Factor 1.

1. Validation set

We included 168 frontal and lateral facial photographs corresponding to 84 patients. The mean age was 3.4 +/- 4.4 and ranged from 0 to 29 years; 54% were girls. Ethnicity was 93% Caucasian, 5% African or Caribbean, and 1% Asian.

The control, AS, CS, MS, PS, and SCS groups comprised 26, 40, 62, 12, 10 and 18 photographs respectively, corresponding to 13, 20, 31, 6, 5 and 9 patients.

Concerning CS and PS, 34% of the genetic variations were located on the extracellular portion of *FGFR2* (IgI, IgII, IgIIIa) with the most frequent variation being p.Cys278Phe (6%); 41% were located on the trans-membrane portion of *FGFR2* (IgIIIc) with the most frequent variation being p.Cys342Tyr (22%); 13% affected a splice donor site, with the most common variant being p.Ala344Ala (9%); finally, 4% affected *FGFR3* (p.Ala391Glu). There were no *FGFR1* and *FGFR2* TK domain variations in the validation set. In AS, 53% of genetic variations were p.Ser252Trp and 47% p.Pro253Arg. In MS, 100% of patients were diagnosed with the characteristic p.Pro250Arg mutation of the *FGFR3* gene. SCS patients were all *TWIST1* mutated, with the most frequent variation being p.Leu149Phe (50%) (Table 2, Supp. Table 2).

	Total	Controls	AS	CS	MS	PS	SCS
N							
Consultations	84	13 (15%)	20 (24%)	31 (37%)	6 (7%)	5 (6%)	9 (11%)
Photographs	168	26 (15%)	40 (24%)	62 (37%)	12 (7%)	10 (6%)	18 (11%)
Patients	84	13 (15%)	20 (24%)	31 (37%)	6 (7%)	5 (6%)	9 (11%)
Gender							
Female	45 (54%)	7 (54%)	13 (65%)	14 (45%)	3 (50%)	3 (60%)	5 (56%)
Male	39 (46%)	6 (54%)	7 (35%)	17 (55%)	3 (50%)	2 (40%)	4 (44%)
Age (years)							
Mean +/- SD	3.4 +/- 4.4	6.3 +/- 4.2	3.1 +/- 2.9	3.2 +/- 5.5	1.9 +/- 1.9	0.1 +/- 0.1	3.3 +/- 3.6
Median	1.6	7.5	2.4	1.7	1.2	0.1	1.2
Minimum	0.0	0.1	0.0	0.2	0.2	0.0	0.4
Maximum	29.6	12.7	8.7	29.6	4.5	0.3	10.5
Ethnicity							
African / Caribbean	5 (6%)	1 (8%)	0 (0%)	4 (13%)	0 (0%)	0 (0%)	0 (0%)
Asian	1 (1%)	1 (8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Caucasian	78 (93%)	11 (85%)	20 (100%)	27 (87%)	6 (100%)	5 (100%)	9 (100%)
Genetic variation							
FGFR1	0 (0%)		0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
FGFR2	51 (61%)		20 (100%)	26 (84%)	0 (0%)	5 (100%)	0 (0%)
FGFR3	11 (13%)		0 (0%)	5 (6%)	6 (100%)	0 (0%)	0 (0%)
TWIST1	9 (11%)		0 (0%)	0 (0%)	0 (0%)	0 (0%)	9 (100%)

Table 2. Clinical description of the validation set. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre-Chotzen Syndrome, SD = Standard Deviation, *FGFR* = Fibroblast Growth Factor, *TWIST1* = Twist Family bHLH Transcription Factor 1.

1. Classification between syndromes

3.1 Phenotype descriptions

Craniofacial phenotype in AS vs controls

In AS, we report significant brachycephaly, hypertelorism, thin and downslanted palpebral fissures, a severe midface retrusion and a small upturned nose. AS patients had labial incompetence compared to controls. The chin appears to be in the same position as in the controls, but the midfacial retrusion resulted in a relative prognathia. (Figures 4, 5). AS ears were similar to controls in shape (Figure 6) but were low set (Figure 5).

Craniofacial phenotype in CS and PS vs controls

Patients with CS and PS also had brachycephaly with hypertelorism, open palpebral slits, a midface retrusion, and low-set ears (Figures 4 and 5). CS and PS patients had labial incompetence compared to controls. Ears were similar to controls in shape and position (Figure 6).

Craniofacial phenotype in MS vs controls

Patients with MS had significant brachycephaly, an increase in size and a flattening of the forehead, temporal bossing, wider face, midface retrusion, and labial incompetence. Palpebral fissures were thin and downslanted without telecanthus (Figures 4 and 5). Ears were similar to controls in shape but anteriorly placed (Figure 5 and 6).

Craniofacial phenotype in SCS vs controls

Brachycephaly was also present in SCS. The palpebral fissures were small and without hypertelorism. The reduction in height of the palpebral fissures probably corresponded to the ptosis characteristic of this syndrome. The hairline was low with a small forehead. There was also a labial incompetence (Figures 4 and 5). The ears appeared rounder, and there was the *crux cymbae* characteristic of this syndrome.

Craniofacial phenotype differences between syndromes

There were very few phenotypic differences between the CS and PS groups (Figures 4, 5, 6). The mean shapes were superimposable on frontal, lateral, and external ear pictures. Patients with PS may have more pronounced brachycephaly.

The AS and MS groups were also very similar phenotypically, but with a less upturned nose, a longer middle third of the face, less low-set ears and a smaller inter-ocular distance in the MS group (Figures 4 and 5). The SCS group was distinguished from the other groups by its normally set eyes, low hairline, and longer middle and lower thirds of the face. The *crux cymbae*, an inconsistent cartilaginous ridge dividing the *cymba conchae* into two parts, and rounded ear also seemed characteristic of SCS.

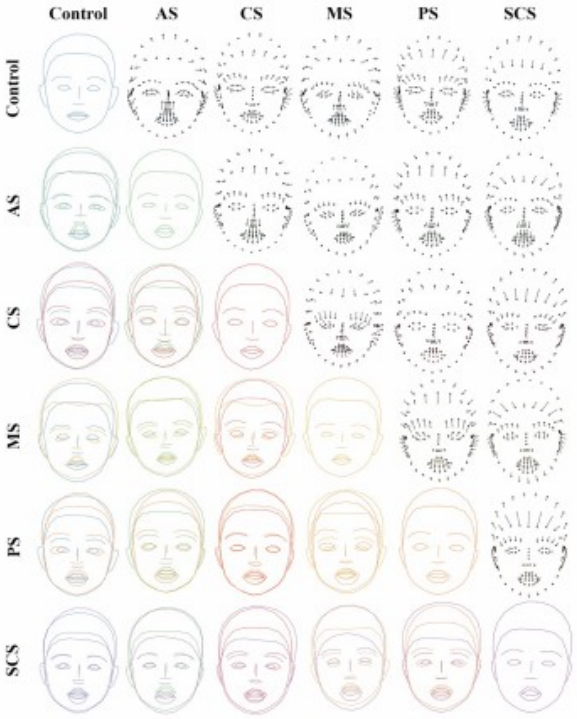


Figure 3. Average shapes by syndrome and comparisons after Procrustes analysis of frontal pictures in patients < 5 y.o. the diagonal corresponds to average faces for each syndromeFace below the diagonal correspond to superimpositions of two average faces from different groups, and faces above the diagonal correspond to vectors of transformation between two groups. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre Chotzen Syndrome.

Figure 4. Average shapes by syndrome and comparisons after Procrustes analysis of lateral pictures in patients < 5 y.o. The elements on the diagonal (from the upper left face to the lower right profile) correspond to the average profiles. The elements below this diagonal correspond to superimpositions of two average profiles from different groups, and the elements above this diagonal correspond to vectors allowing a transformation from a first group to a second group. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre Chotzen Syndrome.

2. ***Binary classifications***

Each of the 5 syndromes was significantly distinguishable from control patients in the validation set, with high AUCs going from 0.983 [0.954 – 1.000] ($p < 0.001$) for CS, to 1.000 [1.000 – 1.000] ($p < 0.001$) for MS and PS. AS was significantly distinguishable from CS (0.955 [0.893 – 1.000], $p < 0.001$) and from SCS (0.906 [0.799 – 1.000], $p < 0.001$), but not from MS (0.658 [0.394 – 0.923], $p = 0.546$) and PS (0.700 [0.470 – 0.930], $p = 0.348$). Interestingly, the model was not able to distinguish CS and PS, with an AUC of 0.548 [0.270 – 0.827] ($p = 0.672$). MS was not discernable from PS and SCS with AUCs of 0.567 [0.157 – 0.977] ($p = 0.841$) and 0.667 [0.370 – 0.963] ($p = 0.642$). Finally, the model was able to distinguish PS from SCS with an AUC of 0.844 [0.606 – 1.000] ($p = 0.021$) (Tables 3, 4).

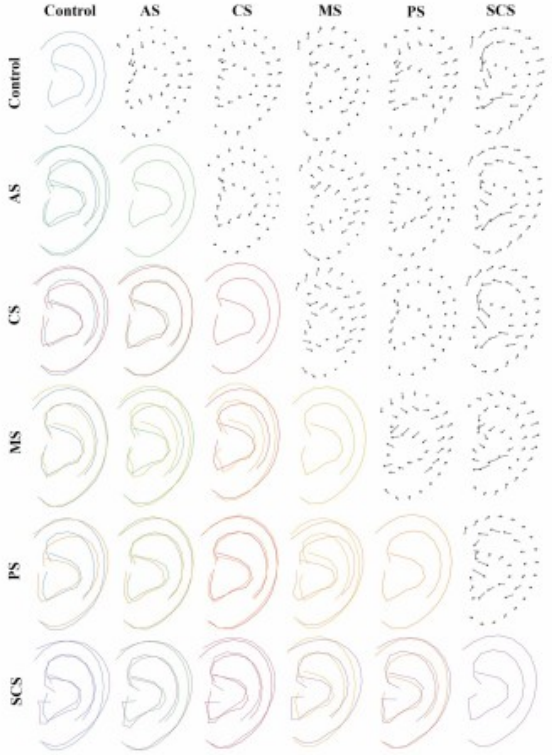


Figure 5. Average shapes by syndrome and comparisons after Procrustes analysis of ears in patients < 5 y.o. The elements on the diagonal (from the upper left face to the lower right ears) correspond to the average ears. The elements below this diagonal correspond to superimpositions of two average ears from different groups, and the elements above this diagonal correspond to vectors allowing a transformation from a first group to a second group. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre Chotzen Syndrome.

	Controls	AS	CS	MS	PS
Controls					
AS	0.992 [0.974 - 1.000] *				
CS	0.983 [0.954 - 1.000] *	0.955 [0.893 - 1.000] *			
MS	1.000 [1.000 - 1.000] *	0.658 [0.394 - 0.923]	0.785 [0.522 - 1.000] *		
PS	1.000 [1.000 - 1.000] *	0.700 [0.470 - 0.930]	0.548 [0.270 - 0.827]	0.567 [0.157 - 0.977]	
SCS	0.940 [0.834 - 1.000] *	0.906 [0.799 - 1.000] *	0.871 [0.704 - 1.000] *	0.667 [0.370 - 0.963]	0.844 [0.606 - 1.000] *

Table 3. AUC for each binary classification on the validation set. AS = Apert Syndrome, CS = Crouzon Syndrome, MS = Muenke Syndrome, PS = Pfeiffer Syndrome, SCS = Saethre-Chotzen Syndrome. * = statistically significant ($p < 0.05$).

2. ***Multi-syndrome classification***

For further analysis, we grouped CS and PS patients together under the name Crouzon-Pfeiffer syndrome (CPS), given the absence of significant phenotypic differences in binary classification.

The number of iterations retained for this multi-class model was 248 and 70.2% [0.593 - 0.797] of patients in the validation set were correctly diagnosed. The No Information Rate (NIR) was 0.429, making this prediction significant at 95% ($p < 0.001$). The best performances were obtained by the control group (balanced accuracy = 0.943 and AUC = 1.000), the CPS group (balanced accuracy = 0.844 and AUC = 0.941) and the AS group (balanced accuracy = 0.786 and AUC = 0.836). The MS and SCS groups were more difficult to predict (Table 4).

Overall (multiclass design)		
Accuracy		0.702 [0.593 - 0.797] *
One-versus-all design		
Balanced accurary		
	Control	0.943
	AS	0.786
	CPS	0.844
	MS	0.692
	SCS	0.667
AUC		
	Control	1.000
	AS	0.836
	CPS	0.941
	MS	0.625
	SCS	0.744

Table 4. Classification performances in a multi-syndrome design. AS = Apert Syndrome, CPS = Crouzon-Pfeiffer Syndrome, MS = Muenke Syndrome, SCS = Saethre-Chotzen Syndrome, AUC = Area Under the Curve.

The confusion matrix revealed 13 correctly predicted controls, 13 AS, 27 CPS, 3 MS and 3 SCS correctly diagnosed (Table 5).

		Reference				
		Control	**AS**	**CPS**	**MS**	**SCS**
Prediction	**Control**	**13**	2	4	1	1
	AS	0	**13**	2	2	1
	CPS	0	3	**27**	0	0
	MS	0	2	3	**3**	4
	SCS	0	0	0	0	**3**

Table 5. Confusion matrix in a multi-syndrome design. AS = Apert Syndrome, CPS = Crouzon-Pfeiffer Syndrome, MS = Muenke Syndrome, SCS = Saethre-Chotzen Syndrome. Bold values: True Positives (TP).

The results were consistent with the ROC curves and the UMAP representation (Figure 6), which showed a clear separation between the control and the CPS groups.

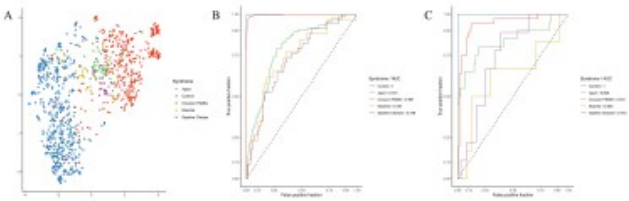


Figure 6. Multi-syndrome classification. A. UMAP representation of the training data according to the different groups. B. Empirical ROC curves (training set) for each group with AUC. C. ROC curves (validation set) for each group, with AUC. AUC = Area Under the Curve.

1. Genotype – phenotype correlations

1. **AS group:** *FGFR2* p.Pro253Arg vs. *FGFR2* p.Ser252Trp

Both genotypes were significantly distinguishable from controls, with an AUC = 1,000 [1,000 - 1,000] ($p < 0.001$). However, our model was unable to correctly classify patients between the two genotypes (AUC = 0.506 [0.215 - 0.797], $p = 0.874$) (Table 6 and Supp. Figures 1-3).

	Controls	**p.Pro253Arg**
Controls		
p.Pro253Arg	1.000 [1.000 - 1.000] *	
p.Ser252Trp	1.000 [1.000 - 1.000] *	0.506 [0.215 - 0.797]

Table 6. AUC for each binary classification in the two genotype groups of AS. * = statistically significant ($p < 0.05$).

1. **CPS group:** comparisons between the 6 genotype groups (IgI-IIIa, IgIIIc, splicing domain, TK, *FGFR3* and *FGFR1*)

There was considerable phenotypic variability within the CPS group (Figure 7).

Figure 7. Phenotypical variability in children with Crouzon-Pfeiffer syndrome. A, G: Genetic variation located on *FGFR2* IgI domain; B, H: Genetic variation located on *FGFR2* IgIIIc domain; C, I: Genetic variation located on a splice donor site of *FGFR2*; D, J: Genetic variation located on *FGFR2* TK domain; E, K: Genetic variation located on *FGFR3*; F/L: Genetic variation located on *FGFR1*.

Genotypes linked to the IgI, IgII, IgIIIc and TK domains of *FGFR2* and variations in *FGFR3* were associated with more severe facial phenotypes than variations in a splice site of *FGFR2*. In the latter group, brachycephaly, ocular phenotype and reduced height of the mid-face seemed to be less severe (Figures 8 and 9). The ears have not been plotted in this case, as Figure 5 does not show any clear differences between the control ears and the CS/PS ears.

A binary classification between the 'IgIIIc' and 'Splicing domain' groups was significant, with an AUC = 0.786 [0.554 - 1.000] ($p < 0.023$) on the validation set. In addition, the controls were indistinguishable from the 'Splicing domain' group (AUC = 0.577 [0.332 - 0.822], p = 0.657) (Table 7).

	Controls	IgI-IIIa	IgIIIc	Splicing	TK	*FGFR3*
Controls						
IgI-IIIa	1.000 [1.000 - 1.000] *					
IgIIIc	0.984 [0.951 - 1.000] *	0.675 [0.444 - 0.907]				
Splicing	0.577 [0.332 - 0.822]	0.705 [0.422 - 0.987]	0.786 [0.554 - 1.000] *			
TK	NA	NA	NA	NA		
FGFR3	0.750 [0.467 - 1.000]	0.602 [0.243 - 0.962]	0.580 [0.223 - 0.938]	0.500 [0.100 - 0.900]	NA	
FGFR1	NA	NA	NA	NA	NA	NA

Table 7. AUC for each binary classification on the validation set. There are no results for the TK and *FGFR1* groups, as there were no patients in the validation group. *FGFR* = Fibroblast Growth Factor Receptor, Ig = Immunoglobulin-like domain, TK = Tyrosine Kinase domain.

Classification between the 'Splicing domain' group and the 4 other genotypes showed a significant AUC on the empirical curve (training set) of 0.706 [0.602 - 0.810] (p = 0.012). This significance was not found on the validation set, with an AUC of 0.629 [0.381 - 0.878] (p = 0.438).

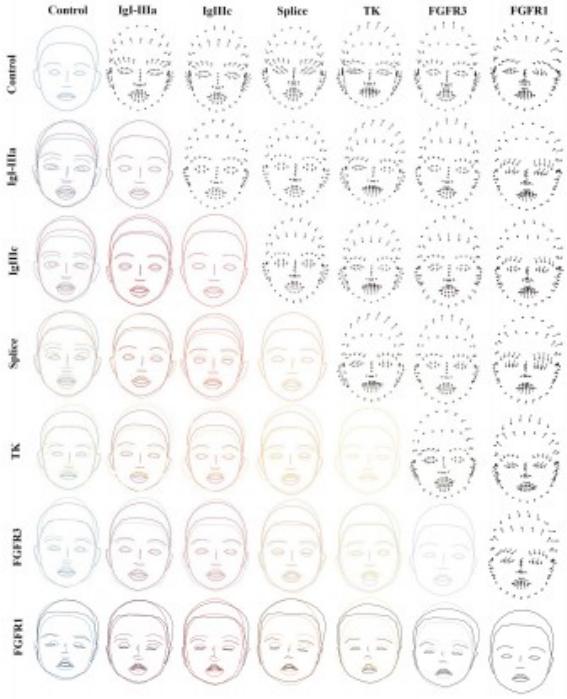


Figure 8. Average shapes by CPS genotype groups and comparisons after Procrustes analysis of frontal pictures in patients < 5 y.o. The elements on the diagonal (from the upper left face to the lower right face) correspond to the average faces. The elements below this diagonal correspond to superimpositions of two average faces from different groups, and the elements above this diagonal correspond to vectors allowing a transformation from a first group to a second group. *FGFR* = Fibroblast Growth Factor Receptor, Ig = Immunoglobulin-like domain, TK = Tyrosine Kinase domain.

Discussion

1. Syndrome diagnosis support

We report the first algorithm for the automatic detection of syndromic craniosynostoses using machine learning and a controlled database, based on frontal and lateral facial pictures.

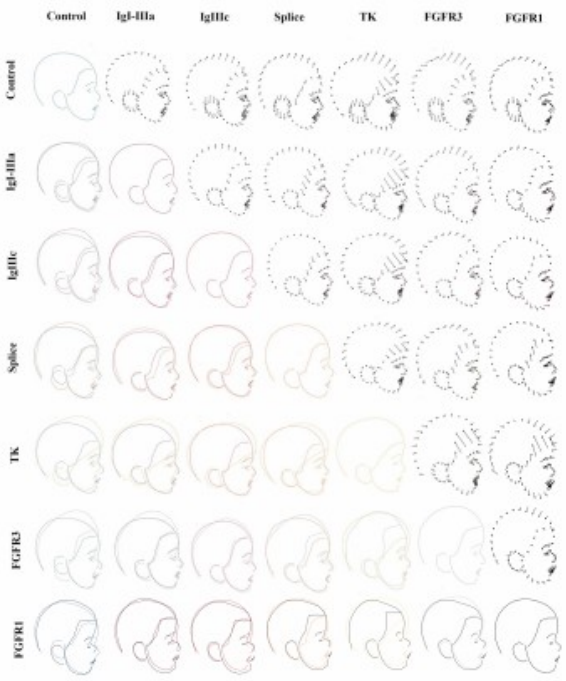


Figure 9. Average shapes by CPS genotype groups and comparisons after Procrustes analysis of profile pictures in patients < 5 y.o. The elements on the diagonal (from the upper left face to the lower right profile) correspond to the average profiles. The elements below this diagonal correspond to superimpositions of two average profiles from different groups, and the elements above this diagonal correspond to vectors allowing a transformation from a first group to a second group. *FGFR* = Fibroblast Growth Factor Receptor, Ig = Immunoglobulin-like domain, TK = Tyrosine Kinase domain.

Face2Gene (FDNA, Inc.), an alternative approach able to tackle similar clinical questions, - is a commercial tool for automatic diagnosis on 2D photographs based on a deep-learning algorithm and automatic landmark detection [5]. However, Face2Gene does not take profiles and ears into account, which contain critical information, such as mandibular position, nose shape, midface retrusion, and position and size of the ears. In addition, unlike Face2Gene, we report a method that records the shape of the cranial vault and the position of the hairline, which are relevant in craniosynostoses. Finally, our approach is unique in that it includes the prediction of a 'control' group.

We were able to correctly diagnose 70.2% [0.593 - 0.797] ($p < 0.001$) of patients in the validation set using our innovative method. The best performances were obtained by the control group (balanced accuracy = 0.943 and AUC = 1.000), the CPS group (balanced accuracy = 0.844 and AUC = 0.941) and the AS group (balanced accuracy = 0.786 and AUC = 0.836).

1. Phenotypical description

We confirm the usual phenotypic descriptions of AS, with brachycephaly, hypertelorism, down-slanting palpebral fissures, severe midface retrusion, open gonial angle and short mandibular ramus height [22,58,59]. CS is most often characterized by brachycephaly, significant hypertelorism, opening of the palpebral fissures due to ocular proptosis, midface retrusion leading to Angle class III occlusion, open gonial angle and short ramus height, and low implantation of the ears [22–25,60,61]. We confirm these findings on the average CS faces computed by our approach. Similarly, we confirm the anomalies classically described in MS [62], i.e. brachycephaly or turribrachycephaly ("tower-shaped" skull), flattening of the forehead, anterior and low-set ears, temporal bossing, short nose and midface retrusion. Finally, we were also able to recognize common features of SCS [28,63,64]: brachycephaly, midface retrusion and low anterior hairline. The small palpebral fissures probably correspond to the characteristic ptosis of SCS. We found also report specific external ear shapes: smaller, rounder, with *crux cymbae*.

Assessment of limb anomalies is an essential element for differentiating between CS, AS, PS, and SCS but was not included into our model. More precisely, the classical anomalies in these syndromes are: syndactyly of fingers 2-3 or duplication of the distal phalanx of the hallux in SCS [65], syndactyly of fingers and toes and lateral deviation of thumbs and great toes in AS [58], and deviation of thumbs and great toes with symphalangism in PS [66].

Mild forms of MS and SCS are often misdiagnosed because of overlapping clinical features [64,65,67]. In our model, we were unable to differentiate MS from SCS in a binary classification (0.667 [0.370 – 0.963], $p = 0.642$). The distinctive features of MS are carpal and tarsal fusions and hearing loss [67], whereas the main distinctive features of SCS are strabismus and ptosis [28].

The round and small pinna with a *crux cymbae*, a prominent crura, contributes to the differential diagnosis. De Heer et al [68] stated that this feature is present in 56% of patients with SCS. Choi et al [65] listed the skeletal differences between MS and SCS: both syndromes result in midface narrowing; however, MS causes posterior rotation of the palatal plane, maxillo-mandibular hyper-divergence resulting in labial incompetence and reduction in the height of the mandibular ramus, while SCS causes posterior rotation of the palatal plane, with no difference in maxillary height, resulting in hyper-divergence and lengthening of the lower third of the face. Choi et al [65]state that *FGFR3* and *TWIST1* act differently on the timing of fusion of circum-maxillary sutures, resulting in differences in facial growth [65].

Several teams consider that CS and PS syndromes correspond to the same entity [60,61,69,70]. We found no significant difference in facial phenotypes between these two syndromes in a binary classification (AUC = 0.548 [0.270 – 0.827], p = 0.672).

2. Genotype-phenotype correlations

AS is caused by 2 mutations only in *FGFR2*: p.Ser252Trp and p.Pro253Arg [71,72]. Jadico et al suggested that patients with the p.Ser252Trp variant have a more severe ocular phenotype [73], while Lajeunie et al [58] stated that p.Pro253Arg leads to more severe cognitive impairement and anomalies of the upper extremities anomalies. Based on a large cohort of patients, we could not show any facial phenotype difference between these two AS genotypes (AUC = 0.506 [0.215 - 0.797], p = 0.874).

Fenwick et al [74] suggested that alterations in normal *FGFR2* splicing may result in a mild CS phenotype. We showed in a binary classification between the 'IgIIIc' and 'Splicing domain' groups in CPS, a significant prediction with an AUC = 0.786 [0.554 - 1.000] (p < 0.023) on the validation set. In addition, the controls were indistinguishable from the 'Splicing domain' group (AUC = 0.577 [0.332 - 0.822], p = 0.657). Classification between the 'Splicing domain' group and the 4 other genotype groups showed a significant AUC on the empirical curve (training set) of 0.706 [0.602 - 0.810] (p = 0.012), but this significance was not confirmed by the validation set, with an AUC of 0.629 [0.381 - 0.878] (p = 0.438), most probably due to limited data volume.

These results are corroborated by average faces, with less hypoplastic midface, milder hypertelorism and brachycephaly in the "Splicing domain" group relative to other genotype groups.

We observed significant phenotype differences between the other CPS genetic groups. Variations localized to the intracellular (TK) domain of *FGFR2*, and variations in *FGFR3* appeared to result in more severe brachycephaly and even turricephaly (Figures 8 and 9). Variations in *FGFR1* and *FGFR3* appear to affect mandibular growth less than *FGFR2*, and the midfacial narrowing results in a class III malocclusion. Morice et al [22] showed a short ramus height in patients with *FGFR2* variant, but not in patients with a *FGFR3* variant. Midfacial involvement appeared extreme for patients with *FGFR1* variation, compared to variations on *FGFR2* and *FGFR3*. These facial and cranial morphological differences between *FGFR2* and *FGFR3* craniosynostoses support the hypothesis that the genes affect the membranous and endochondral processes differently. Cornille et al [75] showed in a CS mouse model with *FGFR3* overactivation, a brachycephaly, an oculo-orbital disproportion, a premature fusion of the anterior-intra-occipital synchondroses, prognathism, and midfacial hypoplasia. The authors explained that the relatively small mandible and the severe maxillary retrusion leads to class III malocclusion.

Conclusion

Here we report a new approach of the automatic detection of syndromic craniosynostoses with AI-based methods, using a database covering forty years of activity of the Necker – Enfants Malades craniofacial center. Performances were validated based on data from another large craniofacial center, Great Ormond St. Hospital, London. All patients in both databases had identified genetic variations.

The novelty of our study is that we account for frontal and profile facial morphology, external ear shape, cranial vault outline, hairline implantation, and prediction within a group of control patients. We were able to detect more than 70% of patients, with excellent performance particularly in the CPS group (84.4% and AUC = 0.941). Finally, we were able to describe the significant phenotypic variability within the Crouzon Pfeiffer group, all to better predict patients' clinical course.

Conflict of interest statement
The authors declare they have no conflicts of interest.

Acknowledgement
We would like to thank Prof. Dominique Renier and Dr. Daniel Marchac for giving us access to their photographic databases and years of hard work in the study of syndromic craniosynostoses. This work was supported by the 'Agence Nationale de la Recherche', 'Investissements d'Avenir' program (ANR-10-IAHU-01) and by France 2030 grant "Face4Kids" (ANR-21-PMRB-0004).

References

1. Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. N Engl J Med *380*, 1347–1358. 10.1056/NEJMra1814259.

2. Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A.E., Pianykh, O.S., Geis, J.R., Pandharipande, P.V., Brink, J.A., and Dreyer, K.J. (2018). Current Applications and Future Impact of Machine Learning in Radiology. Radiology *288*, 318–328. 10.1148/radiol.2018171820.

3. Novoa, R.A., Gevaert, O., and Ko, J.M. (2019). Marking the Path Toward Artificial Intelligence-Based Image Classification in Dermatology. JAMA Dermatol *155*, 1105–1106. 10.1001/jamadermatol.2019.1633.

4. Loftus, T.J., Tighe, P.J., Filiberto, A.C., Efron, P.A., Brakenridge, S.C., Mohr, A.M., Rashidi, P., Upchurch, G.R., and Bihorac, A. (2020). Artificial Intelligence and Surgical Decision-making. JAMA Surg *155*, 148–158. 10.1001/jamasurg.2019.4917.

5. Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P.M., Kamphausen, S.B., Zenker, M., et al. (2019). Identifying facial phenotypes of genetic disorders using deep learning. Nat Med *25*, 60–64. 10.1038/s41591-018-0279-0.

6. Zhang, Q., Ding, Y., Feng, B., Tang, Y., Chen, Y., Wang, Y., Chang, G., Liu, S., Wang, J., Li, Q., et al. (2022). Molecular and Phenotypic Expansion of Alström Syndrome in Chinese Patients. Front Genet *13*, 808919. 10.3389/fgene.2022.808919.

7. Javitt, M.J., Vanner, E.A., Grajewski, A.L., and Chang, T.C. (2022). Evaluation of a computer-based facial dysmorphology analysis algorithm (Face2Gene) using standardized textbook photos. Eye *36*, 859–861. 10.1038/s41433-021-01563-5.

8. Latorre-Pellicer, A., Ascaso, Á., Trujillano, L., Gil-Salvador, M., Arnedo, M., Lucia-Campos, C., Antoñanzas-Pérez, R., Marcos-Alcalde, I., Parenti, I., Bueno-Lozano, G., et al. (2020). Evaluating Face2Gene as a Tool to Identify Cornelia de Lange Syndrome by Facial Phenotypes. Int J Mol Sci *21*, E1042. 10.3390/ijms21031042.

.9. Mishima, H., Suzuki, H., Doi, M., Miyazaki, M., Watanabe, S., Matsumoto, T., Morifuji, K., Moriuchi, H., Yoshiura, K.-I., Kondoh, T., et al. (2019). Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. J Hum Genet *64*, 789–794. 10.1038/s10038-019-0619-z.

10. Rouxel, F., Yauy, K., Boursier, G., Gatinois, V., Barat-Houari, M., Sanchez, E., Lacombe, D., Arpin, S., Giuliano, F., Haye, D., et al. (2022). Using deep-neural-network-driven facial recognition to identify distinct Kabuki syndrome 1 and 2 gestalt. Eur J Hum Genet *30*, 682–686. 10.1038/s41431-021-00994-8.

11. Governale, L.S. (2015). Craniosynostosis. Pediatr Neurol *53*, 394–401. 10.1016/j.pediatrneurol.2015.07.006.

9. 13. Di Rocco, F., Biosse Duplan, M., Heuzé, Y., Kaci, N., Komla-Ebri, D., Munnich, A., Mugniery, E., Benoist-Lasselin, C., and Legeai-Mallet, L. (2014). FGFR3 mutation causes abnormal membranous ossification in achondroplasia. Hum Mol Genet *23*, 2914–2925. 10.1093/hmg/ddu004.

14. Snyder-Warwick, A.K., Perlyn, C.A., Pan, J., Yu, K., Zhang, L., and Ornitz, D.M. (2010). Analysis of a gain-of-function FGFR2 Crouzon mutation provides evidence of loss of function activity in the etiology of cleft palate. Proc Natl Acad Sci U S A *107*, 2515–2520. 10.1073/pnas.0913985107.

15. Betances, E.M., Mendez, M.D., and M Das, J. (2023). Craniosynostosis. In StatPearls (StatPearls Publishing).

16. Singh, R.P., Dhariwal, D., Bhujel, N., Shaikh, Z., Davies, P., Nishikawa, H., Solanki, G., and Dover, M.S. (2010). Role of parental risk factors in the aetiology of isolated non-syndromic metopic craniosynostosis. Br J Oral Maxillofac Surg *48*, 438–442. 10.1016/j.bjoms.2009.06.233.

17. Kimonis, V., Gold, J.-A., Hoffman, T.L., Panchal, J., and Boyadjiev, S.A. (2007). Genetics of craniosynostosis. Semin Pediatr Neurol *14*, 150–161. 10.1016/j.spen.2007.08.008.

18. Magge, S.N., Snyder, K., Sajja, A., DeFreitas, T.A., Hofherr, S.E., Broth, R.E., Keating, R.F., and Rogers, G.F. (2017). Identical Twins Discordant for Metopic Craniosynostosis: Evidence of Epigenetic Influences. J Craniofac Surg *28*, 14–16. 10.1097/SCS.0000000000003368.

19. Ohishi, A., Nishimura, G., Kato, F., Ono, H., Maruwaka, K., Ago, M., Suzumura, H., Hirose, E., Uchida, Y., Fukami, M., et al. (2017). Mutation analysis of FGFR1-3 in 11 Japanese patients with syndromic craniosynostoses. Am J Med Genet A *173*, 157–162. 10.1002/ajmg.a.37992.

20. Boulet, S.L., Rasmussen, S.A., and Honein, M.A. (2008). A population-based study of craniosynostosis in metropolitan Atlanta, 1989-2003. Am J Med Genet A *146A*, 984–991. 10.1002/ajmg.a.32208.

21. Kan, S., Elanko, N., Johnson, D., Cornejo-Roldan, L., Cook, J., Reich, E.W., Tomkins, S., Verloes, A., Twigg, S.R.F., Rannan-Eliya, S., et al. (2002). Genomic Screening of Fibroblast Growth-Factor Receptor 2 Reveals a Wide Spectrum of Mutations in Patients with Syndromic Craniosynostosis. Am J Hum Genet *70*, 472–486.

22. Morice, A., Cornette, R., Giudice, A., Collet, C., Paternoster, G., Arnaud, É., Galliani, E., Picard, A., Legeai-Mallet, L., and Khonsari, R.H. (2020). Early mandibular morphological differences in patients with FGFR2 and FGFR3-related syndromic craniosynostoses: A 3D comparative study. Bone *141*, 115600. 10.1016/j.bone.2020.115600.

23. Bouaoud, J., Hennocq, Q., Paternoster, G., James, S., Arnaud, E., and Khonsari, R.H. (2020). Excessive ossification of the bandeau in Crouzon and Apert syndromes. J Craniomaxillofac Surg *48*, 376–382. 10.1016/j.jcms.2020.02.022.

