GestaltMatcher Database - A global reference for the facial phenotypic variability of rare human diseases

Hellen Lesmann[1,2,*], Alexander Hustinx[2,*], Shahida Moosa[3], Elaine Marchi[4], Pilar Caro[5], Ibrahim M. Abdelrazek[6], Jean Tori Pantel[7,8], Hannah Klinkhammer[2,9], Merle ten Hagen[2], Tom Kamphans[10], Wolfgang Meiswinkel[10], Jing-Mei Li[2], Behnam Javanmardi[2], Alexej Knaus[2], Annette Uwineza[11], Cordula Knopp[12], Tinatin Tkemaladze[13,14], Miriam Elbracht[12], Larissa Mattern[12], Rami Abou Jamra[15], Clara Velmans[16], Vincent Strehlow[15], Himanshu Goel[17], Beatriz Carvalho Nunes[18], Thainá Vilella[18], Isabel Furquim Pinheiro[19], Chong Ae Kim[19], Maria Isabel Melaragno[18], Tahsin Stefan Barakat[20], Amira Nabil[6], Julia Suh[21], Luisa Averdunk[22], Ekanem Ekure[23], Claudio Graziano[24], Prasit Phowthongkum[25,26], Nergis Güzel[27], Tobias B. Haack[28], Theresa Brunet[29], Sabine Rudnik-Schöneborn[30], Konrad Platzer[15], Artem Borovikov[31], Franziska Schnabel[15], Lara Heuft[15], Vera Herrmann[15], Antonio F. Martinez-Monseny[32], Matthias Höller[33], Khoshoua Alaaeldin[6], Aleksandra Jezela-Stanek[34], Amal Mohamed[6], Amaia Lasa-Aranzasti[35,36], John A. Sayer[37,38], Ping Hu[39], Suzanna E. Ledgister Hanchard[40], Gehad Elmakkawy[6], Sylvia Safwat[6], Frédéric Ebstein[41,42], Elke Krüger[43], Sébastien Küry[41,42], Annabelle Arlt[2], Felix Marbach[5], Christian Netzer[16], Sophia Kaptain[2], Hannah Weiland[2], Dong Li[44], Lucie Dupuis[45], Roberto Mendoza-Londono[45], Sofia Douzgou Houge[46], Denisa Weis[47], Brian Hon-Yin Chung[48,49], Christopher C.Y. Mak[49], Koen Devriendt[50], Karen W. Gripp[51], Martin Mücke[7,8], Alain Verloes[52], Christian P. Schaaf[5], Christoffer Nellåker[53], Benjamin D. Solomon[54], Rebekah L. Waikel[54], Markus M. Nöthen[1], Ebtesam Abdalla[6], Gholson J. Lyon[55,56,57], Peter M. Krawitz[2], Tzung-Chien Hsieh[2,#]

[1]Institute of Human Genetics, University of Bonn, Bonn, Germany
[2]Institute for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, Germany
[3]Division of Molecular Biology and Human Genetics, Stellenbosch University and Medical Genetics, Tygerberg Hospital, Stellenbosch, South Africa
[4]New York State Institute for Basic Research in Developmental Disabilities, New York State, Albany, USA
[5]Institute of Human Genetics, Heidelberg University, Heidelberg, Germany

[6]Department of Human Genetics, Medical Research Institute, Alexandria University, Alexandria, Egypt
[7]Institute for Digitalization and General Medicine, University Hospital RWTH Aachen, Aachen, Germany
[8]Centre for Rare Diseases Aachen (ZSEA), University Hospital RWTH Aachen, Aachen, Germany
[9]Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany
[10]GeneTalk GmbH, Bonn, Germany
[11]College of Medicine and Health Sciences, University of Rwanda, and University Teaching Hospital of Kigali, Kigali, Rwanda
[12]Institute for Human Genetics and Genomic Medicine, Medical Faculty, RWTH Aachen University, Aachen, Germany
[13]Department of Molecular and Medical Genetics, Tbilisi State Medical University, Tbilisi, Georgia
[14]Givi Zhvania Pediatric Academic Clinic, Tbilisi State Medical University, Georgia
[15]Institute of Human Genetics, University of Leipzig Medical Center, Leipzig, Germany
[16]Institute of Human Genetics, University of Cologne, Faculty of Medicine and University Hospital Cologne, Cologne, Germany
[17]School of Medicine and Public Health, University of Newcastle, Callaghan NSW, Australia
[18]Genetics Division, Department of Morphology and Genetics, Universidade Federal de São Paulo, São Paulo, Brazil
[19]Genetics Unit, Instituto da Criança, Universidade de São Paulo, São Paulo, Brazil
[20]Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[21]Institute for Human Genetics and Genomic Medicine, Medical Faculty, RWTH Aachen University, Aachen, Germany
[22]Department of Pediatrics, Universtiy Hospital Düsseldorf, Düsseldorf, Germany
[23]Department of Paediatrics, College of Medicine, University of Lagos, Lagos, Nigeria
[24]Medical Genetics Unit, Ausl Romagna, Cesena, Italy
[25]Excellence Center for Genomics and Precision Medicine, King Chulalongkorn

Memorial Hospital, the Thai Red Cross Society, Bangkok, Thailand
[26]Division of Medical Genetics and Genomics, Department of Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
[27]Institut für Humangenetik und Genommedizin, Uniklinik RWTH Aachen, Aachen, Germany
[28]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
[29]Institut für Humangenetik, Klinikum rechts der Isar, Technische Universität München, München, Germany
[30]Institute of Human Genetics, Medical University Innsbruck, Innsbruck, Austria
[31]Research Centre for Medical Genetics (RCMG), Moscow, Russia
[32]Department of Clinical Genetics, SJD Barcelona Children's Hospital, Esplugues del Llobregat (Barcelona), Spain
[33]Institute for Human Genetics, Universitätsklinikum Freiburg, Freiburg, Germany
[34]Department of Genetics and Clinical Immunology, National Institute of Tuberculosis and Lung Diseases, Warsaw, Poland
[35]Medicine Genetics Group, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Hospital Universitari, Barcelona, Spain
[36]Department of Clinical and Molecular Genetics, Vall d'Hebron Barcelona Hospital Campus, Vall d'Hebron Hospital Universitari, Barcelona, Spain
[37]Biosciences Institute, Newcastle University, Central Parkway, Newcastle upon Tyne, UK
[38]Renal Services, The Newcastle Upon Tyne NHS Hospitals Foundation Trust, Freeman Road, Newcastle Upon Tyne, UK
[39]Division of Cancer prevention, National Cancer Institute, Bethesda, USA
[40]Department of Medical Genomics, National Human Genome Research Institute, Bethesda, USA
[41]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[42]Nantes Université, CHU Nantes, Service de Génétique Médicale, F-44000 Nantes, France
[43]Insitute for Medical Biochemistry and Molecular Biology, University of Greifswald, Greifswald, Greifswald, Germany
[44]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, USA

