medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

# GestaltMatcher Database - A global reference for facial phenotypic variability in rare human diseases

Hellen Lesmann[1,2,*], Alexander Hustinx[2,*], Shahida Moosa[3], Hannah Klinkhammer[2,4], Elaine Marchi[5], Pilar Caro[6], Ibrahim M. Abdelrazek[7], Jean Tori Pantel[8,9], Merle ten Hagen[2], Meow-Keong Thong[10], Rifhan Azwani Binti Mazlan[10], Sok Kun Tae[10], Tom Kamphans[11], Wolfgang Meiswinkel[11], Jing-Mei Li[2], Behnam Javanmardi[2], Alexej Knaus[2], Annette Uwineza[12], Cordula Knopp[13], Tinatin Tkemaladze[14,15], Miriam Elbracht[13], Larissa Mattern[13], Rami Abou Jamra[16], Clara Velmans[17], Vincent Strehlow[16], Maureen Jacob[18], Angela Peron[19,20], Cristina Dias[21,22,23,24], Beatriz Carvalho Nunes[25], Thainá Vilella[25], Isabel Furquim Pinheiro[26], Chong Ae Kim[26], Maria Isabel Melaragno[25], Hannah Weiland[2], Sophia Kaptain[2], Karolina Chwiałkowska[27,28], Miroslaw Kwasniewski[28,27], Ramy Saad[22,29], Sarah Wiethoff[30], Himanshu Goel[31], Clara Tang[32], Anna Hau[33], Tahsin Stefan Barakat[34], Przemysław Panek[35], Amira Nabil[7], Julia Suh[13], Frederik Braun[36], Israel Gomy[37], Luisa Averdunk[38], Ekanem Ekure[39], Gaber Bergant[40], Borut Peterlin[41], Claudio Graziano[42], Nagwa Gaboon[43,44], Moisés Fiesco-Roa[45,46], Alessandro Mauro Spinelli[47], Nina-Maria Wilpert[48,49,50], Prasit Phowthongkum[51,52], Nergis Güzel[13], Tobias B. Haack[53], Rana Bitar[54,55], Andreas Tzschach[56], Agusti Rodriguez-Palmero[57], Theresa Brunet[18], Sabine Rudnik-Schöneborn[58], Silvina Noemi Contreras-Capetillo[59], Ava Oberlack[18], Carole Samango-Sprouse[60,61,62], Teresa Sadeghin[63], Margaret Olaya[63], Konrad Platzer[16], Artem Borovikov[64], Franziska Schnabel[16], Lara Heuft[16], Vera Herrmann[16], Renske Oegema[65], Nour Elkhateeb[66], Sheetal Kumar[1], Katalin Komlosi[56], Khoushoua Mohamed[7], Silvia Kalantari[67], Fabio Sirchia[67,68], Antonio F. Martinez-Monseny[69], Matthias Höller[56], Louiza Toutouna[56], Amal Mohamed[7], Amaia Lasa-Aranzasti[70,71], John A. Sayer[72,73], Nadja Ehmke[74], Magdalena Danyel[74], Henrike Sczakiel[74], Sarina Schwartzmann[74], Felix Boschann[74], Max Zhao[74], Ronja Adam[74], Lara Einicke[74], Denise Horn[74], Kee Seang Chew[75], Choy Chen KAM[75], Miray Karakoyun[76], Ben Pode-Shakked[77,78], Aviva Eliyahu[79,80,81], Rachel Rock[82,83], Teresa Carrion[84], Odelia Chorin[85], Yuri A. Zarate[86,87], Marcelo Martinez Conti[88], Mert Karakaya[17], Moon Ley Tung[89,90], Bharatendu Chandra[89,90], Arjan Bouman[34], Aime Lumaka[91], Naveed Wasif[92,93], Marwan Shinawi[94], Patrick R. Blackburn[95], Tianyun Wang[96,97,98], Tim Niehues[99], Axel Schmidt[1], Regina Rita Roth[100], Dagmar Wieczorek[100], Ping Hu[101], Rebekah L. Waikel[101], Suzanna E. Ledgister Hanchard[101], Gehad Elmakkawy[7], Sylvia Safwat[7], Frédéric Ebstein[102,103], Elke Krüger[104], Sébastien Küry[102,103], Stéphane Bézieau[102,103], Annabelle Arlt[2], Eric Olinger[105], Felix Marbach[6], Dong Li[106], Lucie Dupuis[107], Roberto Mendoza-Londono[107], Sofia Douzgou Houge[108], Denisa Weis[109], Brian Hon-Yin Chung[110,111], Christopher C.Y. Mak[111], Hülya Kayserili[112], Nursel Elcioglu[113], Ayca Aykut[114], Peli Özlem Şimşek-Kiper[115], Nina Bögershausen[116], Bernd Wollnik[116,117,118], Heidi Beate Bentzen[119,120], Ingo Kurth[13], Christian Netzer[17], Aleksandra Jezela-Stanek[35], Koen Devriendt[121], Karen W. Gripp[122], Martin Mücke[8,9], Alain Verloes[123], Christian P. Schaaf[6],

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Christoffer Nellåker[124], Benjamin D. Solomon[101], Markus M. Nöthen[1], Ebtesam Abdalla[7], Gholson J. Lyon[125,126,127], Peter M. Krawitz[2], Tzung-Chien Hsieh[2,#]

[1]Institute of Human Genetics, University of Bonn, Bonn, NRW, Germany, [2]Institute for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, NRW, Germany, [3]Division of Molecular Biology and Human Genetics, Stellenbosch University and Medical Genetics, Tygerberg Hospital, Stellenbosch, South Africa, [4]Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, NRW, Germany, [5]New York State Institute for Basic Research in Developmental Disabilities, New York State, Albany, New York, USA, [6]Institute of Human Genetics, Heidelberg University, Heidelberg, Baden-Württemberg, Germany, [7]Department of Human Genetics, Medical Research Institute, Alexandria University, Alexandria, Alexandria, Egypt, [8]Institute for Digitalization and General Medicine, University Hospital RWTH Aachen, Aachen, NRW, Germany, [9]Centre for Rare Diseases Aachen (ZSEA), University Hospital RWTH Aachen, Aachen, NRW, Germany, [10]Department of Paediatrics, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia, [11]GeneTalk GmbH, Bonn, NRW, Germany, [12]College of Medicine and Health Sciences, University of Rwanda, and University Teaching Hospital of Kigali, Kigali, Rwanda, [13]Institute for Human Genetics and Genomic Medicine, Medical Faculty, RWTH Aachen University, Aachen, NRW, Germany, [14]Department of Molecular and Medical Genetics, Tbilisi State Medical University, Tbilisi, Georgia, [15]Givi Zhvania Pediatric Academic Clinic, Tbilisi State Medical University, Georgia, [16]Institute of Human Genetics, University of Leipzig Medical Center, Leipzig, Germany, [17]Institute of Human Genetics, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, NRW, Germany, [18]Institute of Human Genetics, Klinikum rechts der Isar, Technical University of Munich, School of Medicine and Health, Munich, Germany, [19]Medical Genetics, Meyer Children's Hospital IRCCS, Firenze, Italy, [20]Department of Experimental and Clinical Biomedical Sciences "Mario Serio", Università degli Studi di Firenze, Italy, [21]Department of Medical Genetics, Guy's and St. Thomas' NHS Foundation Trust, London, UK, [22]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children, Great Ormond Street, London, UK, [23]Neural Stem Cell Biology Laboratory, The Francis Crick Institute, UK, [24]Department of Medical & Molecular Genetics, School of Basic and Medical Biosciences, Faculty of Life Sciences & Medicine, King's College London, UK, [25]Genetics Division, Department of Morphology and Genetics, Universidade Federal de São Paulo, São Paulo, Brazil, [26]Genetics Unit, Instituto da Criança, Universidade de São Paulo, São Paulo, Brazil, [27]Centre for Bioinformatics and Data Analysis, Medical University of Bialystok, Bialystok, Poland, [28]IMAGENE.ME SA, Bialystok, Poland, [29]Department of Genetics and Genomic Medicine, UCL Institute of Child Health, London UK, [30]Department of Neurology with Institute of Translational Neurology, University Hospital Münster, Münster, NRW, Germany, [31]School of Medicine and Public Health, University of Newcastle, Callaghan NSW, Australia, [32]Kabuki Syndrome Foundation,

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Northbrook, IL, USA, [33]Hunter Genetics, Hunter New England Health Service, Newcastle, Australia, [34]Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, The Netherlands, [35]Department of Genetics and Clinical Immunology, National Institute of Tuberculosis and Lung Diseases, Warsaw, Poland, [36]Institute of Human Genetics, University Hospital Essen, University Duisburg-Essen, Essen, NRW, Germany, [37]Department of Genetics, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Sao Paulo, Brazil, [38]Department of General Pediatrics and Neonatology, University Children's Hospital, Heinrich-Heine-University, Medical Faculty, Düsseldorf, Germany, [39]Department of Paediatrics, College of Medicine, University of Lagos, Lagos, Nigeria, [40]Clinical Institute of Genomic Medicine, University Medical Centre Ljubljana, Ljubljana, Slovenia, [41]Clinical Institute of Genomic Medicine, University Medical Centre Ljubljana, [42]Medical Genetics Unit, Ausl Romagna, Cesena, Italy, [43]Medical Genetics Center, Faculty of Medicine, Ain Shams University, Cairo, Egypt, [44]Medical Genetics Department, Armed Forces College of Medicine, Cairo, Egypt, [45]Programa de Maestría y Doctorado en Ciencias Médicas, Odontológicas y de la Salud, Universidad Nacional Autónoma de México, México City, Mexico, [46]Laboratorio de Citogenética, Instituto Nacional de Pediatría, México City, Mexico, [47]Institute for Maternal and Child Health, IRCCS Burlo Garofolo, Trieste, Italy, [48]NeuroCure Cluster of Excellence; Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, D-10117 Berlin, Germany, [49]Department of Neuropediatrics, Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, D-13353 Berlin, Germany, [50]Berlin Institute of Health at Charité - Universitätsmedizin Berlin, BIH Biomedical Innovation Academy, BIH Charité Junior Clinician Scientist Program, D-10117 Berlin, German, [51]Excellence Center for Genomics and Precision Medicine, King Chulalongkorn Memorial Hospital, the Thai Red Cross Society, Bangkok, Thailand, [52]Division of Medical Genetics and Genomics, Department of Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand, [53]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany, [54]Pediatric Gastroenterology Department, Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates, [55]Khalifa University, Abu Dhabi, United Arab Emirates, [56]Institute of Human Genetics, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany, [57]Paediatric Neurology Unit, Department of Pediatrics, Hospital Universitari Germans Trias i Pujol, Universitat Autònoma de Barcelona, Barcelona, Spain, [58]Institute of Human Genetics, Medical University Innsbruck, Innsbruck, Austria, [59]Universidad Autonoma de Yucatan (University Autonomus of Yucatan), Merida, Yucatan, Mexico, [60]Department of Pediatrics, George Washington University, 2121 I St. NW, Washington D.C. 2005, [61]Department of Human and Molecular Genetics, Florida International University, 11200 SW 8th Street, AHC2 Miami, Florida 22199, [62]Department of Research, The Focus Foundation, 820 W. Central Ave. #190, Davidsonville, MD 21035, [63]Department of Research, The Focus Foundation, 2772 Rutland Road P.O. Box 190, Davidsonville, MD 21035, [64]Research Centre for