24. Reardon, W., Winter, R.M., Rutland, P., Pulleyn, L.J., Jones, B.M., and Malcolm, S. (1994). Mutations in the fibroblast growth factor receptor 2 gene cause Crouzon syndrome. Nat Genet *8*, 98–103. 10.1038/ng0994-98.

25. Meyers, G.A., Orlow, S.J., Munro, I.R., Przylepa, K.A., and Jabs, E.W. (1995). Fibroblast growth factor receptor 3 (FGFR3) transmembrane mutation in Crouzon syndrome with acanthosis nigricans. Nature genetics *11*, 462–464.

Pal, U.S., Gupta, C., and Chellappa, A.A.L. (2012). Crouzon syndrome with primary optic nerve atrophy and normal brain functions: A case report. J Oral Biol Craniofac Res *2*, 116–118. 10.1016/j.jobcr.2012.03.011.

27. Jezela-Stanek, A., and Krajewska-Walasek, M. (2013). Genetic causes of syndromic craniosynostoses. Eur J Paediatr Neurol *17*, 221–224. 10.1016/j.ejpn.2012.09.009.

28. Gallagher, E.R., Ratisoontorn, C., and Cunningham, M.L. (1993). Saethre-Chotzen Syndrome. In GeneReviews®, M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, and A. Amemiya, eds. (University of Washington, Seattle).

29. Rosen, H., Andrews, B.T., Meara, J.G., Stoler, J.M., Mulliken, J.B., and Rogers, G.F. (2011). Audiologic findings in Saethre-Chotzen syndrome. Plast Reconstr Surg *127*, 2014–2020. 10.1097/PRS.0b013e31820cf16a.

30. Garcelon, N., Neuraz, A., Salomon, R., Faour, H., Benoit, V., Delapalme, A., Munnich, A., Burgun, A., and Rance, B. (2018). A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform *80*, 52–63. 10.1016/j.jbi.2018.02.019.

31. Koonce, B. (2021). ResNet 50. In, pp. 63–72. 10.1007/978-1-4842-6168-2_6.

32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. 10.48550/arXiv.1912.01703.

33. Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. 10.48550/arXiv.1412.6980.

34. Alabort-i-Medina, J., Antonakos, E., Booth, J., Snape, P., and Zafeiriou, S. (2014). Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models. In Proceedings of the 22nd ACM international conference on Multimedia (ACM), pp. 679–682. 10.1145/2647868.2654890.

35. Hennocq, Q., Bongibault, T., Bizière, M., Delassus, O., Douillet, M., Cormier-Daire, V., Amiel, J., Lyonnet, S., Marlin, S., Rio, M., et al. An automatic facial landmarking for children with rare diseases. American Journal of Medical Genetics Part A *n/a*. 10.1002/ajmg.a.63126.

36. Lucas, B., and Kanade, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI). In.

37. landmarker.io · The Menpo Project https://www.menpo.org/landmarkerio/.

38. Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychol Rep *19*, 3–11. 10.2466/pr0.1966.19.1.3.

39. Rohlf, F.J., and Slice, D. (1990). Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. Systematic Zoology *39*, 40–59. 10.2307/2992207.

40. Baken, E.K., Collyer, M.L., Kaliontzopoulou, A., and Adams, D.C. (2021). geomorph v4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. Methods in Ecology and Evolution *12*, 2355–2363. 10.1111/2041-210X.13723.

41. Avcı, H., and Karakaya, J. (2023). A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning. Diagnostics (Basel) *13*, 348. 10.3390/diagnostics13030348.

42. Anifah, L., Purnama, I.K.E., Hariadi, M., and Purnomo, M.H. (2013). Osteoarthritis Classification Using Self Organizing Map Based on Gabor Kernel and Contrast-Limited Adaptive Histogram Equalization. The Open Biomedical Engineering Journal *7*, 18. 10.2174/1874120701307010018.

43. Huang, C., Li, X., and Wen, Y. (2021). AN OTSU image segmentation based on fruitfly optimization algorithm. Alexandria Engineering Journal *60*, 183–188. 10.1016/j.aej.2020.06.054.

44. Kiflie, A., Tesema Tufa, G., and Salau, A.O. (2023). Sputum smears quality inspection using an ensemble feature extraction approach. Front Public Health *10*, 1032467. 10.3389/fpubh.2022.1032467.

45. Haralick, R.M., Shanmugam, K., and Dinstein, I. (1973). Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics *SMC-3*, 610–621. 10.1109/TSMC.1973.4309314.

46. Mohanaiah, P., Sathyanarayana, P., and GuruKumar, L. (2013). Image Texture Feature Extraction Using GLCM Approach. *3*.

47. Löfstedt, T., Brynolfsson, P., Asklund, T., Nyholm, T., and Garpebring, A. (2019). Gray-level invariant Haralick texture features. PLoS One *14*, e0212110. 10.1371/journal.pone.0212110.

Mundt, P., Tharmaseelan, H., Hertel, A., Rotkopf, L.T., Nörenberg, D., Riffel, P., Schoenberg, S.O., Froelich, M.F., and Ayx, I. (2023). Periaortic adipose radiomics texture features associated with increased coronary calcium score-first results on a photon-counting-CT. BMC Med Imaging *23*, 97. 10.1186/s12880-023-01058-7.

49. Adelsmayr, G., Janisch, M., Müller, H., Holzinger, A., Talakic, E., Janek, E., Streit, S., Fuchsjäger, M., and Schöllnast, H. (2023). Three dimensional computed tomography texture analysis of pulmonary lesions: Does radiomics allow differentiation between carcinoma, neuroendocrine tumor and organizing pneumonia? Eur J Radiol *165*, 110931. 10.1016/j.ejrad.2023.110931.

50. Peng, B., Wang, K., Xu, R., Guo, C., Lu, T., Li, Y., Wang, Y., Wang, C., Chang, X., Shen, Z., et al. (2023). Preoperative computed tomography-based tumoral radiomic features prediction for overall survival in resectable non-small cell lung cancer. Front Oncol *13*, 1131816. 10.3389/fonc.2023.1131816.

51. Ou, X., Pan, W., and Xiao, P. (2014). In vivo skin capacitive imaging analysis by using grey level co-occurrence matrix (GLCM). Int J Pharm *460*, 28–32. 10.1016/j.ijpharm.2013.10.024.

52. Muenke, M., Adeyemo, A., and Kruszka, P. (2016). An electronic atlas of human malformation syndromes in diverse populations. Genet Med *18*, 1085–1087. 10.1038/gim.2016.3.

53. Burchard, E.G., Ziv, E., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Pérez-Stable, E.J., Sheppard, D., and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med *348*, 1170–1175. 10.1056/NEJMsb025007.

54. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16. (Association for Computing Machinery), pp. 785–794. 10.1145/2939672.2939785.

55. Sachs, M.C. (2017). plotROC: A Tool for Plotting ROC Curves. J Stat Softw *79*, 2. 10.18637/jss.v079.c02.

56. McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 10.48550/arXiv.1802.03426.

57. R Core Team (2020). — European Environment Agency https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006.

58. Lajeunie, E., Cameron, R., El Ghouzzi, V., de Parseval, N., Journeau, P., Gonzales, M., Delezoide, A.L., Bonaventure, J., Le Merrer, M., and Renier, D. (1999). Clinical variability in patients with Apert's syndrome. J Neurosurg *90*, 443–447. 10.3171/jns.1999.90.3.0443.

59. Jadico, S.K., Young, D.A., Huebner, A., Edmond, J.C., Pollock, A.N., McDonald-McGinn, D.M., Li, Y.-J., Zackai, E.H., and Young, T.L. (2006). Ocular abnormalities in Apert syndrome: genotype/phenotype correlations with fibroblast growth factor receptor type 2 mutations. J AAPOS *10*, 521–527. 10.1016/j.jaapos.2006.07.012.

60. Khonsari, R.H., Hennocq, Q., Nysjö, J., Sandy, R., Haber, S., James, S., Britto, J.A., Paternoster, G., and Arnaud, É. (2019). Defining Critical Ages for Orbital Shape Changes after Frontofacial Advancement in Crouzon Syndrome. Plast Reconstr Surg *144*, 841e–852e. 10.1097/PRS.0000000000006162.

61. Staal, F.C.R., Ponniah, A.J.T., Angullia, F., Ruff, C., Koudstaal, M.J., and Dunaway, D. (2015). Describing Crouzon and Pfeiffer syndrome based on principal component analysis. J Craniomaxillofac Surg *43*, 528–536. 10.1016/j.jcms.2015.02.005.

62. Kruszka, P., Rolle, M., Kahle, K.T., and Muenke, M. (1993). Muenke Syndrome. In GeneReviews®, M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. Bean, K. W. Gripp, and A. Amemiya, eds. (University of Washington, Seattle).

63. Howard, T.D., Paznekas, W.A., Green, E.D., Chiang, L.C., Ma, N., Ortiz de Luna, R.I., Garcia Delgado, C., Gonzalez-Ramos, M., Kline, A.D., and Jabs, E.W. (1997). Mutations in TWIST, a basic helix-loop-helix transcription factor, in Saethre-Chotzen syndrome. Nat Genet *15*, 36–41. 10.1038/ng0197-36.

64. Apostolopoulou, D., Kaxira, O.S., Hatzaki, A., Panagopoulos, K.P., Alexandrou, K., Stratoudakis, A., Kollia, P., and Aleporou, V. (2018). Genetic Analysis of Syndromic and Nonsyndromic Patients With Craniosynostosis Identifies Novel Mutations in the TWIST1 and EFNB1 Genes. Cleft Palate Craniofac J *55*, 1092–1102. 10.1177/1055665618760412.

65. Choi, T.M., Lijten, O.W., Mathijssen, I.M.J., Wolvius, E.B., and Ongkosuwito, E.M. (2022). Craniofacial morphology and growth in Muenke syndrome, Saethre-Chotzen syndrome, and TCF12-related craniosynostosis. Clin Oral Investig *26*, 2927–2936. 10.1007/s00784-021-04275-y.

66. M Das, J., and Winters, R. (2023). Pfeiffer Syndrome. In StatPearls (StatPearls Publishing).

67. Doherty, E.S., Lacbawan, F., Hadley, D.W., Brewer, C., Zalewski, C., Kim, H.J., Solomon, B., Rosenbaum, K., Domingo, D.L., Hart, T.C., et al. (2007). Muenke syndrome (FGFR3-related craniosynostosis): expansion of the phenotype and review of the literature. Am J Med Genet A *143A*, 3204–3215. 10.1002/ajmg.a.32078.

68. de Heer, I.M., de Klein, A., van den Ouweland, A.M., Vermeij-Keers, C., Wouters, C.H., Vaandrager, J.M., Hovius, S.E.R., and Hoogeboom, J.M. (2005). Clinical and genetic analysis of patients with Saethre-Chotzen syndrome. Plast Reconstr Surg *115*, 1894–1902; discussion 1903-1905. 10.1097/01.prs.0000165278.72168.51.

69. Khonsari, R.H., Way, B., Nysjö, J., Odri, G.A., Olszewski, R., Evans, R.D., Dunaway, D.J., Nyström, I., and Britto, J.A. (2016). Fronto-facial advancement and bipartition in Crouzon-Pfeiffer and Apert syndromes: Impact of fronto-facial surgery upon orbital and airway parameters in FGFR2 syndromes. J Craniomaxillofac Surg *44*, 1567–1575. 10.1016/j.jcms.2016.08.015.

70. Way, B.L.M., Khonsari, R.H., Karunakaran, T., Nysjö, J., Nyström, I., Dunaway, D.J., Evans, R.D., Hayward, R.D., and Britto, J.A. (2019). Correcting Exorbitism by Monobloc Frontofacial Advancement in Crouzon-Pfeiffer Syndrome: An Age-Specific, Time-Related, Controlled Study. Plast Reconstr Surg *143*, 121e–132e. 10.1097/PRS.0000000000005105.

71. Passos-Bueno, M.R., Sertié, A.L., Zatz, M., and Richieri-Costa, A. (1997). Pfeiffer mutation in an Apert patient: how wide is the spectrum of variability due to mutations in the FGFR2 gene? Am J Med Genet *71*, 243–245.

72. Fernandes, M.B.L., Maximino, L.P., Perosa, G.B., Abramides, D.V.M., Passos-Bueno, M.R., and Yacubian-Fernandes, A. (2016). Apert and Crouzon syndromes-Cognitive development, brain abnormalities, and molecular aspects. Am J Med Genet A *170*, 1532–1537. 10.1002/ajmg.a.37640.

73. Jadico, S.K., Young, D.A., Huebner, A., Edmond, J.C., Pollock, A.N., McDonald-McGinn, D.M., Li, Y.-J., Zackai, E.H., and Young, T.L. (2006). Ocular abnormalities in Apert syndrome: genotype/phenotype correlations with fibroblast growth factor receptor type 2 mutations. J AAPOS *10*, 521–527. 10.1016/j.jaapos.2006.07.012.

74. Fenwick, A.L., Goos, J.A., Rankin, J., Lord, H., Lester, T., Hoogeboom, A.J.M., Ouweland, A.M. van den, Wall, S.A., Mathijssen, I.M., and Wilkie, A.O. (2014). Apparently synonymous substitutions in FGFR2 affect splicing and result in mild Crouzon syndrome. BMC Medical Genetics *15*, 95. 10.1186/s12881-014-0095-4.

75. Cornille, M., Moriceau, S., Khonsari, R.H., Heuzé, Y., Loisay, L., Boitez, V., Morice, A., Arnaud, E., Collet, C., Bensidhoum, M., et al. (2022). FGFR3 overactivation in the brain is responsible for memory impairments in Crouzon syndrome mouse model. J Exp Med *219*, e20201879. 10.1084/jem.20201879.

6.2 Next Generation Phenotyping model for diagnosis and phenotype – genotype correlations in Kabuki syndrome

Kabuki syndrome (KS) is a rare genetic disorder, with an estimated prevalence of 1:86,000 to1:32,000 (145–147). The typical KS face includes long palpebral fissures associated with eversion of the lower third of the lower eyelid; long and heavy lashes giving the impression of made-up eyes; broad, arched and interrupted eyebrows; broad, depressed nasal tip; and prominent, cupped ears (145,146,148). Non-facial anomalies include mild to moderate cognitive impairment, visceral malformations, skeletal dysplasia and immunological manifestations (149). KS has been described in all ethnic groups (150,151). More than 80% of KS patients have a pathogenic variant in the coding regions of *KMT2D* (KS type 1, KS1, OMIM147920), and around 10% of patients have a pathogenic variant in the *KDM6A* gene (KS type 2, KS2, OMIM300128) (152–156). The aim of this study was to test our model for KS diagnosis and distinguish KS1 from KS2.

Ranging between 1998 and 2023, we included 1448 frontal and lateral facial photographs, corresponding to 634 patients. The control group comprised 1084 photographs, corresponding to 527 patients. The KS group comprised 364 photographs, corresponding to 107 patients. 82 (78%) of patients had a variation in the *KMT2D* gene (KS1), 23 (22%) in the *KDM6A* gene (KS2).

We confirmed the usual characteristics described in KS: high and arched eyebrows, long palpebral fissures, and large and prominent ears. The facial phenotype seemed more severe in patients under 3 years of age than in older patients. We were able to distinguish KS vs controls in the independent validation group with an accuracy of 95.8% (78.9 - 99.9%, $p < 0.001$). AUCs were comparable in the training set (0.994) and in the validation set (0.993) (Figure 14).

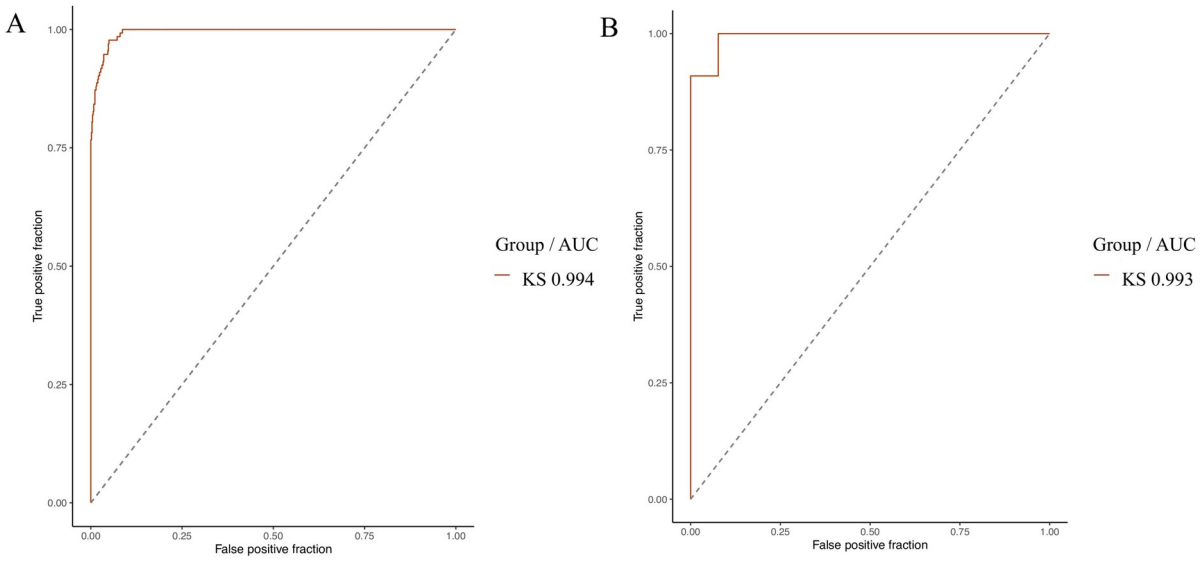


Figure 14. A. Empirical ROC curves (training set) for KS with AUC in design №1. B. ROC curves (validation set) for KS with AUC in design №1. AUC = Area Under the Curve, KS = Kabuki Syndrome.

Ten out of eleven patients were correctly predicted as KS with our model, and this performance was the same using Face2Gene CLINIC. In addition, we were able to predict all control patients (Table 9).

		Reference	
		Control	**KS**
Prediction	**Control**	**13**	1
	KS	0	**10**

Table 9. Confusion matrix for design №1 (KS versus controls) in the validation group. AUC = Area Under the Curve. Bold values: True Positives (TP).

KS2 patients had a rounder face, a shorter nose, a thicker upper lip, anteverted nostrils, and a shorter midface. There was no obvious difference in the eyebrows and eyes. The external ears were more elongated vertically in KS2, with a hypoplastic lobe, and with a counterclockwise rotation. The conch seemed more vertical in KS1. The model was able to distinguish KS1 from KS2 with an empirical AUC of 0.805 (0.729 - 0.880, $p < 0.001$) (Figure 15).

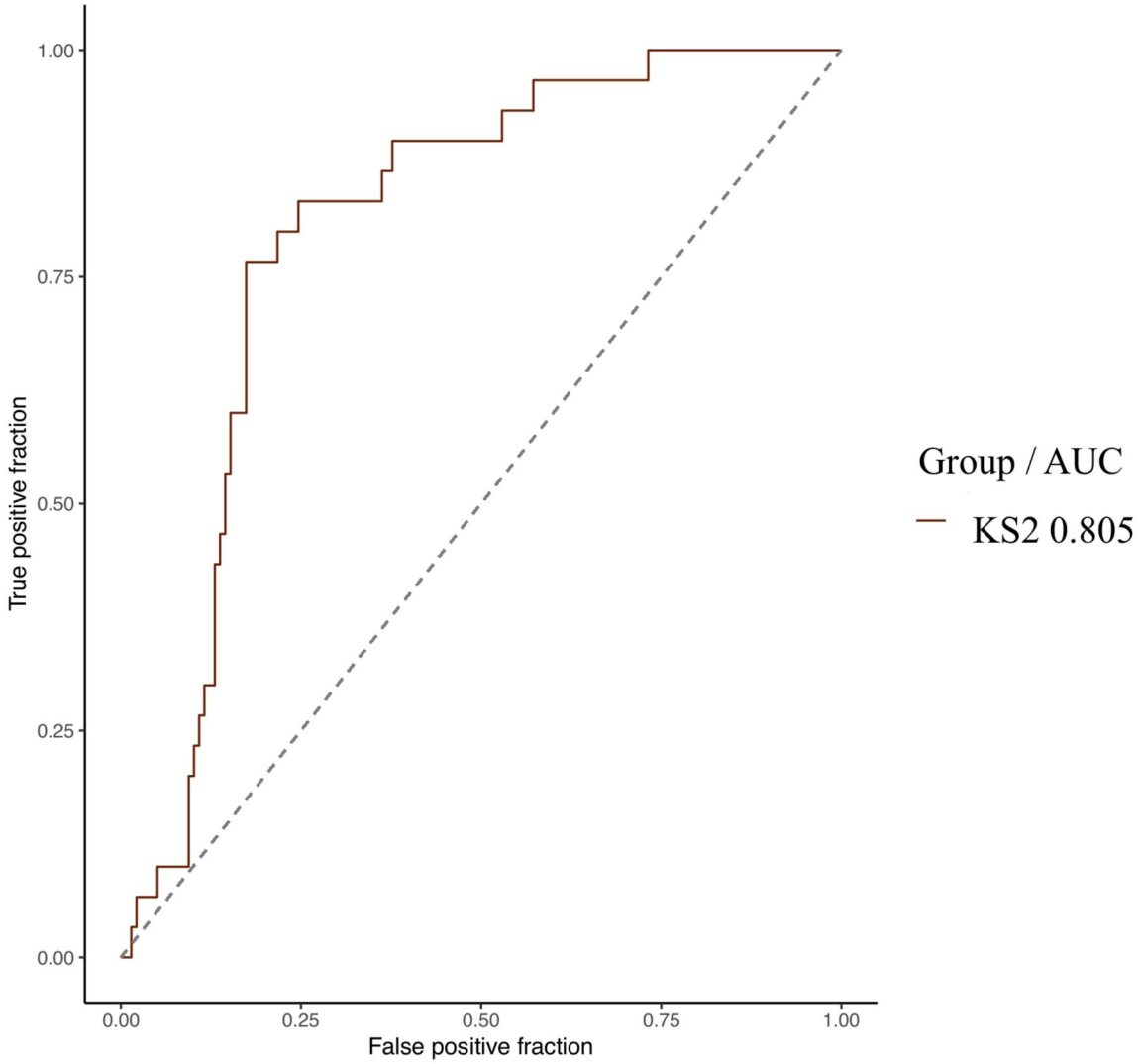


Figure 15. Empirical ROC curve (training set) for KS2 with AUC in design 2. AUC = Area Under the Curve, KS = Kabuki Syndrome.

Rouxel et al (149) showed that the Face2Gene RESEARCH tool distinguished KS1 from KS2 in a cohort of 66 patients with an AUC of 0.722 (p = 0.022). The same team showed a classification accuracy of 61% (20/33) by clinical genetics experts between KS1 and KS2. The performance of our model was at least comparable to Face2Gene RESEARCH and seemed to outperform that of clinical experts.

Rouxel et al (149) explained that KS1 patients had a longer face and nose, a thin upper lip vermilion and a longer midface than KS2 patients, who have a rounder face, a thicker vermilion and anteverted nostrils. Our study reports new phenotypic features not seen on frontal images alone for KS2, such as a particular morphology of the external ear, longer along the vertical axis and with counterclockwise rotation.

The model was unable to detect a difference in facial phenotype between KS1 patients with a Protein-Truncating Variant (PTV) compared to patients with a Protein-Altering Variant (PAV) (0.555 [0.419 – 0.690], $p = 0.786$).

Next Generation Phenotyping for diagnosis and phenotype - genotype correlations in Kabuki syndrome

Submitted to Genetics in Medicine, october 2023

Quentin **Hennocq**[1,2,3*], Marjolaine **Willems**[4], Jeanne **Amiel**[1,5], Stéphanie **Arpin**[6], Tania **Attie-Bitach**[1,5], Thomas **Bongibault**[1,3], Thomas **Bouygues**[1,3], Valérie **Cormier-Daire**[1,5], Pierre **Corre**[7], Klaus **Dieterich**[8], Maxime **Douillet**[1], Jean **Feydy**[9], Eva **Galliani**[2], Fabienne **Giuliano**[10], Stanislas **Lyonnet**[1,5], Arnaud **Picard**[2], Thantrira **Porntaveetus**[11], Marlène **Rio**[1,5], Flavien **Rouxel**[4], Vorasuk **Shotelersuk**[12,13], Annick **Toutain**[6], Kevin **Yauy**[4], David **Geneviève**[4**], Roman H. **Khonsari** [1,2,3**], Nicolas **Garcelon**[1]

[1]Imagine Institute, INSERM UMR1163, 75015 Paris, France

[2]Service de chirurgie maxillo-faciale et chirurgie plastique, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Centre de Référence des Malformations Rares de la Face et de la Cavité Buccale MAFACE, Filière Maladies Rares TeteCou ; Faculté de Médecine, Université de Paris Cité ; 75015 Paris, France

[3]Laboratoire 'Forme et Croissance du Crâne', Faculté de Médecine, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris, Université Paris Cité, Paris, France.

[4]Montpellier University, Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, Génétique clinique, CHU Montpellier, Centre de référence anomalies du développement SOOR, INSERM U1183, Montpellier, France.

[5]Service de médecine génomique des maladies rares, Hôpital Necker – Enfants Malades, Assistance Publique – Hôpitaux de Paris ; Faculté de Médecine, Université de Paris Cité ; 75015 Paris, France

[6]Service de Génétique, CHU Tours, UMR 1253, iBrain, Université de Tours, Inserm, Tours, France.

[7]Nantes Université, CHU Nantes, Service de chirurgie maxillo-faciale et stomatologie, F-44000 Nantes, France ; Nantes Université, Oniris, UnivAngers, CHU Nantes, INSERM, Regenerative Medicine and Skeleton, RMeS, UMR 1229, F-44000 Nantes, France.

[8]Univ. Grenoble Alpes, Inserm, U1209, IAB, CHU Grenoble Alpes, 38000, Grenoble, France.

[9]HeKA team, INRIA, Paris 75012, France.

[10]Division of Genetic Medicine, Department of Medicine, Lausanne University Hospital, Lausanne, Switzerland.

[11]Center of Excellence in Genomics and Precision Dentistry, Department of Physiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand.

[12] Excellence Center for Genomics and Precision Medicine, King Chulalongkorn Memorial Hospital, the Thai Red Cross Society, Bangkok, 10330, Thailand

[13]Center of Excellence in Genomics and Precision Dentistry, Department of Physiology, Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand.

Abstract. Introduction: Kabuki syndrome (KS) is a rare genetic disorder with a well recognizable facial phenotype. More than 80% of KS patients have a pathogenic variant in the coding regions of *KMT2D* (KS type 1, KS1), and around 10% of patients have a pathogenic variant in the *KDM6A* gene (KS type 2, KS2). Improving diagnosis in clinical genetics is a crucial challenge in reducing diagnostic wandering. The field of dysmorphology is changing with Artificial Intelligence (AI) and the development of Next Generation Phenotyping (NGP). The aim of this study was to propose a new NGP model for predicting KS on 2D facial photographs and distinguish KS1 from KS2. **Material and methods:** We included retrospectively and prospectively, from 1998 to 2023, all frontal and lateral pictures of patients with a molecular confirmation of KS. After Region Of Interest (ROI) detection using a Faster RCNN (Faster Region-based Convolutional Neural Network) and automatic placement of landmarks using a patch based AAM (Active Appearance Model), we extracted geometric features using Procrustes superimposition, and textural features using a Gray-Level Co-occurrence Matrix (GLCM). After a dimension reduction step using Principal Component Analysis (PCA) and incorporation of metadata such as age, gender, and ethnicity, we used XGboost (eXtreme Gradient Boosting), a supervised machine learning classifier. The model was tested on an independent validation set of genetically confirmed KS and controls using accuracies and Area Under the Curves (AUC). We then tested for belonging to KS1 and KS2 groups. Finally, we compared the performances of this NGP model with DeepGestalt (Face2Gene), a current commercially available AI-based diagnostic tool. **Results:** The study included 1448 frontal and lateral facial photographs from 6 centers, corresponding to 634 patients (527 controls, 107 KS); 82 (78%) of patients had a variation in the *KMT2D* gene (KS1), 23 (22%) in the *KDM6A* gene (KS2). We were able to distinguish KS from controls in the independent validation group with an accuracy of 95.8% (78.9 - 99.9%, $p < 0.001$), and distinguish KS1 from KS2 with an empirical AUC of 0.805 (0.729 - 0.880, $p < 0.001$). We found no facial shape difference between KS1 patients with a protein-truncating variant (PTV) vs a protein-altering variant (PAV). **Conclusion:** We report an automatic detection model for KS that considers the face, profile and ears, with high performances (AUC 0.993 and accuracy 95.8%). We were able to separate patients with KS1 (*KMT2D*) from KS2 (*KDM6A*), with an AUC of 0.805. These results outperform the current commercial AI-based solutions and expert clinicians. **Keywords:** Next Generation Phenotyping, DeepGestalt, Kabuki syndrome, *KMT2D*, *KDM6A*, dysmorphology.

Introduction. Kabuki syndrome (KS) is a rare genetic disorder, with an estimated prevalence of 1:86,000 to1:32,000 [1–3]. The typical KS face includes long palpebral fissures associated with eversion of the third part of the lower eyelid; long and heavy lashes giving the impression of made-up eyes; broad, arched and interrupted eyebrows; broad, depressed nasal tip; and prominent, cupped ears [1,2,4]. Non-facial features include mild to moderate intellectual disability, visceral malformations, skeletal dysplasia and immunological manifestations [5]. KS has been described in all ethnic groups [6,7]. More than 80% of KS patients have a pathogenic variant in the coding regions of *KMT2D* (KS type 1, KS1, OMIM147920), and around 10% of patients have a pathogenic variant in the *KDM6A* gene (KS type 2, KS2, OMIM300128) [8–12]. Improving diagnosis in clinical genetics is a crucial challenge in reducing diagnostic wandering. In France, the 7,000 rare diseases identified to date represent 4.5% of the population, half of which affect children under the age of 5 with 10% of deaths between 0-5. Around 50% of patients are not diagnosed, and for the remaining 50%, diagnostic wandering reaches an average of 5 years [13]. Diagnostic wandering is defined by the failure to define the precise cause of a disease after having performed all available investigations. Applications of AI (Artificial Intelligence) are increasing in healthcare [14–17]. The field of dysmorphology has been transformed by these new methods, under the name of Next Generation Phenotyping (NGP) [18]. Publications comparing human performances to NGP are flourishing [19–22], and some suggest that digital tools do it better than human experts in terms of diagnosis: Dudding-Byth et al [23] showed a better performance of NGP compared to clinicians in a group of ten genetic syndromes, not including KS; Rouxel et al [5] compared the performance of the DeepGestalt technology [18] using the Face2Gene online tool (FDNA Inc. Boston, MA, USA) to the performances of clinicians trained in the recognition of KS1 and KS2. The aim of this study was to develop a NGP model for KS diagnosis and distinguishing KS1 from KS2. We trained and validated the model on a large national and international multi-centric cohort of patients of all ages and ethnicities. The unique specificity of this approach was the integration of full frontal face, including the shape of the cranial vault and the position of the ears, as well as the frontal face and the morphology of the external ear.

Materials and methods. The study was approved by the CESREES (Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé, №4570023bis), the CNIL (Commission Nationale Informatique et Libertés, №MLD/MFI/AR221900), the Institutional Review Board, Faculty of Medicine, Chulalongkorn University (IRB 264/62), and in accordance with the 1964 Helsinki declaration and its later amendments. Informed and written consents were obtained from the legal representatives of each child or from the patients themselves if they were of age.

Photographic dataset

We included most pictures from the photographic database of the Maxillofacial surgery and Plastic surgery department of *Hôpital Necker – Enfants Malades* (Assistance Publique – Hôpitaux de Paris), Paris, France. This database contains 594,000 photographs from 22,000 patients, and all pictures since 1995 were taken by a professional medical photographer using a Nikon D7000 device in standardized positions.

We included retrospectively and prospectively, from 1995 to 2023, all frontal and lateral pictures of patients diagnosed with KS. The photographs were not calibrated. All patients had genetic confirmation of KS (*KMT2D* or *KDM6A*). We excluded patients with a history of craniofacial surgery. Multiple photographs per patient corresponded to different ages of follow-up. Duplicates were excluded.

Controls were selected among patients admitted for lacerations, trauma, infection and various skin lesions, without any record of chronic conditions. More precisely, follow-up for any type of chronic disease was considered as an exclusion criterion. The reports were retrieved using the local data warehouse Dr Warehouse [24]. For each patient, the best lateral view was included.

Data from five other medical genetics departments were also included according to the same criteria: (1) Montpellier University hospital, (2) Grenoble University hospital, (3) Tours University hospital, (4) King Chulalongkorn Memorial Hospital Bangkok, Thailand, and (5) Lausanne University Hospital, Lausanne, Switzerland.

Validation set

For designs №1 and №2, we randomly selected a group of individuals corresponding to 10% of the number of patients with KS, and the equivalent number of control patients. These patients were removed from the training set. The two sets were therefore independent.

Landmarking

We used three different templates based on 105 landmarks for the frontal views, 73 for the lateral views and 41 for the external ear pictures. We developed an automatic annotation model for each template following a pipeline including: (1) detection of the Region Of Interest (ROI) and (2) automatic placement of the landmarks.

For ROI detection, a Faster RCNN (Faster Region-based Convolutional Neural Network) model was trained after data augmentation (images and their +10° and -10° rotations), with a learning rate of 0.001, a batch size of 4, a gamma of 0.05 and 2000 iterations, optimized and split into two stages: ROI detection and determination of profile laterality.

(1) *ROI detection:* Faster RNN trained on 15633 images, after data augmentation (images and their +10° and -10° rotations): 6186 frontal images (2062 x 3) and 9447 right and left profile images (3159 x 3). The batch size was 2, learning rate was 0.0025 and the maximum number of iterations was 2800.

(2) *Determination of profile laterality:* Pre-trained ResNet50 network [25] using the Pytorch library [26]. The training images included 1,570 left profiles and 1,579 right profiles. The batch size was 16, an Adam optimizer [27] was used with a learning rate of 0.001, a step of 7, a gamma of 0.1, trained over 25 epochs.

For the automatic placement of landmarks, we used a patch-based AAM (Active Appearance Model) using the *menpo* library on Python 3.7 [28]. We have previously reported the relevance of this approach [29]. We used two-scale landmarking*:* the model for frontal pictures was trained on 904 manually annotated photographs, with a first stage of dimensioning (diagonal = 150), a patch shape of [(15, 15), (23, 23)] and 50 iterations, and a second stage without resizing, with a patch shape of [(20, 20), (30, 30)] and 10 new iterations. The model for profile pictures was trained on 1,439 manually annotated photographs, with a first stage of dimensioning (diagonal = 150), a patch shape of [(15, 15), (23, 23)] and 25 iterations, and a second stage without resizing, with a patch shape of [(15, 15), (23, 23)] and 5 new iterations. The model for ears was trained on 1221 manually annotated photographs, with a first stage of dimensioning (diagonal = 100), a patch shape of [(15, 15), (23, 23)] and 50 iterations, and a second stage without resizing, with a patch shape of [(20, 20), (30, 30)] and 20 new iterations. All three models used the Lucas Kanade optimizer [30].

