Published in final edited form as:

Nat Genet. 2023 September ; 55(9): 1598–1607. doi:10.1038/s41588-023-01469-w.

PhenoScore: AI-based phenomics to quantify rare disease and genetic variation

Alexander J M Dingemans[1,2], **Max Hinne**[2], **Kim M G Truijen**[1], **Lia Goltstein**[1], **Jeroen van Reeuwijk**[1], **Nicole de Leeuw**[1], **Janneke Schuurs-Hoeijmakers**[1], **Rolph Pfundt**[1], **Illja J Diets**[1], **Joery den Hoed**[3], **Elke de Boer**[1], **Jet Coenen-van der Spek**[1], **Sandra Jansen**[4], **Bregje W van Bon**[1], **Noraly Jonis**[1], **Charlotte W Ockeloen**[1], **Anneke T Vulto-van Silfhout**[1], **Tjitske Kleefstra**[1], **David A Koolen**[1], **Philippe M Campeau**[5], **Elizabeth E Palmer**[6,7], **Hilde Van Esch**[8], **Gholson J Lyon**[9,10], **Fowzan S Alkuraya**[11], **Anita Rauch**[12], **Ronit Marom**[13], **Diana Baralle**[14], **Pleuntje J van der Sluijs**[15], **Gijs W E Santen**[15], **R Frank Kooy**[16], **Marcel A J van Gerven**[2], **Lisenka E L M Vissers**[1,*], **Bert B A de Vries**[1,*]

[1]Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6500 HB, P.O. Box 9101, Nijmegen, the Netherlands.

[2]Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Thomas van Aquinostraat 4, 6525 GD Nijmegen, the Netherlands

[3]Language and Genetics Department, Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, the Netherlands

[4]Department of Human Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

[5]Department of Pediatrics, University of Montreal, Montreal, QC, Canada.

[6]Faculty of Medicine and Health, UNSW Sydney, Sydney, Australia

[7]Sydney Children's Hospitals Network, Sydney, Australia

[8]Center for Human Genetics, University Hospitals Leuven, University of Leuven, Leuven, Belgium.

[9]Department of Human Genetics and George A. Jervis Clinic, Institute for Basic Research in Developmental Disabilities (IBR), Staten Island, New York, USA

[*]corresponding authors (both authors contributed equally): Lisenka.Vissers@radboudumc.nl, Bert.deVries@radboudumc.nl.

Author information

Conceptualization: A.J.M.D, M.H, L.E.L.M.V, B.B.A.d.V, M.A.J.v.G; Data curation: A.J.M.D, K.M.G.T, L.G, J.v.R, N.d.L, J.S.H, R.P, I.J.D, E.d.B, J.d.H, J.v.d.S, S.J, B.W.v.B, N.J, E.E.P., P.M.C, A.T.V.v.S, T.K, D.A.K, F.K, H.V.E, G.J.L, F.S.A, A.R, R.M, D.B, P.J.v.d.S., G.S, L.E.L.M.V, B.B.A.d.V; Formal Analysis: A.J.M.D, M.H; Funding acquisition: L.E.L.M.V, B.B.A.d.V; Investigation: A.J.M.D, M.H; Modelling: A.J.M.D, M.H.; Software development: A.J.M.D; Writing – original draft: A.J.M.D, M.H, L.E.L.M.V, B.B.A.d.V, M.A.J.v.G; Writing – review and editing: all authors.

Statement of conflict of interest: there is no conflict of interest.

Ethics declaration

In this study, data from the Biobank 'Intellectual Disability', which is part of the Radboud Biobank initiative (for more information, see [79] or https://www.radboudumc.nl/en/research/radboud-technology-centers/radboud-biobank) were used. Within this biobank, phenotypic and molecular data have been systematically captured for individuals with (non-)syndromic ID referred to the Radboud university medical center. The use of this dataset was approved by the ethical committee of the Radboud university medical center (#2020–6151 and #2020–7142). Furthermore, the authors declare no competing interests.

[10]Biology PhD Program, The Graduate Center, The City University of New York, New York, United States of America

[11]Department of Translational Genomics, Center for Genomic Medicine, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

[12]Institute of Medical Genetics, University of Zurich, Schlieren, 8952, Zurich, Switzerland.

[13]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.

[14]Faculty of Medicine, University of Southampton, University Rd, Southampton, SO17 1BJ, UK

[15]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands

[16]Department of Medical Genetics, University of Antwerp, Antwerp, Belgium.

Abstract

Several molecular and phenotypic algorithms exist that establish genotype-phenotype correlations, including facial recognition tools. However, no unified framework that investigates both facial- and phenotypic data directly from individuals exists. We developed PhenoScore: an open source, artificial intelligence-based phenomics framework, combining facial recognition technology with Human Phenotype Ontology data analysis to quantify phenotypic similarity. We prove PhenoScore's ability to recognize distinct phenotypic entities by establishing recognizable phenotypes for 37 of 40 investigated syndromes against clinical features observed in individuals with other neurodevelopmental disorders and show it is a significant improvement on existing approaches. PhenoScore provides predictions for individuals with variants of unknown significance and enables sophisticated genotype-phenotype studies by testing hypotheses on possible phenotypic (sub)groups. PhenoScore confirmed previously known phenotypic subgroups caused by variants in the same gene for *SATB1*, *SETBP1*, and *DEAF1* and provides objective clinical evidence for two distinct *ADNP*-related phenotypes, already established functionally.