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Medical Genetics (RCMG), Moscow, Russia, [65]Department of Genetics, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands, [66]Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK, [67]Department of Molecular Medicine, University of Pavia, Pavia, Italy, [68]Medical Genetics Unit, IRCCS San Matteo Foundation, Pavia, Italy, [69]Department of Clinical Genetics, SJD Barcelona Children's Hospital, Esplugues del Llobregat (Barcelona), Spain, [70]Medicine Genetics Group, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Hospital Universitari, Barcelona, Catalunya, Spain, [71]Department of Clinical and Molecular Genetics, Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Hospital Universitari, Barcelona, Catalunya, Spain, [72]Biosciences Institute, Newcastle University, Central Parkway, Newcastle upon Tyne, UK, [73]Renal Services, The Newcastle Upon Tyne NHS Hospitals Foundation Trust, Freeman Road, Newcastle Upon Tyne, UK, [74]Institute of Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany, [75]Department of Paediatrics, Faculty of Medicine, University Malaya, 59100 Kuala Lumpur, Malaysia, [76]Ege University, Faculty of Medicine, Department of Pediatric Gastroenterology Hepatology and Nutrition, Izmir, Turkey, [77]The Institute of Rare Diseases, Edmond and Lily Safra Children’s Hospital, Sheba Medical Center, Ramat Gan, Israel, [78]The faculty of Medical and Health Sciences, Tel-Aviv University, Tel-Aviv, Israel, [79]The Danek Gertner Institute of Human Genetics, Sheba Medical Center, Tel-Hashomer, Israel, [80]Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel, [81]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel, [82]Metabolic Diseases Clinic, Edmond and Lily Safra Children's Hospital, Sheba Medical Center, [83]National Newborn Screening Program, Public Health Services, Ministry of Health Tel-Hashomer, Israel, [84]Rare diseases Unit, Pediatric Department, Hospital Universitari Son Espases, Palma de Mallorca, Spain, [85]The Institute of Rare Diseases, Edmond and Lily Safra Children’s Hospital, Sheba Medical Center, Tel-Hashomer, Israel, [86]Department of Pediatrics, Section of Genetics and Metabolism, University of Arkansas for Medical Sciences and Arkansas Children’s Hospital, Little Rock, AR, USA, [87]Division of Genetics and Metabolism, University of Kentucky, Lexington, KY, USA, [88]Project Director AI for Health at Foundation 29, Foundation 29, Madrid, Spain, [89]University of Iowa Roy J and Lucille A Carver College of Medicine, Iowa City, IA 52242, USA, [90]Division of Medical Genetics and Genomics, Stead Family Department of Pediatrics, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA, [91]Center for Human Genetics, Faculty of Medicine, University of Kinshasa, Kinshasa, DR Congo, [92]Institute of Human Genetics, University of Ulm, Ulm, Baden-Württemberg, Germany, [93]University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany, [94]Division of Genetics and Genomic Medicine, Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA, [95]Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA, [96]Department of Medical Genetics, Center for Medical Genetics, Peking University Health Science Center,

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Beijing 100191, China, [97]Neuroscience Research Institute, Peking University; Key Laboratory for Neuroscience, Ministry of Education of China & National Health Commission of China, Beijing 100191, China, [98]Autism Research Center, Peking University Health Science Center, Beijing 100191, China, [99]Department of Pediatrics, Helios Klinik Krefeld, Krefeld 47805, Germany, [100]Institute of Human Genetics, Medical Faculty, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Germany, [101]Medical Genomics Unit, Medical Genetics Branch, National Human Genome Research Institute, Bethesda, USA, [102]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France, [103]Nantes Université, CHU Nantes, Service de Génétique Médicale, F-44000 Nantes, France, [104]Insitute for Medical Biochemistry and Molecular Biology, University of Greifswald, Greifswald, Greifswald, Germany, [105]Center for Human Genetics, Cliniques Universitaires Saint-Luc, Brussels, Belgium, [106]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, [107]Department to Paediatrics, Division of Clinical and Metabolic Genetics, The Hospital of Sick Children, Toronto, Ontario, Canada, [108]Department of Medical Genetics, Haukeland University Hospital, Bergen, Norway, [109]Institue for Medical Genetics, Kepler University Hospital, Linz, Austria, [110]Hong Kong Genome Institute, Hong Kong, China, [111]Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong, China, [112]Medical Genetics Department, Koç University School of Medicine (KUSoM), 34010, Istanbul, Türkiye, [113]Department of Pediatric Genetics, Marmara University School of Medicine, Istanbul, Türkiye, [114]Department of Medical Genetics, Ege University Faculty of Medicine, Izmir, Türkiye, [115]Hacettepe University Faculty of Medicine, Department of Pediatric Genetics, Ankara, Türkiye, [116]Institut of Human Genetics, University Medical Center Göttingen, Göttingen, Germany, [117]Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, Göttingen, Germany, [118]German Center for Cardiovascular Research (DZHK), Partner Site Göttingen, Göttingen, Germany, [119]Centre for Medical Ethics, Faculty of Medicine, University of Oslo, Oslo, Norway, [120]Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway, [121]Center for Human Genetics, KU Leuven, Leuven, Belgium, [122]Division of Medical Genetics, A.I. du Pont Hospital for Children/Nemours, USA, Wilmington, Delaware, USA, [123]Department of Clinical Genetics, Robert-Debré Hospital, Paris, France, [124]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK, [125]Department of Human Genetics, New York State Institute for Basic Research in Developmental Disabilities, Staten Island, New York, United States of America, [126]George A. Jervis Clinic, New York State Institute for Basic Research in Developmental Disabilities, Staten Island, New York, United States of America, [127]Biology PhD Program, The Graduate Center, The City University of New York, New York, United States of America

[*]These authors contributed equally

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

[#]Corresponding author

# Abstract

The most important factor that complicates the work of dysmorphologists is the significant phenotypic variability of the human face. Next-Generation Phenotyping (NGP) tools that assist clinicians with recognizing characteristic syndromic patterns are particularly challenged when confronted with patients from populations different from their training data. To that end, we systematically analyzed the impact of genetic ancestry on facial dysmorphism. For that purpose, we established the GestaltMatcher Database (GMDB) as a reference dataset for medical images of patients with rare genetic disorders from around the world. We collected 10,980 frontal facial images – more than a quarter previously unpublished - from 8,346 patients, representing 581 rare disorders. Although the predominant ancestry is still European (67%), data from underrepresented populations have been increased considerably via global collaborations (19% Asian and 7% African). This includes previously unpublished reports for more than 40% of the African patients. The NGP analysis on this diverse dataset revealed characteristic performance differences depending on the composition of training and test sets corresponding to genetic relatedness. For clinical use of NGP, incorporating non-European patients resulted in a profound enhancement of GestaltMatcher performance. The top-5 accuracy rate increased by +11.29%. Importantly, this improvement in delineating the correct disorder from a facial portrait was achieved without decreasing the performance on European patients. By design, GMDB complies with the FAIR principles by rendering the curated medical data findable, accessible, interoperable, and reusable. This means GMDB can also serve as data for training and benchmarking. In summary, our study on facial dysmorphism on a global sample revealed a considerable cross ancestral phenotypic variability confounding NGP that should be counteracted by international efforts for increasing data diversity. GMDB will serve as a vital reference database for clinicians and a transparent training set for advancing NGP technology.

# Introduction

Facial dysmorphism is one of the most complex and informative clinical features in syndromic disorders, and is therefore often crucial in terms of establishing a diagnosis

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

in rare genetic diseases[1,2]. However, the recognition of dysmorphic patterns, is a challenging endeavour, and relies on the skills, knowledge, and experience of the examiner. In certain syndromes, in particular those that are ultra-rare, variability in facial features can pose challenges even for highly experienced clinicians[3]. Facial features can also vary according to sex, age, and ancestry, which further complicates the recognition of a specific dysmorphic pattern[4–6].

Ancestry plays a particularly significant role since considerable inter-ancestral variability exists in facial gestalt[7]. Thus, facial features that are common in certain ancestral groups may be considered dysmorphic in others. For example, while upslanting palpebral fissures are common in healthy Asians, they may be perceived as dysmorphic in other populations[8]. Previous studies have also highlighted differences in facial gestalt between different ancestries in common dysmorphic genetic syndromes such as Down Syndrome, 22q11.2 deletion syndrome, Noonan syndrome, and Williams–Beuren syndrome[4,9,10]. Furthermore, Lumaka et al. have demonstrated that this variability can influence the assessor, with European clinicians failing to recognize dysmorphic features in individuals of African ancestry[11]. This is a growing problem as globalization and migration increasingly blur ancestral and cultural boundaries, and geography is no longer a key determining factor in mating patterns[12]. Hence, in diverse populations, such as those with admixed ancestries, the challenge of accurately diagnosing rare diseases becomes even more pronounced since new phenotypes can evolve via admixture[13].

Ancestry also has a significant impact on the detection of rare dysmorphic disorders via artificial intelligence (AI)[11] because in most healthcare datasets, non-European ancestries are underrepresented[14]. Many next-generation phenotyping (NGP) approaches that predict disorders on the basis of facial image analysis, such as GestaltMatcher[15], have demonstrated high accuracy in patients from the ancestries in which they were predominantly trained and validated, i.e., European and North American[15–19].

Since the significantly higher birth rates in non-European regions account for 80% of the global population and 90% of all annual births (Figure 1a)[20], action is required to include non-European patients currently considered to be underrepresented. So far, few studies exist about the performance of NGP tools where the ancestry composition

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

of individuals in the training and test set differs. Literature suggests that AIs trained on individuals of European ancestry perform better on a test set of Asian rather than African ancestry[21–24] that may be explained by their closer genetic relatedness[25]. This raises the question of whether AIs need to be trained for different ancestries or whether a similar performance can be achieved by sufficiently increasing the ancestral diversity in the joint training set. The latter is indicated by a study conducted on Down syndrome patients of African ancestry[11]. However, comparing these studies is difficult since they were not performed on data compliant with FAIR principles that are findable, accessible, interoperable, and reusable, meaning the results cannot be

The motivation of our work is therefore threefold: 1) scientific, because we wanted to study the effect of inter- and intra-ancestral phenotypic variability on NGP, such as GestaltMatcher, in a systematic manner; 2) clinical, because more diverse training data can presumably increase the performance of NGP on non-European ancestries; and 3) societal, because so far underrepresented populations would benefit from potential performance improvements.