Each automatically annotated photograph was checked by two authors blinded for the diagnosis, QH and MD, and landmarks were manually re-positioned when necessary, using *landmarker.io* [31]. The ICC (Intraclass Correlation Coefficient) was computed between the raters. ICC values greater than 0.9 corresponded to excellent reliability of the manual annotation [32].

Geometric morphometrics

We performed Generalized Procrustes Analysis (GPA) [33] on all landmark clouds using the *geomorph* package on R [34]. Since the data were uncalibrated photographs, ROI sizes were not available: shape parameters only were assessed and not centroid sizes. Procrustes coordinates were processed using Principal Component Analysis (PCA) for dimension reduction. We retained the principal components explaining 99% of the total variance in cumulative sum. The last 1% was considered as negligible information.

Texture extraction

We partitioned the frontal and profile pictures into key areas and applied textural feature extraction methods to each zone, allowing to check the results and determine which zone had contributed most to the diagnosis.

We defined 14 key areas that could potentially contribute to diagnosis: 11 on frontal views (right/left eyes, right/left eyebrows, glabella, forehead, nasal tip, philtrum, right/left cheeks, chin) and 3 on lateral views (pre-auricular region, eye, zygoma relief). Each zone was extracted automatically using the previously placed landmarks.

We used the CLAHE (Contrast Limited Adaptative Histogram Equalization) algorithm for histogram equalization, as previously reported before the use of feature extractors [35,36]. CLAHE enhanced contrast by evenly dispersing gray values [37] and by reducing the influences of illumination during picture capture and of skin color. Kiflie et al recommended CLAHE as a first choice equalization method [38].

Gray-level co-occurrence matrix (GLCM) methods, as proposed by Haralick [39], are based on the estimation of the second-order joint conditional probability density functions, which characterize the spatial relationships between pixels. GLCM is commonly used in texture analysis [40,41], for instance in radiomics on CT-scan or MRI images [42–44] or for skin texture assessment [45].

In GLCM, the co-occurrence matrix contains information on entropy, homogeneity, contrast, energy and correlation between pixels. GLCM includes 28 features, taking into account the average and range for each item of information and for each zone, representing 28 x 14 = 394 textural features for each patient.

Stratification using metadata

The textural features and the geometric principal components were combined for further analysis. To consider associated metadata (age and gender) and the fact that we included more than one photograph per patient (that is the non-independence of the data), a mixed model was designed for each feature. The variables to be explained were the features (geometric and textural), with age, gender and ethnicity considered as explanatory variables. A random effect on age and individuals was introduced. The equation of the mixed model was:

$$\boldsymbol{Features_{i,j}} \sim \alpha + age.\beta_1 + gender.\beta_2 + ethnicity.\beta_3 + age.\beta_{1,i} + \varepsilon_{i,j}$$

where $age.\beta_{1,i}$ corresponded to a random slope for age per individual, and $\varepsilon_{i,j}$ was a random error term. We did not use an interaction term between age and gender and age and ethnicity as it did not increase the likelihood of the model. Age, gender and ethnicity are significant factors in dysmorphology [46,47].

The residuals of each feature were computed to consider potential biases linked to the metadata:

$$\boldsymbol{\varepsilon_{i,j}} = \boldsymbol{Features_{i,j}} - \alpha + age.\beta_1 + gender.\beta_2 + ethnicity.\beta_3 + age.\beta_{1,i}$$

Classification model

The inputs to the model were the residuals from the linear models described above, for each geometric or textural feature. We used XGBoost (eXtreme Gradient Boosting), a supervised machine learning classifier, for all the analyses [48]. We chose a tree-based booster, and the loss function to be minimized was a logistic regression for binary classification. We set several hyperparameters to improve the performance and effect of the machine learning model: learning rate = 0.3, gamma = 0, maximum tree depth = 6. The model with the lowest error rate was chosen for analysis. We separated the dataset into a training set and a testing set, and a 5-fold cross-validation was used to define the ideal number of iterations to avoid overfitting.

The chosen model with the ideal number of iterations was then used on the independent validation set to test performances, by plotting accuracy and AUC. The ROC (Receiver Operating Characteristics) curves were plotted in R using the *plotROC* package [49]. We used the DeepGestalt tool proposed by Face2Gene CLINIC on our validation set, to be able to compare its performance (accuracies).

Uniform Manifold Approximation and Projection (UMAP) representations

The residuals $\varepsilon_{i,j}$ were represented using UMAP for visual clustering, a nonlinear dimension reduction technique [50]. We retained the residuals associated with features with a classification gain (in their cumulative sum) > 0.75 in the importance matrix associated with the XGboost model. A k (local neighborhood size) value of 15 was used. A cosine metric was introduced to compute distances in high dimensional spaces: the effective minimal distance between embedded points was 10^{-6}. The three conditions of UMAP, namely uniform distribution, local constancy of the Riemannian metric and local connectivity were verified. UMAP analyses were performed using the package *umap* on R [51] (Figure 1).

Classification designs

1) *Design №1*, syndrome diagnosis support: KS was tested against controls in a binary classification.
2) *Design №2*, genotype-phenotype correlations: KS1 and KS2 were tested in binary classifications.
3) *Design №3*, genotype-phenotype correlations: KS1 Protein-Altering Variants (PAVs) and Protein-Truncating Variants (PTV) were tested in binary classifications.

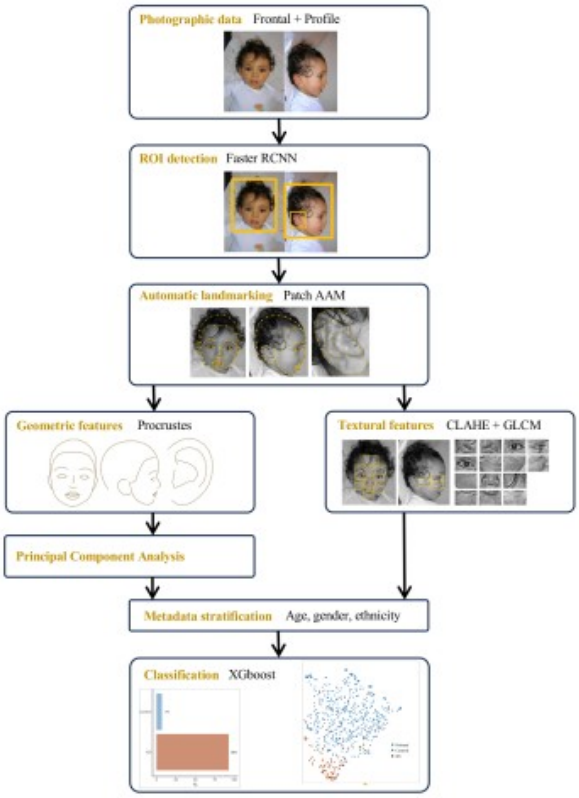


Figure 1. **Analysis pipeline, from the initial photograph to diagnostic probability.** Faster RCNN = Faster Region-based Convolutional Neural Network; CLAHE = Contrast Limited Adaptative Histogram Equalization; GLCM = Gray-level Co-occurrence Matrix; XGboost = eXtreme Gradient Boosting.

Results

Population description

Ranging between 1998 and 2023, we included 1448 frontal and lateral facial photographs, corresponding to 634 patients. The mean age was 7.2 +/- 4.2 and ranged from 0 to 40.2 years; 50% were girls. Ethnicity was 91% Caucasian, 6% African or Caribbean, and 3% Asian.

The control group comprised 1084 photographs, corresponding to 527 patients with a mean age of 7.0 +/- 4.6 years. Fifty-four percent were girls and ethnicities were 93% Caucasian, 5% African / Caribbean, and 2% Asian.

The KS group comprised 364 photographs, corresponding to 107 patients with a mean age of 7.8 +/- 6.7 years. Forty-one percent were girls and ethnicities were 83% Caucasian, 9% African / Caribbean, and 8% Asian. Eighty percent of patients were KS1 (Table 1).

Two patients had a genetically confirmed diagnosis of KS, but we had no information on the causal gene. We thus collected information on genetic variation for 105 KS individuals with 82 (78%) and 23 (22%) with variations in *KMT2D* (KS1) and *KDM6A* (KS2) respectively.

In the KS1 group, 74% of variants were PTVs, with 49% nonsense variants leading to a premature stop codon (24% non-sense, 24% frameshift) and 26% splice donor site variants. Eighteen percent were PAVs, with 17% missense variants and 1% in-frame indel.

In the KS2 group, 78% of variants were PTVs, with 43% nonsense variants leading to a premature stop codon (30% non-sense, 13% frameshift), 30% splice donor site variants and 4% a large deletion. Nine percent were missense PAVs (Supp. table 2).

		Total	Controls	KS
N				
	Consultations	724	542 (75%)	182 (25%)
	Photographs	1448	1084 (75%)	364 (35%)
	Patients	634	527 (83%)	107 (17%)
Gender				
	Female	365 (50%)	291 (54%)	74 (41%)
	Male	359 (50%)	251 (46%)	108 (59%)
Age (years)				
	Mean +/- SD	7.2 +/- 4.2	7.0 +/- 4.6	7.8 +/- 6.7
	Median	6.8	7.1	6.0
	Minimum	0	0.1	0
	Maximum	40.2	22.1	40.2
Ethnicity				
	African / Caribbean	45 (6%)	29 (5%)	16 (9%)
	Asian	25 (3%)	10 (2%)	15 (8%)
	Caucasian	654 (91%)	503 (93%)	151 (83%)
Genetic variation				
	KMT2D **(KS1)**			146 (80%)
	KDM6A **(KS2)**			35 (20%)

Table 1. **Clinical description of the cohort.** SD = Standard Deviation, *KMT2D* = Lysine (K)-specific methyltransferase 2D, *KDM6A* = Lysine (K)-specific demethylase 6A.

Design №1 : KS vs controls

1) *Phenotype*

We confirmed the usual characteristics described in KS: high and arched eyebrows, long palpebral fissures, and large and prominent ears.

1) *Classification*

We were able to distinguish KS vs controls in the independent validation group with an accuracy of 95.8% (78.9 - 99.9%, $p < 0.001$). AUCs were comparable in the training set (0.994) and in the validation set (0.993) (Figure 3).

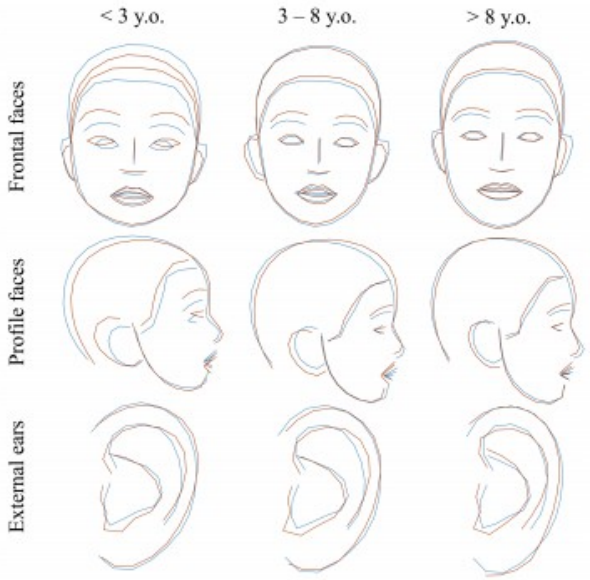


Figure 2. **Average shapes in KS and controls, and comparisons after Procrustes superimposition of frontal views, profile views and external ears for three age groups.** Blue = controls, Dark red = KS.

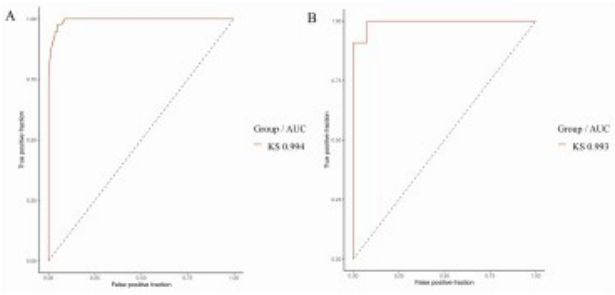


Figure 3. **A. Empirical ROC curves (training set) for KS with AUC in design №1. B. ROC curves (validation set) for KS with AUC in design №1.** AUC = Area Under the Curve, KS = Kabuki Syndrome.

Ten out of eleven patients were correctly predicted as KS with our model, and this performance was the same using Face2Gene CLINIC (Supp. Table 1). In addition, we were able to predict all control patients (Figure 4, Table 4).

Design №1		
Accuracy	0.958 [0.789 - 0.999]	$p < 0.001$*
AUC	0.993 [0.974 - 1.000]	$p < 0.001$*
F1 score	0.963	

Table 3. **Classification performances for design №1 (KS vs controls) in the validation group.** AUC = Area Under the Curve. * = statistically significant ($p < 0.05$).

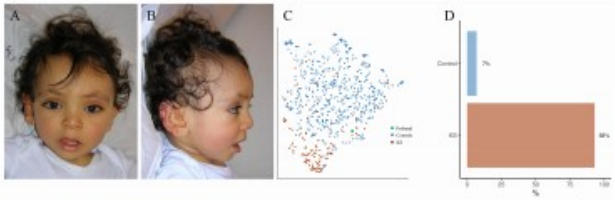


Figure 4. **Classification using design №1 for proband 3 of the validation set. A and B. Frontal and profile views of proband 3. C. UMAP representation of the training data according to the two groups, with positioning of proband 3. D. Histogram of predictions by the model.** This child was also detected as KS by Face2Gene CLINIC. KS = Kabuki Syndrome.

		Reference	
		Control	**KS**
Prediction	**Control**	**13**	1
	KS	0	**10**

Table 4. **Confusion matrix for design №1 (KS versus controls) in the validation group.** AUC = Area Under the Curve. Bold values: True Positives (TP).

Design №2 : KS1 vs KS2

1) *Phenotype*

KS2 individuals had a rounder face (HP:0000311), a shorter nose (HP:0003196), a thicker upper lip (HP:0000215), anteverted nostrils (HP:0000463), and a shorter midface (HP:0011800). There was no obvious difference in the eyebrows and eyes. The external ears were more elongated vertically in KS2 (HP:0400004), with a hypoplastic lobe (HP:0000385), and with a counter-clockwise rotation. The conch seemed more vertical in KS1.

1) *Classification*

The model was able to distinguish KS1 from KS2 with an empirical AUC of 0.805 (0.729 - 0.880, $p < 0.001$) (Figures 6, 7). This trend was found in the validation group, with an accuracy of 70% without reaching the significance threshold (Supp. Tables 2 and 3).

Design №3: PTV vs PAV in KS1

The model was unable to detect a difference in facial phenotype between KS1 patients with a PTV compared to patients with a PAV (0.555 [0.419 – 0.690], $p = 0.786$) (Figure 8).

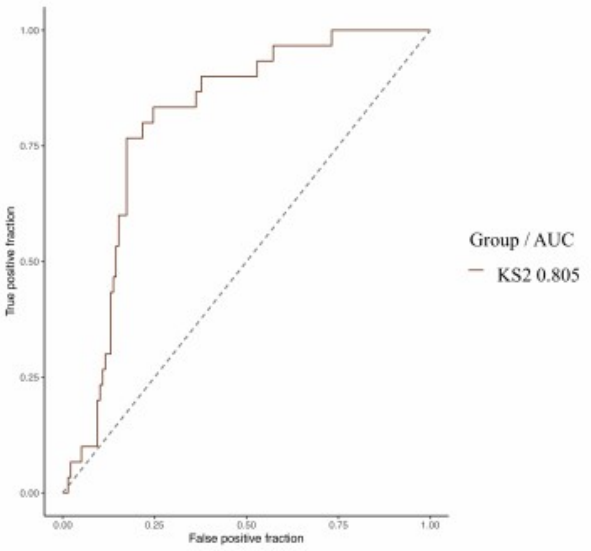


***Figure 6*. Empirical ROC curve (training set) for KS2 with AUC in design 2.** AUC = Area Under the Curve, KS = Kabuki Syndrome.

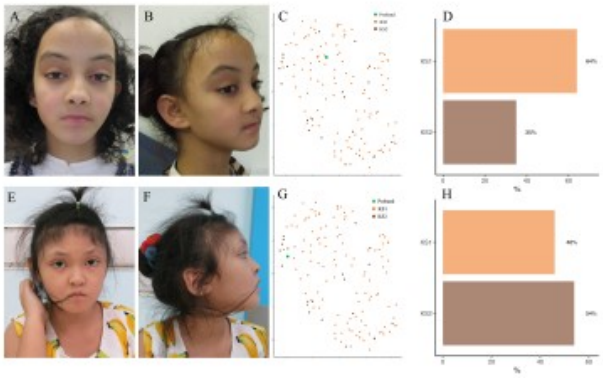


***Figure 7*. Classification using design №2 for two probands of the training set. A, B, E, F. Frontal and profile views of the two probands. C, G. UMAP representations of the training data according to the two groups, with positioning of probands 3. D, H. Histograms of predictions by the model.** The phenotype included a reduced height of the midface, a thicker upper lip, and a vertical elongation of the external ear in the KS2 group (E and F). KS = Kabuki Syndrome.

Discussion. The model we report distinguished KS from controls in the independent validation group with an accuracy of 95.8% (78.9 - 99.9%, $p < 0.001$). Only 1 patient out of 24 was classified as 'control' while she had KS (accuracy 96%). In the KS group, 10 out of 11 patients were correctly classified (accuracy 91%).

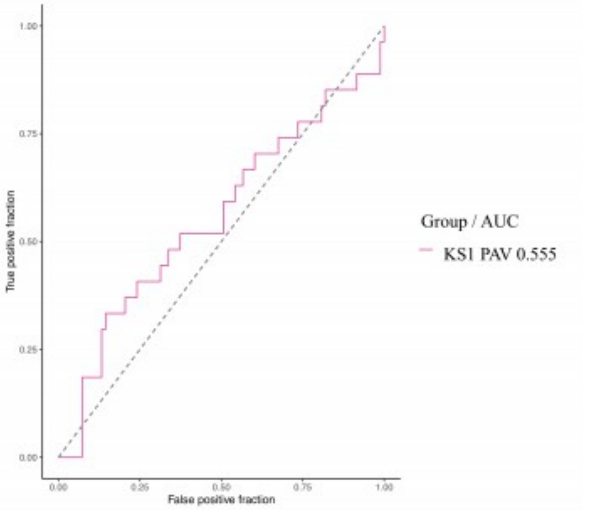


***Figure 8*. Empirical ROC curve (training set) for KS1 PAV with AUC in design №3.** AUC = Area Under the Curve, KS = Kabuki Syndrome, PAV = Protein-Altering Variant.

Using the Face2Gene CLINIC tool on KS patients (because DeepGestalt technology is not able of recognizing non-syndromic patients) 1 patient out of 11 could not be analyzed and could not be classified as KS (accuracy 91%). Performances were therefore comparable. Interestingly, the patient not recognized by our model and by Face2Gene CLINIC was of African ethnicity, highlighting the lack of training data for non-Caucasian patients.

The model we report was also able to distinguish KS1 from KS2 with an empirical AUC of 0.805 (0.729 - 0.880, $p < 0.001$). Rouxel et al [5] showed that the Face2Gene RESEARCH tool distinguished KS1 from KS2 in a cohort of 66 patients with an AUC of 0.722 ($p = 0.022$). The same team showed a classification accuracy of 61% (20/33) by clinical genetics experts between KS1 and KS2. The performance of our model was at least comparable to Face2Gene RESEARCH and seemed to outperform that of clinical experts.
Rouxel et al [5] explained that KS1 patients had a longer face and nose, a thin upper lip vermilion and a longer midface in comparison to KS2 patients, who have a rounder face, a thicker vermilion and anteverted nostrils. Our study reports new phenotypic features not seen on frontal images alone for KS2, such as a particular morphology of the external ear, longer along the vertical axis and with counter-clockwise rotation.

Phenotype-genotype correlation have been reported in KS for extra-facial anomalies. Cardiovascular abnormalities, namely ventricular septal defects, coarctation of the aorta, atrial septal defects, bicuspid aortic valve, patent ductus arteriosus, and hypoplastic left heart syndrome [52,53,53–55] are more prevalent in KS2 compared to KS1 [1,56]. Persistent hypoglycemia due to pituitary hormone deficiency, adrenal insufficiency, growth hormone deficiency and dysregulated insulin secretion by the pancreatic β-cells [57,58] are also more frequent in KS2 [10], probably because the inhibition of *KDM6A* increases the release of insulin from pancreatic islet cells in murine models [1,59]. Urinary tract anomalies, such as horseshoe kidneys and renal hypoplasia, seem to be more frequent in KS1, and genital defects such as cryptorchidism and hypospadias could be more frequent in KS2 [56,60,61]. Rouxel et al [5] underline the lack of Asian patients in their evaluation, and proposed that larger series were needed to better define phenotypical differences between KS1 and KS2, and the general dependance of the phenotype with ethnicity [6,12]. The collaboration with an Asian clinical genetics center (Bangkok) is thus a strong point of this study. The use of textural feature extraction allowed our model account for typical KS characteristics not recognized by geometric analysis (Procrustes) alone. The lateral sparsening of the eyebrows and heavy lashes giving the impression of make-up eyes were thus included into in the classification. In addition, a side effect of diazoxide [62–64] used to treat hyperinsulinism more often in KS2 is hirsutism. This side effect could have contributed to the classification between KS1 and KS2 via the analysis of textures. Barry et al [1] reported a large meta-analysis including 152 articles and 1369 individuals with KS and assessed the prevalence of the different types of pathogenic variation per gene. The majority of *KMT2D* variants were truncating (nonsense 34%, frameshift 34%), then missense (23%) and finally splice site variants (9%). The majority of *KDM6A* variants were truncating (frameshift 36% > nonsense 27%), followed by splice site (20%), and missense (18%). We found similar results, with a higher prevalence of truncating nonsense variants for both genes. There was a higher prevalence of splice donor site variants, with 26% for *KMT2D* and 30% for *KDM6A*. Some authors report a more severe clinical outcomes in patients with non-sense variants than in patients with a frameshift variant [1]. Faundes et al [56] found more severe neurodevelopmental anomalies in patients with protein-truncating mutations in the KS2 group.

Shah et al [65] reported ophthalmological anomalies such as strabismus, blue sclerae, microphthalmia and refractive anomalies that were more severe in patients with a nonsense variant, and less frequent in patients with a frameshift variant. Our model did not find any significant difference in facial phenotype between PTV and PAV.

Conclusion. Here we report an automatic detection model for KS including the face, profiles and ears, with performances (AUC 0.993 and accuracy 95.8%) comparable to those of Face2Gene, on an independent validation set. These performances were achieved using an international cohort of 107 patients with a confirmed molecular diagnosis of KS. Using the same model, we were able to separate patients with KS1 (*KMT2D*) from KS2 (*KDM6A*), with an AUC of 0.805. These results seem to at least outperform Face2Gene and support the possibility of using a phenotype-first strategy to diagnose KS and detect its two causal genes.

Funding

This work was supported by the 'Agence Nationale de la Recherche', 'Investissements d'Avenir' program (ANR-10-IAHU-01), France 2030 grant "Face4Kids" (ANR-21-PMRB-0004), Health Systems Research Institute (66-101, 66-122), Thailand Science Research and Innovation Fund Chulalongkorn University, and National Research Council of Thailand (N42A650229).

Conflict of interest statement

The authors declare they have no conflicts of interest.

References

1. Barry KK, Tsaparlis M, Hoffman D, et al. From Genotype to Phenotype—A Review of Kabuki Syndrome. *Genes (Basel)*. 2022;13(10):1761. doi:10.3390/genes13101761

2. Niikawa N, Kuroki Y, Kajii T, et al. Kabuki make-up (Niikawa-Kuroki) syndrome: a study of 62 patients. *Am J Med Genet*. 1988;31(3):565-589. doi:10.1002/ajmg.1320310312

3. White SM, Thompson EM, Kidd A, et al. Growth, behavior, and clinical findings in 27 patients with Kabuki (Niikawa-Kuroki) syndrome. *Am J Med Genet A*. 2004;127A(2):118-127. doi:10.1002/ajmg.a.20674

Kuroki Y, Suzuki Y, Chyo H, Hata A, Matsui I. A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. *J Pediatr*. 1981;99(4):570-573. doi:10.1016/s0022-3476(81)80256-9

5. Rouxel F, Yauy K, Boursier G, et al. Using deep-neural-network-driven facial recognition to identify distinct Kabuki syndrome 1 and 2 gestalt. *Eur J Hum Genet*. 2022;30(6):682-686. doi:10.1038/s41431-021-00994-8

6. Adam MP, Hudgins L. Kabuki syndrome: a review. *Clin Genet*. 2005;67(3):209-219. doi:10.1111/j.1399-0004.2004.00348.x

7. Bögershausen N, Gatinois V, Riehmer V, et al. Mutation Update for Kabuki Syndrome Genes KMT2D and KDM6A and Further Delineation of X-Linked Kabuki Syndrome Subtype 2. *Hum Mutat*. 2016;37(9):847-864. doi:10.1002/humu.23026

8. Lederer D, Grisart B, Digilio MC, et al. Deletion of KDM6A, a Histone Demethylase Interacting with MLL2, in Three Patients with Kabuki Syndrome. *Am J Hum Genet*. 2012;90(1):119-124. doi:10.1016/j.ajhg.2011.11.021

9. Paděrová J, Holubová A, Simandlová M, et al. Molecular genetic analysis in 14 Czech Kabuki syndrome patients is confirming the utility of phenotypic scoring. *Clin Genet*. 2016;90(3):230-237. doi:10.1111/cge.12754

10. Banka S, Lederer D, Benoit V, et al. Novel KDM6A (UTX) mutations and a clinical and molecular review of the X-linked Kabuki syndrome (KS2). *Clin Genet*. 2015;87(3):252-258. doi:10.1111/cge.12363

11. Bögershausen N, Wollnik B. Unmasking Kabuki syndrome. *Clin Genet*. 2013;83(3):201-211. doi:10.1111/cge.12051

12. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet*. 2010;42(9):790-793. doi:10.1038/ng.646

13. DGOS. Les maladies rares. Ministère de la Santé et de la Prévention. Published September 30, 2023. Accessed September 30, 2023. https://sante.gouv.fr/soins-et-maladies/prises-en-charge-specialisees/maladies-rares/article/les-maladies-rares

14. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi:10.1056/NEJMra1814259

15. Choy G, Khalilzadeh O, Michalski M, et al. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*. 2018;288(2):318-328. doi:10.1148/radiol.2018171820

16. Novoa RA, Gevaert O, Ko JM. Marking the Path Toward Artificial Intelligence-Based Image Classification in Dermatology. *JAMA Dermatol*. 2019;155(10):1105-1106. doi:10.1001/jamadermatol.2019.1633

17. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial Intelligence and Surgical Decision-making. *JAMA Surg*. 2020;155(2):148-158. doi:10.1001/jamasurg.2019.4917

18. Gurovich Y, Hanani Y, Bar O, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med*. 2019;25(1):60-64. doi:10.1038/s41591-018-0279-0

19. Zhang Q, Ding Y, Feng B, et al. Molecular and Phenotypic Expansion of Alström Syndrome in Chinese Patients. *Front Genet*. 2022;13:808919. doi:10.3389/fgene.2022.808919

20. Javitt MJ, Vanner EA, Grajewski AL, Chang TC. Evaluation of a computer-based facial dysmorphology analysis algorithm (Face2Gene) using standardized textbook photos. *Eye*. 2022;36(4):859-861. doi:10.1038/s41433-021-01563-5

21. Latorre-Pellicer A, Ascaso Á, Trujillano L, et al. Evaluating Face2Gene as a Tool to Identify Cornelia de Lange Syndrome by Facial Phenotypes. *Int J Mol Sci*. 2020;21(3):E1042. doi:10.3390/ijms21031042

22. Mishima H, Suzuki H, Doi M, et al. Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. *J Hum Genet*. 2019;64(8):789-794. doi:10.1038/s10038-019-0619-z

23. Dudding-Byth T, Baxter A, Holliday EG, et al. Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. *BMC Biotechnol*. 2017;17(1):90. doi:10.1186/s12896-017-0410-1

24. Garcelon N, Neuraz A, Salomon R, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. *J Biomed Inform*. 2018;80:52-63. doi:10.1016/j.jbi.2018.02.019

25. Koonce B. ResNet 50. In: ; 2021:63-72. doi:10.1007/978-1-4842-6168-2_6

26. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Published online December 3, 2019. doi:10.48550/arXiv.1912.01703

27. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Published online January 29, 2017. doi:10.48550/arXiv.1412.6980

28. Alabort-i-Medina J, Antonakos E, Booth J, Snape P, Zafeiriou S. Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM; 2014:679-682. doi:10.1145/2647868.2654890

29. Hennocq Q, Bongibault T, Bizière M, et al. An automatic facial landmarking for children with rare diseases. *American Journal of Medical Genetics Part A*. n/a(n/a). doi:10.1002/ajmg.a.63126

30. Lucas B, Kanade T. An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI). In: Vol 81. ; 1981.

31. landmarker.io · The Menpo Project. Accessed March 20, 2022. https://www.menpo.org/landmarkerio/

32. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19(1):3-11. doi:10.2466/pr0.1966.19.1.3

33. Rohlf FJ, Slice D. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. *Systematic Zoology*. 1990;39(1):40-59. doi:10.2307/2992207

34. Baken EK, Collyer ML, Kaliontzopoulou A, Adams DC. geomorph v4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution*. 2021;12(12):2355-2363. doi:10.1111/2041-210X.13723

35. Avcı H, Karakaya J. A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning. *Diagnostics (Basel)*. 2023;13(3):348. doi:10.3390/diagnostics13030348

36. Anifah L, Purnama IKE, Hariadi M, Purnomo MH. Osteoarthritis Classification Using Self Organizing Map Based on Gabor Kernel and Contrast-Limited Adaptive Histogram Equalization. *The Open Biomedical Engineering Journal*. 2013;7:18. doi:10.2174/1874120701307010018

37. Huang C, Li X, Wen Y. AN OTSU image segmentation based on fruitfly optimization algorithm. *Alexandria Engineering Journal*. 2021;60(1):183-188. doi:10.1016/j.aej.2020.06.054

38. Kiflie A, Tesema Tufa G, Salau AO. Sputum smears quality inspection using an ensemble feature extraction approach. *Front Public Health*. 2023;10:1032467. doi:10.3389/fpubh.2022.1032467

39. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973;SMC-3(6):610-621. doi:10.1109/TSMC.1973.4309314

40. Mohanaiah P, Sathyanarayana P, GuruKumar L. Image Texture Feature Extraction Using GLCM Approach. 2013;3(5).

41. Löfstedt T, Brynolfsson P, Asklund T, Nyholm T, Garpebring A. Gray-level invariant Haralick texture features. *PLoS One*. 2019;14(2):e0212110. doi:10.1371/journal.pone.0212110

42. Mundt P, Tharmaseelan H, Hertel A, et al. Periaortic adipose radiomics texture features associated with increased coronary calcium score-first results on a photon-counting-CT. *BMC Med Imaging*. 2023;23(1):97. doi:10.1186/s12880-023-01058-7

43. Adelsmayr G, Janisch M, Müller H, et al. Three dimensional computed tomography texture analysis of pulmonary lesions: Does radiomics allow differentiation between carcinoma, neuroendocrine tumor and organizing pneumonia? *Eur J Radiol*. 2023;165:110931. doi:10.1016/j.ejrad.2023.110931

44. Peng B, Wang K, Xu R, et al. Preoperative computed tomography-based tumoral radiomic features prediction for overall survival in resectable non-small cell lung cancer. *Front Oncol*. 2023;13:1131816. doi:10.3389/fonc.2023.1131816

45. Ou X, Pan W, Xiao P. In vivo skin capacitive imaging analysis by using grey level co-occurrence matrix (GLCM). *Int J Pharm*. 2014;460(1-2):28-32. doi:10.1016/j.ijpharm.2013.10.024

46. Muenke M, Adeyemo A, Kruszka P. An electronic atlas of human malformation syndromes in diverse populations. *Genet Med*. 2016;18(11):1085-1087. doi:10.1038/gim.2016.3

47. Burchard EG, Ziv E, Coyle N, et al. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med*. 2003;348(12):1170-1175. doi:10.1056/NEJMsb025007

48. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

49. Sachs MC. plotROC: A Tool for Plotting ROC Curves. *J Stat Softw*. 2017;79:2. doi:10.18637/jss.v079.c02

50. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Published online September 17, 2020. doi:10.48550/arXiv.1802.03426

51. R Core Team (2020). — European Environment Agency. Accessed January 18, 2023. https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006

52. Hughes HE, Davies SJ. Coarctation of the aorta in Kabuki syndrome. *Arch Dis Child*. 1994;70(6):512-514. doi:10.1136/adc.70.6.512

53. Digilio MC, Gnazzo M, Lepri F, et al. Congenital heart defects in molecularly proven Kabuki syndrome patients. *Am J Med Genet A*. 2017;173(11):2912-2922. doi:10.1002/ajmg.a.38417

54. Cheon CK, Ko JM. Kabuki syndrome: clinical and molecular characteristics. *Korean J Pediatr*. 2015;58(9):317-324. doi:10.3345/kjp.2015.58.9.317

55. Yoon JK, Ahn KJ, Kwon BS, et al. The strong association of left-side heart anomalies with Kabuki syndrome. *Korean J Pediatr*. 2015;58(7):256-262. doi:10.3345/kjp.2015.58.7.256

Faundes V, Goh S, Akilapa R, et al. Clinical delineation, sex differences, and genotype-phenotype correlation in pathogenic KDM6A variants causing X-linked Kabuki syndrome type 2. *Genet Med*. 2021;23(7):1202-1210. doi:10.1038/s41436-021-01119-8

57. Yap KL, Johnson AEK, Fischer D, et al. Congenital hyperinsulinism as the presenting feature of Kabuki syndrome: clinical and molecular characterization of 9 affected individuals. *Genet Med*. 2019;21(1):233-242. doi:10.1038/s41436-018-0013-9

58. Gole H, Chuk R, Coman D. Persistent Hyperinsulinism in Kabuki Syndrome 2: Case Report and Literature Review. *Clin Pract*. 2016;6(3):848. doi:10.4081/cp.2016.848

59. Gibson CE, Boodhansingh KE, Li C, et al. Congenital Hyperinsulinism in Infants with Turner Syndrome: Possible Association with Monosomy X and KDM6A Haploinsufficiency. *Horm Res Paediatr*. 2018;89(6):413-422. doi:10.1159/000488347

60. Courcet JB, Faivre L, Michot C, et al. Clinical and molecular spectrum of renal malformations in Kabuki syndrome. *J Pediatr*. 2013;163(3):742-746. doi:10.1016/j.jpeds.2013.02.032

61. Cetinkaya E, Misirlioğlu ED, Vidinlisan S, Baydar Z, Ozhan ZR. Hypospadias in a patient with Kabuki make-up (Niikawa-Kuroki) syndrome. *J Pediatr Endocrinol Metab*. 2001;14(6):803-805. doi:10.1515/jpem.2001.14.6.803

62. Souabni SA, Harvengt A, Legat C, Lysy PA. Congenital hyperinsulinemic hypoglycemia (HH) requiring treatment as the presenting feature of Kabuki syndrome. *Clin Case Rep*. 2023;11(6):e7336. doi:10.1002/ccr3.7336

63. Salguero MV, Chan K, Greeley SAW, et al. Novel KDM6A Kabuki Syndrome Mutation With Hyperinsulinemic Hypoglycemia and Pulmonary Hypertension Requiring ECMO. *J Endocr Soc*. 2022;6(4):bvac015. doi:10.1210/jendso/bvac015

64. Hoermann H, El-Rifai O, Schebek M, et al. Comparative meta-analysis of Kabuki syndrome with and without hyperinsulinaemic hypoglycaemia. *Clin Endocrinol (Oxf)*. 2020;93(3):346-354. doi:10.1111/cen.14267

65. Shah SS, Fulton A, Jabroun M, Brightman D, Simpson BN, Bodamer OA. Insights into the genotype-phenotype relationship of ocular manifestations in Kabuki syndrome. *Am J Med Genet A*. 2023;191(5):1325-1338. doi:10.1002/ajmg.a.63155

7 Results for the secondary objective: effect of a treatment on facial morphology

We will present three cases analyzing the effects of a medical or surgical treatment on facial morphology:

- Efficacy of alpelisib on the facial phenotype in *PIK3CA*-related hemifacial myohyperplasia
- Comparison of the nasal phenotype of two techniques for the surgical treatment of cleft lip and palate
- Comparison of two surgical techniques for the treatment of non-syndromic scaphocephaly

7.1 Hemifacial myohyperplasia is due to somatic muscular PIK3CA gain- of-function mutation and responds to pharmacological inhibition (Appendix 1)

Hemifacial myohyperplasia (HFMH) is a rare cause of facial asymmetry exclusively involving facial muscles, initially reported as 'hypertrophy and asymmetry of the facial muscles' (157). This disorder is reported in very few patients in the literature (36,158–161). The clinical presentation of HFMH patients is strikingly consistent, with unilateral muscular hypertrophy mimicking spasm and orofacial dystonia, leading to diagnostic errors and inadequate management strategies, including aggressive attempts of surgical correction (162). Currently the genetic causes of HFMH are unknown. The recent discovery of the role played by somatic mutation of genes activating the PI3KCA/AKT/mTOR pathway has opened new treatment perspectives for patients (163). Particularly, *PIK3CA* gain-of-function mutations explain most

overgrowth syndromes. The understanding of the genetic bases of overgrowth has allowed to treat patients presenting *PIK3CA*-related overgrowth syndromes (PROS) with a specific PI3K inhibitor (BYL719, alpelisib), initially designed as an anti-neoplastic drug (164). Alpelisib is efficiently tackling soft-tissue overgrowth in PROS (164–167). Bayard et al hypothesized that PIK3CA/AKT/mTOR pathway was abnormally affected in patients with HFMH (168). Here, we report the results of the clinical screening and treatment in 5 patients with HFMH. After 6 months, 2D photographs were performed for all patients. 2D photographs were analyzed as previously described. The pre-treatment facial phenotype was compared to the facial phenotype after a minimum of 6 months of treatment. An associated assessment using 3D photography was also conducted, as detailed in (168), but falls outside the score of this PhD.