Keywords

artificial intelligence; variant of unknown significance; rare disease; personalized medicine; facial recognition; deep phenotyping

1 Introduction

A significant portion of individuals with clinically and genetically heterogeneous rare diseases, such as neurodevelopmental disorders (NDD), has been molecularly diagnosed in the last decade using whole-exome sequencing (WES) [1, 2, 3, 4]. Clinical WES data interpretation relies on filtering and prioritization for rare genetic variants in disease-gene panels, which are subsequently interpreted in the context of the patient's clinical presentation [5]. Whereas this strategy is essential to identify the disease-causing variant(s), it is estimated that, depending on the number of genes included in the panel, dozens of variants are prioritized as diagnostic noise [6] — and this number is expected to rise even

more in the coming years with technological innovations, such as genome sequencing, finding their way into the diagnostic arena [7, 8, 9].

At the molecular level, several computational methods, such as MutationTaster [10], PolyPhen [11], SIFT [12], CADD score [13], have been designed to predict the deleteriousness of variants to effectively prioritize causal variants. These tools use diverse approaches, such as looking at the impact of the variant on protein structure (MutationTaster, PolyPhen), taking conservation into account (MutationTaster, PolyPhen, SIFT) —– or trying to incorporate multiple sources of genomic information (CADD score). At the phenotypic level, headway has been made by introducing Human Phenotype Ontology (HPO), systematically capturing the presence of features observed in individuals with rare diseases [14]. However, equivalent to molecular tools, algorithms using these HPO data to quantify phenotypic HPO similarity between individuals with genetic disorders would provide significant benefits to diagnose rare disease. Such a quantitative phenotypic score could for instance assist with the interpretation of genetic variants of unknown clinical significance (VUS), which constitute 10–30% of all variants clinically assessed [4, 15]. Reducing the number of VUSs is of essence since studies have shown that not all individuals and families respond similarly to the result of a VUS test-result, and usually do not fully comprehend its meaning [16, 17], potentially leading to frustration, and/or distress due to the uncertainty involving a possible diagnosis and course of disease. Importantly, VUSs have also been shown to inflict inappropriate medical decisions [18, 19].

Next to reclassifying VUSs, quantifying phenotypic HPO similarity at the cohort level could also help to provide further steps towards personalized medicine by automatically recognizing distinct phenotypic subtypes leading to more tailored clinical prognosis [20, 21, 22].

A branch of science that could assist in objectively quantifying phenotypic data is artificial intelligence (AI). AI has dramatically reformed the manner clinical data are processed and analyzed in recent years, with the AI revolution in medicine starting in pathology and radiology [23, 24, 25, 26]. In genetics, these new techniques have been employed in assisted interpretation of genomic variants [27, 28, 29] and combining molecular and phenotypic evaluations, mainly looking at methods to use phenotypic data in HPO to automatically prioritize genetic variants [30, 31, 32, 33, 34, 35, 36]. Furthermore, advances in computer vision have led to the application of facial recognition technology in clinical genetics [37, 38, 39, 40, 41, 42], with the current state-of-the-art application GestaltMatcher achieving a top-10 accuracy of 64% [42]. Facial recognition can assist in the recognition of (neuro)developmental syndromes, since the development of the brain and facial shape are closely linked [43, 44, 45, 46] — and therefore, it comes as no surprise that a significant part of genetic disorders have distinct facial features [47]. However, not all genetic syndromes have a clear, recognizable facial gestalt, which hinders methods solely looking at facial features. Moreover, a syndromic phenotype often includes more than ‘just the face’. Whereas tools have previously looked at either combining molecular data with either HPO, or alternatively, with facial features [1, 39], an important area has been left unexplored, which combines the facial- and HPO data into an AI-framework to predict phenotypic similarities without the need for genomic data input. Therefore, we developed PhenoScore: a

next-generation open-source phenomics framework combining facial recognition technology with clinical features, quantitatively collected in Human Phenotype Ontology (HPO) from deep phenotyping.

2 Results

2.1 The PhenoScore framework

PhenoScore is a framework that currently consists of two modules: a component that extracts the facial features from a 2D facial photograph and a second module that takes HPO-based phenotypic similarity into account (Figure 1). The AI-based framework joins these results in three outputs: a Brier score and corresponding *p*-value, defining the individual's clinical similarity to the syndrome assessed; a facial heatmap, highlighting important facial features for the syndrome; and, a visualization of the most important other (non-facial) clinical features. In the training phase of PhenoScore, first an age-, sex-, ethnicity- matched control for every individual with the genetic syndrome of interest is sampled from our in-house database of 1,200 individuals with neurodevelopmental disorders. Next, the facial features are automatically extracted from the facial photographs for both affected individuals and controls and the phenotypic HPO similarity is calculated (with several HPO terms and their child terms first removed from the dataset, as these are either facial HPO terms to be processed by the facial recognition module, or HPO terms that are deemed subjective and therefore at risk for interobserver variability). A support vector machine (SVM), a widely used classification algorithm in machine learning, is trained on these features, resulting in a trained classifier that can be used to generate a score for individuals, suspected to have the syndrome of interest. If we are interested in quantifying phenotypic (sub)groups, a permutation test is added during the training phase, determining whether the trained classifier performs better than random chance — providing evidence whether the two groups are distinguishable by PhenoScore. Finally, to provide insight into what PhenoScore is doing and to learn more about the investigated syndromes, explainable AI is incorporated into PhenoScore as well, using Local Interpretable Model-agnostic Explanations (LIME) [48, 49]. LIME works by generating random perturbed input data and inspecting the change in predictions, thereby obtaining data on the relative importance of each feature. By using LIME for both the facial- and HPO data, PhenoScore can generate facial heatmaps and visualizations on the most important clinical features.