To achieve these goals, we aimed for a FAIR database with an increased number of patients of non-European ancestry with respect to comparable databases[20,26,27]. Therefore, we established the GestaltMatcher Database (GMDB) as a community-driven online framework that facilitates acquiring patient consent and incentivizes data sharing, acknowledging contributions from clinician-scientists as citeable micro-publications (Figure 2)[28–31]. Through this framework, we established global collaborations, enabling the collection of a wide range of data from various ancestries.

GMDB is the first database for medical imaging data of patients with rare genetic disorders from diverse ancestries that is compliant with the FAIR principles[32]. By its machine-readable design, GMDB also enables systematic analyses of the influence of genetic background on NGP performance, which we will report in this study.

# Results:

## Overview of FAIR data in GMDB

Retrospective data from curated publications, along with data provided by clinicians or patients, were made available as FAIR cases in the GMDB (Figure 3, Supplementary Figures 1 and 2)[33]. At the time of the data freeze for this paper on April 6th 2024, we

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

curated the GMDB-FAIR dataset consisting of 10,980 portrait images (Supplementary Figure 3) of 8,346 patients with 581 genetic disorders, including patients curated from 2,224 scientific publications. 2,312 unpublished images were contributed by 138 clinicians from 106 institutions (indicated by location markers in Figure 1a), including novel cases from GMDB micro-publications (micro-publication section in Supplementary Note). For the portrait data, which is the scope of this study, in terms of sex, the data distribution is relatively balanced (Figure 4a). However, age is biased toward patients aged below 10 years (Figure 4b). Figure 4c shows a two-dimensional representation of Human Phenotype Ontology[34] (HPO)-defined symptom groups in GMDB via Uniform Manifold Approximation and Projection (UMAP). While GMDB incorporates cases from all HPO-defined symptom groups across the disease landscape, the HPO-defined symptom group 'facial dysmorphism' is enriched in GMDB. Since each individual can be attributed to several HPO-defined symptom groups according to their features, facial dysmorphism was also present in the other HPO-defined symptom groups, as shown in the heatmap.

**Underrepresented populations benefited from micro-publication case reports in GMDB**

Through our international collaborations (Figure 1a), the representation of non-European ancestral groups is 19% for Asian, 7% for African, and 7% for Others. 67% comprises individuals of European descent (Figure 1b). Moreover, the ancestry distribution varies among different disorders. Some disorders, such as Williams-Beuren syndrome, Hyperphosphatasia with impaired intellectual development syndrome, and Cohen syndrome, have relatively diverse and balanced ancestral distributions (Supplementary Figure 4).

Notably, the proportion of African ancestry was strongly increased by means of GMDB micro-publications which account for 40% of the individuals with African ancestry (Figure 4d). In terms of specific sub-ancestries (Figure 4e), more than 80% of cases with sub-Saharan ancestry and over 20% of cases with North African, Native American, and Latin American ancestries were obtained through GMDB micro-publications.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

## Performance disparities in underrepresented populations

We analyzed the performance of GestaltMatcher on the test set of 882 images of 275 disorders with different ancestries that have not been used for the training of GestaltMatcher. Performance is measured as a top-k accuracy (as described in Methods). We report the top-1 to top-30 accuracies in Table 1. When considering top-1 accuracy, the 'Others' group demonstrated the highest performance at 73.91%, followed by the African group at 62.07%, the Asian group at 53.54%, and the European group at 55.45%. The African group achieved the highest top-5 accuracy (82.76%), the Asian group attained the highest top-10 accuracy (85.04%), while the European group only achieved 75.14% and 82.60% for top-5 and top-10 accuracies, respectively. However, the European group contains more than 50% of the testing images (523 out of 882), covering many more disorders than the other ancestry groups. That includes ultra-rare disorders known to achieve lower performances[19].

To fairly compare the European group to another non-European ancestry, we only looked at the disorders that were present in both ancestry groups. In

**Table *2***, when comparing the African and European groups on the six overlapping disorders, the European group outperformed the African group by achieving +16.96% top-1 accuracy and +11.17% top-10 accuracy. The European group also exhibited higher accuracies compared to the Asian group, with a top-1 accuracy of +6.92% and a top-10 accuracy of +4.15%. However, the European and ‘Others’ groups achieved relatively comparable results. The ‘Others’ group had a higher top-1 accuracy, while the European group performed better on the top-10 accuracy.

We further reported the performance of sex and age groups in Table 1. The distribution of testing images was relatively balanced across different groups, and no significant performance gap was observed between males and females. However, the under-one-year-old group exhibited the lowest performance, while the five- to ten-year-old group demonstrated notably higher top-5 and top-10 accuracies.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

## Diverse ancestry data enhance prediction accuracy for underrepresented populations

To investigate the impact of incorporating ancestry-diverse data on the overall performance of GestaltMatcher across ancestries, we designed two sets of ancestry analysis experiments. First, we investigated the expansion of the training set of GestaltMatcher (as described in Methods), including either European only (EU + EU*) or European and non-European (EU + non-EU) patients. We measured a top-1 accuracy averaged over all ancestral groups of 49.65% for the European only training set (EU + EU*) and 66.90% for the diverse training set (EU + non-EU) (Figure 5a). Similarly, top-5 accuracy of the European training set was 69.95%, and when we trained on the diverse set, the top-5 accuracy increased to 81.24%. Notably, the evaluation performance on images of patients with European ancestry showed only a marginal performance dropdown. Specifically, the top-1 accuracy decreased by 3.82% and the top-5 accuracy by 3.61% when the dataset was augmented with 50% more non-European images. Meanwhile, the top-1 and top-5 performance increased notably for almost every other ancestral group. Figure 5a and Table 3 show further per-ancestry performances.

The training of GestaltMatcher results in a clinical face phenotype space that can be populated by additional cases, which we refer to as the gallery set (as described in Methods). We next investigated the influence of expanding the gallery with ancestry-diverse data by gradually raising the proportion of included non-European data from 10% to 100%. Figure 5b shows that the top-1 accuracy of the non-European groups was clearly increased when we added more non-European data in the gallery. However, the top-1 accuracy of the European group did not change even when we added 100% of the non-European data into the gallery.

## GMDB-FAIR dataset drives the advancement of NGP technology

GMDB-FAIR dataset is the first dataset that can be shared with the research community to train and benchmark their NGP approaches. After the first publication of the GestaltMatcher approach in 2022, for which we initially started the collection of our FAIR data, many researchers have utilized GMDB-FAIR to develop different NGP approaches. Hustinx et al.[19], Sumer et al.[35], and Campbell et al.[36] improved the prediction accuracy of their models significantly by utilizing different loss functions,

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

network architectures, and data augmentation. Recently, Wu et al. proposed combining a large language model with facial image analysis to streamline the rare disorder diagnosis[37]. Furthermore, running facial analysis with an on-premise solution is possible using the FAIR data set to further prioritize genomic variants[38].

Moreover, the GMDB-FAIR dataset can be taken as a validatable control cohort to facilitate the delineation of the facial phenotype of disorders. GestaltMatcher can detect clusters and assess whether, for example, cases with an identical variant or pathogenic variants in the same gene share a similar facial phenotype. For example, Ebstein et al. showed that facial dysmorphism was heterogeneous among the entire *PSMC3* patient cohort, but facial similarities were found in patients sharing the same pathogenic variants[39]. To date, 15 publications have analyzed the facial phenotype of the cohort with the GMDB-FAIR dataset and GestaltMatcher[39–53]. All results can be reproduced in the research platform of GMDB, which we introduce in the Methods section (Figure 2c, Figure 3c and Supplementary Note).

# Discussion

GMDB is a modern, searchable reference and publication medium encompassing diverse populations that is designed for both clinicians and computer scientists engaged in NGP development. The ultimate goal of this study is to drive research in rare genetic disorders to understand the phenotypic variability among ancestries systematically and improve support for underrepresented populations.

GMDB stands out as the sole database compliant with FAIR principles, distinguished by its extensive collection of facial images covering diverse populations. This was mainly possible through the contributions and crowd-sourced annotations by our global collaborators. To increase motivation for data submission in the future, every case in the database has the potential to become a citable micro-publication with a Digital Object Identifier (DOI)[54]. Furthermore, future micro-publications could be indexed in reputable scientific indexing services, such as PubMed, as is the case for some existing micro-publication communication platforms[55]. Active patient involvement and the ability to access, upload and delete their data enhance patient autonomy and facilitate the acquisition of longitudinal patient data, further enriching GMDB's repository of facial images. Similar to other natural history study data, the

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

longitudinal image and associated phenotypic meta data add significant value to the understanding of disease progression in patients with facial dysmorphism[56]. Moreover, micro-publication encourages the recruitment of patients from underrepresented populations. For example, more than 40% of all images obtained for Africans had been previously unpublished. These micro-publications from unpublished images of patients with underrepresented ancestries underscored the importance of GMDB since they cannot be found in any medical journals.

The diverse ancestry data in GMDB further enabled us to investigate the GestaltMatcher performance differences among ancestral groups systematically. In **Table *2***, the performance disparities in the Asian and African groups were observed when compared to the European group. The “Others” group showed a comparable or even higher performance than the European group. The reason could be that Latin Americans in the ‘Others’ group show relatively similar facial phenotypes to the Europeans.

Our findings indicate that increasing the ancestral diversity in FAIR databases will particularly benefit populations currently regarded as underprivileged. We investigated how the top-1 and top-5 accuracies for the different ancestries changed when equally sized groups of European or non-European patients were added to the training set. Overall, the top-5 accuracy for non-European ancestral groups increased significantly when the training set was expanded with non-Europeans (+11.29%). When the training data were extended from only Europeans to Europeans and non-Europeans, only a marginal change in the performance of the European group was observed. Including more non-European patients in the gallery can also improve non-European groups' performances dramatically while European performance remains roughly the same (Figure 5b). The results indicate that recruiting non-European patients to support the underrepresented populations is more effective than recruiting more European patients, which often leads to models’ extreme bias toward European ancestry.

The GMDB-FAIR dataset offers a transparent AI training set, which is crucial for the

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

NGP development because all FAIR data are available to the clinical and scientific community. This transparency, combined with the increased representativeness of the training set, helps minimise the risk of algorithmic bias, which is key for ensuring respect for the fundamental right to non-discrimination[57]. The high quality of the GMDB data allows researchers to train, validate, and test AI in a manner that aligns with the expectation in the EU AI Act and the EU Medical Device Regulation[58]. Finally, the controlled access and consent options as described in the Methods section not only ensures respect for the fundamental right to protection of personal data[57] and EU General Data Protection Regulation (GDPR)[59] compliance, but it also enabled the creation of a more diverse, representative, and larger data set as people are more comfortable with sharing health and genetic data, including images, under controlled conditions and responsible data governance than in open access publications and repositories. By this, the GMDB-FAIR dataset falls in line with other large public datasets, such as ImageNet[60] for object classification or Labeled Faces in the Wild (LFW)[61] for face verification, which have been fundamental for deep-learning technology driving computer vision over the last decade. GMDB-FAIR has been used to develop many NGP approaches[19,35–37] for predicting rare disorders after the first usage in GestaltMatcher in 2022. Moreover, GMDB-FAIR data can be used in the research platform (Supplementary Note) to validate the results shown in the published works[39–53] that provides transparency to the researcher using GestaltMatcher and the probability to extend the existing research with the user's additional data.