Based on 2D photography quantification, we confirmed the lowering of the lip commissure, the widening of the palpebral fissure, the reduction of nose and chin deviation (Figures 16, 17). These results were published in 2023 (168), together with cellular, molecular, radiographic, and 3D photography data demonstrating the efficiency of alpelisib in treating HFMH. The results from 2D photography analyses were nevertheless removed from the final version of the article during the revision process as required by the reviewers of the Journal of Experimental Medicine (168).

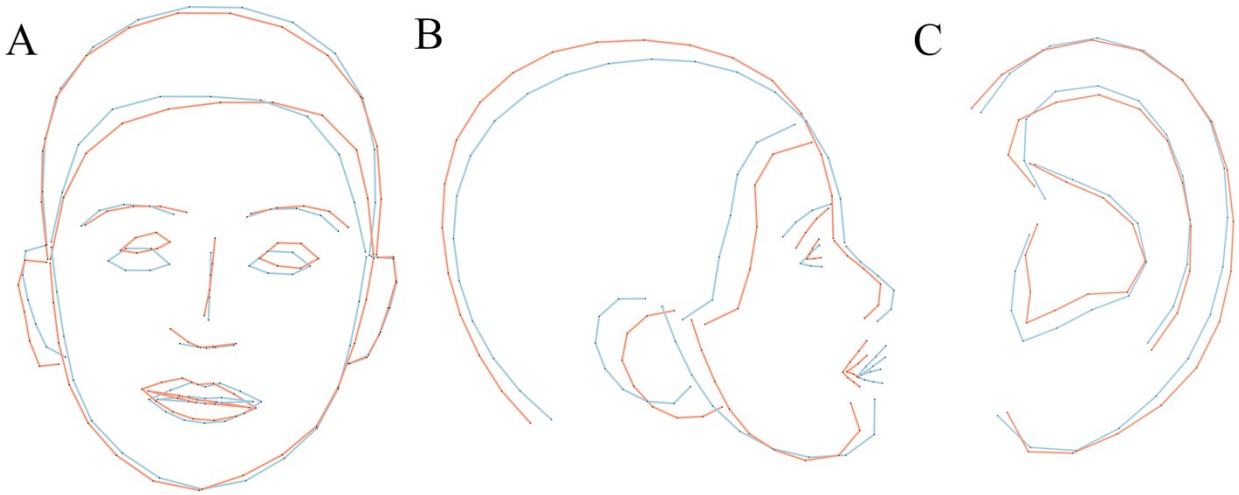


Figure 16. Phenotypical description of Hemifacial myohyperplasia (HFMH). Face (A), profile (B) and external ear (C) assessment for controls (blue) and patients with HFMH (red).

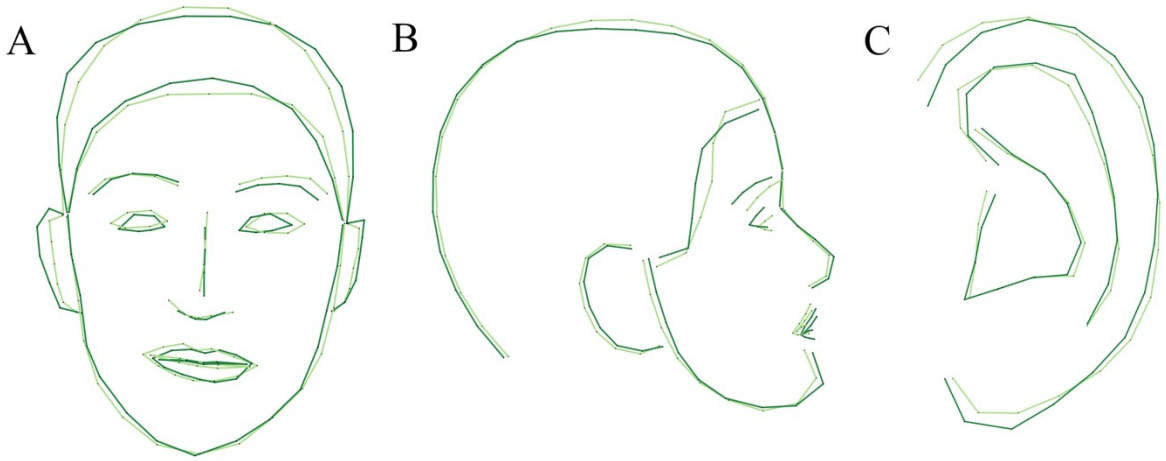


Figure 17. Alpelisib improves hemifacial myohyperplasia phenotype in patients. Face, profile, and external ear assessment based on 2D photographs of the 5 patients before treatment (light green) compared with presentation after 6 months of treatment (dark green).

7.2 Comparison of nasal symmetry in two surgical techniques for cleft lip and palate repair

The aim of the next study was to evaluate the short- and long-term results of two cleft lip and palate repair techniques (Delaire technique (5) vs. Talmant technique (6)) on the symmetry of the nose. We included full-face, low-angle photographs from two referral centers for cleft lip and palate surgery. For each patient, if available, we included a pre-operative photograph and two post-operative photographs, one less than 3 years after the operation (early post-

operative) and the second one more than 3 years after cleft repair (late post-operative). For each patient, we recorded the following additional data: age at surgery, gender, laterality of the cleft, complete / partial cleft, and the identity of the surgeon.

We used a new template based on 61 landmarks placed on both nostrils and the perimeter of the nasal pyramid. Landmarks were manually placed.

To ensure a uniform distribution of landmarks along the curves, anatomical landmarks were transformed into sliding semi-landmarks using the *geomorph* package on R (7). As performed previously to account for associated metadata (age and type of cleft) and the fact that we had included more than one photograph per patient (that is the non-independence of the data), a mixed model was designed for each principal component. The variable to be explained was PC, with age and type of cleft considered as explanatory variables. A random effect on age and individuals was introduced. The equation of the mixed model was:

$$PC_{i,j} \quad \alpha + age \,.\, \beta_1 + type\ of\ cleft \,.\, \beta_2 + \varepsilon_{i,j}$$

where $age \,.\, \beta_{1,i}$ corresponded to a random slope for age per individual and $\varepsilon_{i,j}$ was a random error term. We measured the directional asymmetry (DA) of the nasal dome and the asymmetry of the nostrils for each landmark cloud. The means and standard deviations of DA for each time point (pre-operative, early post-operative, and late post-operative) were measured and compared by Student's t tests, according to the surgical technique groups. We used a Linear Discriminant Analysis (LDA) to measure the classification accuracy between the two surgical techniques at different operative times. A 10-fold cross validation was used to measure the uncertainty. Accuracy and standard deviation at different times were reported.

We included a total of 290 photographs, corresponding to 74 patients, 47% in the Talmant group and 53% in the Delaire group. In the Talmant group, 23% were females, 63% were operated on by surgeon №1 and 37% by surgeon №2. 29% were right clefts, and 86% were complete. The mean age at surgery was 6.4 +/- 2.3 months. 24% were pre-operative

photographs, 44% photographs early post-operative, and 32% late post-operative. In the Delaire group, 18% were females, 51% were operated on by surgeon №1 and 49% by surgeon №2. 38% were right clefts, and 85% were complete. The mean age at surgery was 6.9 +/- 1.7 months. 25% were pre-operative photographs, 41% photographs early post-operative and 35% late post-operative (Table 10).

		Talmant	**Delaire**
N		35 (47%)	39 (53%)
N photos		139 (48%)	151 (52%)
Gender			
	Female	8 (23%)	7 (18%)
	Male	27 (73%)	32 (82%)
Surgeon			
	1	22 (63%)	20 (51%)
	2	13 (37%)	19 (49%)
Laterality			
	R	10 (29%)	15 (38%)
	L	25 (71%)	24 (62%)
Cleft type			
	Complete	30 (86%)	33 (85%)
	Incomplete	5 (14%)	6 (15%)
Age at surgery (months)			
	Mean +/- SD	6.4 +/- 2.3	6.9 +/- 1.7
	Median	6.0	6.6
	Min	3.0	4.7
	Max	17.0	10.2
Time after surgery			
Pre-op.		34 (24%)	37 (25%)
< 3 y.o – early post-op.		61 (44%)	61 (41%)
> 3 y.o – late post-op.		44 (32%)	53 (35%)

Table 10. Description of the cohort. R = Right, L = Left, SD = Standard Deviation.

Before surgery, we found a statistically equivalent asymmetry index between the Delaire and Talmant groups for the nasal dome (p = 0.599) and the nostrils (0.623).

Three years after surgery, the asymmetry index for the nasal dome was higher in the Delaire group (1.41 +/- 0.05 vs. 1.39 +/- 0.07) than in the Talmant group, but this result did not reach the significance threshold (p = 0.072). The asymmetry index for the nostrils was also higher in the Delaire group (1.48 +/- 0.15 vs. 1.44 +/- 0.09) than in the Talmant group, but this result failed to reach the significance threshold (p = 0.078).

After three years, theses asymmetry indexes were equivalent between the two groups for the nasal dome (p = 0.999). The asymmetry index was higher in the Delaire group for the nostrils (1.47 +/- 0.11 vs. 1.45 +/- 0.09) but this result was not statistically significant (p = 0.340) (Table XX).

These numerical results were coherent with the graphical results (Figure XX).

		Delaire	**Talmant**	*p*-value
Nose AI				
	Pre-op.	1.43 +/- 0.07	1.44 +/- 0.09	0.599
	Early post-op.	1.41 +/- 0.05	1.39 +/- 0.07	0.072
	Late post-op.	1.38 +/- 0.07	1.38 +/- 0.09	0.999
Nostrils AI				
	Pre-op.	1.71 +/- 0.07	1.70 +/- 0.10	0.623
	Early post-op.	1.48 +/- 0.15	1.44 +/- 0.09	0.078
	Late post-op.	1.47 +/- 0.11	1.45 +/- 0.09	0.340

Table 11. Comparisons of asymmetry indices for both techniques at different time points. AI = asymmetry index.

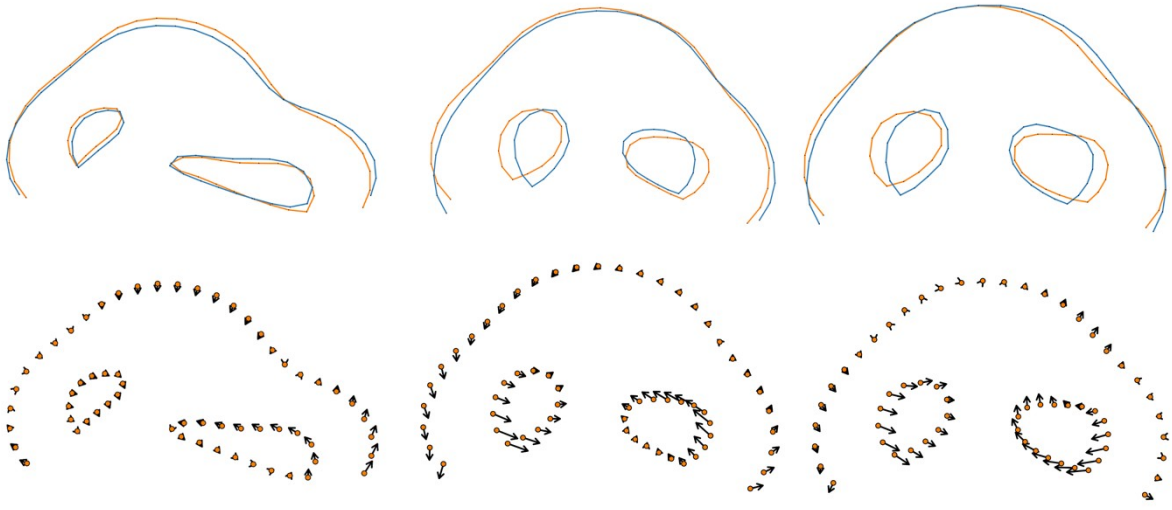


Figure 18. Comparison of average landmark positions for the three different time points, for both surgical methods (Talmant = blue, Delaire = orange) after Procrustes superimposition.

We found a non-significant accuracy of 0.649 +/- 0.178 ($p = 0.386$) in the prediction of the group before surgery. This accuracy became significant in the early post-operative period (< 3 years), with a value of 0.731 +/- 0.112 ($p = 0.023$), and in the late post-operative period (> 3 years), with a value of 0.742 +/- 0.129 ($p = 0.018$).

7.3 Pre- and post-operative phenotyping of non-syndromic isolated scaphocephaly

This section is dedicated to a condition that falls within the category of non-syndromic craniostenoses, i.e., due to the premature closure of a single cranial suture. Scaphocephaly is the most common isolated craniosynostosis (169), linked to the premature closure of the sagittal suture. One of the aims of surgery is morphological, to restore a normal skull shape. In this study, we compared a population of patients with non-syndromic scaphocephaly before and after cranioplasty with a population of control patients.

We also compared the phenotypes of patients operated on using two surgical methods: Renier H cranioplasty and total vault remodeling (TVR). Retrospective data from 1979 to 2023 on patients diagnosed with non-syndromic scaphocephaly were considered.

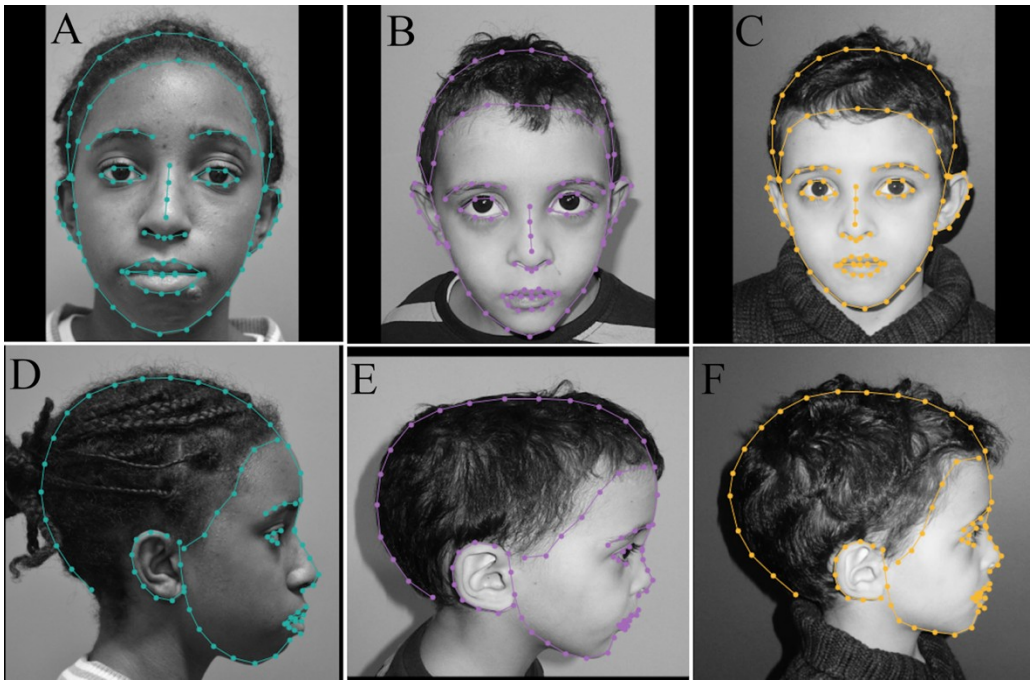


Figure 19. Landmarking for faces (A, B, C) **and profiles** (D, E, F), in a control subject (A and D), and in a child with non-syndromic scaphocephaly, before (B and E) and after cranioplasty (C and F).

We collected 1,720 photographs of the face and profile, corresponding to 860 consultations and 742 patients, between 1979 and 2022: 1,016 (59%) photographs of control patients and 704 (41%) photographs of patients with non-syndromic scaphocephaly, including 466 (27%) pre-operative, 160 (9%) post-H-cranioplasty and 78 post-TVR (5%).

Controls versus pre-op.

In the '< 3 y.o.' group (Figures 20A and 20B), non-operated patients had a more elongated cranial vault, both anteriorly and posteriorly, with a more rounded forehead compared with controls (Figure 20B). The skull appeared narrower with a higher forehead (Figure 20A).

Facial morphology appeared to differ little between the two groups, apart from a lower mandibular height in the 'pre-operative scaphocephaly' group.

Pre-op. versus post-op. H-cranioplasty

In the '< 3 y.o.' group, the procedure was likely to result in a widening of the forehead (Figure 20A), with receding and flattening (Figure 20B), no posterior vault changes, and an increase in vertex height.

Pre-op. versus post-op. TVR

In the '< 3 y.o.' group, the procedure probably resulted in a widening of the forehead (Figure 20A) and an increase in the height of the vertex without any change in the antero-posterior length of the skull on profiles (Figure 20B).

Controls versus post-op. (H-cranioplasty + TVR)

In the '< 3 y.o.' group who underwent surgery, there were several morphological differences compared with controls. The skull was higher when viewed from the front for both types of surgeries (Figure 20A), the vertex was higher in profile (Figure 20B), and the posterior cranial vault was more rounded (Figure 20B), especially for patients operated on using TVR.

Post-op. H cranioplasty versus post-op. TVR

The skull appeared to be longer on profiles and wider from frontal views in the TVR group for children under 3 years of age. The posterior vault appeared more rounded in TVR (Figures 20A/B).

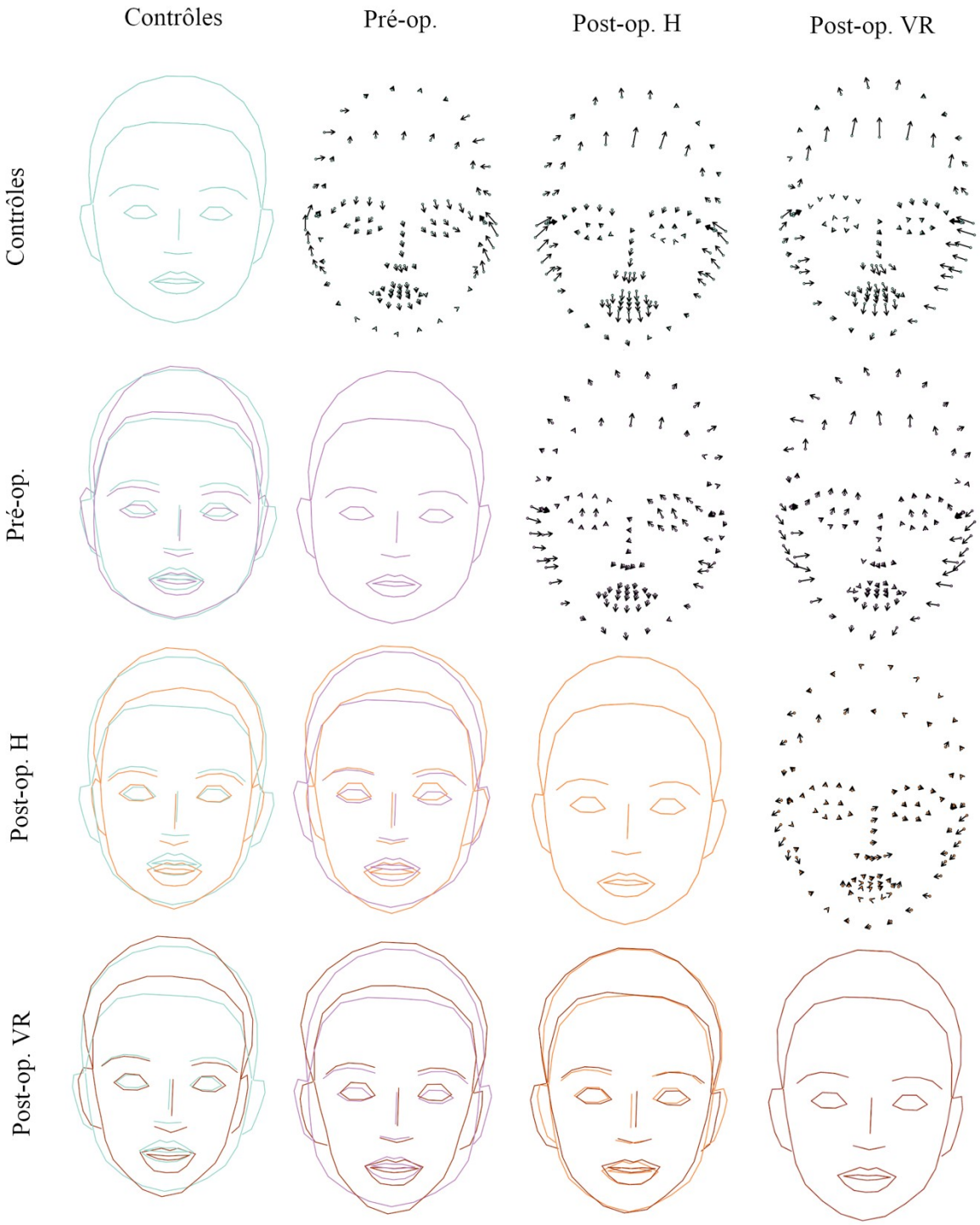


Figure 20A. Average shapes after Procrustes analysis of frontal views, in children under 3 years of age, in the ‘control’, ‘pre-operative scaphocephaly’, ‘post H-cranioplasty’ and ‘post-TVR’ groups. The elements on the diagonal (from the upper left face to the lower right face)

corresponded to average faces. The elements below this diagonal corresponded to superimpositions of two average faces from different groups, and the elements above this diagonal corresponded to vectors of transformation from a first group to a second group.

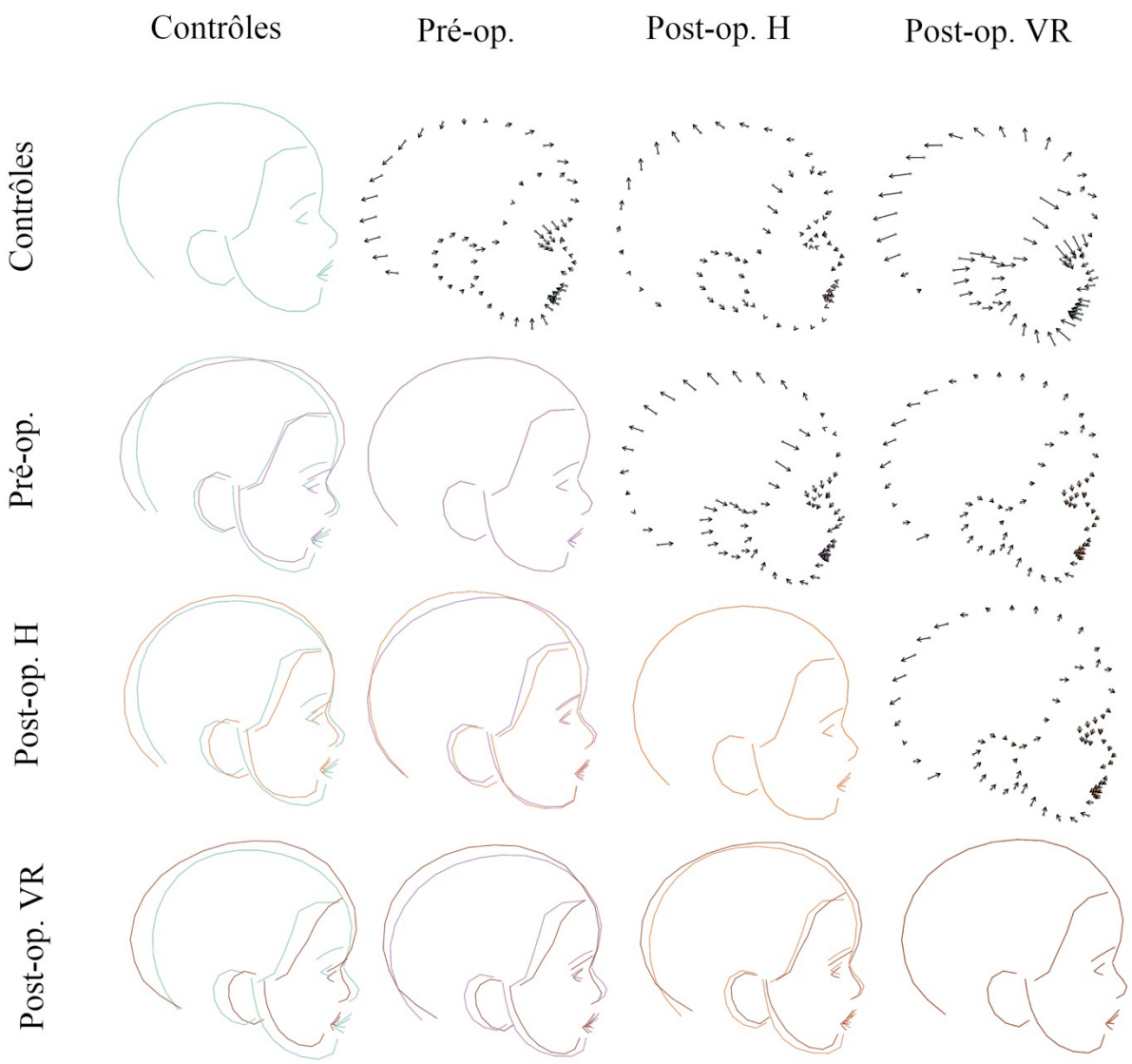


Figure 20B. Average shapes after Procrustes analysis of profiles, in children under 3 years of age, in the 'control', 'pre-operative scaphocephaly', 'post-H-cranioplasty' and 'post-TVR' groups. The elements on the diagonal (from the upper left face to the lower right face)

corresponded to average faces. The elements below this diagonal corresponded to superimpositions of two average faces from different groups, and the elements above this diagonal corresponded to vectors of from a first group to a second group.

An initial binary classification model automatically distinguished the ‘control’ from ‘pre-operative scaphocephaly’ groups with an AUC = 0.905 [0.885 - 0.926] ($p < 0.001$). These performances were slightly improved in a ‘controls’ classification against ‘post-H-cranioplasty’ (AUC = 0.917 [0.888 - 0.946], $p < 0.001$) and against ‘post-TVR’ (AUC = 0.908 [0.860 - 0.957], $p < 0.001$).

The difference remained significant when comparing ‘pre-operative scaphocephaly’ and ‘post H-cranioplasty’ (AUC = 0.648 [0.572 - 0.724], $p < 0.001$). On the other hand, the model was unable to significantly detect TVR patients among pre-operative patients (AUC = 0.481 [0.374 - 0.589], $p = 0.651$). Finally, H-cranioplasty patients were significantly distinguishable from TVR patients (AUC = 0.622 [0.511 - 0.733], $p = 0.023$). The results are detailed in Table 12.

	Controls	**Pre-op.**	**Post-op. H**
Controls			
Pre-op.	0.905 [0.885 - 0.926] *		
Post-op. H	0.917 [0.888 - 0.946] *	0.648 [0.572 - 0.724] *	
Post-op. TVR	0.908 [0.860 - 0.957] *	0.481 [0.374 - 0.589]	0.622 [0.511 - 0.733] *

Table 12. AUC and 95% confidence interval for each binary classification model. * = statistical significance at $p < 0.05$. H = H cranioplasty, TVR = total vault remodeling.

These results precisely described the morphological difference in frontal and profile views between patients with unoperated scaphocephaly and controls, with an antero-posteriorly elongated skull and a more rounded forehead. We also described the effect of surgery on the cranial shape of children with scaphocephaly: flattening and upward elongation of the forehead, with little effect on the posterior cranial vault. We suggest that children who have undergone cranioplasty for scaphocephaly have a facial phenotype in frontal and in profile views that is distinguishable from control children, in particular by a more rounded posterior cranial vault and a higher, flatter forehead. Finally, H-shaped cranioplasty appears to have a greater effect on cranial morphology than TVR, with a greater reduction in anteroposterior diameter. The publication summarizing these results is currently under review.

8 Discussion

8.1 Summary of results

Our algorithm, based on the use of a combination of deep learning, geometric morphometrics, texture analysis, and machine learning methods, used a large high-quality database to detect 10 genetic syndromes and non-syndromic control patients with a top-1 accuracy of 71.3% and a top-3 accuracy of 93.5% on an independent validation set. These performances were excellent for some groups, with 96.1% accuracy for the controls (AUC = 1,000) or 90.7% for Apert (AUC = 0.992) and CHARGE (AUC = 0.912) groups.

We were able to perform phenotype-genotype correlation assessments by describing a significant phenotypic variability within the Crouzon-Pfeiffer (CPS) group. Genotypes linked to the IgI, IgII, IgIIIc and TK domains of *FGFR2* and variations in *FGFR3* were associated with more severe facial phenotypes than variations in a splice site of *FGFR2*. In the latter group, brachycephaly, ocular phenotype and reduced height of the mid-face seemed to be less severe. Our model was also able to distinguish patients with Kabuki syndrome 1 and 2 with an empirical AUC of 0.805 (0.729 - 0.880, $p < 0.001$), providing results comparable to Face2Gene.

Finally, our method has also demonstrated its ability to analyze the effects of drug or surgical treatment on facial morphology and to compare the efficiency of different treatments.

8.2 Competitors and added value

The DeepGestalt method from Face2Gene (FDNA, Inc.) offers an algorithm based on deep learning following automatic landmark detection. The seminal article of Gurovich et al (56) described an algorithm trained on more than 200 syndromes with more than 17,000 photographs, and obtained a top 10 sensitivity of 91% on an independent validation set.

The study of Porras et al (170) described a new tool, not owned by FDNA Inc., also based on deep learning methods with landmark detection, using geometric morphometry parameters known as Statistical Shape Model (SSM). These authors trained their model on 2800 photographs comprising 128 syndromes using data augmentation methods. The authors describe an accuracy of 88% for detecting the presence of a syndrome (90% sensitivity and 86% specificity).

8.2.1 Taking profiles into account

Neither of the above methods took profile analysis into account. Many clinical signs in dysmorphology can only be detected on this incidence:

- Micro- or retrognathism: mandibulofacial dysostoses (MFD), i.e. TC (171), NAFD (172) and MFDM (173), as well as all the syndromes associated with a Pierre Robin sequence, such as 22q11 deletion syndrome or Stickler syndrome (174);
- Zygomatic projection defect: MFD (175), Hypohydrotic Ectodermal Dysplasia (HED) (176), Laron syndrome (26,177), spliceosomopathies (178) or syndromic craniostenoses (142) are just a few examples;
- Nose shape and projection: for example, Williams Beuren (24) and Noonan syndromes (179) lead to a small snub nose, genes associated with Binder maxilla-nasal dysplasia syndrome lead to nasal hypoplasia (180);

- Position and size of the ears: low-set ears are a sign found in frequent syndromes, such as Down (181) and Turner syndromes (182), or in rarer syndromes such as Wiedemann Beckwith (183), Pallister-Killian (184) or Rubinstein-Taybi syndromes (185). MFDs (186) and Meier-Gorlin syndromes (187) lead to microtia, which is only detectable on profiles;
- Hair abnormalities: a characteristic sign of TC syndrome is preauricular hair displacement (171). This clinical sign cannot be seen on a frontal photograph.

8.2.2 Taking the cranial vault and the hairline position into account

Existing tools do not take the entire craniofacial morphology into account.

- Cranial vault: it is an essential indicator of dysmorphology. It is estimated that there are at least 180 genetic syndromes are associated with craniostenosis, simple or complex, and therefore exhibit an anomaly in skull shape (188,189). In addition, microcephaly is a sign in several diseases, such as MFDM (190) and Mowat-Wilson syndromes (191);
- Hairline position: a high anterior hairline is characteristic of Noonan syndrome (192,193), whereas a low hairline may point to Saethre Chotzen syndrome (143,194).

8.2.3 Taking the external ear morphology into account

In a previous paper, we demonstrated that CHARGE syndromes or MFDs can be suspected based on the morphology of the external ear (117), in children and even in fetuses. Some syndromes have a very specific phenotype, such as auriculo-condylar syndrome (195,196) with question-mark ears, or the *crux cymbae* of Saethre Chotzen syndrome (197).

8.2.4 Including controls in the prediction

Part of our database came from the maxillofacial and plastic surgery department at Hôpital Necker - Enfants Malades. Some of the children were being followed for conditions unrelated to a genetic syndrome, such as dental infections, facial trauma, or skin lesions. After checking the clinical records, these photographs were used to create a cohort of non-syndromic control children. The use of control photographs seemed essential to us, and the absence of a genetic syndrome should be a possible prediction for our algorithm.