[45]Department to Paediatrics, Division of Clinical and Metabolic Genetics, The Hospital of Sick Children, Toronto, Canada
[46]Department of Medical Genetics, Haukeland University Hospital, Bergen, Norway
[47]Institue for Medical Genetics, Kepler University Hospital, Linz, Austria
[48]Hong Kong Genome Institute, Hong Kong, China
[49]Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong, China
[50]Center for Human Genetics, KU Leuven, Leuven, Belgium
[51]Division of Medical Genetics, A.I. du Pont Hospital for Children/Nemours, USA, Wilmington, USA
[52]Department of Clinical Genetics, Robert-Debré Hospital, Paris, France
[53]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK
[54]Medical Genomics Unit, Medical Genetics Branch, National Human Genome Research Institute, Bethesda, USA
[55]Department of Human Genetics, New York State Institute for Basic Research in Developmental Disabilities, Staten Island, New York, United States of America
[56]George A. Jervis Clinic, New York State Institute for Basic Research in Developmental Disabilities, Staten Island, New York, United States of America
[57]Biology PhD Program, The Graduate Center, The City University of New York, New York, United States of America

[*]These authors contributed equally
[#]Corresponding author

Abstract

Dysmorphologists sometimes encounter challenges in recognizing disorders due to phenotypic variability influenced by factors such as age and ethnicity. Moreover, the performance of Next Generation Phenotyping Tools such as GestaltMatcher is dependent on the diversity of the training set. Therefore, we developed GestaltMatcher Database (GMDB) - a global reference for the phenotypic variability of rare diseases that complies with the FAIR-principles.

We curated dysmorphic patient images and metadata from 2,224 publications, transforming GMDB into an online dynamic case report journal. To encourage clinicians worldwide to contribute, each case can receive a Digital Object Identifier (DOI), making it a citable micro-publication. This resulted in a collection of 2,312 unpublished images, partly with longitudinal data.

We have compiled a collection of 10,189 frontal images from 7,695 patients representing 683 disorders. The web interface enables gene- and phenotype-centered queries for registered users (https://db.gestaltmatcher.org/). Despite the predominant European ancestry of most patients (59%), our global collaborations have facilitated the inclusion of data from frequently underrepresented ethnicities, with 17% Asian, 4% African, and 6% with other ethnic backgrounds. The analysis has revealed a significant enhancement in GestaltMatcher performance across all ethnic groups, incorporating non-European ethnicities, showcasing a remarkable increase in Top-1-Accuracy by 31.56% and Top-5-Accuracy by 12.64%. Importantly, this improvement was achieved without altering the performance metrics for European patients.

GMDB addresses dysmorphology challenges by representing phenotypic variability and including underrepresented groups, enhancing global diagnostic rates and serving as a vital clinician reference database.

Introduction

Facial dysmorphism is often used as a crucial handle in the diagnosis of many rare genetic disorders, occurring in many genetic syndromes[1,2]. The identification of facial dysmorphism and the recognition of dysmorphic patterns as a lead to diagnosis is often challenging and relies on the skills and experience of the examiner. However, the variability of facial features in certain syndromes, especially ultra-rare ones, can pose challenges even for experienced clinicians[3]. Additionally, facial features also vary based on sex, age, and ethnicity, and this can further complicate the recognition of a diagnostic dysmorphic pattern [4–6]. Ethnicity, in particular, plays a significant role, as there is considerable inter-ethnic variability in facial gestalt[7]. Thus, some common facial features in certain ethnic groups may be considered dysmorphic in others. For example, upslanting palpebral fissures are common in healthy Asians but may be perceived as dysmorphic in other populations[8]. Studies have also highlighted

differences in facial gestalt between different ethnicities in frequent 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 themselves, as European clinicians failed to correctly recognize dysmorphic individuals of African ethnicity[11]. This is a particular problem as globalization and migration are increasingly blurring ethnic and cultural boundaries, and geography is no longer a real determining factor in mating patterns[12]. Hence, in diverse populations, such as those with admixed ethnicities, the challenge of accurately diagnosing rare diseases becomes even more pronounced, as new phenotypes can evolve by admixture[13]. Therefore, the need for reference databases to facilitate inter-case variability comparisons.

Due to these diagnostic difficulties, there is increasing use of next-generation phenotyping (NGP) technology, which offers the automated analysis of frontal images of patients to suggest suspected genetic diagnoses based on patterns of facial dysmorphism[14–16]. Among them is GestaltMatcher, a clustering approach that also supports the prediction of ultra-rare disorders and the matching of undiagnosed patients based on facial phenotype. The scientific validity and technical performance of the GestaltMatcher Artificial Intelligence (AI) has already been demonstrated by compelling evidence of the system's reliability and accuracy[16,17]. Moreover, the clinical utility of GestaltMatcher has been illustrated through individual cases[18–20] as well as through the German, large-scale, national study TNAMSE (*Translate-Nationales Aktionsbündnis Seltene Erkrankungen*). In the TNAMSE study, the prospective evaluation of the GestaltMatcher AI within the national healthcare system revealed a significant improvement in the diagnostic yield of exome sequencing[21].

Despite the increasing interest and technological advances in NGP, properly labeled training data is still the biggest bottleneck in developing NGP applications[22]. Furthermore, the existing data are often siloed, so curation is usually done repeatedly[23]. NGP encounters challenges similar to those of human assessors, with factors like age and ethnicity affecting its performance, as it depends on the quality, diversity and curation of the training data. However, despite legislative efforts such as the National Institute of Health (NIH) Revitalization Act of 1993[24], which mandates the inclusion of women and minorities in research studies, non-European ethnicities

remain underrepresented in literature and genetic disease studies. In 2021, nearly 86% of participants in genome-wide association studies (GWAS) worldwide were of European descent, despite constituting just 16% of the global population[25]. This disparity extends to diagnostic morphological atlases, where non-European populations are notably underrepresented[26]. In contrast, despite a global decline, birth rates remain highest in developing countries. For instance, in 2021, the average number of births per woman was 1.5 in Europe and North America, compared to 4.6 in sub-Saharan Africa[27] (Figure 1).

Hence, it is essential for both clinicians and computer scientists to have access to a reference database that encompasses diverse data. Achieving this diversity can be facilitated through global collaboration and crowd-sourced labeling, enabling the collection of a wide range of data from various ethnicities and populations. However, data deposition and curation are time-consuming tasks that depend on the motivation of the single scientist[28]. Presumably, one major reason is the lack of sufficient benefit from a submission, as there is no comparable recognition to a scientific publication[29]. To make submissions to a database more attractive, micro- or nanopublications can serve as an incentive, as it has already been shown that they contribute significantly to increasing the data submission rate[30,31]. Micropublications are concise scientific statements that can be simple assertions with attributions or comprehensive arguments supported by evidence. They are flexible in format, allowing for minimal presentation with basic provenance or maximal knowledgebases with evidence graphs, applicable across biomedical communication[32].

Therefore, we developed GestaltMatcher Database (GMDB), serving as a global repository for the phenotypic variability of rare human diseases while at the same time providing machine-readable medical image data for NGP. GMDB is the first image database for NGP that complies with the FAIR principles, making data findable, accessible, interoperable, and reusable[33]. It aims to shorten the time to diagnosis in rare disorders with facial dysmorphism by offering users a reference for facial variability in genetic dysmorphic disorders, as well as by improving the performance of NGP by offering the largest data set of its kind. Each case uploaded to GMDB can be transformed into a citable micro-publication with a Digital Object Identifier (DOI), promoting data sharing among collaborators worldwide. Moreover, GMDB can

function as an image repository for preprint servers or journals that prohibit image publication[34].