Due to variability in facial phenotypes secondary to ancestry, diverse reference image databases are crucial in order to enable clinicians to learn about the phenotypic variability in facial dysmorphism within a given disorder. While efforts have been made to create an atlas of human malformations that addresses the issue of ancestral diversity, this remains limited to only a few disorders[20]. With GMDB-FAIR, we created a large-scale dataset that can be searched for disorders or genes of interest in the GMDB gallery view (Figure 2c, Figure 3b), which provides clinicians with a comprehensive selection of patient images from different ancestries at a glance, thereby eliminating the need for extensive literature searches. In addition, it facilitates facial phenotype comparisons within a given disorder among different ancestries (Supplementary Note). GMDB also represents a valuable teaching tool for training students and residents to recognize disorders based on facial features.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

To conclude, GMDB is a medical imaging database for rare disorders that encompasses diverse populations. The FAIR data will serve as reference material for clinicians that facilitates learning about facial dysmorphism across ancestries, and as a transparent training and benchmarking dataset for advancing the NGP approach. While we show improved performance for the underrepresented populations, it is important to point out that the performance is far from the optimum that can be achieved by collecting more diverse data. We envision that the gap between the European ancestral group and the underrepresented ancestries can be mitigated by micro-publications in the future, and this will result in substantially improved support for underrepresented populations.

# Methods

## Implementation of the online GMDB platform

The online platform was built using Ruby on Rails in order to allow users to input images and other patient data. A database was set up using MySQL to store the patient data. GMDB is hosted physically in the University Hospital of Bonn and is maintained by Arbeitsgemeinschaft für Gen-Diagnostik e.V. (AGD), which is a non-profit organization for genomic research. The service is funded by membership fees of the AGD and donations from the Eva-Luise und Horst Köhler Foundation and the Wirtgen Foundation.

## Image data and meta data stored in GMDB

An entry in GMDB consists of a medical image such as a portrait, X-ray, or fundoscopy and machine-readable meta information containing: 1) demographic data (including sex, age, and ancestry); 2) the molecularly confirmed diagnosis (OMIM index[62]); 3) the disease-causing mutation reported in Human Genome Variation Society format[63] (HGVS) or International System for Human Cytogenomic Nomenclature[64] (ISCN) with test method and zygosity; and 4) the clinical feature encoded in HPO terminology[34] (Figure 2b). When submitting data, clinicians are also asked to state their expert opinion concerning the distinctiveness of a phenotype: They are asked to score whether the medical imaging data was supportive (1), important (2), or key (3) in establishing the clinical diagnosis. Computer scientists can use this information to interpret the performance of their AI[15].

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

## Digital consent form and patient-centered data upload

To facilitate faster retrospective patient recruitment, a digital consent form has been implemented, which allows patients to select conditions for storing their data within the database and enables the provision of their signature online. To address the specific requests of patients, this feature was further developed in close collaboration with patient support groups, e.g., the German Smith-Magenis Syndrome patient organization Sirius e.V. Patients can access their own cases and provide or withdraw their consent online. They can also upload images themselves, which greatly simplifies the curation process for longitudinal image data and other prospective data. The fact that documents such as letters from clinicians or laboratory results can also be uploaded, while only being visible to the responsible clinician, makes it possible to obtain molecular and phenotype information on patients recruited retrospectively from patient support groups. This digital consent is developed in such a way that it could also, in principle, be used as a dynamic consent model in the future[65]. The consent form is available in German and English, and other languages will be incorporated in the near future. Please find them in Supplementary Note (Digital consent, and Supplementary Figures 5 and 6) for more details.

## Data curation

The curated data can be broadly categorized as retrospective and prospective. Retrospective refers primarily to data collected from the literature or from similar projects with global consent for data sharing (e.g., Minerva&Me[66]). For cases curated from the literature, the DOI and PubMed ID as well as the contact details of the corresponding author were collected in order to clarify whether reuse is possible while respecting intellectual property rights. Following the provision of written informed consent, our collaboration partners, clinicians from around the world (Figure 1a and the co-authors), also recruited patients with an established diagnosis from within their clinical practice or from patient support groups. Prospective curation refers to the collection of further images or metadata over time. This can be done by the attending clinician after subsequent consultations, or by the patients themselves.

The curation process can be broadly subdivided into three phases. First, medical students in their final year annotated cases from the literature, mainly searched

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

PubMed and Google Scholar for publications with images of patients with facial dysmorphism and monogenic molecular diagnosis.

Second, solved patients were recruited from patient support groups. Included patients were allowed to upload and delete images and findings autonomously and access their data at any time. To develop a patient-centered, user-friendly platform and strengthen patient autonomy, feedback was obtained from the recruited patients during this phase in order to determine whether any adjustments to the process were required.

In the third phase, the database was expanded via international collaborations with clinicians from different continents. Initially, this focused on patients who had already been solved but had not yet been published in order to improve the AI's performance. However, as we progressed, more clinicians shared their unsolved cases with the scientific community. GMDB then started focusing on facial portraits of patients with rare monogenic diseases, and is now dominated by, but not limited to, such cases. Later in the curation process, we also annotated cytogenetic disorders with facial dysmorphism. In addition to these clinicians, the medical students continued to annotate data from the literature.

**Digital Object Identifier assignment**

After data submission, the respective case is immediately published on the website. Subsequently, the author has the option of generating a DOI in order to create a citable micro-publication[54]. To do this, clinicians must, after uploading the required data and metadata, enter their own personal identifier (e.g.,ORCID), specify all other scientists or clinicians involved in this case, and provide a title and an abstract. To ensure the credibility and reliability of the published data, this process will adhere to a rigorous review similar to that described by Raciti et al.[55]. The DOIs are created and managed by the University and State Library of Bonn using the DataCite Application Programming Interface (API) (https://datacite.org).

Additionally, a dedicated landing page will be created for each case, according to the specifications of the DataCite metadata schema (Supplementary Figure 2). The landing page is accessible via the generated DOI, even for individuals without access to GMDB or those who are not logged in. The landing page contains the full citation

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

with the DOI as a link, the abstract, and a description of the case data. No phenotypic information, HPO terms, or images are available. However, the landing page indicates how many images the micropublication contains.

**Main components of the GMDB online platform**

The GMDB consists of three main components that can in principle be utilized by registered users (Figure 2c). 1) Search: Clinicians can use the Gallery view to search the GMDB for disorders or genes of interest and get all patients matching this search criterion displayed in the database at a glance. 2) Analyze: Clinicians and scientists can use the GMDB-FAIR data to perform similarity comparisons of cohorts with GestaltMatcher within the research platform of GMDB. 3) Train: The GMDB-FAIR dataset that can be used by external researchers to train NGP tools. More detailed information on these features can be found in the Supplement Note.

**GMDB datasets**

All analyses performed in this paper are based on GMDB-FAIR data (v1.1.0). But actually, the GMDB consists of the GMDB-FAIR dataset and the GMDB-private set (Supplementary Note and Supplementary Figures 7 and 8). We introduced this distinction because it is known that patient consent to data sharing is higher when not shared with a broad mass, but only for a specific study[67]. However, many patients agree to controlled access for the general scientific community to advance research[67]. For this reason, patients can decide whether they want to be part of only the GMDB-private set for AI training or agree to be part of the FAIR data set.

The website displays the statistics to the public, showing how many patients are in the database and how many disorders and disease genes have been curated. When the user has the link to a specific case in the GMDB (e.g., from a publication in which the original image may not be branched, but a link to the case is given in the GMDB), if the user is not logged in, the landing page for the case will show how many images and metadata are available for the case. Only sex and ancestry, as well as the disease gene, are given. If it is a case report published with a DOI in the GMDB, the corresponding title and abstract of the case can also be viewed. The remaining data can only be viewed after logging in. To visualize the images, the user has to log in to the platform.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

## GMDB-FAIR data set

The FAIR data set (Supplementary Figure 7b) is accessible to the scientific community. Data comes from publications and from clinicians or patients themselves. However, the case is accessible in the Gallery view for all registered users of the GMDB, and the data sheet with all relevant data and metadata can be viewed. It is also available to all users of the GMDB to perform similarity comparisons of cohorts in the research platform (Supplementary Note). The data is used for the GestaltMatcher training and test set but can also be made available to other scientists to train and test their AI after they have applied to us with an Institutional Review Board (IRB)-approved study and proposal.

## Data Governance and Ethical, Legal and Social Implications of GMDB

Ethical approval for the GMDB was granted by the IRB of the University of Bonn, and all patients have given informed written consent to participate. During the GestaltMatcher consent procedure, patients can also indicate whether they agree to the use of the images in presentations, teaching activities, or in publications in other journals. This differentiation from other journals is important since patients/parents show less willingness to consent to publication in open-access journals than to publication in access-controlled databases that are not publicly accessible[67].

The GMDB has four different levels of data access (Supplementary Figure 8): 1) The public data, which includes a summary of the GMDB statistics on the website and a landing page for case reports with DOI (Supplementary Figure 2), requires no login and is openly accessible. 2) The FAIR data, which can be viewed with a GMDB user account, and in principle, downloaded by external AI researchers. 3) The restricted data, which is not accessible to GMDB users and external AI researchers and can only be used to train the GestaltMatcher AI. 4) Patient-shared data: Patients can only view their own case and upload data if they are invited to do so by the attending clinician.

External scientist in the field of AI can apply to download of GMDB-FAIR data for the development of NGP approaches. Prerequisites for this are IRB approval and submission of a proposal to info@gestaltmatcher.org. In addition, external scientists must sign and adhere to the GDPR. The Advisory Board will conduct a thorough review of all applications. If the majority of the members of the Board approve the application,

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

access (under the extent permissable by law) will be granted to applicants within two to three weeks.

## Advisory Board

Advisory Board comprises the following co-authors: Benjamin D. Solomon, Koen Devriendt, Shahida Moosa, Christian Netzer, Martin Mücke, Christian Schaaf, Alain Verloes, Christoffer Nellåker, Markus M. Nöthen, Gholson J. Lyon, Aleksandra Jezela-Stanek, and Karen W. Gripp.