In fact, the commercial tools mentioned above do not include control databases (Figure 21).

Figure 21. Examples of use of the Face2Gene CLINIC tool on 3 non-syndromic control children from our base. A. The suggested diagnosis was Stickler syndrome in a 6-year-old boy being followed up for a dental abscess. B. Angelman syndrome was proposed in this 1-

year-old girl followed for a gingival trauma. C. Fragile X syndrome was suggested in this 2-year-old girl who is being followed for a palpebral nevus.

8.2.5 Limitations of deep learning

The Face2Gene solution uses diagnostic methods based on deep learning (56). Our AI algorithm is based on supervised learning, i.e., it is possible to dissect each step that led to a prediction. The alternative approach is unsupervised learning, or deep learning. The prediction process is then no longer transparent, leading to a 'black-box effect' (198,199). He et al state that '*AI technologies will need transparency to justify a particular diagnosis, treatment recommendation, or outcome prediction*' (198,200). One concern that may occur regarding automatic diagnostic approaches on 2D photographs is the recognition of extra-facial features: For example, the recognition of a cochlear implant in a patient with TC syndrome or a nasogastric tube in a patient with Goldenhar syndrome associated with esophageal atresia. This is why we chose to use a fully supervised classification method.

Figure 22. Risks of the black box effect. A. Figure from Kumps et al (201) describing the average face of patients with Spondylo-Dysplastic Ehlers-Danlos Syndrome Type 3,

generated by DeepGestalt, Face2Gene's algorithm. Presence of a pair of glasses on the synthetic face. DeepGestalt therefore used the presence or absence of glasses for its diagnosis. B. Patient in our database followed for a TC syndrome. The risk here would be the recognition of his cochlear implant, which is obviously not specific to this syndrome. C. Patient in our database followed for Goldenhar syndrome with esophageal atresia. The risk here is uncontrolled recognition of the nasogastric tube.

8.3 Prospects for improvement

In our opinion, it was essential to add information about the shape of the skull, profiles and the ears. However, to progress towards a closer match with the mental processes of dysmorphologists, many other criteria should be taken into account:

- Clinical data associated with the patient: height, weight, head circumference, medical history and visceral malformations, family history are just a few examples.
- An automatic analysis of other areas of the body, such as the hands, feet or chest.

The number of syndromes present in our multi-class model obviously needs to be increased, to make our results comparable with those of Gurovich et al (56) and Porras et al (170). This process is time-consuming, as we need to check the clinical and genetic records looking for any clearly identified genetic variation or surgery that may have altered the facial phenotype.

Another area of development that could not be started during this thesis work is the generation of average and synthetic faces, by syndrome, age and ethnicity, for educational purposes. We should be addressing this issue in a few months' time.

8.4 Ethical considerations

Using AI for diagnosing rare diseases from patient photographs has the potential to revolutionize healthcare, but it also opens significant ethical challenges. Ensuring patient privacy, addressing bias and fairness issues, and maintaining transparency and accountability are essential to responsibly deploy such technologies in the healthcare sector. Ethical considerations should be at the forefront of AI development and implementations to maximize benefits while minimizing potential harms.

We still need to assess and address potential disparities in the AI performance across different ethnic groups to ensure equitable access to accurate diagnoses in a human-AI collaboration. We also need to emphasize that AI should augment the capabilities of healthcare professionals rather than replace practitioners. The final diagnosis and treatment decisions still require human expertise.

9 Conclusion

Here we present a new Next Generation Phenotyping model, which can be used to detect a genetic syndrome on 2D photographs of the face and profile, in order to help practitioners in their diagnostic process. This work was made possible thanks to the incredible quantity and quality of the *Necker – Enfants Malades Hospital* photographic database. The added value compared with existing tools are the analysis of profile views, taking into account the shape of the cranial vault and the possibility of detecting non-syndromic patients.

This model will never replace clinical expertise, but will enable us to reconsider phenotype-first approaches, the ultimate aim being to reduce the average length of diagnostic wandering in rare diseases. This thesis is just the introduction to a great project, with the recruitment of several engineers and researchers, in order to integrate several hundred genetic syndromes into the algorithm, and to be used in everyday practice by doctors.

10 References

1. Battaglia A, Carey JC. Reflections on observing faces in art. Am J Med Genet C Semin Med Genet. 2021 Jun;187(2):144–7.

2. Dequeker J. Paget's disease in a painting by Quinten Metsys (Massys). BMJ. 1989 Dec 23;299(6715):1579–81.

3. Douzgou S, Chervinsky E, Gyftodimou Y, Kitsiou-Tzeli S, Shalev S, Kanavakis E, et al. Dysmorphology services: a snapshot of current practices and a vision for the future. Clin Genet. 2016 Jan;89(1):27–33.

4. Hunter AGW. Medical genetics: 2. The diagnostic approach to the child with dysmorphic signs. CMAJ. 2002 Aug 20;167(4):367–72.

5. Bhat M. The human face: genes, embryological development and dysmorphology. The International Journal of Developmental Biology. 2020 Jun 23;64(4-5–6):383–91.

6. Diliberti JH. Use of computers in dysmorphology. J Med Genet. 1988 Jul;25(7):445–53.

7. Moeschler JB, Shevell M, American Academy of Pediatrics Committee on Genetics. Clinical genetic evaluation of the child with mental retardation or developmental delays. Pediatrics. 2006 Jun;117(6):2304–16.

8. Muenke M, Adeyemo A, Kruszka P. An electronic atlas of human malformation syndromes in diverse populations. Genet Med. 2016 Nov;18(11):1085–7.

9. Lumaka A, Cosemans N, Lulebo Mampasi A, Mubungu G, Mvuama N, Lubala T, et al. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. Clin Genet. 2017 Aug;92(2):166–71.

10. Veerapandiyan A, Abdul-Rahman OA, Adam MP, Lyons MJ, Manning M, Coleman K, et al. Chromosome 22q11.2 deletion syndrome in African-American patients: a diagnostic

challenge. Am J Med Genet A. 2011 Sep;155A(9):2186–95.

11. Kruszka P, Addissie YA, McGinn DE, Porras AR, Biggs E, Share M, et al. 22q11.2 deletion syndrome in diverse populations. Am J Med Genet A. 2017 Apr;173(4):879–88.

12. Limwongse C. Medical genetic services in a developing country: lesson from Thailand. Curr Opin Pediatr. 2017 Dec;29(6):634–9.

13. Introducing the Necker-Enfants Malades hospital [Internet]. Hôpital Necker-Enfants malades. 2015 [cited 2023 Sep 4]. Available from: https://hopital-necker.aphp.fr/introducing-necker-enfants-malades-hospital

14. McDonald-McGinn DM, Minugh-Purvis N, Kirschner RE, Jawad A, Tonnesen MK, Catanzaro JR, et al. The 22q11.2 Deletion in African-American Patients: An Underdiagnosed Population? Am J Med Genet A. 2005 Apr 30;134(3):242–6.

15. Hart TC, Hart PS. Genetic studies of craniofacial anomalies: clinical implications and applications. Orthod Craniofac Res. 2009 Aug;12(3):212–20.

16. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347–58.

17. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current Applications and Future Impact of Machine Learning in Radiology. Radiology. 2018 Aug;288(2):318–28.

18. Novoa RA, Gevaert O, Ko JM. Marking the Path Toward Artificial Intelligence-Based Image Classification in Dermatology. JAMA Dermatol. 2019 Oct 1;155(10):1105–6.

19. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial Intelligence and Surgical Decision-Making. JAMA Surg [Internet]. 2019 Dec 11 [cited 2019 Dec 14]; Available from: https://jamanetwork.com/journals/jamasurgery/fullarticle/2756311

20. Thevenot J, Lopez MB, Hadid A. A Survey on Computer Vision for Assistive Medical

Diagnosis From Faces. IEEE J Biomed Health Inform. 2018 Sep;22(5):1497–511.

21. Zhao Q, Rosenbaum K, Okada K, Zand DJ, Sze R, Summar M, et al. Automated Down syndrome detection using facial photographs. Annu Int Conf IEEE Eng Med Biol Soc. 2013;2013:3670–3.

22. Burçin K, Nabiyev V. Down syndrome recognition using local binary patterns and statistical evaluation of the system. Expert Syst Appl. 2011 Jul 1;38:8690–5.

23. Kruszka P, Porras AR, Sobering AK, Ikolo FA, La Qua S, Shotelersuk V, et al. Down syndrome in diverse populations. Am J Med Genet A. 2017 Jan;173(1):42–53.

24. Kruszka P, Porras AR, de Souza DH, Moresco A, Huckstadt V, Gill AD, et al. Williams-Beuren syndrome in diverse populations. Am J Med Genet A. 2018 May;176(5):1128–36.

25. Dalal AB, Phadke SR. Morphometric analysis of face in dysmorphology. Comput Methods Programs Biomed. 2007 Feb;85(2):165–72.

26. Schaefer GB, Rosenbloom AL, Guevara-Aguirre J, Campbell EA, Ullrich F, Patil K, et al. Facial morphometry of Ecuadorian patients with growth hormone receptor deficiency/Laron syndrome. J Med Genet. 1994 Aug;31(8):635–9.

27. Miller RE, Learned-Miller EG, Trainer P, Paisley A, Blanz V. Early diagnosis of acromegaly: computers vs clinicians. Clin Endocrinol (Oxf). 2011 Aug;75(2):226–31.

28. Kong X, Gong S, Su L, Howard N, Kong Y. Automatic Detection of Acromegaly From Facial Photographs Using Machine Learning Methods. EBioMedicine. 2018 Jan;27:94–102.

29. Schneider HJ, Kosilek RP, Günther M, Roemmler J, Stalla GK, Sievers C, et al. A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification. J Clin Endocrinol Metab. 2011 Jul;96(7):2074–80.

30. Kosilek RP, Frohner R, Würtz RP, Berr CM, Schopohl J, Reincke M, et al. Diagnostic

use of facial image analysis software in endocrine and genetic disorders: review, current results and future perspectives. Eur J Endocrinol. 2015 Oct;173(4):M39-44.

31. Learned-Miller E, Lu Q, Paisley A, Trainer P, Blanz V, Dedden K, et al. Detecting acromegaly: screening for disease with a morphable model. Med Image Comput Comput Assist Interv. 2006;9(Pt 2):495–503.

32. Kosilek RP, Schopohl J, Grunke M, Reincke M, Dimopoulou C, Stalla GK, et al. Automatic face classification of Cushing's syndrome in women - a novel screening approach. Exp Clin Endocrinol Diabetes. 2013 Oct;121(9):561–4.

33. Chen S, Pan ZX, Zhu HJ, Wang Q, Yang JJ, Lei Y, et al. Development of a computer-aided tool for the pattern recognition of facial features in diagnosing Turner syndrome: comparison of diagnostic accuracy with clinical workers. Sci Rep. 2018 Jun 18;8(1):9317.

34. Chen Y, Liu W, Zhang L, Yan M, Zeng Y. Hybrid facial image feature extraction and recognition for non-invasive chronic fatigue syndrome diagnosis. Comput Biol Med. 2015 Sep;64:30–9.

35. Buretić-Tomljanović A, Petaros A, Suc EG, Nincoletti S, Rubesa G, Tomljanović D, et al. Craniofacial morphologic and anthropometric features of Croatian schizophrenia patients and non-psychiatric controls--a pilot study. Anthropol Anz. 2012;69(4):379–97.

36. Ferry Q, Steinberg J, Webber C, FitzPatrick DR, Ponting CP, Zisserman A, et al. Diagnostically relevant facial gestalt information from ordinary photos. Tollman S, editor. eLife. 2014 Jun 24;3:e02020.

37. Boehringer S, Vollmar T, Tasse C, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, et al. Syndrome identification based on 2D analysis software. Eur J Hum Genet. 2006 Oct;14(10):1082–9.

38. Shukla P, Gupta T, Saini A, Singh P, Balasubramanian R. A Deep Learning Frame-Work for Recognizing Developmental Disorders. 2017 IEEE Winter Conference on

Applications of Computer Vision (WACV). 2017 Mar;705–14.

39. Loos HS, Wieczorek D, Würtz RP, von der Malsburg C, Horsthemke B. Computer-based recognition of dysmorphic faces. Eur J Hum Genet. 2003 Aug;11(8):555–60.

40. Vollmar T, Maus B, Wurtz RP, Gillessen-Kaesbach G, Horsthemke B, Wieczorek D, et al. Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces. Eur J Med Genet. 2008;51(1):44–53.

41. Cerrolaza JJ, Porras AR, Mansoor A, Zhao Q, Summar M, Linguraru MG. Identification of dysmorphic syndromes using landmark-specific local texture descriptors. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). 2016 Apr;1080–3.

42. Özdemir ME, Telatar Z, Eroğul O, Tunca Y. Classifying dysmorphic syndromes by using artificial neural network based hierarchical decision tree. 2018 Jun 1 [cited 2023 Sep 4]; Available from: https://gcris.etu.edu.tr/handle/20.500.11851/713

43. Dudding-Byth T, Baxter A, Holliday EG, Hackett A, O'Donnell S, White SM, et al. Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. BMC Biotechnol. 2017 Dec 19;17(1):90.

44. Kuru K, Niranjan M, Tunca Y, Osvank E, Azim T. Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. Artif Intell Med. 2014 Oct;62(2):105–18.

45. Saraydemir S, Taşpınar N, Eroğul O, Kayserili H, Dinçkan N. Down syndrome diagnosis based on Gabor Wavelet Transform. J Med Syst. 2012 Oct;36(5):3205–13.

46. Knaus A, Pantel JT, Pendziwiat M, Hajjir N, Zhao M, Hsieh TC, et al. Characterization of glycosylphosphatidylinositol biosynthesis defects by clinical features, flow cytometry, and automated image analysis. Genome Med. 2018 Jan 9;10(1):3.

47. Valentine M, Bihm DCJ, Wolf L, Hoyme HE, May PA, Buckley D, et al. Computer-

Aided Recognition of Facial Attributes for Fetal Alcohol Spectrum Disorders. Pediatrics. 2017 Dec;140(6):e20162028.

48. Chen S, Pan ZX, Zhu HJ, Wang Q, Yang JJ, Lei Y, et al. Development of a computer-aided tool for the pattern recognition of facial features in diagnosing Turner syndrome: comparison of diagnostic accuracy with clinical workers. Sci Rep. 2018 Jun 18;8(1):9317.

49. Rohlf FJ, Slice D. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. Systematic Zoology. 1990;39(1):40–59.

50. Peters JR, Campbell RM, Balasubramanian S. Characterization of the age-dependent shape of the pediatric thoracic spine and vertebrae using generalized procrustes analysis. J Biomech. 2017 Oct 3;63:32–40.

51. Douglas TS, Mutsvangwa TEM. A review of facial image analysis for delineation of the facial phenotype associated with fetal alcohol syndrome. Am J Med Genet A. 2010 Feb;152A(2):528–36.

52. Shen L, Bai L. A review on Gabor wavelets for face recognition. Pattern Anal Appl. 2006 Oct 1;9(2–3):273–92.

53. Wiskott L, von der Malsburg C. Recognizing faces by dynamic link matching. Neuroimage. 1996 Dec;4(3 Pt 2):S14-18.

54. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002 Jul;24(7):971–87.

55. Saraydemir S, Taşpınar N, Eroğul O, Kayserili H, Dinçkan N. Down syndrome diagnosis based on Gabor Wavelet Transform. J Med Syst. 2012 Oct;36(5):3205–13.

56. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019 Jan;25(1):60–4.

57. Zarate YA, Smith-Hicks CL, Greene C, Abbott MA, Siu VM, Calhoun ARUL, et al.

Natural history and genotype-phenotype correlations in 72 individuals with SATB2-associated syndrome. Am J Med Genet A. 2018 Apr;176(4):925–35.

58. Martinez-Monseny A, Cuadras D, Bolasell M, Muchart J, Arjona C, Borregan M, et al. From gestalt to gene: early predictive dysmorphic features of PMM2-CDG. J Med Genet. 2019 Apr;56(4):236–45.

59. Pascolini G, Calvani M, Grammatico P. First Italian experience using the automated craniofacial gestalt analysis on a cohort of pediatric patients with multiple anomaly syndromes. Ital J Pediatr. 2022 Jun 13;48(1):91.

60. Stamberger H, Hammer TB, Gardella E, Vlaskamp DRM, Bertelsen B, Mandelstam S, et al. NEXMIF encephalopathy: an X-linked disorder with male and female phenotypic patterns. Genet Med. 2021 Feb;23(2):363–73.

61. Porras AR, Rosenbaum K, Tor-Diez C, Summar M, Linguraru MG. Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study. Lancet Digit Health. 2021 Oct;3(10):e635–43.

62. Mensah MA, Ott CE, Horn D, Pantel JT. A machine learning-based screening tool for genetic syndromes in children. The Lancet Digital Health. 2022 May 1;4(5):e295.

63. Todd ES, Weinberg SM, Berry-Kravis EM, Silvestri JM, Kenny AS, Rand CM, et al. Facial phenotype in children and young adults with PHOX2B-determined congenital central hypoventilation syndrome: quantitative pattern of dysmorphology. Pediatr Res. 2006 Jan;59(1):39–45.

64. Abell K, Hopkin RJ, Bender PL, Jackson F, Smallwood K, Sullivan B, et al. Mandibulofacial dysostosis with microcephaly: an expansion of the phenotype via parental survey. Am J Med Genet A. 2021 Feb;185(2):413–23.

65. Tüysüz B, Güneş N, Geyik F, Yeşil G, Celkan T, Vural M. Investigation of (epi)genotype causes and follow-up manifestations in the patients with classical and atypical

phenotype of Beckwith-Wiedemann spectrum. Am J Med Genet A. 2021 Jun;185(6):1721–31.

66. Legendre M, Abadie V, Attié-Bitach T, Philip N, Busa T, Bonneau D, et al. Phenotype and genotype analysis of a French cohort of 119 patients with CHARGE syndrome. American Journal of Medical Genetics Part C: Seminars in Medical Genetics. 2017;175(4):417–30.

67. Gordon CT, Petit F, Oufadem M, Decaestecker C, Jourdain AS, Andrieux J, et al. EFTUD2 haploinsufficiency leads to syndromic oesophageal atresia. J Med Genet. 2012 Dec;49(12):737–46.

68. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018 Apr;80:52–63.

69. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep. 1966 Aug;19(1):3–11.

70. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001 [Internet]. Kauai, HI, USA: IEEE Comput. Soc; 2001 [cited 2022 Mar 17]. p. I-511-I–518. Available from: http://ieeexplore.ieee.org/document/990517/

71. Bradski G. The OpenCV Library. Dr Dobb's Journal: Software Tools for the Professional Programmer [Internet]. 2000 [cited 2022 Mar 17];25(11). Available from: https://www.elibrary.ru/item.asp?id=4934581

72. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017 Jun;39(6):1137–49.

73. Wu Y, Kirillov A, Massa F, Lo WY, Girshick R. Detectron2. 2019.

74. Wada K. Labelme: Image Polygonal Annotation with Python [Internet]. 2022 [cited

2022 Mar 17]. Available from: https://github.com/wkentaro/labelme

75. Pawara P, Okafor E, Schomaker L, Wiering M. Data Augmentation for Plant Classification. In 2017.

76. Padilla R, Netto SL, da Silva EAB. A Survey on Performance Metrics for Object-Detection Algorithms. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). 2020. p. 237–42.

77. landmarker.io · The Menpo Project [Internet]. [cited 2022 Mar 20]. Available from: https://www.menpo.org/landmarkerio/

78. Alabort-i-Medina J, Antonakos E, Booth J, Snape P, Zafeiriou S. Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models. In: Proceedings of the 22nd ACM international conference on Multimedia [Internet]. Orlando Florida USA: ACM; 2014 [cited 2022 Feb 24]. p. 679–82. Available from: https://dl.acm.org/doi/10.1145/2647868.2654890

79. Matthews I, Baker S. Active Appearance Models Revisited. International Journal of Computer Vision. 2004 Nov;60(2):135–64.

80. Active Appearance Model · The Menpo Project [Internet]. [cited 2022 Mar 20]. Available from: https://www.menpo.org/menpofit/aam.html

81. Xiong X, De la Torre F. Supervised Descent Method and Its Applications to Face Alignment. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013. p. 532–9.

82. Antonakos E, Alabort-i-Medina J, Tzimiropoulos G, Zafeiriou SP. Feature-Based Lucas–Kanade and Active Appearance Models. IEEE Trans on Image Process. 2015 Sep;24(9):2617–32.

83. Cristinacce D, Cootes TF. Feature Detection and Tracking with Constrained Local Models. In: Procedings of the British Machine Vision Conference 2006 [Internet]. Edinburgh:

British Machine Vision Association; 2006 [cited 2022 Jan 28]. p. 95.1-95.10. Available from: http://www.bmva.org/bmvc/2006/papers/024.html

84. Cristinacce D, Cootes T. Automatic feature localisation with constrained local models. Pattern Recogn. 2008 Oct 1;41(10):3054–67.

85. Deng J, Roussos A, Chrysos G, Ververas E, Kotsia I, Shen J, et al. The Menpo Benchmark for Multi-pose 2D and 3D Facial Landmark Localisation and Tracking. Int J Comput Vis. 2019 Jun;127(6–7):599–624.

86. Belhumeur PN, Jacobs DW, Kriegman DJ, Kumar N. Localizing Parts of Faces Using a Consensus of Exemplars. :8.

87. Zhu X, Ramanan D. Face Detection, Pose Estimation, and Landmark Localization in the Wild.

88. Le V, Brandt J, Lin Z, Bourdev L, Huang TS. Interactive facial feature localization: 12th European Conference on Computer Vision, ECCV 2012. Computer Vision, ECCV 2012 - 12th European Conference on Computer Vision, Proceedings. 2012 Oct 30;679–92.

89. Messer K, Kittler J, Sadeghi M, Marcel S, Marcel C, Bengio S, et al. Face Verification Competition on the XM2VTS Database. In 2003. p. 964–74.

90. Kostinger M, Wohlhart P, Roth PM, Bischof H. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) [Internet]. Barcelona, Spain: IEEE; 2011 [cited 2022 Jan 28]. p. 2144–51. Available from: http://ieeexplore.ieee.org/document/6130513/

91. Zhou Y, Zaferiou S. Deformable Models of Ears in-the-Wild for Alignment and Recognition. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). 2017. p. 626–33.

92. Avcı H, Karakaya J. A Novel Medical Image Enhancement Algorithm for Breast

Cancer Detection on Mammography Images Using Machine Learning. Diagnostics (Basel). 2023 Jan 18;13(3):348.

93. Anifah L, Purnama IKE, Hariadi M, Purnomo MH. Osteoarthritis Classification Using Self Organizing Map Based on Gabor Kernel and Contrast-Limited Adaptive Histogram Equalization. The Open Biomedical Engineering Journal. 2013;7:18.

94. Huang C, Li X, Wen Y. AN OTSU image segmentation based on fruitfly optimization algorithm. Alexandria Engineering Journal. 2021 Feb 1;60(1):183–8.

95. Kiflie A, Tesema Tufa G, Salau AO. Sputum smears quality inspection using an ensemble feature extraction approach. Front Public Health. 2023 Jan 25;10:1032467.

96. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics. 1973 Nov;SMC-3(6):610–21.

97. Mohanaiah P, Sathyanarayana P, GuruKumar L. Image Texture Feature Extraction Using GLCM Approach. 2013;3(5).

98. Löfstedt T, Brynolfsson P, Asklund T, Nyholm T, Garpebring A. Gray-level invariant Haralick texture features. PLoS One. 2019 Feb 22;14(2):e0212110.

99. Mundt P, Tharmaseelan H, Hertel A, Rotkopf LT, Nörenberg D, Riffel P, et al. Periaortic adipose radiomics texture features associated with increased coronary calcium score-first results on a photon-counting-CT. BMC Med Imaging. 2023 Jul 26;23(1):97.

100. Adelsmayr G, Janisch M, Müller H, Holzinger A, Talakic E, Janek E, et al. Three dimensional computed tomography texture analysis of pulmonary lesions: Does radiomics allow differentiation between carcinoma, neuroendocrine tumor and organizing pneumonia? Eur J Radiol. 2023 Aug;165:110931.

101. Peng B, Wang K, Xu R, Guo C, Lu T, Li Y, et al. Preoperative computed tomography-based tumoral radiomic features prediction for overall survival in resectable non-small cell lung cancer. Front Oncol. 2023;13:1131816.

102. Ou X, Pan W, Xiao P. In vivo skin capacitive imaging analysis by using grey level co-occurrence matrix (GLCM). Int J Pharm. 2014 Jan 2;460(1–2):28–32.

103. Veena DK, Jatti A, Joshi R, Deepu KS. Characterization of dental pathologies using digital panoramic X-ray images based on texture analysis. Annu Int Conf IEEE Eng Med Biol Soc. 2017 Jul;2017:592–5.

104. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901 Nov 1;2(11):559–72.

105. Jolliffe IT. Principal Component Analysis. Springer Science & Business Media; 2002. 524 p.

106. Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, et al. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med. 2003 Mar 20;348(12):1170–5.

107. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2023 Jul 4]. p. 785–94. (KDD '16). Available from: https://dl.acm.org/doi/10.1145/2939672.2939785

108. Sachs MC. plotROC: A Tool for Plotting ROC Curves. J Stat Softw. 2017 Aug;79:2.

109. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2020 [cited 2022 Aug 30]. Available from: http://arxiv.org/abs/1802.03426

110. R Core Team (2020). — European Environment Agency [Internet]. [cited 2023 Jan 18]. Available from: https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006

111. Hennocq Q, Bongibault T, Bizière M, Delassus O, Douillet M, Cormier-Daire V, et al.

An automatic facial landmarking for children with rare diseases. American Journal of Medical Genetics Part A [Internet]. [cited 2023 Jan 30];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajmg.a.63126

112. Koonce B. ResNet 50. In 2021. p. 63–72.

113. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library [Internet]. arXiv; 2019 [cited 2023 Jul 17]. Available from: http://arxiv.org/abs/1912.01703

114. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization [Internet]. arXiv; 2017 [cited 2023 Aug 10]. Available from: http://arxiv.org/abs/1412.6980

115. Lucas B, Kanade T. An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI). In 1981.

116. RESERVES IUTD. Orphanet: Diagnosis of Mandibulofacial Dysostosis, Guion Almeida type EFTUD2 gene [Internet]. [cited 2023 Sep 6]. Available from: https://www.orpha.net/consor/cgi-bin/ClinicalLabs_Search.php?lng=FR&data_id=136281&search=ClinicalLabs_Search_Simple&data_type=Test&title=Diagnosis%20of%20Mandibulofacial%20Dysostosis,%20Guion-Almeida%20type%20EFTUD2%20gene&MISSING%20CONTENT=Diagnosis-of-Mandibulofacial-Dysostosis--Guion-Almeida-type--EFTUD2-gene-

117. Hennocq Q, Bongibault T, Marlin S, Amiel J, Attie-Bitach T, Baujat G, et al. AI-based diagnosis in mandibulofacial dysostosis with microcephaly using external ear shapes. Front Pediatr. 2023;11:1171277.

118. Andresen C, Matias A, Merz E. Fetal face: the whole picture. Ultraschall Med. 2012 Oct;33(5):431–40.

119. Bartel-Friedrich S, Wulke C. Classification and diagnosis of ear malformations. GMS Curr Top Otorhinolaryngol Head Neck Surg. 2008 Mar 14;6:Doc05.

120. Dudarewicz L, Kałuzewski B. Prenatal screening for fetal chromosomal abnormalities using ear length and shape as an ultrasound marker. Med Sci Monit. 2000;6(4):801–6.

121. Nuñez-Castruita A, López-Serna N. Low-set ears and associated anomalies in human foetuses. Int J Pediatr Otorhinolaryngol. 2018 Jan;104:126–33.

122. Sacchini C, El-Sheikhah A, Cicero S, Rembouskos G, Nicolaides KH. Ear length in trisomy 21 fetuses at 11-14 weeks of gestation. Ultrasound Obstet Gynecol. 2003 Nov;22(5):460–3.

123. Lehalle D, Gordon CT, Oufadem M, Goudefroye G, Boutaud L, Alessandri JL, et al. Delineation of EFTUD2 haploinsufficiency-related phenotypes through a series of 36 patients. Hum Mutat. 2014 Apr;35(4):478–85.

124. Sanlaville D, Etchevers HC, Gonzales M, Martinovic J, Clément-Ziza M, Delezoide AL, et al. Phenotypic spectrum of CHARGE syndrome in fetuses with CHD7 truncating mutations correlates with expression during human development. J Med Genet. 2006 Mar;43(3):211–7.

125. Haute Autorité de Santé [Internet]. [cited 2023 Aug 11]. Protocole type d'examen autopsique foetal ou néonatal. Available from: https://www.has-sante.fr/jcms/c_1747470/fr/protocole-type-d-examen-autopsique-foetal-ou-neonatal

126. Governale LS. Craniosynostosis. Pediatr Neurol. 2015 Nov;53(5):394–401.

127. Lajeunie E, Crimmins DW, Arnaud E, Renier D. Genetic considerations in nonsyndromic midline craniosynostoses: a study of twins and their families. J Neurosurg. 2005 Oct;103(4 Suppl):353–6.

128. Di Rocco F, Biosse Duplan M, Heuzé Y, Kaci N, Komla-Ebri D, Munnich A, et al. FGFR3 mutation causes abnormal membranous ossification in achondroplasia. Hum Mol Genet. 2014 Jun 1;23(11):2914–25.

129. Snyder-Warwick AK, Perlyn CA, Pan J, Yu K, Zhang L, Ornitz DM. Analysis of a gain-of-function FGFR2 Crouzon mutation provides evidence of loss of function activity in the etiology of cleft palate. Proc Natl Acad Sci U S A. 2010 Feb 9;107(6):2515–20.

130. Betances EM, Mendez MD, M Das J. Craniosynostosis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 Jun 25]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK544366/

131. Singh RP, Dhariwal D, Bhujel N, Shaikh Z, Davies P, Nishikawa H, et al. Role of parental risk factors in the aetiology of isolated non-syndromic metopic craniosynostosis. Br J Oral Maxillofac Surg. 2010 Sep;48(6):438–42.

132. Kimonis V, Gold JA, Hoffman TL, Panchal J, Boyadjiev SA. Genetics of craniosynostosis. Semin Pediatr Neurol. 2007 Sep;14(3):150–61.

133. Magge SN, Snyder K, Sajja A, DeFreitas TA, Hofherr SE, Broth RE, et al. Identical Twins Discordant for Metopic Craniosynostosis: Evidence of Epigenetic Influences. J Craniofac Surg. 2017 Jan;28(1):14–6.

134. Ohishi A, Nishimura G, Kato F, Ono H, Maruwaka K, Ago M, et al. Mutation analysis of FGFR1-3 in 11 Japanese patients with syndromic craniosynostoses. Am J Med Genet A. 2017 Jan;173(1):157–62.

135. Boulet SL, Rasmussen SA, Honein MA. A population-based study of craniosynostosis in metropolitan Atlanta, 1989-2003. Am J Med Genet A. 2008 Apr 15;146A(8):984–91.

136. Kan S hsin, Elanko N, Johnson D, Cornejo-Roldan L, Cook J, Reich EW, et al. Genomic Screening of Fibroblast Growth-Factor Receptor 2 Reveals a Wide Spectrum of Mutations in Patients with Syndromic Craniosynostosis. Am J Hum Genet. 2002 Feb;70(2):472–86.

137. Morice A, Cornette R, Giudice A, Collet C, Paternoster G, Arnaud É, et al. Early mandibular morphological differences in patients with FGFR2 and FGFR3-related syndromic

craniosynostoses: A 3D comparative study. Bone. 2020 Dec;141:115600.

138. Bouaoud J, Hennocq Q, Paternoster G, James S, Arnaud E, Khonsari RH. Excessive ossification of the bandeau in Crouzon and Apert syndromes. J Craniomaxillofac Surg. 2020 Apr;48(4):376–82.

139. Reardon W, Winter RM, Rutland P, Pulleyn LJ, Jones BM, Malcolm S. Mutations in the fibroblast growth factor receptor 2 gene cause Crouzon syndrome. Nat Genet. 1994 Sep;8(1):98–103.

140. Meyers GA, Orlow SJ, Munro IR, Przylepa KA, Jabs EW. Fibroblast growth factor receptor 3 (FGFR3) transmembrane mutation in Crouzon syndrome with acanthosis nigricans. Nature genetics. 1995;11(4):462–4.

141. Pal US, Gupta C, Chellappa AAL. Crouzon syndrome with primary optic nerve atrophy and normal brain functions: A case report. J Oral Biol Craniofac Res. 2012;2(2):116–8.

142. Jezela-Stanek A, Krajewska-Walasek M. Genetic causes of syndromic craniosynostoses. Eur J Paediatr Neurol. 2013 May;17(3):221–4.

143. Gallagher ER, Ratisoontorn C, Cunningham ML. Saethre-Chotzen Syndrome. In: Adam MP, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2023 Jul 31]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1189/

144. Rosen H, Andrews BT, Meara JG, Stoler JM, Mulliken JB, Rogers GF. Audiologic findings in Saethre-Chotzen syndrome. Plast Reconstr Surg. 2011 May;127(5):2014–20.

145. Barry KK, Tsaparlis M, Hoffman D, Hartman D, Adam MP, Hung C, et al. From Genotype to Phenotype—A Review of Kabuki Syndrome. Genes (Basel). 2022 Sep 29;13(10):1761.

146. Niikawa N, Kuroki Y, Kajii T, Matsuura N, Ishikiriyama S, Tonoki H, et al. Kabuki

make-up (Niikawa-Kuroki) syndrome: a study of 62 patients. Am J Med Genet. 1988 Nov;31(3):565–89.

147. White SM, Thompson EM, Kidd A, Savarirayan R, Turner A, Amor D, et al. Growth, behavior, and clinical findings in 27 patients with Kabuki (Niikawa-Kuroki) syndrome. Am J Med Genet A. 2004 Jun 1;127A(2):118–27.

148. Kuroki Y, Suzuki Y, Chyo H, Hata A, Matsui I. A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. J Pediatr. 1981 Oct;99(4):570–3.

149. Rouxel F, Yauy K, Boursier G, Gatinois V, Barat-Houari M, Sanchez E, et al. Using deep-neural-network-driven facial recognition to identify distinct Kabuki syndrome 1 and 2 gestalt. Eur J Hum Genet. 2022 Jun;30(6):682–6.

150. Adam MP, Hudgins L. Kabuki syndrome: a review. Clin Genet. 2005 Mar;67(3):209–19.

151. Bögershausen N, Gatinois V, Riehmer V, Kayserili H, Becker J, Thoenes M, et al. Mutation Update for Kabuki Syndrome Genes KMT2D and KDM6A and Further Delineation of X-Linked Kabuki Syndrome Subtype 2. Hum Mutat. 2016 Sep;37(9):847–64.

152. Lederer D, Grisart B, Digilio MC, Benoit V, Crespin M, Ghariani SC, et al. Deletion of KDM6A, a Histone Demethylase Interacting with MLL2, in Three Patients with Kabuki Syndrome. Am J Hum Genet. 2012 Jan 13;90(1):119–24.