2.2 Proof-of-Concept using PhenoScore for Koolen-de Vries syndrome

First, we investigated whether using our combined PhenoScore was actually an improvement on solely using either facial- or phenotypic data. The SVM was trained on both separate feature sets alone (e.g. HPO and facial features) and subsequently compared with the classification performance of PhenoScore. To measure classification performance, the Brier score [50] was chosen as the performance measure to focus on: it is defined as the mean squared difference between predicted outcome and observed actual outcome (lower is better). Next to that, we also report the area under the receiving operator curve (AUC; higher is better).

To demonstrate the power of the PhenoScore framework, we first performed a proof-of-concept study using 63 individuals with Koolen-de Vries syndrome (KdVS, OMIM #610443, Figure 2), caused by either proven pathogenic loss-of-function variants in *KANSL1* (n=11) or the 17q21.31 microdeletion (n=52). KdVS most prominent features reported in literature include hypotonia, intellectual disability, and joint laxity [51, 52, 53], for which the interdependence in our modelling is preserved using the graph structure of the HPO terms (Figure 2). Running PhenoScore on the 63 individuals with KdVS, we confirm the improvement on overall predictive performance when using both facial and clinical features compared to using either one alone (Brier score 0.09 or AUC 0.94 for PhenoScore, in contrast to 0.13/0.91 when using only facial data and 0.10/0.92 when using only phenotypic data, Table 1).

We next randomly excluded four individuals (facial images shown in Figure 2) from the training dataset and retrained PhenoSscore, allowing us to test the performance of PhenoScore when treating them as if diagnoses of KdVS were unknown. We then used PhenoScore to predict the similarity of these four individuals when comparing them with 59 remaining individuals with KdVS in the training set. PhenoScore output was displayed using LIME, providing heatmaps of prioritized facial information according to PhenoScore (Figure 2). In addition, the most important clinical features according to PhenoScore to be predictive for KdVS were summarized by numerically scoring and ranking them. According to PhenoScore, the nose and eyes are the most important facial parts when recognizing KdVS — while the presence of hypotonia, moderate intellectual disability, EEG abnormalities, strabismus, pes planus and motor delay are the clinical features of interest. This is completely consistent with expert opinion and the literature [51, 52, 53] and shows that the prediction is based on the extracted facial features from 2D photos and phenotypic data in HPO — harnessing the power of both and outperforming the separate predictions.

2.3 Expanding PhenoScore to 40 syndromes

After our proof-of-concept using KdVS, we next assessed the performance of PhenoScore for the classification of other genetic syndromes too. Hereto, we selected 39 further syndromes (Table 1 and Supplemental Table 1) including both clinically well-recognizable syndromes based on facial gestalt, such as Kleefstra syndrome (OMIM #610253, caused by pathogenic variants in *EHMT1*), Helsmoortel-van der Aa syndrome (OMIM #615873, caused by pathogenic variants in *ADNP*) and Coffin-Siris syndrome (OMIM #135900, *ARID1B*), but also more recently identified syndromes for which facial gestalt is less prominent, including IDDAM (OMIM #615032, *CHD8*) and IDDFBA (OMIM #618089, *FBXO11*).

Analyzing all these syndromes, we demonstrate that PhenoScore is a statistically significant improvement on using either feature set alone, and therefore, the whole is more than the sum of its parts in this case (median Brier score 0.24 for facial features on the whole dataset, 0.14 for HPO data and 0.13 for PhenoScore, p <0.001; median AUC 0.58 for facial features, 0.89 for HPO data and 0.91 for PhenoScore, p <0.001, Table 1). Furthermore, our post-hoc checks show that there was no overfitting using the internal control dataset (Supplemental Table 2 and Supplemental methods). To compare the performance of PhenoScore to other

approaches, we generated predictions for all individuals with a genetic syndrome in the data set using Phenomizer [30, 54] and LIRICAL [36]. Phenomizer correctly included the correct diagnosis in its output in 29% of the individuals and LIRICAL in 39%, while PhenoScore did so in 84% of individuals (p <0.001 for both, Supplemental Figure 1, Supplemental Table 3).

For 37 of 40 syndromes (93%), PhenoScore was able to identify predictive features that characterized these syndromes and recognized a distinct phenotypic entity (Table 1, Supplemental Figure 2). As expected, and visualized in the LIME heatmaps (Figure 3), these features corresponded remarkably well with those described in the literature. For instance, for Helsmoortel-van der Aa syndrome (*ADNP*, OMIM #615873), the eye- and forehead regions are prioritized in the predictions, as seen in the generated heatmap (Figure 3) — corresponding with the known dysmorphic characteristics for this syndrome.

Moreover, for a genetic syndrome which lacks explicit facial features, like IDDAM (caused by pathogenic variants in *CHD8*), apparent overgrowth symptoms, such as macrocephaly and tall stature, were identified as significant predictors, while no relevant facial features were extracted, as displayed in the heatmap and summarized ranking scores. A similar case is made for the genetic disorder associated with pathogenic variants in *DYRK1A*: while the classifier based only on the facial features does not provide any meaningful predictions, the addition of other phenotypic data in HPO did allow PhenoScore to distinguish this syndrome as a phenotypic entity. These data suggest that PhenoScore objectively extracts, distinguishes, and visualizes the specific clinical features for genetic syndromes and highlights that the addition of non-facial phenotypic data in HPO is essential.