## HPO-defined symptom groups

In one of our previous works[68], twelve distinct and non-overlapping categories of HPO terms were defined by clinical experts ("HPO defined symptom groups"). All GMDB cases for which HPO terms were annotated were then assigned to each of those groups, if at least one of the HPO terms in this group was annotated; i.e., each GMDB case can be assigned to several HPO-defined symptom groups. For each case, the most pronounced HPO-defined symptom group was defined as the single group comprising the largest number of the case's annotated HPO terms. The HPO-defined symptom group "Others" was only assigned as the leading HPO-defined symptom group if no other HPO-defined symptom group was present for the case.

Phenotypic similarity between cases was calculated using the R-package ontologySimilarity (version 2.5). Pairwise similarities were calculated for the combined data set of GMDB cases with HPO terms (n=4,474), the TRANSLATE-NAMSE exome sequencing data set (n=1,577), and data on known diseases and their clinical features downloaded from the HPO website (n=7,765, https://hpo.jax.org/app/download/annotation, file: genes_to_phenotype.txt, downloaded on 10 April 2021). The resulting distance matrix was projected in a four-dimensional space via Uniform Manifold Approximation and Projection (UMAP). The first two dimensions were plotted using ggplot2 (version 3.4.4). To analyze which HPO-defined symptom groups occur jointly, the proportion of patients assigned to the first group that were also assigned to the second group was assessed. All analyses were conducted in R (version 4.3.2).

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

## GestaltMatcher Algorithm

DeepGestalt[17] is a deep learning-based NGP tool using frontal face photos to classify up to 216 syndromes it has seen during training. However, it needed a lot of training data to achieve a reasonable performance on these syndromes. That also meant it could not classify unseen syndromes during training (ultra-rare syndromes). This led to the development of GestaltMatcher[15], which uses a clustering approach. As such, if at least one image of the sought-after syndrome is in the gallery set, a test image can be matched to/clustered with that image using some similarity metric. Later this approach was further enhanced by Hustinx et al.[19], using a more recent architecture (iResNet) and training loss (ArcFace Loss), as well as test-time augmentation and a model ensemble to improve robustness. That is also the approach we used for our experiments. Thus, for fine-tuning we utilized the Adam optimizer, cross-entropy loss, and class weighting to deal with the imbalance in data availability between disorders.

The overall idea behind the methodology is to train a classifier on a more frequent subset of the syndromes, achieving a model that generalizes well on those seen syndromes. In practice, the authors of both papers decided to use syndromes with at least seven patients as the training set for this classifier. Thereafter, everything up to the penultimate layer of the classifier is used as an encoder, obtaining feature embeddings of images of interest. These could be images for the gallery set or images for the test set.

The aforementioned gallery set is the set of images (and their feature embeddings) with known syndromes. This can include the syndromes used for training (seen) and syndromes with too few images to train on (unseen). The theory is that similar facial phenotypes form clusters in the feature space, which is spanned by the feature embeddings in 512 dimensions and which we refer to as clinical face phenotype space. The similarity between images and clusters is computed using the cosine distance, where a lower distance implies a higher similarity. Contrary to the approach by Gurovich et al.[17], this approach can easily increase support for ultra-rare syndromes. The quality and diversity of the gallery set is crucial for this approach to match test images to clusters in the gallery set.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

### Performance metric (top-k accuracy)

The applied performance metric was top-k accuracy. Top-1 indicates that the disorder was correctly classified as the first guess, while top-5 indicates the correct class was in the first five guesses. We reported top-k accuracies (k=1, 5, 10, and 30) as the performance readout.

### Ancestry analysis

The genetic ancestry of each individual was documented as precisely as possible using self-reported data. For instance, if an individual was born in Germany and all of the respective grandparents also originated from there, this individual was assigned to Germany (country) and Europe (continent). The same approach was used for all individuals with no self-reported migration history in previous generations. For individuals with mixed ancestry, the respective ancestries were combined. For example, an individual with a father from Gambia and a mother from Eastern Europe was assigned European-African mixed ancestry.

The performance of GestaltMatcher is highly dependent on the training set and the gallery set. To investigate the impact of incorporating diverse ancestry on the performance, we have therefore conducted two sets of experiments for those two components, respectively. First, we analyzed the influence on the models' performance when including only European versus both European and non-European data into the training set. And second, we analyzed the same performance when iteratively increasing the amount of non-European data into the gallery set.

In the first experiment, a subset of images of European patients (EU) was extended by either the inclusion of a different subset of images of European patients (EU*), or a subset of patients with non-European ancestries (non-EU) (Supplementary Figure 9). Random sampling of these subsets was performed five times. EU consisted of on average 3,139.2 images, and EU* comprised on average 1,567.6 images. First, the model was trained on the EU + EU* set containing on average 4,706.8 images of patients of solely European ancestry. For EU + non-EU, a subset containing on average 1,567.6 images of patients with any non-European ancestry was used, totaling to 4,706.8 images. The experiment design ensured the maintenance of the same distribution of disorders as that found in the training data.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

The model was fine-tuned for 50 epochs on subsets EU + EU* and EU + non-EU of GMDB (v1.1.0). All other hyperparameters were left unchanged. It is important to note that the model was not tasked with learning to classify the ancestry, only with learning to classify the disorder.

Post-training, the models' performances were measured on the same evaluation set, containing images of patients with diverse ancestral backgrounds. This evaluation set consisted of 649 images and was sampled in such a manner that there was no overlap between patients or images in any subset. Top-k accuracy was averaged over each ancestry rather than each image in order to address the imbalance in ancestry frequency. As such, the performance of any infrequent group weighed equally with those of the more frequent groups.

In the second set of experiments, we trained the models of Hustinx et al.[19] using the GMDB-FAIR training set, including different proportions of non-EU data for the gallery set. We compared the performance of the syndromes our models have seen during training. For completeness, Table 1 shows the top-k accuracy (over all images) for different categories (sex, ancestry, and age range) using the entire gallery set (100% EU + 100% non-EU). For the experiments, we computed the performance when including different proportions of non-EU data, extending the gallery set by +10% per iteration. This experiment was repeated tenfold, randomly sampling patients with different ancestries and all their photos for the gallery set. As such, at 0%, we include only data from EU patients in the gallery set, and at 100%, we include all patient data for the relevant syndromes.

We further computed the performance on syndromes that occur in both the European-group and each non-European group to more accurately reflect the performance differences, avoiding the imbalance between offered support for each ancestral group.

## Data and code availability

GMDB-FAIR can be downloaded in GMDB after the application is approved by the advisory board. Please find more details in the Data Governance and ELSI section. Code is available in the GitHub repository (github.com/igsb/GestaltMatcher-Arc/tree/gmdb).

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

# Acknowledgments

This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, the United States of America. Sofia Douzgou Houge was supported by the Norwegian National Advisory Unit on Rare Disorders (grant number #43066). Tahsin Stefan Barakat was supported by the Netherlands Organisation for Scientific Research (ZonMw Vidi, grant 09150172110002). Heidi Beate Bentzen is supported by EU grant 101071203 and Research Council of Norway grants 322672 and 324278. Nina-Maria Wilpert was supported by the DFG Research Unit 2841 “Beyond the Exome” and is a participant in the BIH Charité Junior Clinician Scientist Program funded by the Charité - Universitätsmedizin Berlin, the Berlin Institute of Health at Charité (BIH), the Alliance4Rare, and the Berliner Sparkassenstiftung Medizin. Cristina Dias was supported by the Wellcome Trust [grant number 209568/Z/17/Z]. The authors thank the Asia Pacific Society of Human Genetics, the Wirtgen Foundation, the Eva Luise und Horst Köhler Foundation, Kabuki Syndrome Foundation, Kleefstra support group, German Smith-Magenis Syndrome patient organization Sirius e.V. and the Focus Foundation for their support.

# Figures

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

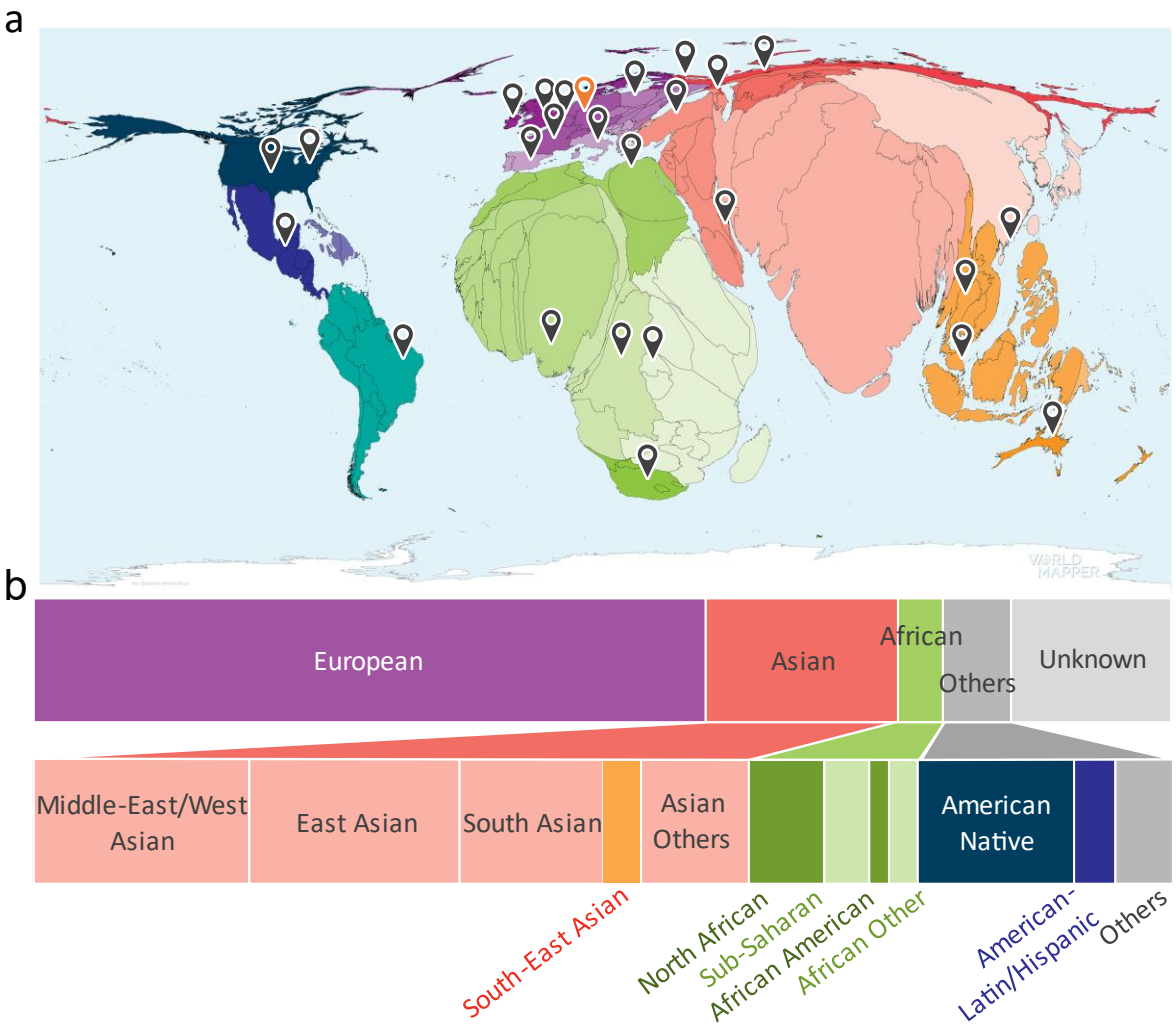



**Figure 1: a)** Birth rate distribution worldwide. The size of country is scaled in accordance with the respective birth rate. The map indicates countries from which unpublished images were obtained (source: https://worldmapper.org/faq/, modified). **b)** Distribution of ancestry groups in GestaltMatcher Database. 16% of the patients without ancestral information were categorized as Unknown. The breakdown of ancestries in the dataset with known ancestry is as follows: European 67%, Asian 19%, African 7%, and Others 7%.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