153. Paděrová J, Holubová A, Simandlová M, Puchmajerová A, Vlčková M, Malíková M, et al. Molecular genetic analysis in 14 Czech Kabuki syndrome patients is confirming the utility of phenotypic scoring. Clin Genet. 2016 Sep;90(3):230–7.

154. Banka S, Lederer D, Benoit V, Jenkins E, Howard E, Bunstone S, et al. Novel KDM6A (UTX) mutations and a clinical and molecular review of the X-linked Kabuki syndrome (KS2). Clin Genet. 2015 Mar;87(3):252–8.

155. Bögershausen N, Wollnik B. Unmasking Kabuki syndrome. Clin Genet. 2013 Mar;83(3):201–11.

156. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010 Sep;42(9):790–3.

157. Lee S, Sze R, Murakami C, Gruss J, Cunningham M. Hemifacial myohyperplasia: description of a new syndrome. Am J Med Genet. 2001 Nov 1;103(4):326–33.

158. Taucher SC, Vargas RAP, Le�n DC. Hemifacial myohyperplasia: An additional case. American Journal of Medical Genetics. 2002;116A(1):103–4.

159. Pereira-Perdomo DF, Vélez-Forero J, Prada-Madrid R. Hemifacial myohyperplasia sequence. Am J Med Genet A. 2010 Jul;152A(7):1770–3.

160. Miranda RT, Barros LM, Santos LAND, Bonan PRF, Martelli H. Clinical and imaging features in a patient with hemifacial hyperplasia. J Oral Sci. 2010 Sep;52(3):509–12.

161. Siponen M, Sándor GKB, Ylikontiola L, Salo T, Tuominen H. Multiple orofacial intraneural perineuriomas in a patient with hemifacial hyperplasia. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2007 Jul;104(1):e38-44.

162. Zissman S, Cooperman Y, Leshem D, Gur E. Progressive Surgical Management of Hemifacial Myohyperplasia for Improved Functional and Aesthetic Results. Plast Reconstr Surg Glob Open. 2020 Jul 21;8(7):e2724.

163. Canaud G, Hammill AM, Adams D, Vikkula M, Keppler-Noreuil KM. A review of mechanisms of disease across PIK3CA-related disorders with vascular manifestations. Orphanet J Rare Dis. 2021 Jul 8;16(1):306.

164. Venot Q, Blanc T, Rabia SH, Berteloot L, Ladraa S, Duong JP, et al. Targeted therapy in patients with PIK3CA-related overgrowth syndrome. Nature. 2018 Jun;558(7711):540–6.

165. Ladraa S, Zerbib L, Bayard C, Fraissenon A, Venot Q, Morin G, et al. PIK3CA gain-

of-function mutation in adipose tissue induces metabolic reprogramming with Warburg-like effect and severe endocrine disruption. Sci Adv. 2022 Dec 9;8(49):eade7823.

166. Delestre F, Venot Q, Bayard C, Fraissenon A, Ladraa S, Hoguin C, et al. Alpelisib administration reduced lymphatic malformations in a mouse model and in patients. Sci Transl Med. 2021 Oct 6;13(614):eabg0809.

167. Morin G, Degrugillier-Chopinet C, Vincent M, Fraissenon A, Aubert H, Chapelle C, et al. Treatment of two infants with PIK3CA-related overgrowth spectrum by alpelisib. J Exp Med. 2022 Mar 7;219(3):e20212148.

168. Bayard C, Segna E, Taverne M, Fraissenon A, Hennocq Q, Periou B, et al. Hemifacial myohyperplasia is due to somatic muscular PIK3CA gain-of-function mutations and responds to pharmacological inhibition. J Exp Med. 2023 Nov 6;220(11):e20230926.

169. Fairman D, Horrax G. Classification of craniostenosis. J Neurosurg. 1949 Jul;6(4):307–13.

170. Porras AR, Rosenbaum K, Tor-Diez C, Summar M, Linguraru MG. Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study. Lancet Digit Health. 2021 Oct;3(10):e635–43.

171. Katsanis SH, Jabs EW. Treacher Collins Syndrome. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2022 Sep 8]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1532/

172. Zhao J, Yang L. Broad-spectrum next-generation sequencing-based diagnosis of a case of Nager syndrome. J Clin Lab Anal. 2020 Sep;34(9):e23426.

173. Guion-Almeida ML, Vendramini-Pittoli S, Passos-Bueno MRS, Zechi-Ceide RM. Mandibulofacial syndrome with growth and mental retardation, microcephaly, ear anomalies with skin tags, and cleft palate in a mother and her son: autosomal dominant or X-linked

syndrome? Am J Med Genet A. 2009 Dec;149A(12):2762–4.

174. Stoll C, Alembick Y, Roth MP. Associated anomalies in Pierre Robin sequence. Am J Med Genet A. 2023 Sep;191(9):2312–23.

175. Smigiel R, Bezniakow N, Jakubiak A, Błoch M, Patkowski D, Obersztyn E, et al. Phenotype analysis of Polish patients with mandibulofacial dysostosis type Guion-Almeida associated with esophageal atresia and choanal atresia caused by EFTUD2 gene mutations. J Appl Genet. 2015 May;56(2):199–204.

176. Callea M, Bignotti S, Semeraro F, Cammarata-Scalisi F, El-Feghaly J, Morabito A, et al. Extended Overview of Ocular Phenotype with Recent Advances in Hypohidrotic Ectodermal Dysplasia. Children (Basel). 2022 Sep 6;9(9):1357.

177. Sarıkaya Özdemir B, Çetinkaya S, Güleray Lafcı N, Şakar M, Karacan Küçükali G, Elmaoğullari S, et al. Mild phenotype in two siblings with a missense GHR variant. J Pediatr Endocrinol Metab. 2021 Oct 26;34(10):1349–54.

178. Beauchamp MC, Alam SS, Kumar S, Jerome-Majewska LA. Spliceosomopathies and neurocristopathies: Two sides of the same coin? Dev Dyn. 2020 Aug;249(8):924–45.

179. Madej O, Kiff R, Kubba H. How common are ear, nose and throat disorders in children with Noonan syndrome and other RASopathies? Int J Pediatr Otorhinolaryngol. 2023 Mar;166:111480.

180. Mazzone E, Cos Sanchez T, Persico N, Cannie MM, Jani J. Binder syndrome: a phenotype rather than a definitive diagnosis? Ultrasound Obstet Gynecol. 2019 Jan;53(1):131–2.

181. Krmpotic-Nemanic J, Schulz J. Low set ears and small oral cavity with relative macroglossia in Down's syndrome: an explanation of the phenomenon. Chic Med Sch Q. 1970;30(1):41–5.

182. Horowitz S, Morishima A, Vinkka H. The position of the external ear in Turner's

syndrome. Clin Genet. 1976 Mar;9(3):333–40.

183. Bilgin B, Kabaçam S, Taşkıran E, Şimşek-Kiper PÖ, Alanay Y, Boduroğlu K, et al. Epigenotype and phenotype correlations in patients with Beckwith-Wiedemann syndrome. Turk J Pediatr. 2018;60(5):506–13.

184. Srinivasan A, Wright D. Pallister-Killian syndrome. Am J Case Rep. 2014;15:194–8.

185. Allanson JE. Rubinstein-Taybi syndrome: the changing face. Am J Med Genet Suppl. 1990;6:38–41.

186. McElrath AD, Winters R. Mandibulofacial Dysostosis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Mar 30]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK562230/

187. Nazarenko MS, Viakhireva IV, Skoblov MY, Soloveva EV, Sleptcov AA, Nazarenko LP. Meier-Gorlin Syndrome: Clinical Misdiagnosis, Genetic Testing and Functional Analysis of ORC6 Mutations and the Development of a Prenatal Test. Int J Mol Sci. 2022 Aug 17;23(16):9234.

188. Cinalli G, Spennato P, Sainte-Rose C, Arnaud E, Aliberti F, Brunelle F, et al. Chiari malformation in craniosynostosis. Childs Nerv Syst. 2005 Oct;21(10):889–901.

189. Nah H. Suture biology: Lessons from molecular genetics of craniosynostosis syndromes. Clin Orthod Res. 2000 Feb;3(1):37–45.

190. Guion-Almeida ML, Zechi-Ceide RM, Vendramini S, Ju Nior AT. A new syndrome with growth and mental retardation, mandibulofacial dysostosis, microcephaly, and cleft palate. Clin Dysmorphol. 2006 Jul;15(3):171–4.

191. Han X, Zhang Q, Wang C, Han B. Clinical Characteristics and Novel ZEB2 Gene Mutation Analysis of Three Chinese Patients with Mowat-Wilson Syndrome. Pharmgenomics Pers Med. 2023;16:777–83.

192. Miyamoto JJ, Yabunaka T, Moriyama K. Cervical characteristics of Noonan

syndrome. European Journal of Orthodontics. 2014 Apr 1;36(2):226–32.

193. Roberts AE, Allanson JE, Tartaglia M, Gelb BD. Noonan syndrome. Lancet. 2013 Jan 26;381(9863):333–42.

194. Choi TM, Lijten OW, Mathijssen IMJ, Wolvius EB, Ongkosuwito EM. Craniofacial morphology and growth in Muenke syndrome, Saethre-Chotzen syndrome, and TCF12-related craniosynostosis. Clin Oral Investig. 2022 Mar;26(3):2927–36.

195. Masotti C, Oliveira KG, Poerner F, Splendore A, Souza J, Freitas R da S, et al. Auriculo-condylar syndrome: mapping of a first locus and evidence for genetic heterogeneity. Eur J Hum Genet. 2008 Feb;16(2):145–52.

196. Romanelli Tavares VL, Zechi-Ceide RM, Bertola DR, Gordon CT, Ferreira SG, Hsia GSP, et al. Targeted molecular investigation in patients within the clinical spectrum of Auriculocondylar syndrome. Am J Med Genet A. 2017 Apr;173(4):938–45.

197. Lacroix G, Karnoub MA, Vinchon M, Wolber A, Martinot V, Guerreschi P. Auricles Anomalies in Patients With a TCF12 Gene Mutation. J Craniofac Surg. 2023 Feb 1;34(1):302–5.

198. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019 Jan;25(1):30–6.

199. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56.

200. The National Artificial Intelligence Research and Development Strategic Plan.

201. Kumps C, Campos-Xavier B, Hilhorst-Hofstee Y, Marcelis C, Kraenzlin M, Fleischer N, et al. The Connective Tissue Disorder Associated with Recessive Variants in the SLC39A13 Zinc Transporter Gene (Spondylo-Dysplastic Ehlers-Danlos Syndrome Type 3): Insights from Four Novel Patients and Follow-Up on Two Original Cases. Genes (Basel). 2020 Apr 14;11(4):420.

Appendix 1

JEM Journal of Experimental Medicine

ARTICLE

Hemifacial myohyperplasia is due to somatic muscular PIK3CA gain-of-function mutations and responds to pharmacological inhibition

Charles Bayard[1,2,3*], Eleonora Segna[4*], Maxime Taverne[5*], Antoine Fraissenon[2,6,7,8*], Quentin Hennocq[5], Baptiste Periou[9], Lola Zerbib[1,2,3], Sophia Ladraa[1,2], Célia Chapelle[1,2], Clément Hoguin[2,3], Sophie Kaltenbach[10], Patrick Villarese[10], Vahid Asnafi[1,2,10], Christine Broissand[11], Ivan Nemazanyy[12], Gwennhael Autret[13], Nicolas Goudin[14], Christophe Legendre[1,2,15], François-Jérôme Authier[9], Thomas Viel[13], Bertrand Tavitian[1,13], Cyril Gitiaux[1,16], Sylvie Fraitag[17], Jean-Paul Duong[1,17], Clarisse Delcros[1,2], Bernard Sergent[4], Arnaud Picard[1,4], Michael Dussiot[18], Laurent Guibaud[2,6], Roman Khonsari[1,4,5**], and Guillaume Canaud[1,2,3**]

Hemifacial myohyperplasia (HFMH) is a rare cause of facial asymmetry exclusively involving facial muscles. The underlying cause and the mechanism of disease progression are unknown. Here, we identified a somatic gain-of-function mutation of *PIK3CA* in five pediatric patients with HFMH. To understand the physiopathology of muscle hypertrophy in this context, we created a mouse model carrying specifically a *PIK3CA* mutation in skeletal muscles. *PIK3CA* gain-of-function mutation led to striated muscle cell hypertrophy, mitochondria dysfunction, and hypoglycemia with low circulating insulin levels. Alpelisib treatment, an approved PIK3CA inhibitor, was able to prevent and reduce muscle hypertrophy in the mouse model with correction of endocrine anomalies. Based on these findings, we treated the five HFMH patients. All patients demonstrated clinical, esthetical, and radiological improvement with proof of target engagement. In conclusion, we show that HFMH is due to somatic alteration of *PIK3CA* and is accessible to pharmacological intervention.

Introduction

Hemifacial myohyperplasia (HFMH) is a rare cause of facial asymmetry exclusively involving facial muscles, initially reported as "hypertrophy and asymmetry of the facial muscles" (Lee et al., 2001). This disorder is reported in very few patients in the literature (Castillo Taucher et al., 2003; Pereira-Perdomo et al., 2010; Miranda et al., 2010; Siponen et al., 2007; Zissman et al., 2020). The clinical presentation of HFMH patients is strikingly consistent, with unilateral muscular hypertrophy mimicking spasm and orofacial dystonia, leading to diagnostic errors and inadequate management strategies, including aggressive attempts of surgical correction (Zissman et al., 2020). The genetic causes of HFMH are currently unknown.

The recent discovery of the role played by somatic mutation of genes activating the PIK3CA/AKT/mTOR pathway has opened new treatment perspectives for patients (Canaud et al., 2021). Particularly, *PIK3CA* gain-of-function mutations explain the vast majority of overgrowth syndromes (Canaud et al., 2021). *PIK3CA* encodes the 110-kD catalytic α-subunit of PI3K, a lipid kinase

[1]Université Paris Cité, Paris, France; [2]Institut national de la santé et de la recherche médicale U1151, Institut Necker-Enfants Malades, Paris, France; [3]Unité de Médecine Translationnelle et Thérapies Ciblées, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [4]Service De Chirurgie Maxillo-Faciale et Chirurgie Plastique, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [5]Laboratoire Forme et Croissance du Crâne, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [6]Service d'Imagerie Pédiatrique, Hôpital Femme-Mère-Enfant, Hospices Civils de Lyon, Bron, France; [7]CREATIS Unité mixte de recherche 5220, Villeurbanne, France; [8]Service de Radiologie Mère-Enfant, Hôpital Nord, Saint Etienne, France; [9]Service d'anatomie Pathologique, Hôpital Henri Mondor, Assistance Publique Hôpitaux de Paris, Paris, France; [10]Laboratoire d'Oncohématologie, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [11]Pharmacie, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [12]Platform for Metabolic Analyses, Structure Fédérative de Recherche Necker, Institut national de la santé et de la recherche médicale Paris, France; [13]Plateforme Imageries du Vivant, Université Paris Cité, Paris Cardiovascular Research Center, Institut national de la santé et de la recherche médicale, Paris, France; [14]Necker Bio-Image Analysis, Institut national de la santé et de la recherche médicale, Paris, France; [15]Service de Néphrologie, Transplantation Adultes, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [16]Service de Neurophysiologie Clinique Pédiatrique, Centre de Référence des Pathologies Neuromusculaires, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [17]Service d'Anatomie Pathologique, Hôpital Necker-Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France; [18]Institut national de la santé et de la recherche médicale U1163, Laboratory of Cellular and Molecular Mechanisms of Hematological Disorders and Therapeutic Implications, Laboratoire d'Excellence GR-Ex, Paris, France.

*C. Bayard, E. Segna, M. Taverne, and A. Fraissenon contributed equally to this paper; **R. Khonsari and G. Canaud contributed equally to this paper. Correspondence to Guillaume Canaud: guillaume.canaud@inserm.fr; Roman Khonsari: roman.khonsari@aphp.fr.

that controls signaling pathways involved in cell proliferation, motility, survival, and metabolism (Madsen and Vanhaesebroeck, 2020). Post-zygotic mosaic gain-of-function *PIK3CA* mutations result in protein activation, leading to abnormal cellular proliferation, tissue hyperplasia, and organ overgrowth. The understanding of the genetic bases of overgrowth has enabled the treatment of patients presenting *PIK3CA*-related overgrowth syndromes (PROS) with a specific PI3K inhibitor (BYL719, alpelisib), initially designed as an antineoplastic drug (Venot et al., 2018). Alpelisib is efficiently tackling soft-tissue overgrowth in PROS (Venot et al., 2018; Ladraa et al., 2022; Delestre et al., 2021; Morin et al., 2022).

We hypothesized that the PIK3CA/AKT/mTOR pathway was abnormally affected in patients with HFMH. Here, we report the results of the clinical screening, mechanistic investigations, and treatment of five patients with HFMH.

Results

PIK3CA gain-of-function mutations explain HFMH

We identified five pediatric patients, including three girls, ranging from 4 to 15 yr old with clinical presentation of HFMH (Fig. 1 and Fig. S1). In more detail, 5/5 patients had chin skin dimpling and small nasal vestibule, 4/5 had chin deviation, dysmorphic ear, narrow palpebral fissure, and eyebrow ptosis, 3/5 had lip commissure canting, and finally, 3/5 had nasal deviation. All patients had magnetic resonance imaging (MRI) demonstrating muscular hypertrophy affecting in various combinations facial, masticatory (pterygoids and masseter), and neck (scalene and sternocleidomastoid) muscles (Fig. 1). All patients had benefited from computerized tomography (CT) scan imaging before HFMH diagnosis that showed bone anomalies in 4/5 cases including mild chin deviation in 3/4 and major skeletal deformation with homolateral maxillary and mandibular growth impairment in 1/5. All patients had available electroneuromyography assessments that showed normal results, including normal blink reflex responses and needle electromyogram of facial muscles.

2/5 patients had been treated with botulinum toxin with injections performed every 6 mo. Patient 5 benefited from three injections in mentalis, zygomaticus minor, and major and orbicularis oculi muscles (130 then twice 200 I.U.), and patient 3 benefited from nine injections in the same three muscles (I.U. injected from 50 to 400). For both patients, no relevant esthetic or functional results were observed.

We performed muscular biopsies from the affected zones (oral mucosa, buccinator muscle, and masseter muscle) in 5/5 patients, under general or local anesthesia depending on age, through an intraoral approach. Standard histology was normal. Genotyping of the biopsies revealed the presence of *PIK3CA* variants in all tested samples. Pathogenic *E545K* variant was found in patient 1 and 5 with a variant allelic frequency (VAF) of 15% and 14%, respectively. The *E542K* variant was detected in patient 3 (VAF 12%) and 4 (VAF 21%). Lastly, the *H1047R* variant was identified in patient 2 (VAF 25%; Table 1). All variants were already characterized as pathogenic (Canaud et al., 2021).

Thus, we concluded that *PIK3CA* gain-of-function mutations explained all five HFMH cases.

A mouse model of PIK3CA gain-of-function mutations in skeletal muscles

To gain insight into the mechanisms of *PIK3CA*-related muscle overgrowth, we designed a mouse model carrying specifically a *PIK3CA* gain-of-function mutation in striated muscle. To this aim, we interbred the R26StopFLP110* mouse strain with HSA Cre mice to generate $PIK3CA^{HSA\text{-}CreER}$ animals that express a constitutively overactivated form of PIK3CA upon tamoxifen administration. To follow Cre recombination, $PIK3CA^{HSA\text{-}CreER}$ mice were then interbred with Gt(ROSA)26Sortm4(ACTB-tdTomato-EGFP)Luo/J mice. In all tissues, these mice express a cell-membrane-localized tdTomato fluorescent protein that is replaced by GFP after *Cre* recombination.

To overcome developmental issues, we used 6-wk-old mice that were treated with a daily dose of 40 mg.kg-1 tamoxifen for 5 d to induce Cre recombination. We observed that starting 3 wk after *Cre* recombination, $PIK3CA^{HSA\text{-}CreER}$ mice progressively gained weight compared with their wild-type littermates ($PIK3CA^{WT}$; Fig. 2 A). This was the case for both males and females. Around 11 wk after induction in males and 24 wk in females, body weight of control mice became higher than in $PIK3CA^{HSA\text{-}CreER}$ (Fig. 2 A). Whole-body MRI at 15 wk after tamoxifen revealed skeletal muscle hypertrophy in $PIK3CA^{HSA\text{-}CreER}$ mice that was approximately twice as in controls (Fig. 2, B and C). Fat content in $PIK3CA^{HSA\text{-}CreER}$ was significantly reduced compared with controls (Fig. 2, B and C). Skeletal muscle hypertrophy was associated with gain in muscle strength compared with controls (Fig. 2 D).

We sacrificed 20 controls and 20 $PIK3CA^{HSA\text{-}CreER}$ mice 15 wk after tamoxifen administration. Necropsy examination confirmed diffuse skeletal muscle hypertrophy with adipose tissue shrinkage compared with control littermates (Fig. 2 E). We confirmed that mutant p110α (p110*) and GFP were expressed in striated muscles of $PIK3CA^{HSA\text{-}CreER}$ mice (Fig. 2, F and G). Histological examination revealed hypertrophic striated cells (Fig. 2 H and Fig. S2, A–D). We did not observe any other histological anomalies. Western blot and immunofluorescence studies showed AKT/mTOR pathway activation in striated muscles of $PIK3CA^{HSA\text{-}CreER}$ mice (Fig. 2, I and J). Mechanistically, PIK3CA is involved in cell growth and proliferation but KI67 immunostaining did not show hyperplasia (Fig. 2 K). However, using the Amnis ImageStream system, we confirmed that cells isolated from striated muscle of $PIK3CA^{HSA\text{-}CreER}$ mice were hypertrophic compared with controls (Fig. 2 L). Blood cell count did not reveal any anomalies (Fig. S2 E). However, we observed that $PIK3CA^{HSA\text{-}CreER}$ mice were hypoglycemic compared with controls (Fig. 2 M). We recently identified an endocrine loop leading to chronic hypoglycemia in another mouse model carrying specifically *PIK3CA* gain-of-function mutation in adipose tissue (Ladraa et al., 2022). Indeed, *PIK3CA* gain-of-function mutation led to permanent glucose transporters addressing the cell membrane, allowing glucose entry and subsequent hypoglycemia. The latter reduced insulin secretion, which in turn dramatically increased insulin-like growth factor binding protein production followed by liver IGF-1 sequestration and a reduction in IGF-1 circulating levels similar to clinical observations. We obtained similar results in

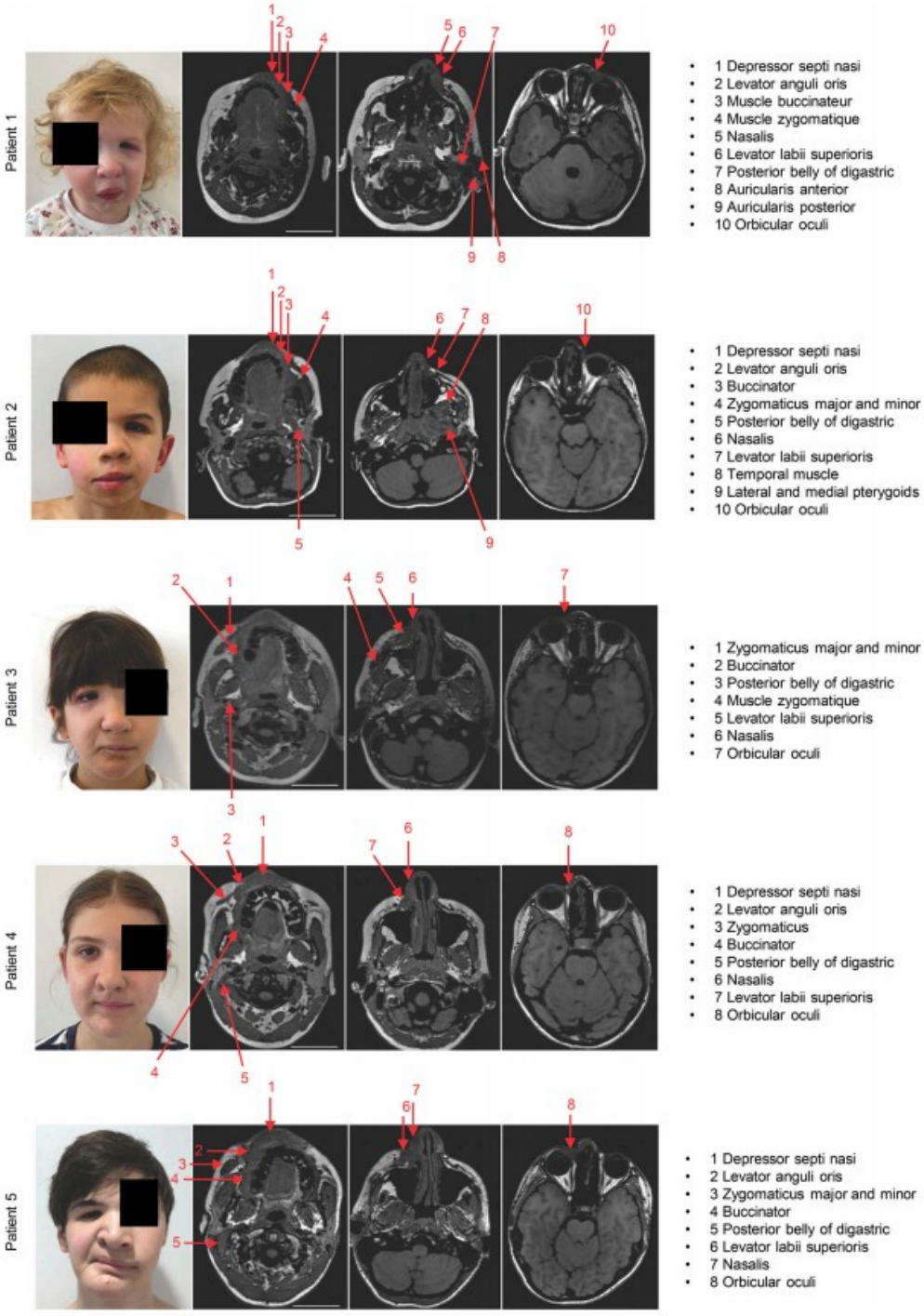


Figure 1. **Patients with hemifacial myohyperplasia.** Representative photographs of the five patients and affected muscles assessed with MRI. Scale bar: 5 cm.

Table 1. **Patient characteristics**

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Age, year	4	6	9	14	15
Specimen used for genotyping	Muscle	Muscle	Muscle	Muscle	Muscle
DNA changes	c.1633G>A	c.3140A>G	c.1624G>A	c.1624G>A	c.1633G>A
VAF (%)[a]	15	25	12	21	14
Amino acid change	p.Glu545Lys	p.His1047Arg	p.Glu542Lys	p.Glu542Lys	p.Glu545Lys
COSMIC[b] genomic mutation ID	COSV55873239	COSV55873195	COSV55873227	COSV55873227	COSV55873239
Treatments prior to alpelisib	-	-	Botulinum toxin injections	-	Botulinum toxin injections

[a]Corresponds to the percentage of alternate or mutant reads to total reads detected by next-generation sequencing.
[b]Catalogue of Somatic Mutations in Cancer.

PIK3CA$^{HSA\text{-}CreER}$ mouse model with low insulin and IGF1 levels (Fig. 2, M and O; and Fig. S2 F) but conserved insulin secretion capacities as demonstrated using oral glucose tolerance test (GTT; Fig. 2 P). Consistently, positron emission tomography and computed tomography (PET-CT) scan showed higher fluoro-D-glucose (FDG) muscular uptake in *PIK3CA*$^{HSA\text{-}CreER}$ mice compared with controls (Fig. 2 Q).

Mitochondria play a major role in energy for striated muscle metabolism. PI3K has been involved in mitochondrial function (Li et al., 2019); we therefore explored the impact of *PIK3CA* gain-of-function on mitochondria number and activity in the *PIK3CA*$^{HSA\text{-}CreER}$ mouse model. Flow cytometry of isolated striated muscle cells revealed alteration in the mitochondrial transmembrane potential as assessed by tetramethylrhodamine ethyl ester (TMRE) staining with reduced Mitotracker expression (Fig. 2, R and S). 10-N-nonyl acridine orange (NAO) staining showed reduced mitochondrial mass in striated muscle cells of *PIK3CA*$^{HSA\text{-}CreER}$ compared with controls (Fig. 2 T). Plasma metabolites metabolomics analysis revealed energy production with increased ATP production with mitochondria and anabolic activities (Fig. S3).

Hence, we created a mouse model of PIK3CA-related skeletal muscle overgrowth affecting the gross morphology, cellular structure, and function of striated muscles.

Alpelisib improves PIK3CA$^{HSA\text{-}CreER}$ mouse model

We recently identified alpelisib (BYL719), a PIK3CA inhibitor, as a promising treatment for patients with PROS, and very recently this drug has been approved by the US FDA for patients with PROS (over 2 yr of age; Venot et al., 2018). We decided to test whether this molecule was efficient at improving skeletal muscle overgrowth in our mouse model. To this end, we used two different approaches. We first administered alpelisib daily starting 48 h after *Cre* induction as a preventive study. We observed that alpelisib-treated *PIK3CA*$^{HSA\text{-}CreER}$ mice had an overtly normal appearance during the 22 wk of treatment and a body weight increase similar to that of controls (Fig. 3 A). MRI performed 8 wk after alpelisib initiation showed no skeletal muscle overgrowth compared with vehicle *PIK3CA*$^{HSA\text{-}CreER}$-treated mice (Fig. 3, B and C). Mice were then sacrificed and a necropsy examination confirmed that alpelisib-treated *PIK3CA*$^{HSA\text{-}CreER}$ mice had no muscle hypertrophy and were indistinguishable from controls. Histology showed no difference between *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with preventive alpelisib and *PIK3CA*^{WT} mice (Fig. 3 D). Western blot and immunofluorescent studies showed blunted AKT and S6RP phosphorylation (Fig. 3, E and F). *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with preventive alpelisib demonstrated normal 12-h fasted glycemia compared with *PIK3CA*$^{HSA\text{-}CreER}$ vehicle–treated mice (Fig. 3 G). Consistently, insulin and IGF-1 circulating levels were corrected (Fig. 3, H and I).

Next, we administered alpelisib to *PIK3CA*$^{HSA\text{-}CreER}$ mice 2 wk after *Cre* induction when global muscle hypertrophy was already prominent for 20 additional weeks, as a therapeutic study. Following alpelisib introduction, we noticed a rapid body weight decrease in alpelisib-treated *PIK3CA*$^{HSA\text{-}CreER}$ mice compared with control (Fig. 3 A). MRI performed at 8 wk showed significant reduction in muscular volume compared with *PIK3CA*$^{HSA\text{-}CreER}$ vehicle mice (Fig. 3, B and C; and Fig. S2, B–D). Similar to previous experiments with early alpelisib introduction, we observed that the histology of the striated muscle of *PIK3CA*$^{HSA\text{-}CreER}$ mice treated in the therapeutic study was conserved (Fig. 3 D), with reduced phosphorylation of the AKT and S6RP proteins (Fig. 3, E and F). Blood glucose, insulin, and IGF-1 circulating levels were increased in the therapeutic alpelisib group (Fig. 3, G–I).

Alpelisib treatments were associated with a partial correction of the different metabolomic anomalies (Fig. S4, A and B).

We then concluded that alpelisib was efficient at both preventing and treating PIK3CA-related striated muscle hypertrophy.

Alpelisib improves patients with PIK3CA-related skeletal muscle overgrowth

Based on the preclinical data, we decided to treat the five HFMH pediatric patients with alpelisib. We obtained the authorization from the French regulatory agency (Agence nationale de sécurité du médicament et des produits de santé) and administered a daily dose of 50 mg per day for each patient. Following alpelisib introduction, we noticed in the five patients clinical meaningful hemifacial volume reduction with softer tissues within 3–4 wk (Fig. 4 A). Based on 2D photography quantification, we confirmed the lowering of the lip commissure, the widening of the palpebral fissure, the reduction of nose and chin deviation, and a

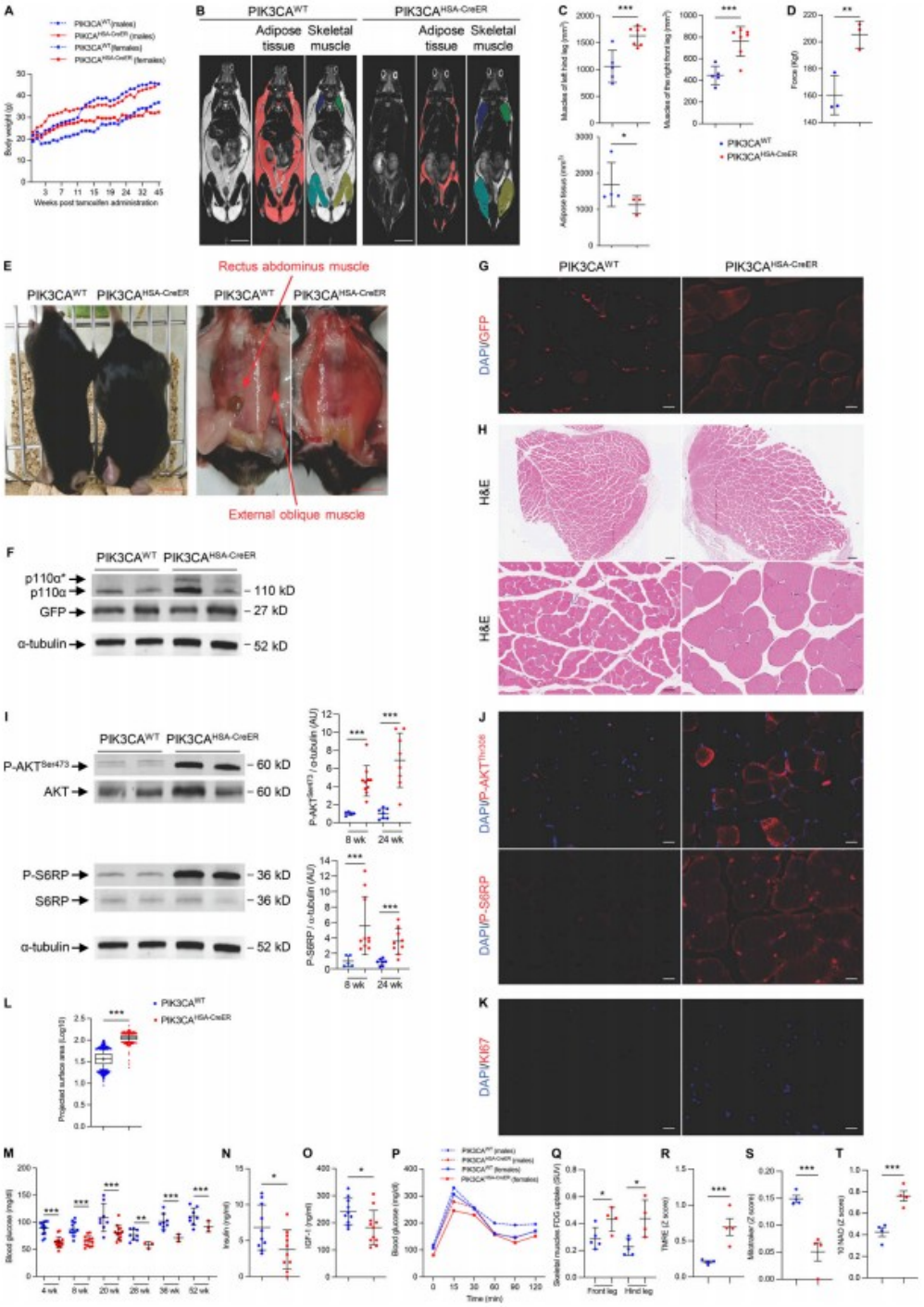


Figure 2. **A mouse model of *PIK3CA*-related skeletal muscle overgrowth. (A)** Male and female body weights of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ (n = 15 per group) mice following *Cre* recombination. **(B)** Coronal whole-body T2-weighted magnetic resonance images of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Scale bar:

1 cm. **(C)** Adipose tissue and skeletal muscle volume quantification. **(D)** Strength measured using grip test (n = 3 per group). **(E)** Representative pictures of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Scale bar: 1 cm. **(F)** Western blot of p110α and GFP in skeletal muscles of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 3 per group). **(G)** Representative GFP immunofluorescence in skeletal muscles of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Scale bar: 10 µm. **(H)** Representative H&E staining of striated muscles of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Scale bar: 20 µm. **(I)** Western blot of P-AKT^{Ser473}, total AKT, P-S6RP, and S6RP in skeletal muscles of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice and quantification at 8 and 24 wk of age (n = 5–10 mice per group). **(J)** Representative immunofluorescence of P-AKT^{Thr308} and P-S6RP in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Scale bar: 10 µm. **(K)** Representative immunofluorescence of KI67 in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Scale bar: 10 µm. **(L)** Quantification of skeletal muscle cell area of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 5 mice per group). **(M)** 12-h fasted glycemia in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 5–13 mice per group). **(N)** Insulin circulating levels in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 10 per group). **(O)** Circulating IGF-1 levels in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 10–11 per group). **(P)** Oral tolerance test (GTT) in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 6 mice per group). **(Q)** 18F-FDG uptake in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 4 per group). **(R–T)** (R) TMRE, (S) Mitotracker, and (T) 10 NAO staining in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice (n = 4 per group). Data are shown as mean ± SEM. *P < 0.05, **P < 0.01, ***P < 0.001 (two-tailed unpaired t test). Each dot represents one mouse. Data are representative of at least two independent experiments. Of note, all blots from this figure originate from the same gel. Source data are available for this figure: SourceData F2.

better skin texture with a decrease of dimpling effects. 3D photography follow-up before and after treatment initiation was available in three patients (1, 3, and 5). Two-blocks partial least-squares regressions (PLS2B) analysis showed a strong covariation between facial shape and treatment duration in two out of three patients (Fig. 4 B and Fig. S5). In 3/3 patients, the hypertrophic regions were preferentially improved by treatment (Fig. 4 B and Fig. S5).