Finally, we demonstrate that the performance of PhenoScore is stable over different age- and population of origin subgroups (Supplemental Table 4), by evaluating the predictive performance using the predictions of all individuals included in this study when divided into subgroups based on their age (<two years old, between two and six, between six and 12, 12 and 17, and 18 and older) and population of origin. While the performance is slightly inferior for the included adults (a Brier score of 0.13), there seems to be no clear difference for the other groups (Brier scores between 0.09 and 0.12, p=0.38). Although only 10% of individuals included in this study being are of non-Caucasian/non-Western descent, the subgroups for the population of origin analysis do not seem lead to overt differences in predictive performance between ethnicities.

2.4 PhenoScore is scalable as it requires only a low number of individuals for training

Most genetic disorders are individually rare, with sometimes only 3–5 individuals reported world-wide. We therefore next investigated how many data sets PhenoScore requires for accurate classification of a specific syndrome. We checked the performance of PhenoScore while increasing the number of individuals in the complete dataset of 40 genetic syndromes with the combination of facial- and HPO features, starting with only 2 individuals. This analysis revealed that, with five individuals to train on, the median classification performance for the investigated syndromes is already clinically acceptable (AUC 0.80; Figure 4). The classification performance can be further improved when the training sets

increase in size (median AUC 0.89 with 10 individuals, while with 20 individuals the median AUC is 0.92).

2.5 Use case 1: Objective clinical quantification for the interpretation of molecular VUS

To display the power of PhenoScore in the clinical interpretation of variants at an individual level, we reassessed reported VUSs (ACMG class 3) in the Radboudumc department of Human Genetics. These individuals were not included in the training of PhenoScore and can therefore be considers real out-of-sample cases. In total, we identified 22 individuals in whom a class 3 variant was reported in either of 16 of the 40 syndromes (Supplemental Table 5). PhenoScores were calculated, and when using thresholds of ≤0.30 (for 'no phenotypic match') and ≥0.70 (for 'phenotypic match'), PhenoScore was able to classify 13/22 (59%) of the cases as either match (n=3) or no match (n=10). The other 9 cases had an inconclusive PhenoScore result (scores >0.30 but <0.70). Interestingly, for 9/13 cases for which PhenoScore was conclusive, the clinician made a decision for the VUS based on the phenotype — PhenoScore was essential for the other four cases.

For most VUSs, pathogenicity during clinical follow-up was not clear at the time of writing, but for six individuals, additional (genetic) testing has led to a change in pathogenicity class. Two variants in *ARID1B* were both regarded as benign: one after methylation analysis (negative), the other variant since the individual was diagnosed with fragile X syndrome at a later stage. PhenoScore agrees with both assessments with a low prediction probability of phenotypic similarity (0.03 for both). Next to that, a splice variant in *CHD8* with a high PhenoScore of 0.93 was deemed pathogenic after RNA analysis was performed. Finally, a variant in *EHMT1* was deemed pathogenic after methylation analysis. This is the only variant in which PhenoScore disagrees with the outcome of a functional test, with a low score (0.04) — probably due to the phenotype not particularly matching. Furthermore, for two variants in *SMARCC2*, PhenoScore is inconclusive, while methylation analysis reclassified these variants as benign.

2.6 Use case 2: Next-generation phenomics for the generation of sophisticated genotype-phenotype correlations

Genotype-phenotype studies for rare diseases are often performed to gain insight into the clinical spectrum, which allows clinicians to provide a more accurate counseling of individuals with rare diseases. Molecularly, the toolkit to gain in-depth insight into aspects of pathogenicity is generally applied in a research setting, and thus often not readily available for diagnostic follow-up. From a clinical perspective, analyses are often limited to cluster analysis and/or principle component analysis, but without being able to determine what aspects clinically distinguish subtypes, if identified. We tested whether PhenoScore can improve these hypothesis-driven approaches to distinguish, or discover, clinical subtypes.

For four genes in our dataset, i.e. *ADNP*, *DEAF1*, *SATB1* and *SETBP1*, it has previously been determined that there are (at least) two molecular subtypes. For *SATB1* for instance, it has been acknowledged that individuals with missense variants and those with loss-of-function variants are clinically and molecularly distinctive (OMIM #619228 and #619229). As a proof-of-concept, PhenoScore convincingly distinguished two groups for *SATB1* (Brier

score 0.18, AUC 0.81, $p = 0.02$), confirming the original results [55]. For *DEAF1*, it has been demonstrated that there are two phenotypic entities based on mode of inheritance, with one being being autosomal recessive (OMIM #615828) and the other autosomal dominant (OMIM #617171) [56]. Next to that, genetic variants in *SETBP1* can lead to either Schinzel-Giedeon-Syndrome (OMIM #269150, missense gain-of-function variants) [57] or MRD29 (OMIM #616078, loss-of-function variants leading to haploinsufficiency) [58]. Analysing both these subgroups shows that PhenoScore distinguishes theses groups (for *SETBP1*, Brier score of 0.02 and AUC of 1.0, $p<0.001$; *DEAF1* leads to a Brier of 0.13 and AUC of 0.94 $p<0.001$, Supplemental Figure 5), suggesting that PhenoScore can readily identify clinical entities associated to the same gene.