Ethnic diversity is crucial for the database, not only for referencing by clinicians but also for training the GestaltMatcher AI. Although it has demonstrated promising performance in a Nigerian cohort of Cornelia de Lange patients[35], the performance of the GestaltMatcher AI has not yet been comprehensively evaluated across diverse populations on a broader scale.

Results:

Overview of FAIR data in GMDB

The GMDB aims to meet the needs of clinicians for a dysmorphology reference database and a more modern publication medium while achieving data curation with crowd-sourced labeling for machine learning at the same time. Retrospective data from publications curated by our data curators, along with retrospective and prospective data provided by clinicians or patients are made available as FAIR cases in the GMDB (Figure 2a). At the time of publication, we were able to collect a total of 10,789 portrait images of 7,695 patients with 683 genetic disorders. This includes a considerable number of images collected from scientific publications, as our curators were able to gather images from 2,224 scientific publications. Additionally, our global collaboration involved 55 clinicians contributing a valuable set of 2,312 unpublished images, enhancing the diversity and depth of our database.

An entry in GMDB consists of a medical image such as a portrait, X-ray, or fundoscopy, as well as demographic data like sex, age and ethnicity and machine-readable meta information such as the diagnosed disorder (OMIM index), disease-causing mutation reported in Human Genome Variation Society format (HGVS) or International System for Human Cytogenomic Nomenclature (ISCN) nomenclature with test method and zygosity and the clinical feature encoded in HPO terminology[36–39] (Figure 2b). Clinicians are also asked to provide their expert opinion about 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[16].

GMDB is a modern publication medium for citable micro-publication case reports

As an incentive for clinicians to share their patients' data and thereby build a more extensive and diverse dataset, the GMDB allows clinicians to assign a DOI to their cases, transforming them into a citable micro-publication. This approach has enabled us to publish 2,312 images of cases in our database over 12 months through global collaborations that have not been published elsewhere.

As an example of a published case with a DOI assigned, we describe a case of admixed ancestry (Gambian and European) that presents as an atypical case of Noonan syndrome with multiple lentigines with atypical facial dysmorphisms (hypotelorism and upslanting palpebral fissures instead of hypertelorism and downslanting palpebral fissures), which excluded this diagnosis by the assessing clinician only for it to be later established by exome sequencing. Interestingly, the analysis with GestaltMatcher AI was nevertheless able to identify a match with another Noonan syndrome patient as the 12th most similar patient. We uploaded this case to the GMDB as a FAIR case and assigned it a DOI to share with the scientific community (Figure 3a). We added metadata, multimodal data (profile images, skin, hand, and feet), and longitudinal data to this case (Figure 3b). The DOI transforms the case into a micro-publication. For this purpose, the case is also assigned a title and abstract. This information (title, abstract, DOI) is displayed on the landing page of a case (Supplementary Figure 1) and is available publicly without access to the GMDB. To protect the patient’s data, phenotype and image data are only accessible after registration to the GMDB. For registered users, the whole case can be seen, including all FAIR data and metadata (Supplementary Figure 2). As the case is a FAIR case, it can also be seen in the Gallery view (Figure 3c) and can be used for cohort analysis (Figure 3d).

Diverse ancestry of GestaltMatcher training set

In most healthcare datasets, non-European ethnicities are notably underrepresented[25]. This contrasts with migration, globalization, and significantly higher birth rates in these regions, where 80 % of the world's population lives and 90 % of births take place[53] (Figure 1a). When using NGP tools, this imbalance can lead to an ethnic bias. Therefore, it is a central objective of GMDB to ensure the representation of all ethnic

groups and to establish a training dataset for the GestaltMatcher AI that is as diverse as possible. Despite the persisting skewed representation, with a significantly higher proportion of patients of European descent, our international collaborations (Figure 1a - location marker) have notably increased the representation of non-European ethnic groups[26]. Specifically, we have increased the representation to 17% for Asian and 4% for African ethnicities in the FAIR dataset (Figure 1b).

Diverse ancestry data enhance prediction accuracy for underrepresented populations

The scientific and clinical efficacy of the GestaltMatcher AI has been substantiated in prior studies[16,17,21]. Nonetheless, specific assessments of its performance across diverse ethnic groups have been limited. While an analysis of GestaltMatcher's performance in a Nigerian cohort has been conducted[35], a comprehensive evaluation across various populations on a broader scale has not yet been performed.

To investigate the impact of incorporating ethnically diverse training data on the overall performance of GestaltMatcher across ethnic groups on a large scale, we set up an experiment where a subset of 2,625 images of European patients was extended by images of A) more European patients, or B) patients with other ethnic backgrounds. For (A), we train our model on a subset of GMDB containing 3,843 images of patients with European ancestry only. While in (B), we used a subset containing 3,843 images of patients with other ethnic backgrounds.

We measured a top-1 accuracy averaged over all ethnic groups of 42% and 59% for the European and diverse sets, respectively, and a top-5 accuracy of 57% and 74% for the European and diverse sets, respectively. Notably, the evaluation performance on images of patients with European ancestry is only marginally different (3% and 4% for top-1 and top-5, respectively) while nearly doubling the amount of images of Europeans. Meanwhile, the top-1 and top-5 performance increases significantly for almost every other ethnic group. Figure 5 shows further per-ethnicity performance.

FAIR AI-Training set

Data sharing is essential to drive scientific progress in the development of AI. FAIR data sets enable the reproducibility and transparency of these scientific studies. For

this reason, we also share the FAIR image data and metadata in the GMDB with other computer scientists to train their AI. The continuous expansion of the database has not only enabled an improved version of the GestaltMatcher AI to be optimized[17] but also enabled further AI projects of other research groups to be carried out[17,41,42]. Furthermore, it is possible to run the facial analysis on a local machine using the FAIR data set. A protocol for this has already been published, describing how the Gestalt scores can ultimately be used to prioritize genomic variants[43].

Empowering clinicians: Image search and visualization for rare disorder learning

The recognition of facial dysmorphic patterns in human rare disease relies heavily on comparison with individuals with a genetic confirmation of the diagnosis. A literature search is often time-consuming, and clinical facial images are often siloed or behind paywalls often inaccessible to users from developing countries. The "Gallery view" feature of GMDB makes it easy to search for the candidate gene or disease and allows for immediate visualization of all relevant portrait images (at-a-glance). Moreover, the user can search for HPO terms or even for the PubMed ID (PMID) and Digital Object Identifier (DOI) of a specific publication. This search enables clinicians to compare the facial dysmorphic pattern of a patient with the images available in the GMDB and get an overview of the heterogeneity of the diagnostic facial characteristics of many disorders (inter-case variability). Since several images for a case can be uploaded to the database (e.g., multimodal or longitudinal data), intra-case variability (e.g. over different ages) can also be depicted.

Facilitating NGS analysis with the GestaltMatcher API and Research Platform

The GestaltMatcher database also includes an Application Programming Interface (API) for the GestaltMatcher AI[16,17]. This disorder prediction tool can be helpful for clinicians considering differential diagnoses, through the suggestion of syndromes and the GestaltMatcher score (https://api.gestaltmatcher.org). E.g. Brand et al. describe how the results of the analysis with the GestaltMatcher disorder prediction tool helped to solve a typical phenotype of Koolen-de Vries syndrome (KdVS) with an unusual disease-causing mutation, by revealing a high gestalt-score for KdVS[19]. Moreover, the direct matching of affected individuals via facial similarity is possible. Thus, not only matches with already solved cases can lead to a diagnosis. For instance, Marbach et

al. describe two patients with a previously unknown genetic disorder caused by the same *de novo* mutation in *LEMD2*. GestaltMatcher was used to demonstrate the similarity of the two cases to each other, which supported the assumption that they represented a new phenotype[18].