MRI performed 6 mo after drug introduction showed a modest mean muscular volume reduction of 2.56% (Fig. 4 C). During the study, the drug was well tolerated with no adverse events reported by patients and parents. Since we used a standard dose of alpelisib (50 mg per day) in the five patients while their body weights and ages were different, we decided to explore AKT activity in the affected skeletal muscles after 6 mo on the treatment. We obtained a biopsy in four out of five patients (patients 2–5). Immunofluorescence studies revealed increased AKT and S6RP phosphorylation in the biopsies from 4/4 patients compared with healthy controls (Fig. 4, D–F). Following drug introduction, we observed that AKT and S6RP phosphorylation were blunted demonstrating that alpelisib penetrated into skeletal muscles at this dose in 4/4 patients (Fig. 4, D–F).

In parallel, we performed an unbiased approach of circulating plasma metabolites before and 6 mo following alpelisib introduction. We observed significant changes in key metabolite levels involved in mitochondria and anabolic activities (Fig. 5).

Discussion

Here, we report for the first time that HFMH is due to a somatic gain-of-function mutation of *PIK3CA* in striated muscles and indeed, belongs to the *PIK3CA*-related disorder spectrum. We further created a mouse model harboring a *PIK3CA* gain-of-function mutation in striated muscles that demonstrates muscle overgrowth. We finally showed that mouse and patient striated muscles are sensitive to PIK3CA inhibition using alpelisib.

HFMH has long been standing without molecular explanation. All 5/5 patients in our cohort carried a somatic variant of *PIK3CA*. From a disease mechanism point of view, these mutations make sense since PIK3CA is a major actor of cell growth (Madsen and Vanhaesebroeck, 2020). This was confirmed by the *PIK3CA*$^{HSA\text{-}CreER}$ mice that demonstrated increased skeletal muscle volumes and strength. PIK3CA activation in mice skeletal muscles induced anabolic activities and mitochondrial changes. Identification of *PIK3CA* mutations opens new therapeutic perspectives for HFMH patients since surgical management, botulinum toxin injections, and physical therapy are currently the only treatment solutions and seem to be associated with little to no efficiency. Interestingly, two patients were previously unsuccessfully treated with botulinum toxin injection and demonstrated some degree of response to alpelisib. Since alpelisib benefited from an accelerated US FDA approval for patients with PROS, our work identifies a new patient population that may benefit from the drug. In these potential new alpelisib indications, it is important to determine the affected tissues that are sensitive to PIK3CA inhibition such as we recently reported in lymphatic vessels (Delestre et al., 2021; Morin et al., 2022) and adipocytes (Ladraa et al., 2022). Here, for the first time, to monitor drug efficacy, we performed secondary biopsies 6 mo after drug introduction, and tissue analyses revealed AKT inhibition in all tested patients demonstrating that low doses of alpelisib were sufficient to achieve target inhibition.

One limitation of this study is that the transgene utilized in our mouse model is not found in human pathology. However, it possesses significant potency in activating the AKT/mTOR pathway and overcoming mouse resistance to the *PIK3CA* human variant (Venot et al., 2018).

In conclusion, in this study, we precise the clinical presentation of a little-known syndrome and determine its molecular bases. We further expand the phenotypic spectrum of PROS, we detail the effects of PIK3CA hyperactivity on skeletal muscles using a mouse model, and we show that PIK3CA inhibition seems to be a promising option for treating patients with HFMF, based both on shape quantification and functional assessments. Further investigations are required to establish a more precise determination of the frequency of *PIK3CA* mutation in HFMF, and the encouraging clinical results obtained using alpelisib will have to be confirmed in larger studies.

Materials and methods

Animals

Following previously described procedures (Delestre et al., 2021), we interbred homozygous *R26StopFLP110** (Stock# 012343) mice with *HSA Cre-ER* mice (Stock# 025750), both obtained from The Jackson Laboratory. We obtained *R26StopFLP110***+/−* x *HSA Cre-ER+* (henceforth *PIK3CA*$^{HSA\text{-}CreER}$) and *R26StopFLP110***+/−* x *HAS Cre-ER−* (henceforth *PIK3CA*^{WT}) littermate mice. The p110* protein

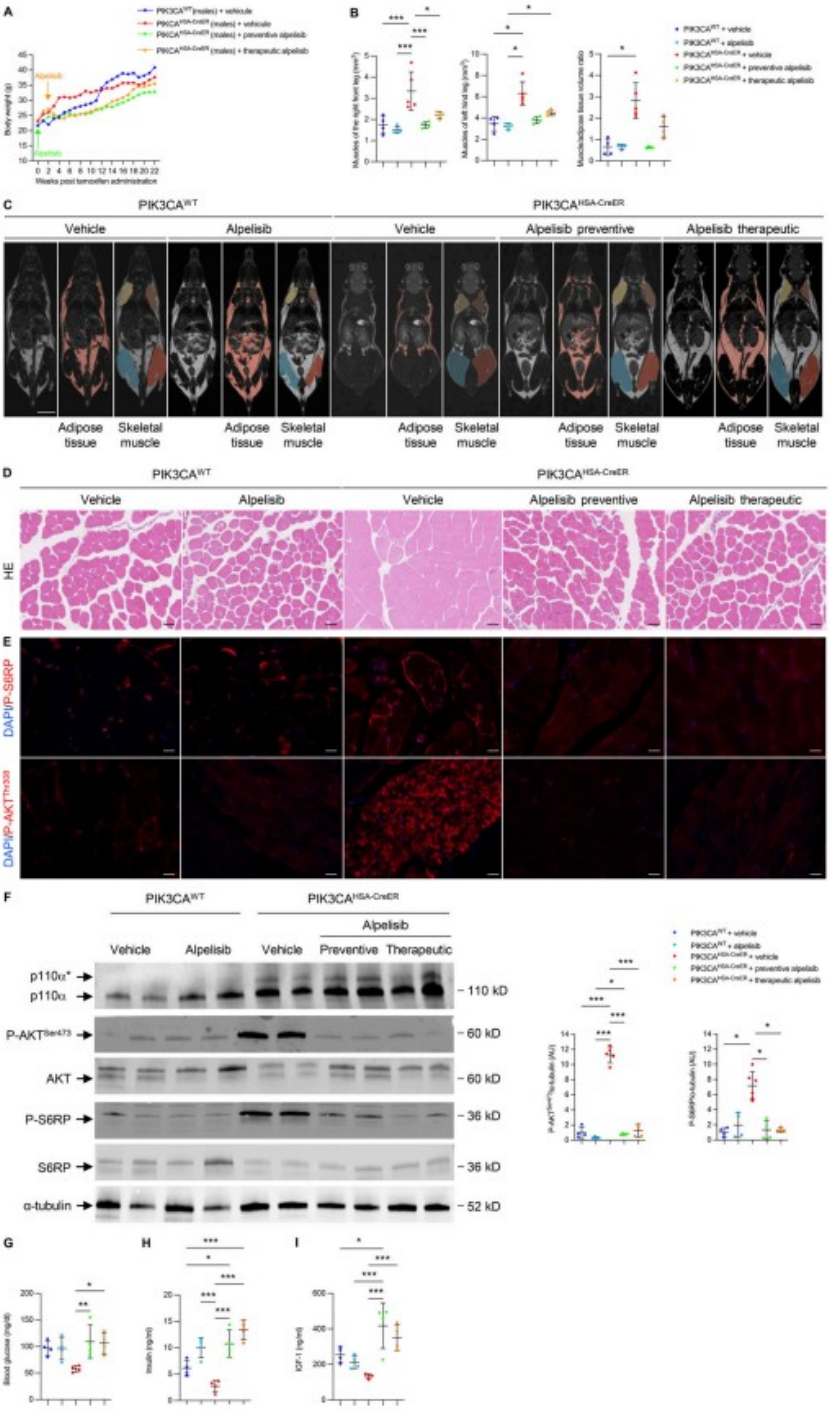


Figure 3. **Alpelisib prevents and reverses skeletal muscle overgrowth in *PIK3CA*$^{HSA\text{-}CreER}$ mice. (A)** Male body weights of *PIK3CA*^{WT} (n = 6) and *PIK3CA*$^{HSA\text{-}CreER}$ treated with either vehicle (n = 12) or preventive (n = 12) or curative alpelisib (n = 12). **(B)** Skeletal muscle and adipose tissue volume quantification. **(C)** Whole-body

T2-weighted magnetic resonance images of *PIK3CA*^{WT} treated either with vehicle (n = 4) or alpelisib (n = 4), *PIK3CA*$^{HSA\text{-}CreER}$ vehicle-treated (n = 5), *PIK3CA*$^{HSA\text{-}CreER}$-treated with preventive alpelisib (n = 4), and *PIK3CA*$^{HSA\text{-}CreER}$-treated with therapeutic alpelisib (n = 4) mice. Scale bar: 1 cm. **(D)** Representative H&E staining of skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or preventive or curative alpelisib. Scale bar: 10 µm. **(E)** Representative immunofluorescence of P-AKT^{Thr308} and P-S6RP in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or alpelisib. Scale bar: 10 µm. **(F)** Western blot and quantification of P110, P-AKT^{Ser473}, AKT, P-S6RP, and S6RP in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or alpelisib (n = 4–5 mice per group). **(G)** 12-h fasted glycemia in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with vehicle, preventive, or curative alpelisib (n = 4 mice per group). **(H and I)** (H) Insulin and (I) IGF-1 circulating levels in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with vehicle, preventive, or curative alpelisib (n = 4 mice per group). Data are shown as mean ± SEM. *P < 0.05, **P < 0.01, ***P < 0.001 (ANOVA, followed by the Tukey–Kramer post hoc test). Each dot represents one mouse. Data are representative of at least two independent experiments.

expressed by *R26StopFLP110** mice is a constitutively active chimera that contains the iSH2 domain of p85 fused to the NH2-terminus of p110 via a flexible glycine linker (Klippel et al., 1996). To generate tissue-specific p110*-transgenic mice, a cloned loxP-flanked neoR-stop cassette was inserted into a modified version of pROSA26-1, followed by cDNA encoding p110* and then a frt-flanked IRES-EGFP cassette and a bovine polyadenylation sequence (*R26StopFLP110**; Srinivasan et al., 2009). To follow *Cre* recombination, *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice were then interbred with *Gt(ROSA)26Sor*$^{tm4(ACTB\text{-}tdTomato,\text{-}EGFP)Luo/J}$ mice (Muzumdar et al., 2007; Delestre et al., 2021). These mice express a cell membrane–localized tdTomato fluorescent protein in all tissues that is replaced by GFP after *Cre* recombination. All mice used were on a C57BL/6 background.

Animals were fed ad libitum and housed at a constant ambient temperature in a 12-h light cycle. Animals were fed regular chow food (2018 Teklad global 18% protein rodent diets, 3.1 Kcal/g; Envigo). Animal procedures were approved by the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (APAFIS N°20439-2018121913526398 and 2021110914486827). All appropriate procedures were followed to ensure animal welfare.

At the age of 6 wk, *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice received a daily dose of 40 mg.kg^{-1} tamoxifen for 5 consecutive days to induce Cre recombination. Tamoxifen was administered through oral gavage.

PIK3CA^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice were treated with the PIK3CA inhibitor alpelisib (MedChem Express; 50 mg.kg^{-1} in 0.5% carboxymethylcellulose [Sigma-Aldrich], daily oral gavage) or vehicle (0.5% carboxymethylcellulose [Sigma-Aldrich], daily oral gavage). Treatment was started either immediately (preventive study) or 6 wk (therapeutic study) following *Cre* induction. The last dose of alpelisib or vehicle was administered ~3 h before sacrifice.

All mice were fasted for 12 h before blood glucose measurement (Accuchek Performa; Roche Diagnostic), MRI, and sacrifice.

Grip test

Grip strength performance of the mice was evaluated using a grip strength dynamometer obtained from Bioseb. To assess hindlimb muscle strength, the mice were positioned on a grid surface. Gentle traction was applied to the mice's tails in the opposite direction, and the maximum strength exerted by each mouse before releasing its grip was measured five times. A recovery period of 30 s was provided between each measurement. The average value from the five measurements was calculated as the indicator of hindlimb grip strength.

MRI evaluation

All images were acquired with a 4.7-T small-animal MRI system (BioSpec USR47/40; Bruker BioSpin) on the Plateforme Imageries du Vivant, Université de Paris, Paris Cardiovascular Research Center, Institut national de la santé et de la recherche médicale, Paris, France. Mice underwent whole-body MRI using 3D T2-weighted sequences with and without fat saturation. Volumetric evaluation by MRI was performed with 3D Slicer software (Fedorov et al., 2012) and IntelliSpace Portal software (Philips Healthcare). Whole-body adipose tissue segmentation was obtained by thresholding on a T2-weighted sequence and then removing hypersignal related to water on a T2-weighted sequence with fat saturation. Muscles of each limb were manually segmented according to their low signal intensity and anatomy on a T2-weighted sequence. For both adipose and muscle tissues, volume was calculated by summing images based on 2D contours and slice thickness.

FDG PET-CT imaging

Mice were fasted overnight with free access to water. Mice were then anesthetized (2 ± 0.5% isoflurane in dioxygen) and weighed, and glycemia was measured in blood drawn from the caudal ventral artery using an Accu-Chek Aviva Nano A (Accu-Chek). A 29G needle catheter (Thermo Fischer Scientific) connected to 5 cm polyethylene tubing (Tygon Microbore Tubing, 0.010" × 0.030" OD; Thermo Fisher Scientific) was inserted into the caudal vein for radiotracer injection. 9.2 ± 1.5 MBq of 2′-deoxy-2′-[18F]FDG (Advanced Applied Applications) in 0.2 ml saline was injected via the catheter. Mice were left awake in their cage for 45 min and then installed into the PET-CT dedicated bed. Respiration and body temperature were registered. Body temperature was maintained at 34 ± 2°C and anesthesia was controlled according to the breathing rate throughout the entire PET-CT examination. CT was acquired in a PET-CT scanner (nanoScan PET-CT; Mediso Medical Imaging Systems) using the following acquisition parameters: semicircular mode, 50 kV tension, 720 projections full scan, 300 ms per projection, and binning 1:4. CT projections were reconstructed by filtered retroprojection (filter: cosine; cutoff: 100%) using Nucline 3.00.010.0000 software (Mediso Medical Imaging Systems). 55 min after tracer injection, PET data were collected for 10 min in list mode and binned using a 5-ns time window with a 400–600 keV energy window and a 1:5 coincidence mode. Data were reconstructed using the Tera-Tomo reconstruction engine (3D-OSEM-based manufactured customized algorithm) with expectation maximization iterations, scatter, and attenuation correction. Volumes of interest were delineated on the organs

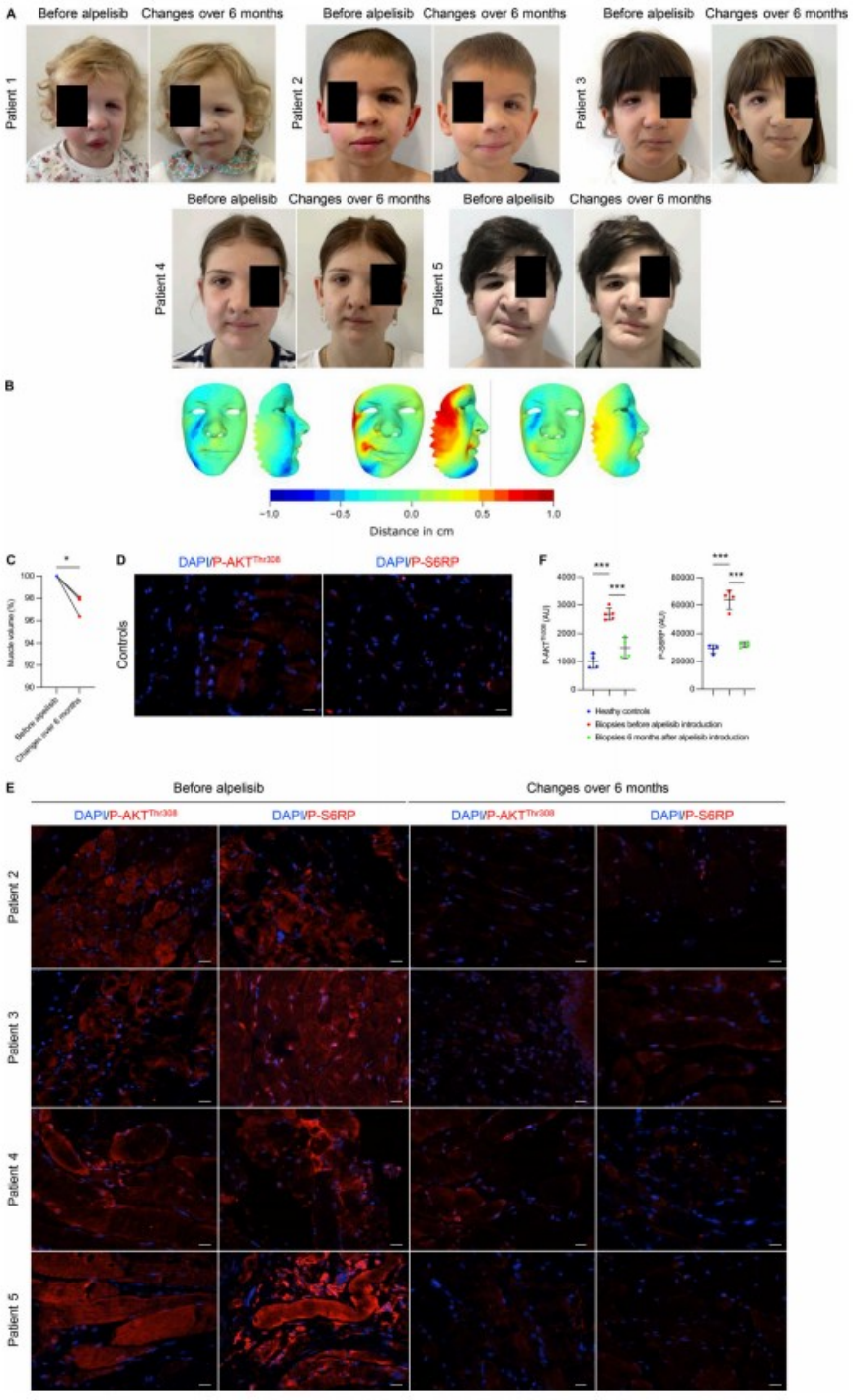


Figure 4. **Alpelisib improves hemifacial myohyperplasia phenotype in patients. (A)** Photographs of the five patients before and 6 mo after alpelisib introduction. **(B)** 2D mapping of facial asymmetry in 3/5 patients. Blue colors are associated with expanded regions relative to contralateral control. Red colors

are associated with regions that display shrinkage compared to contralateral control. Distances (cm) indicate the absolute distance between each node of the affected side and the mirrored non-affected side. Affected side: right. **(C)** Volumetric changes of facial muscles in patients before and after drug introduction. **(D and E)** (D) Representative immunofluorescence of P-AKT^{Thr308} and P-S6RP in healthy contralateral skeletal muscle and (E) in the affected skeletal muscle before and 6 mo after alpelisib introduction. **(F)** Quantification of immunofluorescence intensity. AU: Arbitrary units. Scale bar: 10 µm. Data are shown as mean ± SEM. *P < 0.05 and ***P < 0.001 (two-tailed unpaired *t* test for C and ANOVA, followed by the Tukey–Kramer post hoc test for F). Each dot represents one mouse. Data are representative of at least two independent experiments.

or anatomical structure of interest on PET/CT fusion slices using the PMOD software package (PMOD Technologies Ltd.). FDG accumulation was quantified as the standard uptake value, which measures the ratio of the radioactivity concentration in volume of interest to the whole-body concentration of the injected radioactivity.

Blood and plasma analysis

At the end of each experiment, blood samples were collected from the mice in EDTA-coated tubes. To measure blood count, fresh blood samples were analyzed on a hematology analyzer (ProCyte Dx; IDEXX Laboratories) and centrifuged at 500 × *g* for 15 min. The collected plasma concentration was used to determine insulin (U-PLEX mouse insulin Assay [Meso Scale Discovery], ref# K1526HK) and IGF-1 (ref# MG100; Novus Biologicals) circulating levels using enzymatic methods from commercially available kits.

Oral glucose test

Mice were fasted overnight (12 h) with access to drinking water. All body weights were measured and tails were carefully cut for blood glucose determination (time point 0). During GTT, freshly prepared glucose solution was administered orally (1 g glucose/kg body weight of 20% glucose solution in water; ref# G8270; Sigma-Aldrich). After 15, 30, 45, 60, 90, and 120 min, blood glucose was measured again at all time points.

Morphological analysis

Mouse tissues were fixed in 4% paraformaldehyde and paraffin embedded. Tissue sections (4 µm thick) were stained with hematoxylin and eosin (H&E).

Measurement of muscle fiber size

H&E slides were scanned with a NanoZoomer 2.0HT (Hamamatsu) and analyzed with Qupath-0.2.3 (Bankhead et al., 2017). Muscles were segmented using deep-learning Cellpose algorithm (Stringer et al., 2021). For both, areas were then measured with Fiji (Schindelin et al., 2012).

Immunohistochemistry and immunofluorescence

Paraffin-embedded tissue sections (4 µm) were submitted to antigen retrieval protocols using high temperature (120°C) and high pressure in citrate buffer and a pressure cooker. Sections were then incubated with primary antibodies (Table S1). For the immunofluorescence procedure, appropriate Alexafluor-conjugated secondary antibodies (Thermo Fischer Scientific) were incubated on the samples and analyzed using an LSM 700 confocal microscope (Zeiss) or Eclipse Ni-E (Nikon). Immunohistochemistry revelation was performed with appropriate horseradish peroxidase linked secondary antibodies and analyzed with E800 (Nikon).

In human skin biopsies, P-S6RP and P-AKT^{T308} staining were segmented with Ilastik v1.3.3post3, a machine-learning pixel classification open-source software. For each immunofluorescence staining, mean intensity area and quantity were measured with Fiji v2.3.0/1.53f51 (Quantity = Mean intensity × Area) and normalized by tissue area (Berg et al., 2019).

Western blot

Tissues were crushed and then lysed in radioimmunoprecipitation assay lysis buffer supplemented with phosphatase and protease inhibitors. Protein concentrations were determined through the bicinchoninic acid method (Pierce). Then, protein extracts were resolved by SDS-PAGE before being transferred onto the appropriate membrane and incubated with the primary antibody (Table S1) followed by the appropriate peroxidase-conjugated secondary antibody (dilution 1:10,000). Chemiluminescence was acquired using Chemidoc MP and bands were quantitated using Image Lab Software (Bio–Rad Laboratories).

Tissue digestion

Muscles of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice were harvested and rinsed with PBS 1X (Gibco). After cutting them into small pieces, digestion buffer was added with DNAse (0.1 mg/ml), Dispase I (0.8 mg/ml), and Collagenase P (0.2 mg/ml) in 10 ml of RPMI (Gibco) and incubated for 40 min at 37°C on GentleMACS (Miltenyi) with the appropriate program. Following dissociation, tissues were filtered (70 µm; Clearline), centrifuged 5 min at 250 *g*, and resuspended in PBS.

Mitotracker cytometry

Following treatment with either vehicle or alpelisib, muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice were collected and digested. Cells were resuspended in 100 µl of PBS in a 96-well round-bottomed plate (Thermo Fisher Scientific). Cells were incubated with fluorescent buffer in the dark for 10 min at 37°C (Table S1) and rinsed with PBS-FCS 2%. Cells were analyzed using Sony Spectral ID7000 and all flow data were processed with Sony ID7000 software and Kaluza software.

Imaging flow cytometry (ImageStream)

Skeletal muscles from *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice were isolated, rinsed in PBS 1X, cut into small pieces, and digested as detailed below. After dissociation, cell suspensions were filtered (70 µm; Clearline), centrifuged 10 min at 350 *g*, and resuspended in PBD solution 1X supplemented with 2% fetal bovine serum

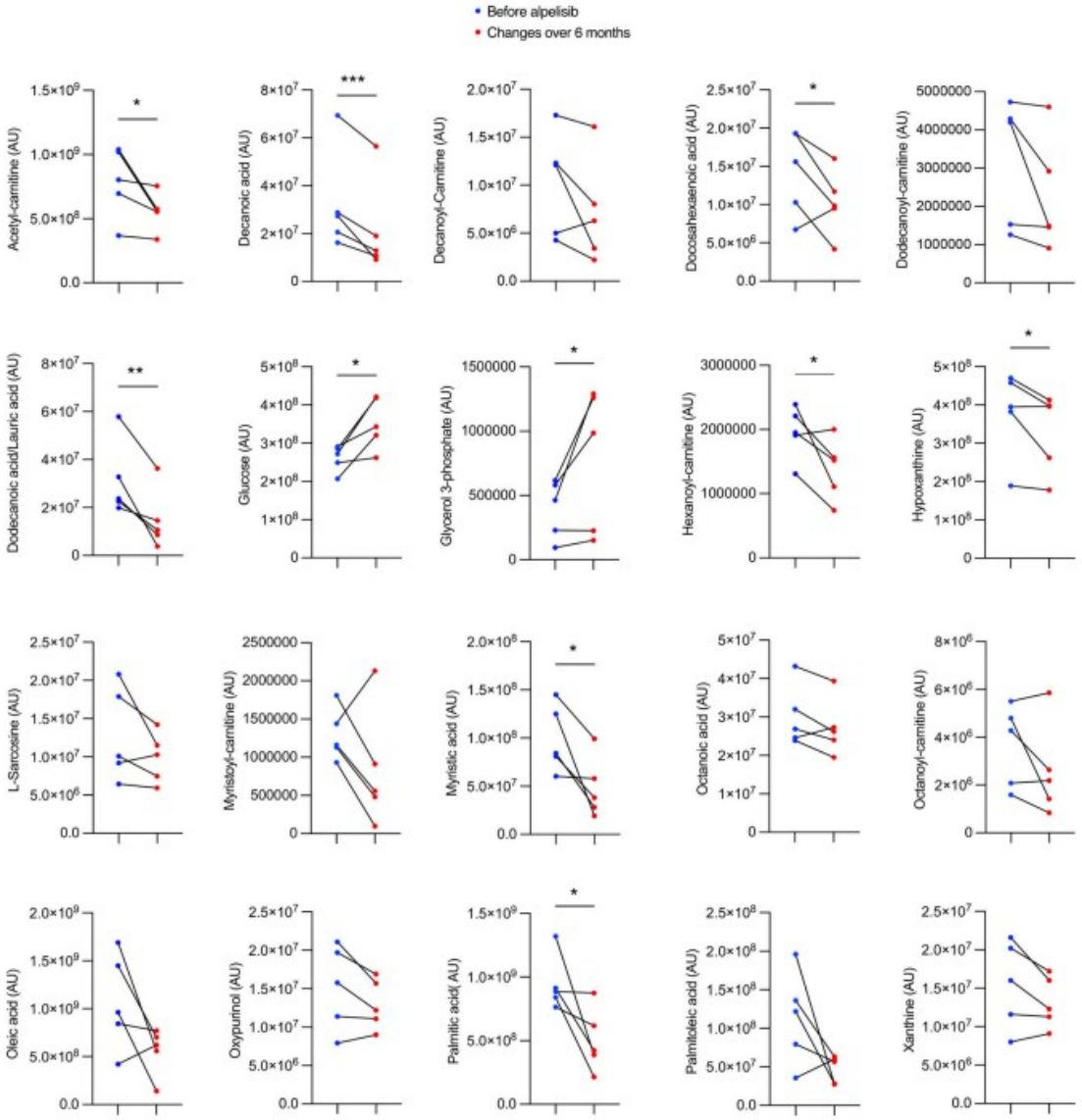


Figure 5. **Metabolic changes observed in patients before and 6 mo after alpelisib introduction.** Graphic example of metabolite level changes observed. AU: Arbitrary units. Data are shown as mean ± sem. *P < 0.05, **P < 0.01, ***P < 0.001 (two-tailed unpaired *t* test).

and EDTA at 0.5 m/mol. Cells were then transferred in a microtube. Samples were run on an ImageStream ISX mkII (Amnis part of Luminex) that combines flow cytometry with detailed cell imaging. Magnification (40×) was used for all acquisitions. Data were acquired with INSPIRE software (Amnis part of Luminex) and analyzed with IDEAS software (v.6.2; Amnis part of Luminex). The antibodies used for flow cytometry are availble in Table S1.

Targeted liquid chromatography–mass spectrometry (LC–MS) metabolites analyses

Plasma and serum were obtained after centrifugation of the blood at 500 *g* for 10 min. Blood samples were obtained in EDTA tubes for plasma analysis and EDTA-free tubes for serum analysis. Cells, plasma, and serum samples were immediately snap-frozen in liquid nitrogen. For the LC–MS analyses, metabolites were extracted as previously described (Ladraa et al.,

2022). Briefly, the extraction solution was composed of 50% methanol, 30% acetonitrile, and 20% water. The volume of the added extraction solution was adjusted to the cell number (1 ml per 1 million cells) or plasma and serum volume (200 µl per 10 µl of plasma or serum). After the addition of extraction solution, samples were vortexed for 5 min at 4°C and then centrifuged at 16,000 *g* for 15 min at 4°C. The supernatants were collected and stored at −80°C until the analyses were performed. LC–MS analyses were conducted using a QExactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific) equipped with an Ion Max source and a HESI II probe coupled to a Dionex UltiMate 3,000 UPLC system (Thermo Fisher Scientific). External mass calibration was performed using the standard calibration mixture every 7 d as recommended by the manufacturer. 5 µl of each sample was injected onto Zic-pHilic (150 mm × 2.1 mm, 5 µm) with the guard column (20 mm × 2.1 mm, 5 µm; Millipore) for liquid chromatography separation. Buffer A was 20 mM ammonium carbonate and 0.1% ammonium hydroxide (pH 9.2); buffer B was acetonitrile. The chromatographic gradient was run at a flow rate of 0.200 µl/min as follows: for 0–20 min, linear gradient from 80 to 20% B; for 20–20.5 min, linear gradient from 20% to 80% B; for 20.5–28 min, hold at 80% B. The mass spectrometer was operated in full-scan polarity switching mode with the spray voltage set to 2.5 kV and the heated capillary held at 320°C. The sheath gas flow was set to 20 U, the auxiliary gas flow was set to 5 U, and the sweep gas flow was set to 0 U. The metabolites were detected across a mass range of 75–1,000 m/z at a resolution of 35,000 (at 200 m/z) with the automated gain control target set to 106 and a maximum injection time of 250 ms. Lock masses were used to ensure mass accuracy below 5 parts per million. Data were acquired with Thermo Xcalibur 4.0.27.131 software (Thermo Fisher Scientific). The peak areas of metabolites were determined using Thermo TraceFinder 3.3 SP1 software (Thermo Fisher Scientific) and identified by the exact mass of each singly charged ion and by known retention time in the high performance liquid chromatography column.

Patients

HFMH was defined as congenital facial asymmetry with thickening of soft tissues leading to a set of specific clinical features: narrow palpebral fissure, eyebrow ptosis, nose deviation with small alar rim, external ear asymmetry and displacement with prominent concha, chin deviation with skin dimpling, and lip commissure canting. Previous treatment attempts were recorded.