For *ADNP*, it was recently shown that individuals with pathogenic variants in *ADNP* show one of two distinct methylation signatures (type 2, when variant affects position between c.2000 and c.2340; or type 1, when the variant occurs outside of this interval), suggesting the possibility of two syndromes associated with this gene [59]. Clinically, however, these individuals could not be conclusively distinguished [60]. Prior to determining PhenoScores, we categorized the individuals as having either a type 1 or type 2 *ADNP* signature. Initially, we assessed the performance of PhenoScore using only individuals ($n=33$) for whom both facial photographs and clinical features were available, but failed to identify a statistically significant difference between the groups (Brier 0.30, AUC 0.52, $p = 0.35$). However, using the *ADNP* Human Disease Genes website (https://humandiseasegenes.info/ADNP), we could collect HPO-only data of more individuals. Using this dataset, we obtained clinical features in HPO of 58 individuals (29 in each group), and on these data PhenoScore did show evidence for two phenotypically different entities (Brier 0.24, AUC of 0.71, $p = 0.01$). Inspecting the generated PhenoScore explanations for clinically relevant differences (Figure 5), it seems that recurrent infections and gastrointestinal problems (reflux, constipation, feeding difficulties) are 2–3 times more common in type 2 than in type 1.

Finally, to further explore the classification of VUSs in genetic syndromes that are phenotypically alike (such as the previously named phenotypic subgroups), we generated predictions for each phenotypic subgroup as it were a VUS for the model created for the other phenotypic subgroup of the same gene. For instance, the individuals of the first *ADNP* methylation group were classified using the model that was trained on *ADNP* methylation group 2 individuals and NDD controls. Depending on the similarity in phenotype between the two subgroups, there are no (for *SETBP1*) phenotypic matches, to almost all individuals that are classified as phenotypic matches (for *ADNP*), since these individuals are (much) more phenotypically alike the investigated syndrome than the control population (Supplemental Table 6).

3 Discussion

PhenoScore provides a significant step in the advancement of AI in clinical genetics: a novel machine learning phenomics framework unifying facial and phenotypic features using high-quality data directly from affected individuals instead of generic phenotypic descriptions of a syndrome. Others have introduced AI in this domain of healthcare, with for instance the application of using HPO terms to prioritize genetic variants while comparing individuals to

the known phenotype of disorders in the literature [30, 31, 61, 36]. The utilization of facial recognition technology to assist clinicians in diagnosing individuals has been successful too, with most, unfortunately, relying on proprietary commercial algorithms [37, 38, 39, 40, 35, 41, 42]. We now show a next step, with an open-source framework that takes the complete phenotype into account, including both facial- and phenotypic features directly from affected individuals, and uses AI to provide a score on how well the patient's phenotype (as a whole) matches individuals with a known syndrome.

PhenoScore detected a recognizable phenotype in the large majority of investigated genetic syndromes (37/40; 93%), is a significant improvement over existing algorithms such as Phenomizer and LIRICAL, and only needed as little as five individuals for acceptable classification performance. In this manner, PhenoScore assists clinicians and molecular biologists in quantifying phenotypic similarity, at both an individual- and group level for theoretically all OMIM-listed disorders. One of the disorders for which PhenoScore failed to identify a phenotype was for variants in *ACTL6A*. Interestingly, this is the only of 40 syndromes that has not been recognized by OMIM as a genetic disorder, due to lack of (phenotypic) evidence. For the other two genetic syndromes that PhenoScore failed to identify (MRD29 caused by pathogenic variants in *SETBP1* and MRD56, *CLTC*), some clinical features could be recognized - but apparently not enough to establish a definitive phenotypic entity, probably due to the low number of individuals with these syndromes included. PhenoScore did distinguish MRD56 from Schinzel-Giedeon-Syndrome (both associated with pathogenic variants in *SETBP1*) when compared directly. Apparently, individuals with MRD56 are hard to distinguish from controls with neurodevelopmental disorders — but individuals with Schinzel-Giedeon-Syndrome are phenotypically different from these controls (Figure 3) and therefore PhenoScore is able to differentiate the two phenotypic subgroups in *SETBP1*. Further investigating these phenotypic subgroups and generating predictions for each subgroup with a model that is trained on the other subgroups and controls (Supplemental Table 6) shows that PhenoScore indeed investigates phenotypic similarity. However, this indicates as well that a clinician should be careful in interpreting the results of the VUS prediction if it is possible that the investigated individual has another, but phenotypically similar, disorder than the suspected disorder because of the VUS — as the rate of false positive results could be elevated in that scenario.

Assisting variant classification of VUSs is an obvious use-case for PhenoScore. Of course, several *in vitro* functional assays are available to assess variant pathogenicity, but so far these are mostly used for genes involved in oncogenetic disorders [62, 63]. For neurodevelopmental disorders, these assays are scarce since they need to be developed on a gene-per-gene basis, and for these rare disorders, this is usually not cost-effective and solely done for research purposes. Other methods to assess genetic variants include protein structural analysis [64], which however still relies on the availability of relevant protein structures. Our approach theoretically works for any (genetic) condition with a recognizable phenotype, provided there are sufficient individuals for training the algorithm, and that HPO data and 2D-facial photos are available. Indeed, PhenoScore is as good as its input data. In the field of rare diseases, however, major efforts are put in obtaining these high-quality quantitative phenotypic data, as for instance shown by collections of datasets by the Human Disease Genes website series [65], GeneReviews, DECIPHER and OMIM [66, 67, 68].

Here, the use of HPO terms, in combination with the use of Resnik scores (ensuring that the use of similar HPO terms leads to comparable results), minimizes the effect of variation in clinical terminology used between clinicians, and thus deriving the most benefits from the AI-based quantitative phenotypic framework. Although these measures should minimize any difference in predictive performance when applying PhenoScore in other institutions, further prospective clinical validation studies, preferably in a multi-center prospective design also including institutions from non-Western countries, are needed to confirm this.