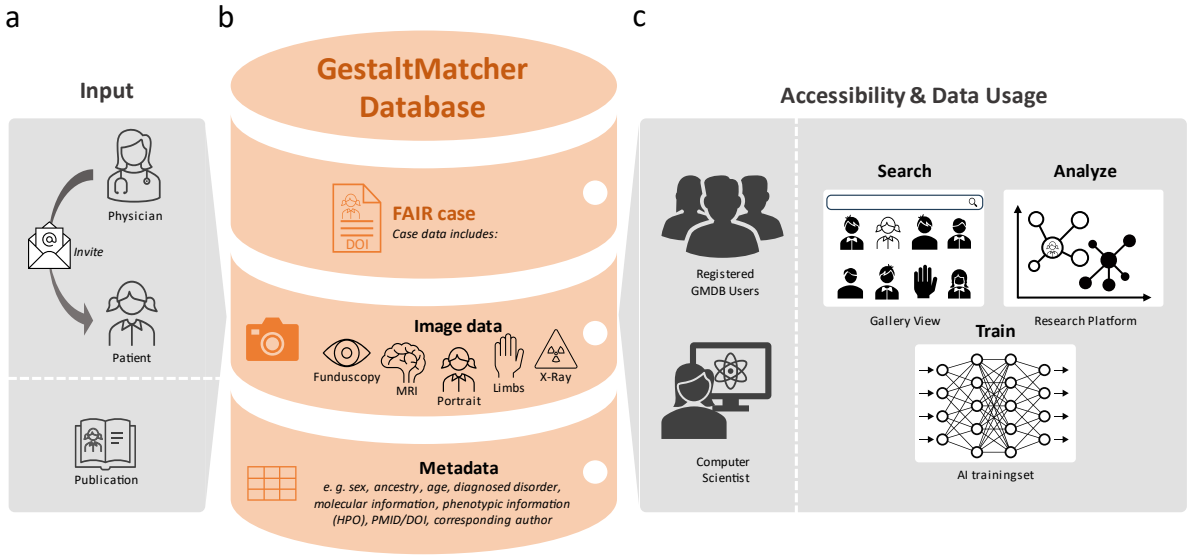



**Figure 2: GestaltMatcher Database (GMDB) Architecture and Dataflow. a)** Retrospective data are collected from the literature and annotated by data curators or are uploaded by collaborating attending clinician. Patients can also upload images of their own cases, incorporate prospective data, and view their own data at any time. **b)** The data (multimodal image data, including portrait images as well as magnetic resonance imaging, X-ray, fundscopy and extremity images) are stored in the GMDB (MySQL database) together with the relevant meta information (such as sex, age, ancestry, molecular, and phenotypic information). **c)** Registered users can view and search the FAIR data in the GMDB Gallery. The patient image can also be analyzed using the Next-Generation Phenotyping tool GestaltMatcher within the Research Platform. In addition, once their application has been approved by the Advisory Board, external computer scientists can use the GMDB-FAIR data set for training purposes for their projects.

**Figure 3: An example case presentation of a FAIR case with a Digital Object Identifier (DOI). a)** FAIR cases in the GestaltMatcher Database (GMDB) are displayed to GMDB users via the data sheet. Each FAIR case can also be assigned a DOI in order to render it a citable micro-publication. This micro-publication contains the image data and metadata, including demographic, molecular, and phenotype information. The dynamic nature of the GMDB case report enables longitudinal image data storage even after initial publication, which is not possible in conventional journals. **b)** After uploading, case reports can be viewed and searched by other users in the Gallery

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

view. **c)** The image data can also be used for inter-cohort comparisons of the gestalt scores within the research platform.

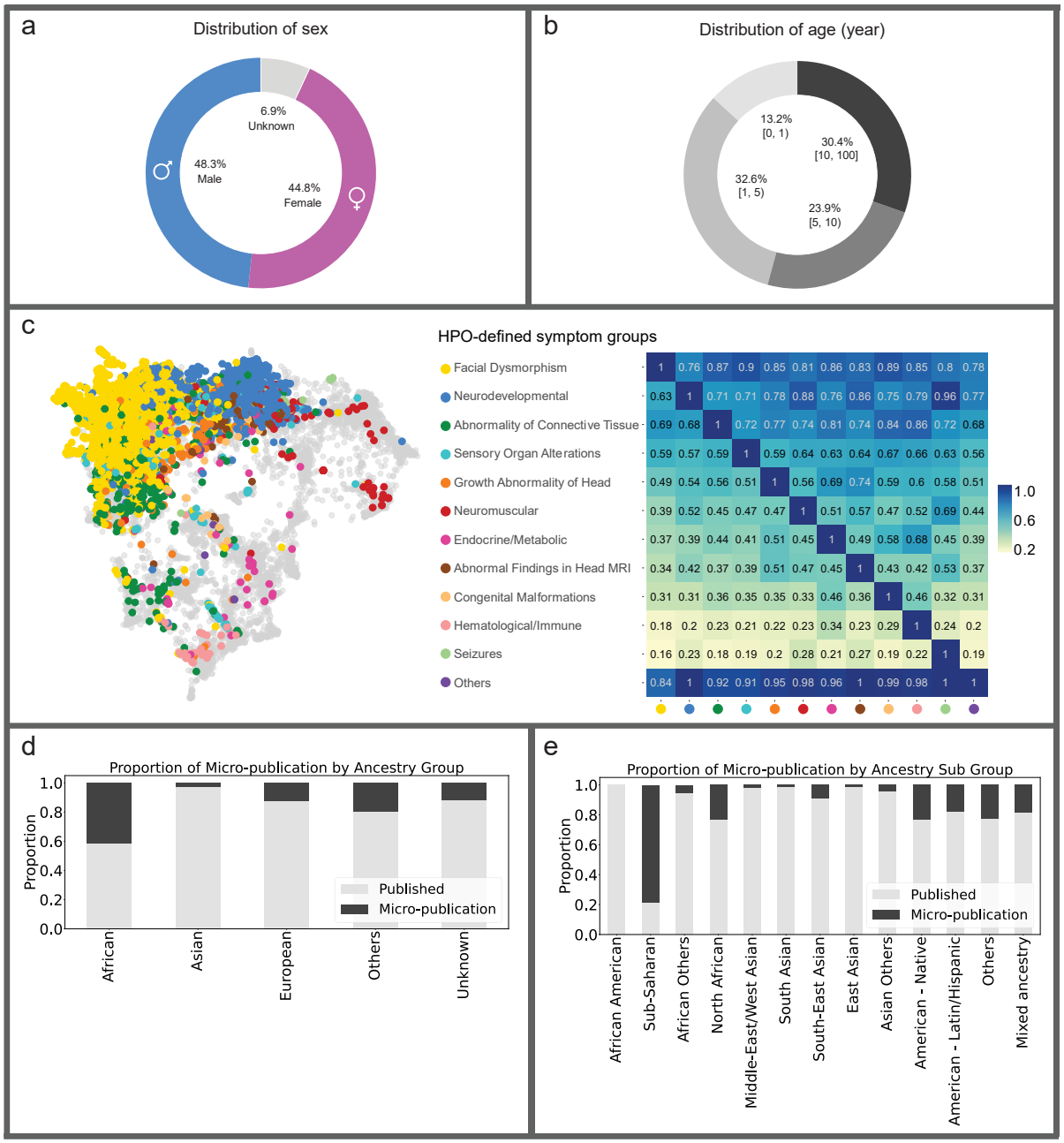
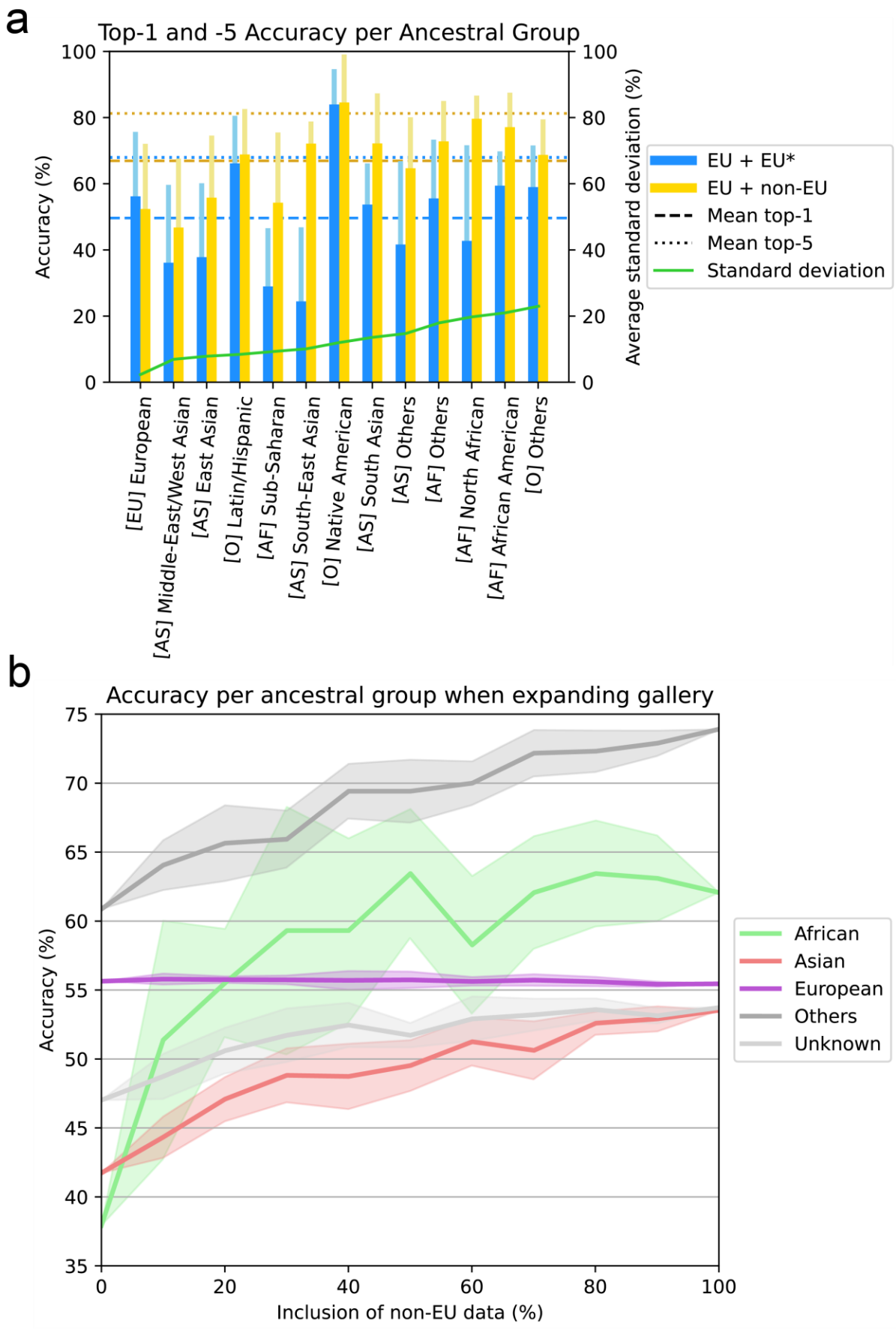