Beyond the function of the disorder prediction tool, entire patient cohorts can also be analyzed within the Research platform in the GMDB. This research platform can, therefore, meet the known needs of the research community in genetics. It is possible to quantify the similarity of the individual patients in the cohort by generating a similarity matrix. This approach can detect clusters and assess whether, for example, cases with an identical variant or pathogenic variants in the same gene, cluster together. 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[34]. Deploying this method within our research platform has facilitated the quantification of similarity across 18 cohorts, with 15 already published[34,44–57].

Discussion

GMDB serves as a modern, searchable reference and publication medium, accessible to clinicians and researchers globally while simultaneously facilitating the compilation and retrieval of labeled data for deep learning in NGP through crowd-sourced labeling. The ultimate goal is to drive research in rare genetic disorders and shorten the time to diagnosis.

Due to the variability of facial phenotypes influenced, e.g., by age, gender, and ethnicity, clinicians find value in reference image databases. While a great effort has already been made to create an atlas addressing the ethical diversity issue, it is still limited to very few disorders[53]. In contrast, GMDB provides clinicians with a comprehensive selection of patient images of different ethnicities at a glance, eliminating the need for extensive literature searches. In addition, GMDB's gallery view allows for easy comparison of phenotypes and can serve as a valuable teaching tool for training students and residents to recognize disorders due to facial features.

GMDB's dynamic character sets it apart from traditional journals, allowing cases to be updated following patient consultations or new findings. This flexibility is crucial, as

symptoms of disorders may develop over time, and facial phenotypes can change with age[58]. Unlike static case reports, GMDB enables the storage of longitudinal data, including patient images at different ages, which can help to better recognize a patient's dysmorphism at any age.

Additionally, GMDB's extensive collection of facial images makes it a unique FAIR database in size. This was mainly possible through the numerous contributions and crowd-sourced labeling from collaborators worldwide. To further increase the motivation of data submission in the future, every case in the database can potentially become a citable micro-publication with a DOI[32]. Furthermore, in the future, our micro-publications could be indexed in reputable scientific indexing services such as PubMed, following the example of some existing micro-publication communication platforms[59]. Active patient involvement, with the ability to access and upload their data, enhances patient autonomy and facilitates the acquisition of longitudinal patient data, even further enriching GMDB's repository of facial images.

The GestaltMatcher API and research platform in GMDB aim to shorten the time to diagnosis by providing suggestions for underlying disorders, facilitating direct patient matching and cohort analysis to see whether the same or a related genetic mechanism underlies the disorder. Tools for matching patients on genotype level with sequencing data already exist (e.g. GeneMatcher[60]). They are connected through the MatchMaker Exchange Network API[61]. As the GestaltMatcher API not only allows disorders to be attributed, but also identifies the patients who are most similar to the analyzed patient regardless of the diagnosis, this means that undiagnosed patients can also be matched. Therefore, GMDB can also be seen as a photo version of GeneMatcher and will also become a node of the MME network**.**

Even in the era of ‘genotype-first’ diagnostic approaches, accurate or ‘deep phenotyping’ is crucial in classifying genomic variants, good phenotyping (*‘deep phenotyping’*) is still crucial and improves variant filtering[62,63]. It is possible to link the GMDB with, for example, PEDIA, an AI-based approach that uses portrait images to interpret clinical exome data, improving the performance of bioinformatics pipelines for exome analysis[64]. The Gestalt scores could even already be incorporated into variant classification[2].

The GMDB not only supports the use of GestaltMatcher AI within the database, but also offers an enormous AI training set. The most crucial advantage of GMDB is transparency. All FAIR data is available and can be shared, whereas the quality and quantity of DeepGestalt's dataset is unknown[65]. However, this is important for evaluating the results, as syndromes not present in the data set cannot be supported by DeepGestalt[65,66]. Influencing factors such as gender, age, and ethnicity cannot be assessed.

Ethnicity has a significant impact on the detection of rare dysmorphic disorders for both clinicians and AI[11], with NGP tools demonstrating high accuracy in patients of European and North American ancestry, on which they are mainly trained and validated on. However, their performance in populations with different ancestry has not been adequately studied, and the influence of ancestry on facial phenotypes associated with genetic disorders is not well understood[67]. The analyses of Face2Gene's performance across different populations revealed varying degrees of success in identifying congenital dysmorphic syndromes. Mishima et al. demonstrated that Face2Gene correctly identified the syndrome with a Top-10 accuracy of 86% in a Japanese cohort[65]. Narayanan et al. found that Face2Gene predicted the correct diagnosis in 70% (Top-10 accuracy) of Indian children with recognizable facial dysmorphism [68].

Conversely, Elmas & Gogus reported a lower success rate of only 48% in a cohort of Turkish patients[69]. Also, Hennocq et al.[70] highlighted the inability of Face2Gene to classify a patient of African ethnicity with Kabuki syndrome, emphasizing the importance of encouraging international collaborations to improve the performance of next-generation phenotyping tools. Some authors describe higher accuracy rates due to more tailored training datasets, e.g. in Thai and Italian cohorts[8,71]. Additionally, Lumaka et al. demonstrated that augmenting datasets with individuals of the same ethnicity can significantly improve performance (+ 57,9 % for Down Syndrome)[11].

The performance of GestaltMatcher across different ethnicities has not yet been evaluated on a large scale. For this reason, we investigated how the top-1 and top-5 accuracy for the different ethnic groups changes when equally sized groups of European or non-European patients are added to the training set. Overall, the top-5 accuracy for most individual non-European ethnic groups increases significantly when

extending our training set with non-Europeans (~12%). The European group's performance only changed marginally when extending the training data to Europeans or non-Europeans. Due to the low number of images of patients belonging to the less frequent ethnic groups in our dataset, simply averaging over all ethnic groups may lead to an incomplete representation of the results.

There are several limitations to this approach. The curation of data by different clinicians and affected individuals introduces an individual variability that is difficult to quantify or investigate and may affect data quality. A standardized portrait image without irritating confounding factors, such as different facial expressions, camera angles, or even patients' items such as glasses, can lead to a distortion of the information for the AI. Moreover, it is not possible to form a completely balanced data set because the disorders vary so much in prevalence.

Even though we have already been able to significantly diversify our data set through the annotations of international collaborations, it is still heavily dominated by European cases. However, the ratio of Asian individuals in GMDB is significantly higher than e.g. in gnomAD (South Asian: 5,65 %), and the ratio of African ethnicities is comparable to that in gnomAD (5,65 %) (https://gnomad.broadinstitute.org/stats). Nonetheless, our dataset still does not reflect the diversity of non-European ethnicities yet, particularly concerning the African continent, which harbors the most significant genetic variability due to historical migration and admixture[74]. The labeling of ethnicity can also lead to problems and influence how ethnicity is perceived. Therefore, Foster and Sharp emphasize the need to demonstrate scientific utility for such classifications, a goal we pursue with our approach[75]. Therefore, guidelines were proposed that also entail ethical and privacy concerns that we adhere to[26].