PhenoScore also helped to objectively obtain genotype-phenotype correlations, by training on suspected phenotypic subgroups combined by permutation testing to quantify statistical significance. We replicated earlier findings in *SATB1*, *DEAF1* and *SETBP1*, quantitatively underscoring that different molecular mechanisms or inheritance patterns lead to a significantly different, but recognizable, phenotype. Whereas for these genes the associated different phenotypes were also subjectively identifiably from expert opinion, the power of PhenoScore was shown by demonstrating the existence of two distinct phenotypes associated with pathogenic variants in *ADNP*. Molecularly, two different methylation signatures have been published, which were discriminated by the mutation location in *ADNP* [69, 59, 60], but for which clinically, no differences were observed. PhenoScore was not only able to prove the existence of clinically distinctive groups, but also provided insight into which clinical features separates the two clinical entities. For instance, neurodevelopmental problems are more common in the *ADNP*-type 1, while gastrointestinal symptoms, recurrent infections and short stature are 2–3 times more common in *ADNP*-type 2. These clinical features have a significant impact on an individual's quality of life, hence, by identifying these subgroups, PhenoScore directly impacts clinical care, prognosis and recommendations for these individuals and families.

These subgroup analyses could in theory be performed for every (genetic) syndrome caused by different types of SNVs or CNVs — which is the case in a significant portion of the currently ~1600 known NDD genes. While recognizing specific novel subgroups is a first step towards personalized medicine and provides improved clinical prognosis and recommendations (as shown for the subgroups in *ADNP* and *SATB1*), not finding a distinct difference is useful too: it helps to assess whether two types of genetic variation have the same effect (i.e. whether missense variants cause haploinsufficiency). By quantifying the complete phenotypic similarity and visualizing differences between (sub)groups, PhenoScore empowers detailed genotype-phenotype studies, leading to new insights on both the genetic- and phenotypic level.

The discriminating clinical features for the two *ADNP*-related disorders were not represented in a different facial gestalt, emphasizing the importance of adding HPO data across all organ systems. In addition, given that these two phenotypic subgroups were not identified from more subjective clinical analysis, using a predefined structured AI method of phenotypic data analysis provides novel insights. To facilitate easy use in routine clinical care, it is, however, also of paramount importance to be able to intuitively understand the AI output. We therefore also provided graphical output such as facial heatmaps to visualize which (facial) features specified PhenoScore output.

Detailed genotype-phenotype analysis could in theory be performed for every (genetic) syndrome, suggesting that PhenoScore may be a valueable tool to also foster novel molecular insights. That is, for many of the 1,600 known genes associated to an NDD phenotype, multiple types of genetic variants (e.g. SNVs and CNVs) may cause the disorder. Whereas the molecular mechanism for CNVs often relate to dosage-sensitivity, such as haploinsufficiency, the mechanisms for SNVs leading to missense variants in those genes, are often more variable. PhenoScore may assess phenotypic differences between individuals with the same syndrome, but caused by either CNVs ('group 1') or missense variants ('group 2') and help to establish whether those missense variants are also haploinsufficient. Similarly, PhenoScore could be utilized to find phenotypic outliers, of which the molecular mechanism leading to disease might be novel.

In conclusion, PhenoScore bridges a gap between the fields of AI and Clinical Genetics by quantifying phenotypic similarity, assisting not only in genetic variant interpretation, but also facilitating objective genotype-phenotype studies. We showcased its use for individuals with NDD, whose phenotypes were captured using HPO. PhenoScore can, however, also easily be used beyond the field of rare disease, as adjustments to use other (graph-based) ontologies, such as for instance SNOMED [70], can readily be integrated. The PhenoScore AI-based framework is thus easily extended to other domains of (clinical) genetics, or even to completely different branches of medicine, due to its open-source modular design.

4 Materials and Methods

4.1 Inclusion of individuals

The literature was searched for clinical studies which included facial photographs for 40 randomly selected genetic syndromes associated with NDD. The photographs were collected and clinical features, if available, were converted to HPO terms. Currently, PhenoScore is trained using data of 711 non-familial individuals diagnosed with one of the 40 different genetic syndromes, collected from 105 different publications (Table 1 includes the complete overview of the demographics per genetic syndrome and Supplemental Table 1 all publications used as sources for the data used in this study). The phenotypic data were uploaded to the specific gene website in the HDG website series [65] to ensure their public availability. The use of these data was approved by the ethical committee of the Radboud university medical center (#2020–6151 and #2020–7142).

4.2 Data processing

To obtain a representative control group for our machine learning models, for each syndrome with n individuals, n age-, sex- and ethnicity matched controls with a neurodevelopmental disorder seen at our outpatient clinic at the Radboud university medical center were selected as described previously [39] from our internal control database with over 1200 individuals with both facial image and quantitative phenotypic data available (for a complete overview of the workflow of this study, Figure 1). When no matched control was available, that particular individual was excluded from our analysis. Next to that, when individuals were related to each other, one individual was chosen (based on the quality of the picture) from that family.

For each syndrome, nested cross-validation was used to assess the performance of the classifiers. The number of folds during the outer-loop of the nested cross-validation varied due to the considerable variation in dataset size: for every syndrome with at least five individuals, 5-fold cross-validation was used — otherwise, leave-one-out cross-validation was chosen. The hyperparameters of the model (section 4.5) were then tuned during the inner-loop of the nested cross-validation procedure. All performance metrics reported in this study, whether it be AUC, Brier score or accuracy, are calculated based on the predictions during the outer-loop.