**Figure 4: Overview of the GestaltMatcher Database (GMDB)-FAIR dataset. a)** Sex distribution. Number of images shown in brackets. **b)** Distribution of patient age in years. **c)** Left: Two-dimensional representation of phenotypic similarities between patients, as calculated on the basis of Human Phenotype Ontology (HPO) terms via Uniform Manifold Approximation and Projection (UMAP). HPO terms were annotated for 4,474 individuals in the GMDB, and expert clinicians defined twelve distinct HPO-

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

defined symptom groups. Based on the annotated HPO terms, each case was assigned to one or more HPO-defined symptom groups. All OMIM diseases were included, using their HPO annotations (gray background dots) as a reference. GMDB cases are color-coded according to their most pronounced HPO-defined symptom group, i.e., the group that includes the majority of their HPO terms. The dataset is dominated by two major clusters (facial dysmorphism in yellow and neurodevelopmental in blue) but shows cases from across the complete disease landscape. Right: Heatmap of the proportion of GMDB individuals within the HPO-defined symptom group on the X-axis who are also assigned to the HPO-defined symptom group on the Y-axis. Notably, facial dysmorphism is present in at least 70% of the cases of each HPO-defined symptom group. **d)** Proportion of the unpublished and published images in each ancestry group. **e)** Proportion of the unpublished and published images in each sub-ancestry group.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

**Figure 5: Performance of ancestry analysis. a)** Top-1 and top-5 accuracy of GestaltMatchers' disorder classification accuracy per ancestral group**.** Top-1 and top-5 accuracy of the models' disorder classification accuracy per ancestral group, where (blue) belongs to the EU only subset, and (yellow) belongs to the diverse subset. Each wide, darker bar and each light, thinner bar indicate the top-1 and top-5 accuracy per ancestral group, respectively. The horizontal dashed lines and dotted lines indicate the top-1 and top-5 overall accuracy averaged over all ancestral groups, respectively. The order of the ancestry group in the x-axis is ranked according to standard deviation between top-1 accuracies of the 5-fold experiment. **b)** Top-1 accuracy of GestaltMatcher when including different proportion of non-European patients in the gallery. The x-axis is the proportion of non-European data included in the gallery. The y-axis is the top-1 accuracy. The colored region along the line indicates the standard deviation.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

# Tables

**Table 1: Performance of GestaltMatcher on different categories of sex, ancestry, and age.** The top-1, top-5, top-10, and top-30 accuracy are reported. For the top-1 to top-30 columns, the best performance in each category is boldfaced. In the ancestry category, the sampling influences European and other ancestry groups' performance due to the significant difference in the test image size. They may evaluate the different sets of disorders. We, therefore, presented the performance of the overlapped disorders in

**Table *2***. In the age category, the notation [x, y) represents a half-open interval, which includes the starting point x but excludes the endpoint y. For example, [0, 1) years range from birth but do not include one year old.

| Category | | Test images | Top-1 | Top-5 | Top-10 | Top-30 |
|---|---|---|---|---|---|---|
| **Overall** | | 882 | 56.58% | 76.08% | 82.61% | 90.36% |
| **Ancestry** | African | 29 | 62.07% | **82.76%** | 82.76% | 86.21% |
| | Asian | 127 | 53.54% | 78.74% | **85.04%** | 89.76% |
| | European | 523 | 55.45% | 75.14% | 82.60% | 90.25% |
| | Others | 69 | **73.91%** | 81.16% | 81.16% | **92.75%** |
| | Unknown | 134 | 53.73% | 73.13% | 81.34% | 91.04% |
| **Sex** | Male | 419 | 55.37% | 74.22% | 80.67% | 88.78% |
| | Female | 393 | 55.98% | 75.83% | 83.21% | 91.09% |
| | Unknown | 70 | **67.14%** | **88.57%** | **91.43%** | **95.71%** |
| **Age** | [0, 1) years | 53 | 52.83% | 71.70% | 79.25% | 90.57% |
| | [1, 5) years | 137 | 56.20% | 75.91% | 81.02% | 90.51% |
| | [5, 10) years | 115 | 57.39% | **83.48%** | **86.09%** | 90.43% |
| | [10, ∞) years | 165 | **58.18%** | 71.51% | 77.58% | 85.45% |
| | Unknown | 412 | 56.31% | 76.46% | 84.71% | **92.23%** |

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

**Table 2: Performance comparison between European and other ancestry groups on the overlapping disorders.** This table is an extension of the ancestry section in Table 1, taking the overlapped disorders between European and other ancestry groups. Each category compares European and non-European ancestry groups' performance on the same set of disorders. The number of overlapped disorders is reported in the 'Disorders' column. In comparing African and European groups, six disorders exist in the test sets of both ancestry groups. The top-1, top-5, top-10, and top-30 accuracy are reported. For the top-1 to top-30 columns, the best performance in each category is boldfaced.

| Category | Disorders | | Test images | Top-1 | Top-5 | Top-10 | Top-30 |
|---|---|---|---|---|---|---|---|
| **[African, European]** | 6 | African | 14 | 64.29% | 85.71% | 85.71% | 92.86% |
| | | European | 32 | **81.25%** | **96.88%** | **96.88%** | **100.00%** |
| **[Asian, European]** | 36 | Asian | 83 | 57.83% | 79.52% | 84.34% | 87.95% |
| | | European | 139 | **64.75%** | **82.73%** | **88.49%** | **91.37%** |
| **[Others, European]** | 20 | Others | 53 | **81.13%** | **90.57%** | 90.57% | **100.00%** |
| | | European | 115 | 69.56% | 84.35% | **93.91%** | 96.52% |
| **[Unknown, European]** | 32 | Unknown | 77 | 59.74% | **81.81%** | 88.31% | **96.10%** |
| | | European | 170 | **62.35%** | 81.18% | **89.41%** | 94.12% |

**Table 3: Training accuracy with EU + non-EU and EU + EU* datasets.** Within the European training row, numbers annotated with * in brackets indicate the training images from EU + EU. Higher top-1 and top-5 accuracies between EU + EU* and EU + non-EU training are denoted in bold.

| | Number of images | | Performance EU + non-EU | | Performance EU + EU* | |
|---|---|---|---|---|---|---|
| | Training | Testing | Top-1 | Top-5 | Top-1 | Top-5 |
| **European** | (4706.2 ± 24.4)* 3139.2 ± 15.1 | 444.6 ± 22.2 | 52.35 ± 2.30% | 72.05 ± 2.66% | **56.17 ± 2.27%** | **75.66 ± 2.70%** |
| **East Asian** | 283.2 ± 5.0 | 31 ± 6.2 | **55.78 ± 10.25%** | **74.56 ± 5.90%** | 37.77 ± 5.45% | 60.13 ± 5.77% |
| **Latin/Hispanic** | 257.8 ± 7.0 | 28.4 ± 4.7 | **68.86 ± 8.92%** | **82.56 ± 7.77%** | 66.16 ± 7.89% | 80.51 ± 6.58% |
| **Middle-East/ West Asian** | 211.2 ± 6.8 | 30 ± 5.8 | **46.76 ± 7.01%** | **67.59 ± 7.52%** | 36.10 ± 6.81% | 59.67 ± 3.68% |

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

| | | | | | | |
|---|---|---|---|---|---|---|
| **South Asian** | 200.2 ± 5.4 | 18.8 ± 2.7 | **72.15 ± 12.24%** | **87.32 ± 10.04%** | 53.70 ± 14.86% | 66.13 ± 13.40% |
| **Asian Others** | 170.6 ± 2.4 | 16.4 ± 4.1 | **64.66 ± 10.93%** | **80.06 ± 13.87%** | 41.64 ± 18.51% | 66.84 ± 11.87% |
| **Sub-Saharan** | 119 ± 2.7 | 18.2 ± 3.4 | **54.23 ± 7.32%** | **75.49 ± 15.27%** | 28.91 ± 11.18% | 46.55 ± 11.71% |
| **North African** | 64.8 ± 3.2 | 7.4 ± 1.6 | **79.64 ± 20.19%** | **86.64 ± 14.62%** | 42.71 ± 19.34% | 71.64 ± 18.38% |
| **Native American** | 63.2 ± 6.8 | 14.8 ± 1.6 | **84.55 ± 12.77%** | **99.09 ± 1.82%** | 83.94 ± 11.18% | 94.65 ± 8.62% |
| **African Others** | 53.2 ± 2.0 | 6.2 ± 2.2 | **72.78 ± 18.29%** | **85.00 ± 13.33%** | 55.56 ± 17.57% | 73.33 ± 22.61% |
| **South-East Asian** | 51.4 ± 2.0 | 5.4 ± 1.3 | **72.12 ± 13.36%** | **78.81 ± 24.61%** | 24.40 ± 6.76% | 46.83 ± 17.97% |
| **Others** | 54.6 ± 2.3 | 6 ± 2.9 | **68.71 ± 23.67%** | **79.43 ± 22.38%** | 59.00 ± 22.35% | 71.57 ± 21.09% |
| **African American** | 38.4 ± 4.3 | 3.8 ± 2.6 | **77.08 ± 18.04%** | **87.50 ± 21.65%** | 59.38 ± 24.00% | 69.79 ± 18.49% |
| **Overall** | 4706.8 ± 26.7 | 631 ± 23.8 | **66.90%** | **81.24%** | 49.65% | 67.95% |

# References


1. Hart, T. C. & Hart, P. S. Genetic studies of craniofacial anomalies: clinical implications and applications. *Orthod. Craniofac. Res.* **12**, 212–220 (2009).
2. Lesmann, H., Klinkhammer, H. & Dr. med. Dipl. Phys. Peter M. Krawitz. The future role of facial image analysis in ACMG classification guidelines. *Med. Genet.* **35**, 115–121 (2023).
3. Tekendo-Ngongang, C. *et al.* Rubinstein-Taybi syndrome in diverse populations. *Am. J. Med. Genet. A* **182**, 2939–2950 (2020).
4. Kruszka, P., Tekendo-Ngongang, C. & Muenke, M. Diversity and dysmorphology. *Curr. Opin. Pediatr.* **31**, 702–707 (2019).
5. Hadj-Rabia, S. *et al.* Automatic recognition of the XLHED phenotype from facial images. *Am. J. Med. Genet. A* **173**, 2408–2414 (2017).
6. Martínez-Abadías, N. *et al.* Facial biomarkers detect gender-specific traits for bipolar disorder. *FASEB J.* **35**, (2021).
7. Fang, F., Clapham, P. J. & Chung, K. C. A systematic review of interethnic variability in facial dimensions. *Plast. Reconstr. Surg.* **127**, 874–881 (2011).
8. Vorravanpreecha, N., Lertboonnum, T., Rodjanadit, R., Sriplienchan, P. & Rojnueangnit,

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

K. Studying Down syndrome recognition probabilities in Thai children with de-identified computer-aided facial analysis. *Am. J. Med. Genet. A* **176**, 1935–1940 (2018).