Conclusion

Overall, using advanced technologies such as NGP can be essential to rationalize the diagnostic process, especially in regions with limited resources and access to appropriate medical care[35]. Biomedical data curation is the best approach for sharing, managing, integrating, and analyzing existing and emerging datasets, even if costly and time-consuming[59]. The exponential growth of images and cases in the GMDB highlights its ability to foster global collaboration and combat data siloing, ensuring that

data remains accessible, reusable, and interoperable for both human and machine learning.

Methods

Implementation of the online GestaltMatcher Database platform

We first built an online platform with Ruby on Rails to allow users to input images and other patient data. For the back end, we set up a database by MySQL to store the patient data.

Data curation

The data curated can be roughly categorized into retrospective and prospective data. Retrospective data primarily refers to data collected through curating data from the literature or similar projects that obtain global consent to share the data (e.g. Minerva&Me[23]). For cases curated from the literature, we collected the DOI and PMID as well as the contact details of the corresponding author. We then clarify whether reuse is possible while respecting intellectual property rights. Our collaboration partners, clinicians around the world, also recruited patients with an established diagnosis within their clinical practice. Patients from patient support groups were also included after informed consent was obtained. Patients can upload images or laboratory findings themselves following an invitation from the informing physician. By prospective data we mean the further collection of data over time. This can be done by the clinicians who see their patients in the course of another patient consultation, or by the patients themselves, who can use their access to upload further images or metadata throughout their clinical pathway.

The curation process can be roughly subdivided into three phases. First, we started having medical students annotate cases from the literature, mainly by searching Pubmed and Google Scholar for publications with images of patients with facial dysmorphism and monogenic molecular diagnosis.

Second, we started to recruit solved patients from patient support groups. As we aimed to develop a patient-centered platform and strengthen patient autonomy, we collected feedback from the recruited patients during this phase to provide patients with a user-friendly experience. Patients are allowed to upload images and findings autonomously and access their data at any time.

To facilitate the retrospective recruitment of patients, we have also implemented digital consent forms, which allow patients to decide under which conditions they consent to store their data in the database and enable direct signature online. We also further developed this feature in close cooperation with patient support groups, e.g. the German Smith-Magenis Syndrome patient organization Sirius e.V., to cover the patients' requests precisely. Patients can access their own case and provide or withdraw their consent online. They can also upload images themselves, which greatly simplifies the curation process of longitudinal data. The fact that documents (doctor's letters or laboratory results) can also be uploaded (only 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 a way that it can in principle also be used as a dynamic consent model in the future[76]. The consent form is available in German and English, other languages will be incorporated soon.

In the last phase, we expanded our database through international collaborations with clinicians from different continents. Initially, we also focused on the patients who had already been solved but had not yet been published to improve the AI's performance. However, as we progressed, more clinicians shared their unsolved cases with the scientific community. GMDB started focusing on facial portraits of patients with rare monogenic diseases and is currently mainly populated by those cases, but not limited. Later in the curation process, we also annotated cytogenetic disorders with facial dysmorphism. In addition to these clinicians, we also employ paid curators who continue to annotate data from the literature.

FAIR data set

The GMDB aims to support clinicians in their work as dysmorphologists while advancing science in the field of AI and preventing data siloing. For this reason, we have compiled a FAIR data set of all patients in our database who have consented to public use in the online platform and sharing data with other AI research groups.

GestaltMatcher training set

In addition to the FAIR cases, we utilize a distinct set of cases specifically for training the GestaltMatcher AI. These cases involve patients who have provided consent solely for the 'private' utilization of their data. Consequently, while the case remains visible

only to the uploading clinician, it is inaccessible to other users of the online platform and cannot be disseminated to external groups. It is, therefore, a data set that is used exclusively for further training of the GestaltMatcher AI in order to improve its performance further.

Data Governance and ELSI

The GMDB is only accessible to the scientific community. To protect the patient data, the database underlies strict access control. Registration is only possible after receiving an invitation link from an existing user.

During the process of obtaining informed consent, patients can decide whether their data should be part of the FAIR data set or the private GestaltMatcher training set. Patients who do not wish to make their data publicly available to scientists and clinicians in the database can also upload their data 'privately'. This means that their data will only be used to train the GestaltMatcher AI and will not be accessible to GMDB online platform users or other researcher groups (Supplementary Figure 3). It is also possible to upload only individual images or documents 'privately'. In our GM consent, patients can also indicate whether they agree to the use of the images for presentations or teaching, or to publication in other journals. This differentiation from other journals is important, as patients/parents show less willingness to consent for publication in open-access journals than for access-controlled databases that are not publicly accessible[77].

The download of GMDB FAIR data for the development of NGP approaches can be made possible for scientists in the field of AI. Therefore, IRB approval and submission of a proposal is required. Additionally, signing of the General Data Protection Regulation (GDPR) consent form is required. The Advisory Board, consisting of the esteemed co-authors: 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, will conduct a thorough review of all applications. Access will be granted to applicants within 2 to 3 weeks if a majority of the members of the Board of Directors are in favour of the application.

The GestaltMatcher Database (GMDB) is hosted physically in the University Hospital of Bonn and guarded 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 Eva-Luise und Horst Köhler Stiftung and Wirtgen Stiftung.

Digital Object Identifier (DOI) assignment

Upon submission of data, the respective case will be promptly published on the website. After submission of the data, 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 from the case[32]. To do this, clinicians must upload the required data and metadata, enter their own personal identifier (e.g. ORCID), specify the other persons involved and write a title and an abstract. This process will adhere to a rigorous review similar to Raciti, Daniela et al.[59], ensuring the credibility and reliability of the published data. The DOIs are created and managed by the Bonn University and State Library by using the DataCite 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 1). The landing page is accessible via the generated DOI, also for individuals without access to GMDB or those who are not logged in. It contains the full citation with the DOI as a link, the abstract, and a description of the case data. Phenotypic information is not available, neither HPO terms nor images. However, it is indicated how many images the micropublication contains. After logging in, the data can be fully accessed.

Ethnicity Analysis

The genetic ancestry of each individual was documented as fine-grained as possible by self-reported data. E.g. if an individual was born in Germany and all its grandparents originated from this country we assigned this individual to Germany (country) and Europe (continent). Likewise, for all individuals without migration history in the previous generations, for individuals with mixed ancestry, e.g. a father from Gambia and a mother from Eastern Europe, we assigned European-African mixed ancestry. To investigate the impact of incorporating ethnically diverse training data on the overall performance of GestaltMatcher across ethnic groups on a large scale, we set up an experiment where a subset of images of European patients was extended by images

of A) more European patients, or B) patients with other ethnic backgrounds. For (A), we train our model on a subset of GMDB containing 3,843 images of patients with European ancestry only. While for (B), we used a subset containing 3,843 images of patients with any ethnic backgrounds. The experiment was set up to ensure the same training data distribution of disorders was maintained.