As the selection of the randomly selected controls might significantly influence the performance, for each genetic syndrome, different controls were sampled during five random restarts and the mean AUC and Brier scores of these five iterations were noted. Furthermore, to confirm the source of the data did not significantly influence our results, we performed post-hoc analyses by using not only the individuals from our internal control dataset. This included analyses with the other syndromes as controls, but also included additional analyses excluding the Koolen-de Vries individuals who were seen at our clinic at the Radboudumc Nijmegen (see Supplemental methods).

4.3 Extraction of facial features

The facial features were extracted using VGGFace2 [71, 72], a state-of-the-art facial recognition method that utilizes a deep neural network. To avoid overfitting, we did not retrain VGGFace2, but used its pretrained weights instead. The images were then processed by VGGFace2 and the representation in the penultimate layer of the network was obtained. This representation was then used as the facial feature vector. The process was performed as described previously: for the entire (technological) methodology, please see [73].

4.4 Phenotypic similarity

To create a homogeneous dataset, the phenotype of every individual in this study was manually converted into HPO terms [14]. A selection of HPO terms and all their child nodes were removed to eliminate any subjectivity in assessing an individual. These were *Behavioral abnormality (HP:0000708)*, *Abnormality of the face (HP:0000271)*, *Abnormal digit morphology (HP:0011297)*, *Abnormal ear morphology (HP:0031703)*, *Abnormal eye morphology (HP:0012372)*, and every node which is a child node of either of these. We chose these terms as these are either facial features (to be assessed by our facial recognition model) or are suspected to vary across clinicians doing the assessment of an individual. In this manner, 3,810 HPO terms were excluded with 12,259 terms remaining (although it was investigated what the consequences of including all HPO terms were, Supplemental Table 7). To further reduce possible inter-observer variability, the phenotypic similarity between individuals was calculated using the Resnik score [74], since it takes the semantic similarity between symptoms into account. The Resnik score utilizes the information content (IC) of a symptom. In an ontology akin to the HPO, the IC of a specific term can be seen as a measure of the rarity of a term. Naturally, terms closer to the root of the HPO tree have a lower IC. For instance, *Abnormality of the nervous system (HP:0000707)* has an IC of 0.60. In contrast, *Focal impaired awareness motor seizure with dystonia (HP:0032717)*, significantly further down the HPO tree, has an IC of 8.97. This corresponds to our intuition:

rare features provide more information than common features — since the prior probability of an individual reporting a rare symptom is, by definition, smaller. The Resnik score uses this property by defining the similarity between two HPO terms as the IC of their most informative (that is, with the highest IC) common ancestor in the HPO tree. Since terms lower in the tree have a higher IC, the most informative common ancestor corresponds to the last HPO term, which has both compared HPO terms as child nodes when traversing the tree downwards. As an example: for the HPO terms *Reflex seizure (HP:0020207)* and *Focal motor seizure (HP:0011153)*, the most informative common ancestor is *Seizure (HP:0001250)*, which has an IC of 1.70. The Resnik similarity score for *Reflex seizure (HP:0020207)* and *Focal motor seizure (HP:0011153)* is therefore 1.70. Next, we used the best-match average (BMA) to calculate the similarity between two individuals (who usually report multiple HPO terms), in which the average is taken over all best-matched pairwise semantic similarities, as previous studies determined it to be most effective [75]. The idea is similar to that discussed above: if two individuals share a rare symptom *(Focal impaired awareness motor seizure with dystonia (HP:0032717)*, for instance), they are more similar than two individuals who only share a common symptom such as *Abnormality of the nervous system (HP:0000707)*. The Resnik similarity score was calculated for every individual and control and then averaged for both groups. In the end, this led to a nx2 matrix for the HPO features: an average similarity score for each individual versus affected individuals and a score for each individual versus the control group. We calculated the BMA Resnik score between the individuals using the phenopy library in Python 3.9 [76].

4.5 Construction of machine learning model

Finally, the data were used to train a binary classifier. We selected a support vector machine (SVM) as our classifier, known for its excellent overall performance in classification tasks. The SVM was trained using the standard radial basis function kernel and a hyperparameter grid search for C, with values investigated being 1e-5, 1e-3, 1, 1e3 and 1e5. For smaller datasets (less than five individuals), a logistic regression model was chosen, because the SVM does not support probability scores by default and needs an additional internal cross-validation procedure to provide those (further reducing the dataset). All experiments were run on a machine with two graphical processing units (both an NVIDIA RTX2080). It is possible to train PhenoScore on a standard laptop without a designated graphical processing unit, however, if facial heatmaps are required, the process may take several hours per syndrome.

After determining the predictive performance of the model, we determined how many data the classifier needed for an acceptable classification performance in clinical practice. Per syndrome, we started with randomly selecting two individuals and two matched controls, training the model on those, and using the rest of the individuals ($n - 2$, as one individual is used as training data) and matched controls as a test set (two individuals that were not used in the first iteration as the grid search in the SVM classifier needs at least two training samples). We ran five random restarts, randomly selecting another individual and matched control in each iteration. In each restart, cross-validation was employed as in the general training of PhenoScore. The Brier score and AUC were noted and averaged over the five restarts. Next, the size of the training set was increased by one patient, and one matched

control. By increasing the training set by one individual and matched control each time and recording the performance, the model's performance with an increasing number of individuals is assessed (Figure 4).

The Wilcoxon signed-rank test was used to determine statistically significant differences in the performance of the classifiers since it is a non-parametric test and, therefore, suitable — as these data are not normally distributed.