Following procedures previously described (Delestre et al., 2021), the study was conducted on five pediatric patients, including three females who were followed at Hôpital Necker Enfants Malades. This protocol was approved by the Agence Nationale de Sécurité du Médicament et des Produits de Santé and the local ethical committee (Comité d'Éthique de Necker—Enfants Malades). Written informed consent was obtained from adult patients and the parents of pediatric patients. Alpelisib was compassionately offered by Novartis. Patients received 50 mg/d (Venot et al., 2018). Alpelisib was taken orally every morning during breakfast. Patients were assessed at regular intervals as previously reported (Venot et al., 2018). At each visit, 2D photographs were performed for all patients. 3D photographs were performed on three patients (EVA 3D Scanner; Artec 3D). For 3D facial reconstruction, we applied non-rigid and dense registrations between a template and patient faces (Amberg, B. 2007). Each 3D image was rigidly aligned with the template, which was a standardized, controlled, symmetrical mesh of an average human head, using Wrap v2020.12.2 (R3DS) by placing landmarks to control relative positioning and by computing translations, rotations, and scaling (Fig. S1). The template was then warped around the 3D face by computing a non-rigid iterative closest point alignment involving 50 iterations. The deformed template mesh had the target mesh shape and preserved its original node topology. The 3D coordinates of all nodes were extracted and processed in R (R Core Team, 2018), where the node coordinates were aligned based on Procrustes superimposition without scaling using ProcSym from Morpho library (Schlager, 2017). The superimposition included only the non-affected side of the face to emphasize the variations located on the affected side. PLS2B were computed to screen for covariation patterns described by Procrustes coordinates and time since treatment introduction using PLS2B from the Morpho library (Schlager, 2017). The theoretical morphological changes associated with each statistically significant covariation axis were visualized using tps3d from the Morpho (Schlager, 2017) and shade3d from the rgl library (Murdoch and Adler, 2023). Non-affected sides were mirrored based on sagittal planes using Wrap and the Procrustes distances between the affected and the mirrored non-affected sides were computed, leading to the generation of facial heatmaps (Fig. S1) using meshDist from the Morpho library (Schlager, 2017). 2D photographs were analyzed as previously described (Hennocq et al., 2023). After manual landmarking, clouds of landmarks were extracted from the frontal views (n = 105), profile views (n = 73), and external ears (n = 41) and compared after Procrustes superimposition and principal component analysis. The pretreatment facial phenotype was compared to the facial phenotype after a minimum of 6 mo of treatment. Analyses were performed in R (R Core Team, 2018) using the geomorph package.

In addition, we assessed the volume of skeletal muscle overgrowth using MRI for each patient. MRI examination was performed using T1, T2, and fat suppression, and T2-weighted imaging sequences were performed before alpelisib (day 0) introduction and again 6 mo after. Volumetric evaluation of skeletal muscle malformation was determined by thresholding and manually delineating hypersignal T2 lesions. Volume was calculated by summing images based on the 2D contours and slice thickness.

Biopsies of the buccinator and/or masseter muscles of the affected and control sides were performed in 5/5 HFMH patients, for diagnosis, and four out of five patients after 6 mo of treatment.

Control buccinator and masseter muscles were harvested on pediatric patients admitted for facial lacerations and/or mandibular trauma, that required surgical approaches to this anatomical region.

Data analysis and statistics

Data are expressed as the means ± SEM. Differences between the experimental groups were evaluated using ANOVA, followed by the Tukey–Kramer post hoc test when the results were significant ($P < 0.05$). When only two groups were compared, Mann–Whitney tests were used. Statistical analyses were performed using GraphPad Prism software (version 10.0.0).

Online supplemental material

Fig. S1 shows the workflow for the assessment of 3D pictures. Fig. S2 shows the mouse model characterization and quantitative histological analysis of tibialis anterior muscle changes in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or alpelisib. Fig. S3 shows metabolic changes observed in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$. Fig. S4 shows metabolic changes observed in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ treated with either vehicle or alpelisib. Fig. S5 shows the covariation between treatment duration and facial asymmetry using two-blocks partial least-squares regressions. Table S1 lists the antibodies, materials, and buffers used in the study.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the supplementary materials.

Acknowledgments

We are very grateful to our generous donors.

This study was supported by the European Research Council (CoG 2020 grant number 101000948 awarded to G. Canaud), the Agence Nationale de la Recherche—Programme d'Investissements d'Avenir (ANR-18-RHUS-005 to G. Canaud), and the Agence Nationale de la Recherche—Programme de Recherche Collaborative (19-CE14-0030-01 to G. Canaud). This work was also supported by the CLOVES Syndrome Community (West Kennebunk, USA), Association Syndrome de CLOVES (Nantes, France), Fondation d'entreprise IRCEM (Roubaix, France), Fonds de dotation Emmanuel Boussard (Paris, France), the Fondation Day Solvay (Paris, France), the Fondation TOURRE (Paris, France) to G. Canaud, the Fondation Bettencourt Schueller (Paris, France) to G. Canaud, the Fondation Simone et Cino Del Duca (Paris, France), the Fondation Line Renaud-Loulou Gaste (Paris, France), the Fondation Schlumberger pour l'Education et la Recherche (Paris, France), the Fondation Maladies Rares, the Association Robert Debré pour la Recherche Médicale awarded to G. Canaud, WonderFIL Smiles (Norway), Institut national de la santé et de la recherche médicale, Assistance Publique Hôpitaux de Paris and Université Paris Cité. In vivo preclinical imaging was performed at the Life Imaging Facility of the University of Paris (Plateforme Imageries du Vivant), supported by France Life Imaging (grant ANR-11-INBS-0006) and Infrastructures BiologieSanté.

Author contributions: C. Bayard, B. Periou, L. Zerbib, and S. Ladraa performed the experiments, analyzed the data, and elaborated figures. E. Segna, C. Broissand, C. Legendre, C. Gitiaux, B. Sergent, and A. Picard followed patients and participated in data analysis. M. Taverne and Q. Hennocq analyzed 2D and 3D photographs. A. Fraissenon, L. Guibaud, G. Autret, and B. Tavitian designed and analyzed MRI experiments. C. Chapelle and C. Hoguin were in charge of mice experiments including genotyping, breeding, tamoxifen administration, and sacrifice. V. Asnafi, S. Kaltenbach, and P. Villarese performed molecular diagnosis in patients. F.-J. Authier, S. Fraitag, and J.-P. Duong analyzed mice and human tissues section. I. Nemazanyy performed the metabolomic analysis. C. Delcros helped with some of the in vivo experiments. T. Viel performed the FDG uptake experiments. N. Goudin analyzed and quantified immunofluorescence expression in tissues. M. Dussiot provided help for flow cytometry experiments. R. Khonsari and G. Canaud followed the patients, provided the conceptual framework, designed the study, supervised the project, and wrote the paper.

Disclosures: B. Periou reported benefitting from fellowship from the Agence Nationale pour la Recherche (RHU CARMMA, ANR-15-RHUS-0003). F.-J. Authier reported benefitting from research grants from the Association Française contre les Myopathies via TRANSLAMUSCLE (PROJECT 19507 and 22946) and the Agence Nationale pour la Recherche (RHU CARMMA, ANR-15-RHUS-0003). G. Canaud reported personal fees from Novartis, Vaderis, Ipsen, BridgeBio, and Alkermes outside the submitted work; in addition, G. Canaud had a patent to WO2017140828A1 licensed to Novartis. No other disclosures were reported.

Submitted: 29 May 2023
Revised: 19 July 2023
Accepted: 9 August 2023

References

Amberg, B. 2007. Conference on Computer Vision and Pattern Recognition. IEEE, Minneapolis, MN, USA. 1–8.

Bankhead, P., M.B. Loughrey, J.A. Fernández, Y. Dombrowski, D.G. McArt, P.D. Dunne, S. McQuaid, R.T. Gray, L.J. Murray, H.G. Coleman, et al. 2017. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* 7:16878. https://doi.org/10.1038/s41598-017-17204-5

Berg, S., D. Kutra, T. Kroeger, C.N. Straehle, B.X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, et al. 2019. Ilastik: Interactive machine learning for (bio)image analysis. *Nat. Methods.* 16:1226–1232. https://doi.org/10.1038/s41592-019-0582-9

Canaud, G., A.M. Hammill, D. Adams, M. Vikkula, and K.M. Keppler-Noreuil. 2021. A review of mechanisms of disease across PIK3CA-related disorders with vascular manifestations. *Orphanet J. Rare Dis.* 16:306. https://doi.org/10.1186/s13023-021-01929-8

Castillo Taucher, S., R.A. Pardo Vargas, and D. Caivi León. 2003. Hemifacial myohyperplasia: An additional case. *Am. J. Med. Genet. A.* 116A:103–104. https://doi.org/10.1002/ajmg.a.10839

Delestre, F., Q. Venot, C. Bayard, A. Fraissenon, S. Ladraa, C. Hoguin, C. Chapelle, J. Yamaguchi, R. Cassaca, L. Zerbib, et al. 2021. Alpelisib administration reduced lymphatic malformations in a mouse model and in patients. *Sci. Transl. Med.* 13:eabg0809. https://doi.org/10.1126/scitranslmed.abg0809

Fedorov, A., R. Beichel, J. Kalpathy-Cramer, J. Finet, J.C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, et al. 2012. 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging.* 30:1323–1341. https://doi.org/10.1016/j.mri.2012.05.001

Hennocq, Q., T. Bongibault, M. Bizière, O. Delassus, M. Douillet, V. Cormier-Daire, J. Amiel, S. Lyonnet, S. Marlin, M. Rio, et al. 2023. An automatic facial landmarking for children with rare diseases. *Am. J. Med. Genet. A.* 191:1210–1221. https://doi.org/10.1002/ajmg.a.63126

Klippel, A., C. Reinhard, W.M. Kavanaugh, G. Apell, M.A. Escobedo, and L.T. Williams. 1996. Membrane localization of phosphatidylinositol 3-kinase

is sufficient to activate multiple signal-transducing kinase pathways. *Mol. Cell. Biol.* 16:4117–4127. https://doi.org/10.1128/MCB.16.8.4117

Ladraa, S., L. Zerbib, C. Bayard, A. Fraissenon, Q. Venot, G. Morin, A.P. Garneau, P. Isnard, C. Chapelle, C. Hoguin, et al. 2022. PIK3CA gain-of-function mutation in adipose tissue induces metabolic reprogramming with Warburg-like effect and severe endocrine disruption. *Sci. Adv.* 8: eade7823. https://doi.org/10.1126/sciadv.ade7823

Lee, S., R. Sze, C. Murakami, J. Gruss, and M. Cunningham. 2001. Hemifacial myohyperplasia: Description of a new syndrome. *Am. J. Med. Genet.* 103: 326–333. https://doi.org/10.1002/1096-8628(20011101)103:4<326::AID-AJMG1578>3.0.CO;2-Z

Li, M.E., H.P.M.M. Lauritzen, B.T. O'Neill, C.-H. Wang, W. Cai, B.B. Brandao, M. Sakaguchi, R. Tao, M.F. Hirshman, S. Softic, et al. 2019. Role of p110a subunit of PI3-kinase in skeletal muscle mitochondrial homeostasis and metabolism. *Nat. Commun.* 10:3412. https://doi.org/10.1038/s41467-019-11265-y

Madsen, R.R. and B. Vanhaesebroeck. 2020. Cracking the context-specific PI3K signaling code. *Sci Signal*, 13:eaay2940. https://doi.org/10.1126/scisignal.aay2940

Miranda, R.T., L.M. Barros, L.A. Santos, P.R. Bonan, and H. Martelli Jr. 2010. Clinical and imaging features in a patient with hemifacial hyperplasia. *J. Oral Sci.* 52:509–512. https://doi.org/10.2334/josnusd.52.509

Morin, G., C. Degrugillier-Chopinet, M. Vincent, A. Fraissenon, H. Aubert, C. Chapelle, C. Hoguin, F. Dubos, B. Catteau, F. Petit, et al. 2022. Treatment of two infants with PIK3CA-related overgrowth spectrum by alpelisib. *J. Exp. Med.* 219:e20212148. https://doi.org/10.1084/jem.20212148

Murdoch, D., and D. Adler. 2023. Rgl: 3D visualization using openGL. https://birs.ok.ubc.ca/wjbraun/DS550/lecture7.html

Muzumdar, M.D., B. Tasic, K. Miyamichi, L. Li, and L. Luo. 2007. A global double-fluorescent Cre reporter mouse. *Genesis.* 45:593–605. https://doi.org/10.1002/dvg.20335

Pereira-Perdomo, D.F., J. Vélez-Forero, and R. Prada-Madrid. 2010. Hemifacial myohyperplasia sequence. *Am. J. Med. Genet. A.* 152A:1770–1773. https://doi.org/10.1002/ajmg.a.33428

R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al. 2012. Fiji: An open-source platform for biological-image analysis. *Nat. Methods.* 9: 676–682. https://doi.org/10.1038/nmeth.2019

Siponen, M., G.K.B. Sándor, L. Ylikontiola, T. Salo, and H. Tuominen. 2007. Multiple orofacial intraneural perineuriomas in a patient with hemifacial hyperplasia. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* 104:e38–e44. https://doi.org/10.1016/j.tripleo.2006.12.030

Schlager, S. 2017. Morpho and rvcg – shape analysis in R. In Statistical Shape and Deformation Analysis. G. Zheng, S. Li, and G. Szekely, editors. Academic Press, Cambridge, MA. 217–256. https://doi.org/10.1016/B978-0-12-810493-4.00011-0

Srinivasan, L., Y. Sasaki, D.P. Calado, B. Zhang, J.H. Paik, R.A. DePinho, J.L. Kutok, J.F. Kearney, K.L. Otipoby, and K. Rajewsky. 2009. PI3 kinase signals BCR-dependent mature B cell survival. *Cell.* 139:573–586. https://doi.org/10.1016/j.cell.2009.08.041

Stringer, C., T. Wang, M. Michaelos, and M. Pachitariu. 2021. Cellpose: A generalist algorithm for cellular segmentation. *Nat. Methods.* 18: 100–106. https://doi.org/10.1038/s41592-020-01018-x

Venot, Q., T. Blanc, S.H. Rabia, L. Berteloot, S. Ladraa, J.P. Duong, E. Blanc, S.C. Johnson, C. Hoguin, O. Boccara, et al. 2018. Targeted therapy in patients with PIK3CA-related overgrowth syndrome. *Nature.* 558: 540–546. https://doi.org/10.1038/s41586-018-0217-9

Zissman, S., Y. Cooperman, D. Leshem, and E. Gur. 2020. Progressive surgical management of hemifacial myohyperplasia for improved functional and aesthetic results. *Plast. Reconstr. Surg. Glob. Open.* 8:e2724. https://doi.org/10.1097/GOX.0000000000002724

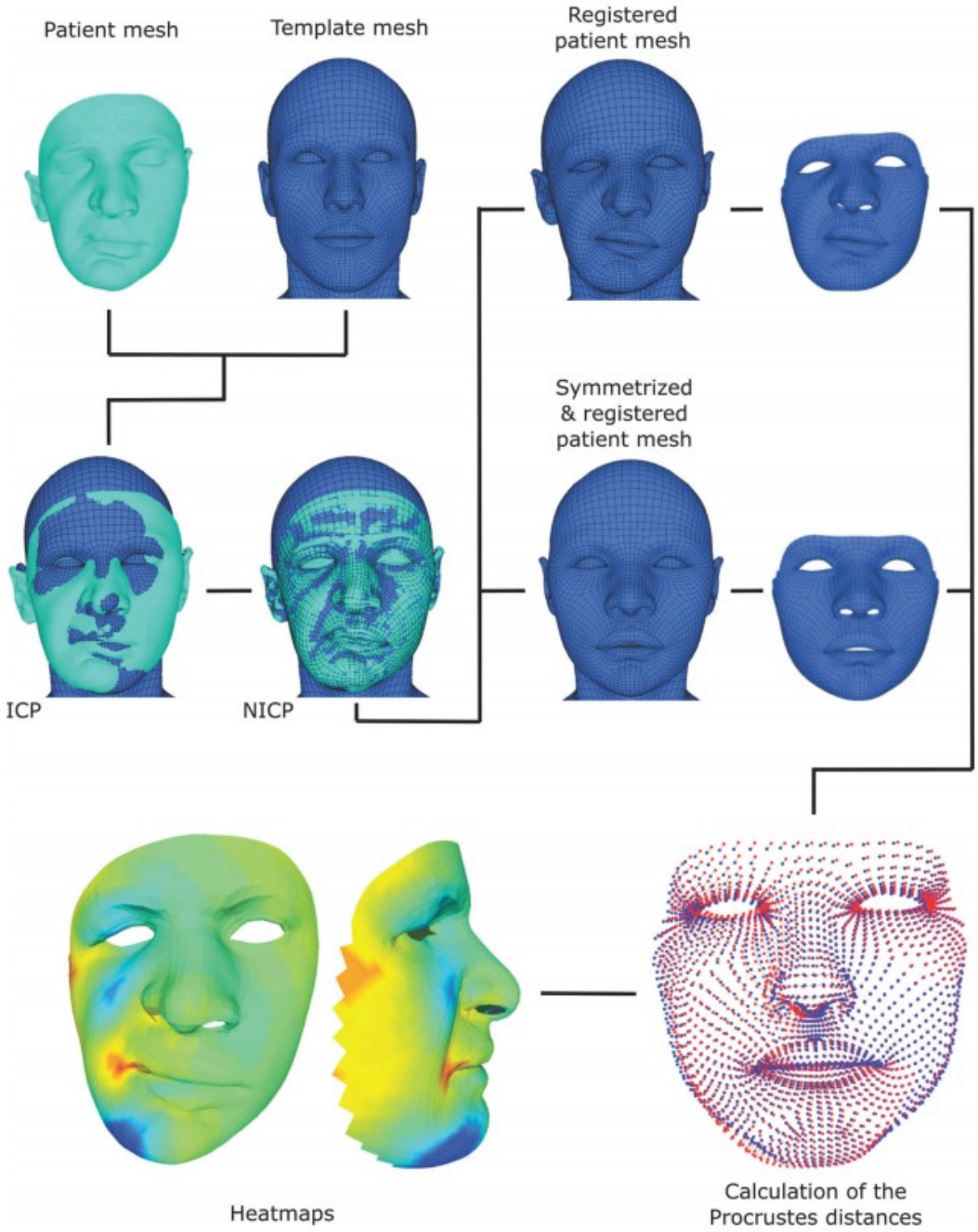


Figure S1. **Workflow for the assessment of 3D pictures.** First, the template (blue mesh) was rigidly matched with the "target" head surface of each patient (cyan mesh) by an ICP (iterative closest point) alignment. The template was then non-rigidly matched with the target surface by means of an NICP (non-rigid iterative closest point) alignment, which consists of the local deformation of the template following thin-plate splines to adopt the conformation of the target. This resulted in a registered template. The process was repeated for each subject at each age. Then, the facial region was extracted and symmetrized against the non-affected facial side. A Procrustes superimposition was performed to align all faces to tackle undesirable effects of position and orientation. Finally, all subsequent analyses including the generation of heatmaps were computed on the Procrustes coordinates exclusively.

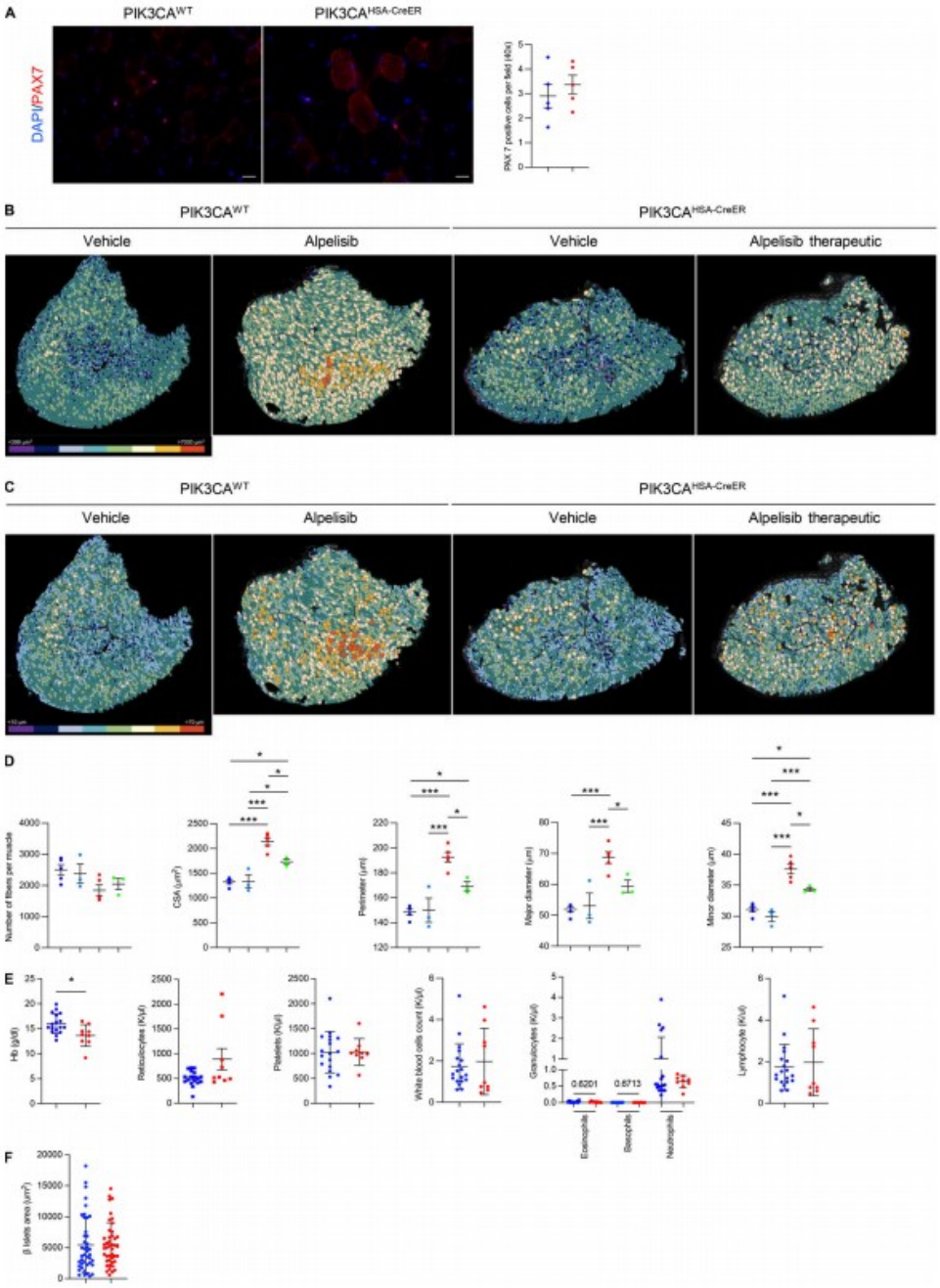


Figure S2. **Mouse model characterization and quantitative histological analysis of tibialis anterior (TA) muscle changes in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or alpelisib. (A)** Representative immunofluorescence of PAX7 satellite cells and quantification. **(B and C)** Digital pictures of TA muscle cross-section area allowing an automated image quantification of (B) major and minor (C) myofiber diameters in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or alpelisib. **(D)** Quantification. **(E)** Complete blood count in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. **(F)** β Islet area in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice. Data are shown as mean ± SEM. *P < 0.05 and ***P < 0.001 (two-tailed unpaired *t* test for A, E, and F, and ANOVA, followed by the Tukey–Kramer post hoc test for D). Each dot represents one mouse. Data are representative of at least two independent experiments.

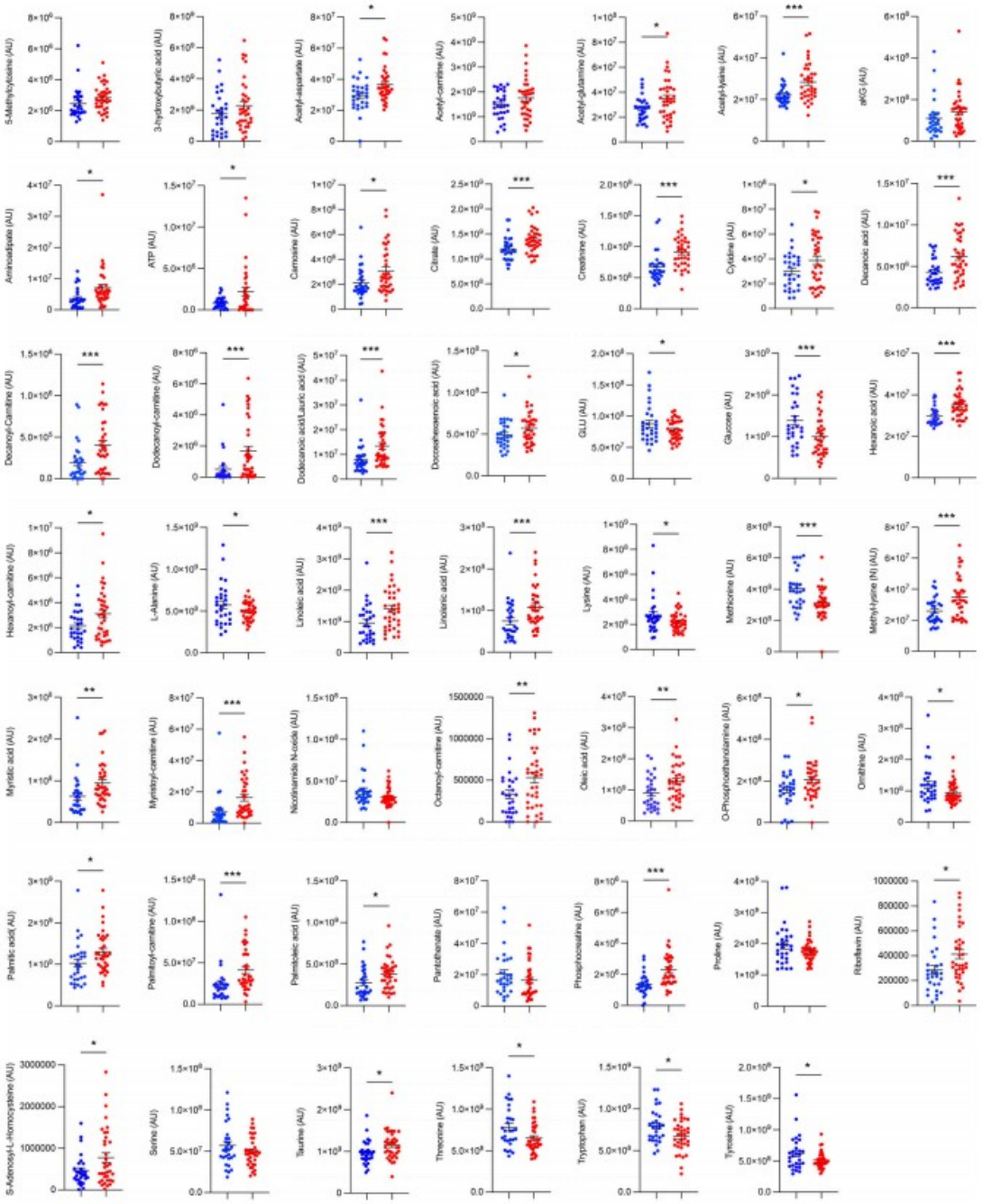


Figure S3. **Metabolic changes observed in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$.** Graphic example of metabolite level changes observed. AU: Arbitrary units. Data are shown as mean ± SEM. *P < 0.05, **P < 0.01, ***P < 0.001 (two-tailed unpaired *t* test).

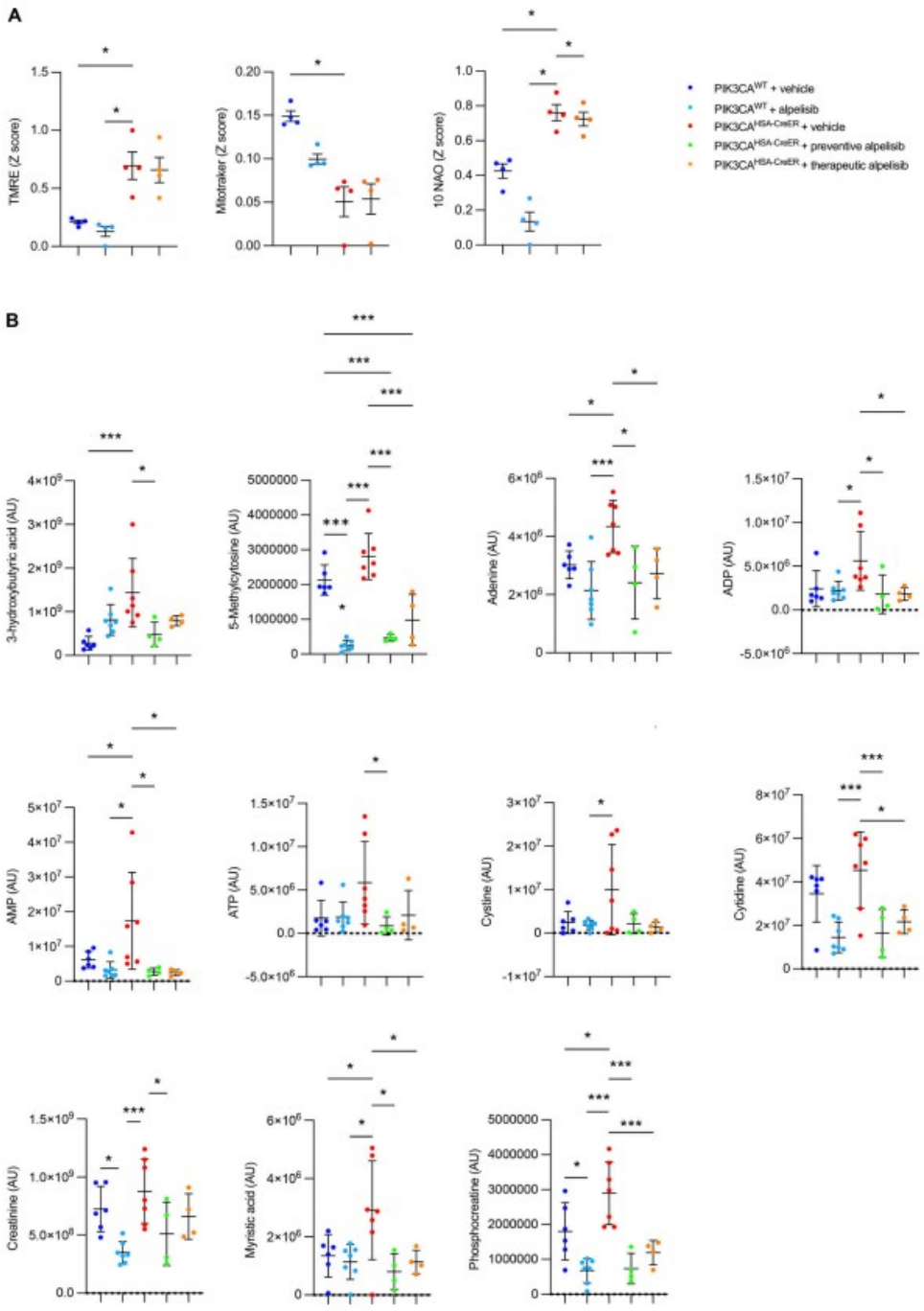


Figure S4. **Metabolic changes observed in *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ treated with either vehicle or alpelisib. (A)** TMRE, Mitotraker, and 10 NAO staining in skeletal muscle of *PIK3CA*^{WT} and *PIK3CA*$^{HSA\text{-}CreER}$ mice treated with either vehicle or alpelisib (n = 4 per group). **(B)** Graphic example of metabolite level changes observed (n = 4–7 per group). AU: Arbitrary units. Data are shown as mean ± sem. *P < 0.05 and ***P < 0.001 (ANOVA, followed by the Tukey–Kramer post hoc test).

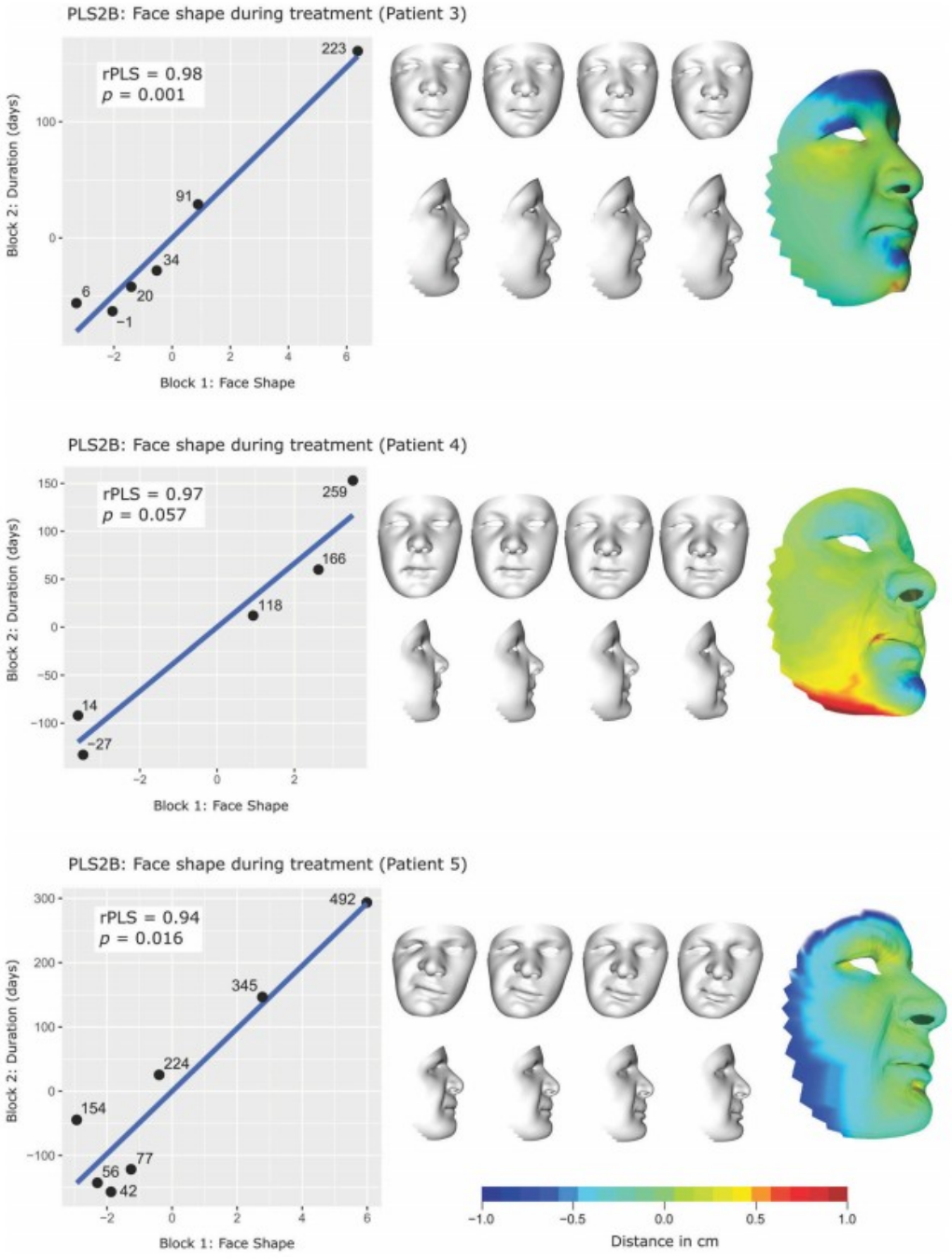


Figure S5. **Covariation between treatment duration and facial asymmetry using PLS2B.** PLS2B assessing the covariation between facial morphology (Block 1) and treatment duration (Block 2) in 3/5 patients. rPLS: coefficient of covariation; p: P value; each point represents a 3D face mesh annotated with the duration, in days, since treatment onset. White faces represent the theoretical face deformation along with the covariation axis (from the left to the right: from the start of the treatment, until the most recent 3D picture). Rightmost is a heatmap representation of the main facial regions corrected by the treatment. Colder colors pinpoint expanded regions, whereas hotter colors pinpoint shrunk regions.