Further, to improve comparability between subsets (A) and (B), subset (B) contains 2,625 images of Europeans sampled from (A). For this purpose, we used the architecture and training process described by Hustinx et al.[19], using a single improved resnet-50 model pre-trained on the GLINT360K face verification dataset using ArcFace loss. The model was fine-tuned for 50 epochs on subsets (A) and (B) of GMDB (v1.0.9) mentioned earlier. All other hyperparameters were left unchanged, using the Adam optimizer, cross-entropy loss, and class weighting to deal with the imbalance in data availability between disorders. Each model was trained five times with different weight initializations, and their results were averaged to obtain a more robust representation of the performance. It is important to note that the model was not tasked with learning to classify the ethnicity, only with learning to classify the disorder.

After training, the models' performance was measured on the same evaluation set, containing images of patients with diverse ethnic backgrounds. This evaluation set consists of 649 images, and was sampled such that there is no overlap between patients or images in any subset. The performance metric used is the top-*n* accuracy. Top-1 indicates the disorder was correctly classified as the first guess, while top-5 indicates the correct class was in the first five guesses. To address the imbalance between ethnic group frequencies, the accuracy was averaged over each ethnic group, rather than each image. As such, the performance of any infrequent group weighs equally with that of the more frequent groups.

Advisory Board

All the applications for acquiring access to download GMDB public data for developing NGP approach will be reviewed by the advisory board consisting of the following coauthors: 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. Once the majority of the board agrees with the application, the applicant will be granted download access.

Acknowledgment

This research was supported [in part] by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. We thank the Asia Pacific Society of Human Genetics for their support. Sofia Douzgou Houge is supported by grant #43066 of the Norwegian National Advisory Unit on Rare Disorders. Tahsin Stefan Barakat was supported by the Netherlands Organisation for Scientific Research (ZonMw Vidi, grant 09150172110002).

Figures

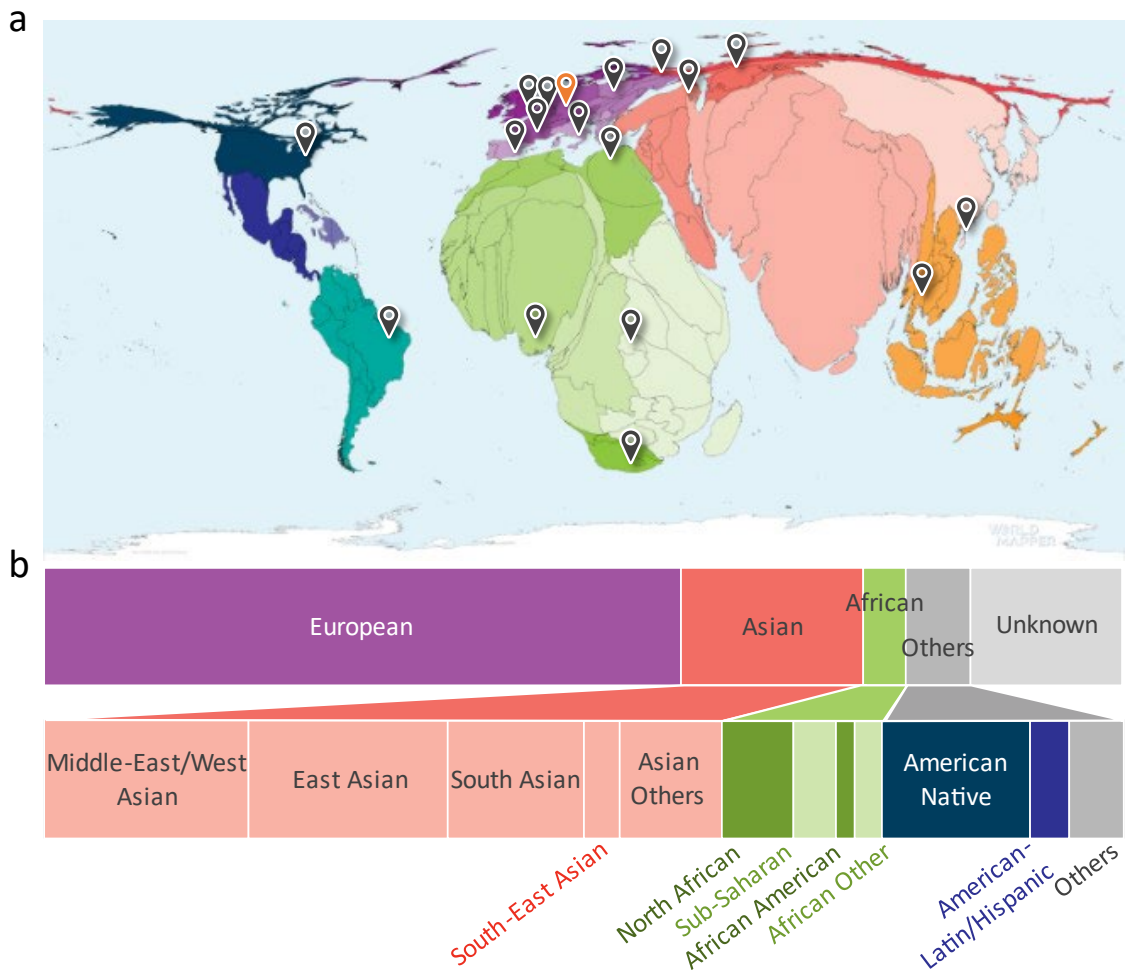

Figure 1: a) World birth rate distribution. The size of country is scaled by the birth rate. **b)** The distribution of the ancestry groups in GMDB.

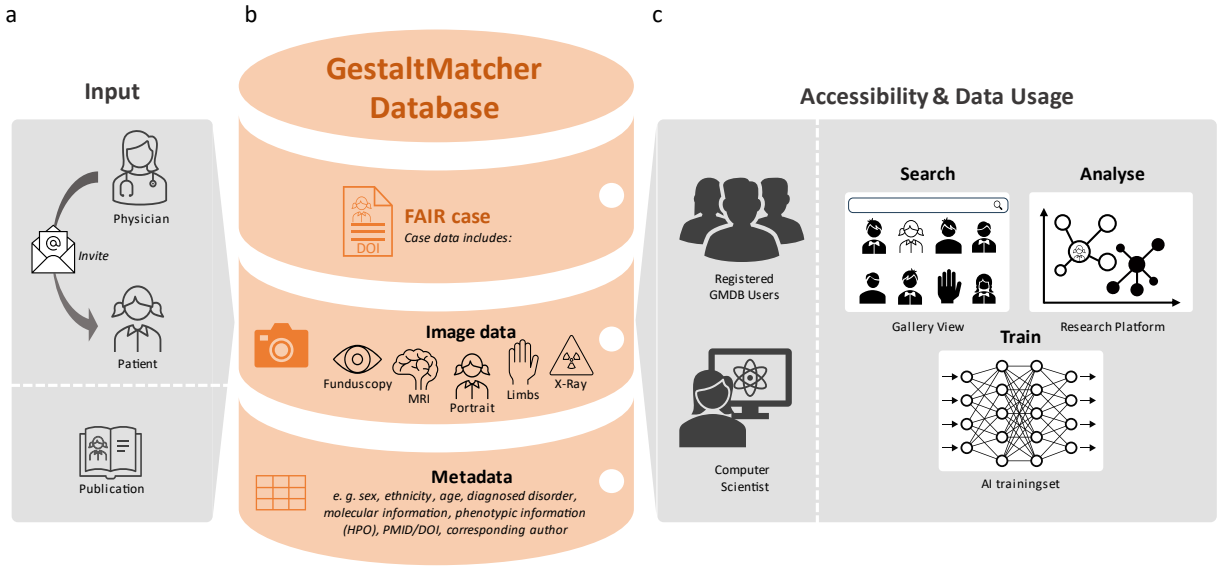

Figure 2: GestaltMatcher Database Architecture and Dataflow. a) The retrospective data is collected from the literature and annotated by data curators, or it is uploaded by the attending clinicians. Patients can also upload images of their own cases, incorporating prospective data, and view their own data at any time. **b)** The data (multimodal image data, including portrait images as well as MRI, X-ray, funduscopy and extremity images) are stored in the GMDB (MySQL database) together with the relevant meta information (such as sex, age, ethnicity, molecular and phenotypic information). **c)** The FAIR data can be viewed and searched in the GMDB by registered users in the Gallery. They can also be analyzed using the Next-Generation Phenotyping tool GestaltMatcher within the Research Platform. In addition, after a confirmed application process, computer scientists can also use the data set for training purposes for their projects.