9. Kruszka, P. *et al.* Down syndrome in diverse populations. *Am. J. Med. Genet. A* **173**, 42–53 (2017).
10. Porras, A. R., Summar, M. & Linguraru, M. G. Objective differential diagnosis of Noonan and Williams-Beuren syndromes in diverse populations using quantitative facial phenotyping. *Mol Genet Genomic Med* **9**, e1636 (2021).
11. Lumaka, A. *et al.* Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clin. Genet.* **92**, 166–171 (2017).
12. Burchard, E. G. *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
13. Martínez-Abadías, N. *et al.* Phenotypic evolution of human craniofacial morphology after admixture: a geometric morphometrics approach. *Am. J. Phys. Anthropol.* **129**, 387–398 (2006).
14. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
15. Hsieh, T.-C. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat. Genet.* **54**, 349–357 (2022).
16. Dudding-Byth, T. *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.* **17**, 90 (2017).
17. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).
18. Porras, A. R., Rosenbaum, K., Tor-Diez, C., Summar, M. & Linguraru, M. G. 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) doi:10.1016/S2589-7500(21)00137-0.
19. Hustinx, A. *et al.* Improving Deep Facial Phenotyping for Ultra-rare Disorder Verification

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Using Model Ensembles. in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 5007–5017 (IEEE, 2023).

20. Muenke, M., Adeyemo, A. & Kruszka, P. An electronic atlas of human malformation syndromes in diverse populations. *Genet. Med.* **18**, 1085–1087 (2016).
21. Mishima, H. *et al.* Evaluation of Face2Gene using facial images of patients with congenital dysmorphic syndromes recruited in Japan. *J. Hum. Genet.* **64**, 789–794 (2019).
22. Narayanan, D. L. *et al.* Computer-aided Facial Analysis in Diagnosing Dysmorphic Syndromes in Indian Children. *Indian Pediatr.* **56**, 1017–1019 (2019).
23. Elmas, M. & Gogus, B. Success of Face Analysis Technology in Rare Genetic Diseases Diagnosed by Whole-Exome Sequencing: A Single-Center Experience. *Mol. Syndromol.* **11**, 4–14 (2020).
24. Hennocq, Q. *et al.* Next generation phenotyping for diagnosis and phenotype-genotype correlations in Kabuki syndrome. *Sci. Rep.* **14**, 2330 (2024).
25. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
26. Winter, R. M. & Baraitser, M. The London Dysmorphology Database. *J. Med. Genet.* **24**, 509–510 (1987).
27. Murdoch Children's Research Institute. POSSUMweb. *POSSUMweb* https://www.possum.net.au/.
28. Patrinos, G. P. Chapter 6 - Incentives for Human Genome Variation Data Sharing. in *Human Genome Informatics* (eds. Lambert, C. G., Baker, D. J. & Patrinos, G. P.) 109–129 (Academic Press, 2018).
29. Mons, B. *et al.* The value of data. *Nat. Genet.* **43**, 281–283 (2011).
30. Patrinos, G. P. *et al.* Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Hum. Mutat.* **33**, 1503–1512 (2012).
31. Giardine, B. *et al.* Systematic documentation and analysis of human genetic variation in

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

hemoglobinopathies using the microattribution approach. *Nat. Genet.* **43**, 295–301 (2011).

32. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
33. Lesmann, H. & Weiland, H. Atypical presentation of a case with Noonan syndrome with multiple lentigines (Version 1). (2024) doi:10.60723/10693.
34. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
35. Sümer, Ö., Hellmann, F., Hustinx, A., Hsieh, T.-C. & Krawitz, P. Few-Shot Meta-Learning for Recognizing Facial Phenotypes of Genetic Disorders. in *Caring is Sharing – Exploiting the Value in Data for Health and Innovation* 932–936 (IOS Press, 2023).
36. Campbell, J., Dawson, M., Zisserman, A., Xie, W. & Nellåker, C. Deep Facial Phenotyping with Mixup Augmentation. in *Medical Image Understanding and Analysis* 133–144 (Springer Nature Switzerland, 2024).
37. Wu, D. *et al.* Multimodal Machine Learning Combining Facial Images and Clinical Texts Improves Diagnosis of Rare Genetic Diseases. *arXiv [q-bio.QM]* (2023).
38. Hsieh, T.-C., Lesmann, H. & Krawitz, P. M. Facilitating the Molecular Diagnosis of Rare Genetic Disorders Through Facial Phenotypic Scores. *Curr Protoc* **3**, e906 (2023).
39. Ebstein, F. *et al.* PSMC3 proteasome subunit variants are associated with neurodevelopmental delay and type I interferon production. *Sci. Transl. Med.* **15**, eabo3189 (2023).
40. Asif, M. *et al.* De novo variants of CSNK2B cause a new intellectual disability-craniodigital syndrome by disrupting the canonical Wnt signaling pathway. *HGG Adv* **3**, 100111 (2022).
41. Kampmeier, A. *et al.* PHIP-associated Chung-Jansen syndrome: Report of 23 new individuals. *Front Cell Dev Biol* **10**, 1020609 (2022).
42. Lyon, G. J. *et al.* Expanding the phenotypic spectrum of NAA10-related neurodevelopmental syndrome and NAA15-related neurodevelopmental syndrome. *Eur.*

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

*J. Hum. Genet.* **31**, 824–833 (2023).

43. Aerden, M. *et al.* The neurodevelopmental and facial phenotype in individuals with a TRIP12 variant. *Eur. J. Hum. Genet.* **31**, 461–468 (2023).
44. Blackburn, P. R. *et al.* Loss-of-function variants in CUL3 cause a syndromic neurodevelopmental disorder. *medRxiv* (2023) doi:10.1101/2023.06.13.23290941.
45. Oppermann, H. *et al.* CUX1-related neurodevelopmental disorder: deep insights into phenotype-genotype spectrum and underlying pathology. *Eur. J. Hum. Genet.* **31**, 1251–1260 (2023).
46. Blackburn, P. R. *et al.* Loss-of-function variants in*CUL3*cause a syndromic neurodevelopmental disorder. *medRxiv* (2023) doi:10.1101/2023.06.13.23290941.
47. Averdunk, L. *et al.* Biallelic variants in CRIPT cause a Rothmund-Thomson-like syndrome with increased cellular senescence. *Genet. Med.* **25**, 100836 (2023).
48. Oppermann, H. *et al.* CUX1-related neurodevelopmental disorder: deep insights into phenotype-genotype spectrum and underlying pathology. *Eur. J. Hum. Genet.* (2023) doi:10.1038/s41431-023-01445-2.
49. Schmetz, A. *et al.* Delineation of the adult phenotype of Coffin-Siris syndrome in 35 individuals. *Hum. Genet.* **143**, 71–84 (2024).
50. Küry, S. *et al.* Unveiling the crucial neuronal role of the proteasomal ATPase subunit gene PSMC5 in neurodevelopmental proteasomopathies. *medRxiv* (2024) doi:10.1101/2024.01.13.24301174.
51. Li, D. *et al.* Spliceosome malfunction causes neurodevelopmental disorders with overlapping features. *J. Clin. Invest.* **134**, (2024).
52. Rigter, P. M. F. *et al.* Role of CAMK2D in neurodevelopment and associated conditions. *Am. J. Hum. Genet.* **111**, 364–382 (2024).
53. Laugwitz, L. *et al.* ZSCAN10 deficiency causes a neurodevelopmental disorder with characteristic oto-facial malformations. *Brain* (2024) doi:10.1093/brain/awae058.
54. Clark, T., Ciccarese, P. N. & Goble, C. A. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J. Biomed.*

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

*Semantics* **5**, 28 (2014).

55. Raciti, D., Yook, K., Harris, T. W., Schedl, T. & Sternberg, P. W. Micropublication: incentivizing community curation and placing unpublished data into the public domain. *Database* **2018**, (2018).
56. Liu, J. *et al.* Natural History and Real-World Data in Rare Diseases: Applications, Limitations, and Future Perspectives. *J. Clin. Pharmacol.* **62 Suppl 2**, S38–S55 (2022).
57. European Union. Charter of Fundamental Rights of the European Union, 2016. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12016P%2FTXT (2016).
58. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745.
59. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679.
60. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
61. Huang, G. B., Ramesh, M., Berg, T. & Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. http://vis-www.cs.umass.edu/lfw/. (2007).
62. Boyadjiev, S. A. & Jabs, E. W. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin. Genet.* **57**, 253–266 (2000).
63. den Dunnen, J. T. *et al.* HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum. Mutat.* **37**, 564–569 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2023.06.06.23290887; this version posted May 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

64. Stevens-Kroef, M., Simons, A., Rack, K. & Hastings, R. J. Cytogenetic Nomenclature and Reporting. in *Cancer Cytogenetics: Methods and Protocols* (ed. Wan, T. S. K.) 303–309 (Springer New York, New York, NY, 2017).

65. Kaye, J. *et al.* Dynamic consent: a patient interface for twenty-first century research networks. *Eur. J. Hum. Genet.* **23**, 141–146 (2015).

66. Nellåker, C. *et al.* Enabling Global Clinical Collaborations on Identifiable Patient Data: The Minerva Initiative. *Front. Genet.* **10**, 611 (2019).

67. Schoeman, L., Honey, E. M., Malherbe, H. & Coetzee, V. Parents' perspectives on the use of children's facial images for research and diagnosis: a survey. *J. Community Genet.* **13**, 641–654 (2022).

68. Schmidt, A. *et al.* Next-generation phenotyping integrated in a national framework for patients with ultra-rare disorders improves genetic diagnostics and yields new molecular findings. *medRxiv* 2023.04.19.23288824 (2023) doi:10.1101/2023.04.19.23288824.