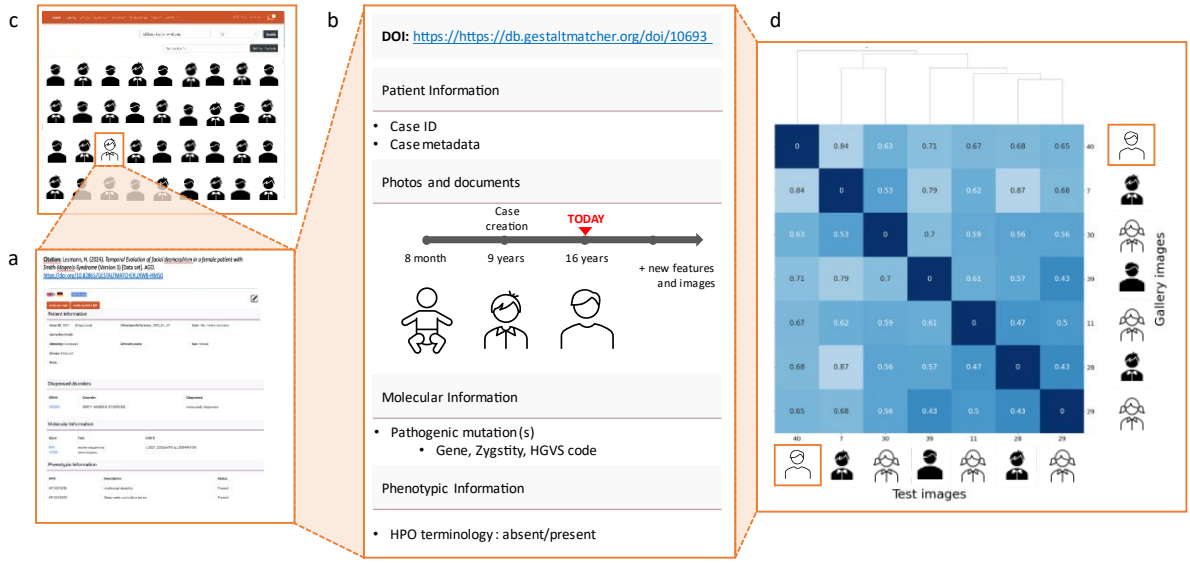

Figure 3: Case presentation of a FAIR case with a Digital Object Identifier (DOI). **a)** A FAIR case in the GestaltMatcher Database (GMDB) is displayed to GMDB users via the data sheet. Each FAIR case can also be assigned a DOI so that it becomes a citable micro-publication. **b)** Included in this micro-publication will be the image data and metadata including demographic, molecular and phenotype information. The dynamic nature of the GMDB case reports enables longitudinal image data storage even after initial publication, which is not possible in conventional journals. **c)** After uploading, case reports can be viewed and searched by other users in the Gallery. **d)** The image data can also be used for inter-cohort comparisons of the gestalt scores within the research platform.

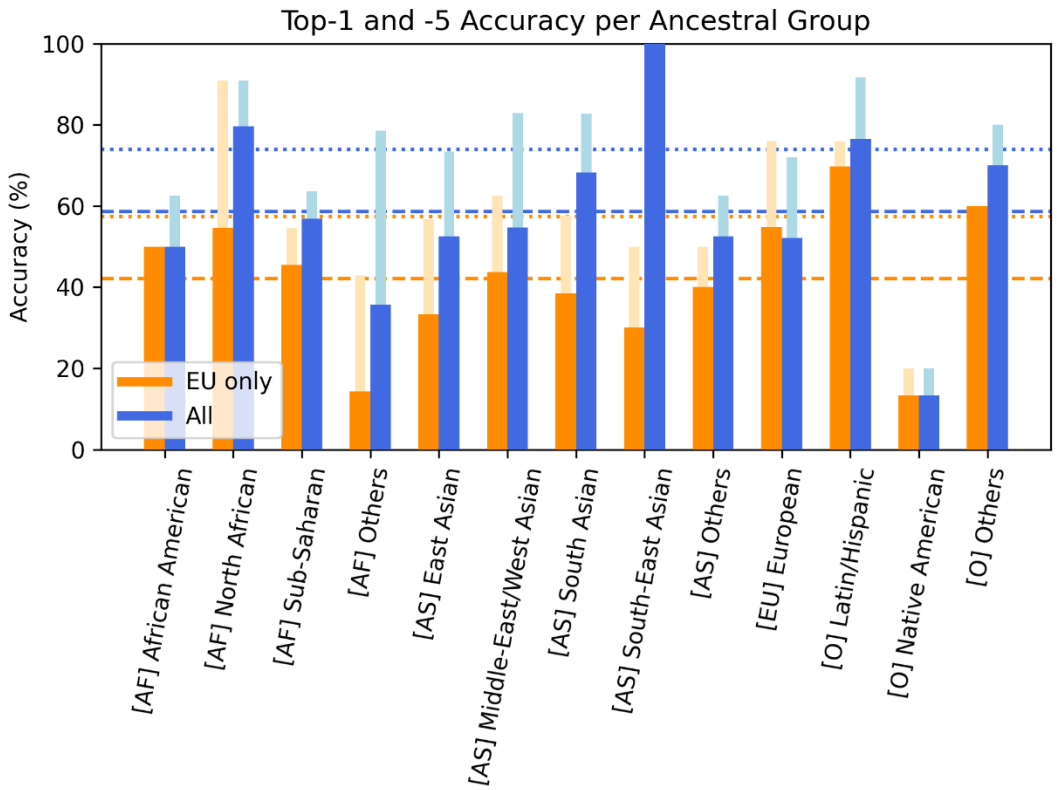

Figure 4: Top-1 and top-5 accuracy of GestaltMatchers' disorder classification accuracy per ethnic group. Top-1 and top-5 accuracy of our models' disorder classification accuracy per ethnic group, where (orange) belongs to the EU only subset, and (blue) belongs to the diverse subset. Each wide, darker bar and each light, thinner bar indicate the top-1 and top-5 accuracy per ethnic group, respectively. The horizontal dashed lines and dotted lines indicate the top-1 and top-5 overall accuracy averaged over all ethnic groups, respectively.

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, 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. 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).
15. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).

16. Hsieh, T.-C. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat. Genet.* **54**, 349–357 (2022).
17. Hustinx, A. *et al.* Improving Deep Facial Phenotyping for Ultra-rare Disorder Verification Using Model Ensembles. in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 5007–5017 (IEEE, 2023).
18. Marbach, F. *et al.* The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping. *Am. J. Hum. Genet.* **104**, 749–757 (2019).
19. Brand, F. *et al.* Next-generation phenotyping contributing to the identification of a 4.7 kb deletion in KANSL1 causing Koolen-de Vries syndrome. *Hum. Mutat.* **43**, 1659–1665 (2022).
20. Forwood, C. *et al.* Integration of EpiSign, facial phenotyping, and likelihood ratio interpretation of clinical abnormalities in the re-classification of an ARID1B missense variant. *Am. J. Med. Genet. C Semin. Med. Genet.* (2023) doi:10.1002/ajmg.c.32056.
21. 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.
22. Hennekam, R. C. M. & Biesecker, L. G. Next-generation sequencing demands next-generation phenotyping. *Hum. Mutat.* **33**, 884–886 (2012).
23. Nellåker, C. *et al.* Enabling Global Clinical Collaborations on Identifiable Patient Data: The Minerva Initiative. *Front. Genet.* **10**, 611 (2019).
24. Institute of Medicine (US) Committee on Ethical and Legal Issues Relating to the Inclusion of Women, Mastroianni, A. C., Faden, R. & Federman, D. *NIH Revitalization Act of 1993 Public Law 103-43*. (National Academies Press (US), 1994).
25. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
26. Koretzky, M. *et al.* Towards a more representative morphology: clinical and ethical considerations for including diverse populations in diagnostic genetic atlases. *Genet.*

Med. **18**, 1069–1074 (2016).

27. UNITED NATIONS DEPARTMENT FOR ECONOMIC AND SOCIAL AFFAIRS. *World Population Prospects 2022: Summary of Results*. (United Nations Fund for Population Activities, 2023).
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 hemoglobinopathies using the microattribution approach. *Nat. Genet.* **43**, 295–301 (2011).
32. Clark, T., Ciccarese, P. N. & Goble, C. A. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *J. Biomed. Semantics* **5**, 28 (2014).
33. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
34. Ebstein, F. *et al.* PSMC3 proteasome subunit variants are associated with neurodevelopmental delay and type I interferon production. *Sci. Transl. Med.* **15**, eabo3189 (2023).
35. Arlt, A. *et al.* Next-generation phenotyping in Nigerian children with Cornelia de Lange Syndrome. *medRxiv* 2024.02.15.24302695 (2024) doi:10.1101/2024.02.15.24302695.
36. 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).
37. den Dunnen, J. T. *et al.* HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum. Mutat.* **37**, 564–569 (2016).

38. 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, 2017).
39. 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).
40. Muenke, M., Adeyemo, A. & Kruszka, P. An electronic atlas of human malformation syndromes in diverse populations. *Genet. Med.* **18**, 1085–1087 (2016).
41. 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).
42. Wu, D. *et al.* Multimodal Machine Learning Combining Facial Images and Clinical Texts Improves Diagnosis of Rare Genetic Diseases. *arXiv [q-bio.QM]* (2023).
43. 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).
44. 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).
45. Kampmeier, A. *et al.* PHIP-associated Chung-Jansen syndrome: Report of 23 new individuals. *Front Cell Dev Biol* **10**, 1020609 (2022).
46. Lyon, G. J. *et al.* Expanding the phenotypic spectrum of NAA10-related neurodevelopmental syndrome and NAA15-related neurodevelopmental syndrome. *Eur. J. Hum. Genet.* **31**, 824–833 (2023).
47. Aerden, M. *et al.* The neurodevelopmental and facial phenotype in individuals with a TRIP12 variant. *Eur. J. Hum. Genet.* **31**, 461–468 (2023).
48. 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.
49. 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).

50. 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.
51. Averdunk, L. *et al.* Biallelic variants in CRIPT cause a Rothmund-Thomson-like syndrome with increased cellular senescence. *Genet. Med.* **25**, 100836 (2023).
52. 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.
53. Schmetz, A. *et al.* Delineation of the adult phenotype of Coffin-Siris syndrome in 35 individuals. *Hum. Genet.* **143**, 71–84 (2024).
54. 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.
55. Li, D. *et al.* Spliceosome malfunction causes neurodevelopmental disorders with overlapping features. *J. Clin. Invest.* **134**, (2024).
56. Rigter, P. M. F. *et al.* Role of CAMK2D in neurodevelopment and associated conditions. *Am. J. Hum. Genet.* **111**, 364–382 (2024).
57. Laugwitz, L. *et al.* ZSCAN10 deficiency causes a neurodevelopmental disorder with characteristic oto-facial malformations. *Brain* (2024) doi:10.1093/brain/awae058.
58. Bongers, E. M. *et al.* Meier-Gorlin syndrome: report of eight additional cases and review. *Am. J. Med. Genet.* **102**, 115–124 (2001).
59. 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).
60. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum. Mutat.* **36**, 928–930 (2015).
61. Philippakis, A. A. *et al.* The Matchmaker Exchange: a platform for rare disease gene

discovery. *Hum. Mutat.* **36**, 915–921 (2015).

62. Vissers, L. E. L. M. & Veltman, J. A. Standardized phenotyping enhances Mendelian disease gene identification. *Nature genetics* vol. 47 1222–1224 (2015).
63. Solomon, B. D. *et al.* Perspectives on the future of dysmorphology. *Am. J. Med. Genet. A* (2022) doi:10.1002/ajmg.a.63060.
64. Hsieh, T.-C. *et al.* PEDIA: prioritization of exome data by image analysis. *Genet. Med.* **21**, 2807–2814 (2019).
65. 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).
66. Marwaha, A., Chitayat, D., Meyn, M. S., Mendoza-Londono, R. & Chad, L. The point-of-care use of a facial phenotyping tool in the genetics clinic: Enhancing diagnosis and education with machine learning. *Am. J. Med. Genet. A* **185**, 1151–1158 (2021).
67. Echeverry-Quiceno, L. M. *et al.* Population-specific facial traits and diagnosis accuracy of genetic and rare diseases in an admixed Colombian population. *Sci. Rep.* **13**, 6869 (2023).
68. Narayanan, D. L. *et al.* Computer-aided Facial Analysis in Diagnosing Dysmorphic Syndromes in Indian Children. *Indian Pediatr.* **56**, 1017–1019 (2019).
69. 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).
70. Hennocq, Q. *et al.* Next generation phenotyping for diagnosis and phenotype-genotype correlations in Kabuki syndrome. *Sci. Rep.* **14**, 2330 (2024).
71. Carrer, A. *et al.* Application of the Face2Gene tool in an Italian dysmorphological pediatric clinic: Retrospective validation and future perspectives. *Am. J. Med. Genet. A* **194**, e63459 (2024).
72. Porras, A. R. *et al.* Facial analysis technology for the detection of Down syndrome in the Democratic Republic of the Congo. *Eur. J. Med. Genet.* **64**, 104267 (2021).

73. 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* **3**, e635–e643 (2021).
74. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
75. Foster, M. W. & Sharp, R. R. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res.* **12**, 844–850 (2002).
76. Kaye, J. *et al.* Dynamic consent: a patient interface for twenty-first century research networks. *Eur. J. Hum. Genet.* **23**, 141–146 (2015).
77. 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).