4.6 Explainability of predictions

To see which features contained important information for our model, we generated Local Interpretable Model-agnostic Explanations (LIME) [48, 49]. The main idea of this method is to train a relatively simple local surrogate model to approximate the predictions of the model of interest. Next, the original input data is perturbed, and the corresponding change in predictions is inspected to obtain the relative importance of individual features. A key advantage of LIME is that it is applicable to any model and can therefore be used directly on top of our pipeline.

When using LIME for image data, it is common practice to divide the image into several segments, called superpixels. Therefore, we generated a raster of 25×25 pixel squares for each facial image, randomly offset for each of 100 runs. Each pixel's relative importance was averaged over these runs to obtain a higher resolution visualization of their significance. For the clinical data, the original HPO features were perturbed to obtain the most significant ones in predictions. In this case, LIME uses input data in which some HPO features are added and some are removed from the input data, to see what the effect on the prediction is.

LIME explanations were generated for the individuals with that genetic syndrome and the five highest prediction scores in each iteration of sampling controls, so 25 times in total, for both the facial heatmaps and the phenotypic explanations. These explanations were then averaged, to obtain an overall explanation representative for that specific genetic syndrome. To ensure only real important features were recovered, only HPO terms that were identified in at least 15 individuals (out of 25 in total) were used in this analysis.

4.7 Hypothesis testing

To see whether we could extend the use of our classifier to other applications than the reclassification of VUSs, we designed a random permutation test for the performance of our model. This enables the testing of a specific hypothesis for facial features, phenotypes, or both. An example would be determining whether a newly discovered genetic syndrome consists of several (phenotypic/facial) subtypes. Using our framework, we trained a classifier on the labels of the suspected subgroups. By performing a random permutation test, a p-value is calculated, so that the appearance of the subgroups can be quantified. For a complete overview of the exact methodology of this permutation test, please see the Supplemental Methods.

4.8 Benchmarking PhenoScore

To determine whether our approach is an improvement over existing methods, we used the Phenomizer algorithm [30, 77] and LIRICAL [36] (considered as state-of-the-art [78]) to generate predictions for all individuals with a genetic syndrome in our data set (except for the genetic syndrome associated with *ACTL6A*, as absence of an OMIM number prohibits Phenomizer and LIRICAL to generate predictions). Since Phenomizer does not output a prediction score, but rather a *p*-value, we counted a prediction as positive if the specific genetic syndrome was included in the list of possible diagnoses with an uncorrected *p*-value smaller than 0.05 — otherwise it was seen as a negative prediction. Furthermore, since Phenomizer and LIRICAL do not process facial images, we included the previously excluded HPO terms (Behavioral abnormality (HP:0000708), Abnormality of the face (HP:0000271), Abnormal digit morphology (HP:0011297), Abnormal ear morphology (HP:0031703), Abnormal eye morphology (HP:0012372)) as well as all the corresponding child nodes in the input for Phenomizer and LIRICAL. The number of positive- and negative predictions for Phenomizer (using 0.5 as a cut-off for its predictions), LIRICAL (with a pretest probability of 0.5 to mimic a VUS prediction) and PhenoScore were counted, and a possible statistical significant difference was assessed using a chi-squared test. Other thresholds for the *p*-value of Phenomizer and the scores of LIRICAL and PhenoScore were investigated as well to see the influence on the results (Supplemental Table 3).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to all families and clinicians who agreed to participate and provide clinical and genotypic information. R.F.K acknowledges financial support of the Research Fund of the University of Antwerp (Methusalem-OEC grant – “GENOMED”). The work of G.J.L. is supported by New York State Office for People with Developmental Disabilities (OPWDD) and NIH NIGMS R35-GM-133408. E.E.P is supported by a National Health and Medical Research Council Investigator Grant (Award number: 2021/GNT2008166). Furthermore, we are grateful to the Dutch Organisation for Health Research and Development: ZON-MW grants 912–12-109 (to B.B.A.d.V. and L.E.L.M.V.), Donders Junior researcher grant 2019 (B.B.A.d.V. and L.E.L.M.V.) and Aspasia grant 015.014.066 (to L.E.L.M.V.). The aims of this study contribute to the Solve-RD project (to L.E.L.M.V.), which has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 779257.

Data and code availability

The code of PhenoScore created during this study is freely available at https://github.com/ldingemans/PhenoScore, to enable anyone to apply PhenoScore to their own dataset. Included in PhenoScore are two examples: the data for the *SATB1* subgroups (positive example) and random data (negative example). The used dataset in this study is not publicly available due to both IRB and General Data Protection Regulation (EU GDPR) restrictions since the data might be (partially) traceable. However, access to the data may be requested from the data availability committee by contacting the corresponding author.

References

[1]. Vissers Lisenka E L M et al. “A de novo paradigm for mental retardation”. en. In: Nat. Genet. 42.12 (Dec. 2010), pp. 1109–1112. [PubMed: 21076407]

[2]. de Ligt J et al. “Diagnostic exome sequencing in persons with severe intellectual disability”. In: N. Engl. J. Med. 367.20 (Nov. 2012), pp. 1921–1929. [PubMed: 23033978]

[3]. Rauch Anita et al. “Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study”. en. In: Lancet 380.9854 (Nov. 2012), pp. 1674–1682. [PubMed: 23020937]

[4]. Gilissen Christian et al. “Genome sequencing identifies major causes of severe intellectual disability”. In: Nature 511.7509 (July 2014), pp. 344–347. [PubMed: 24896178]

[5]. Richards Sue et al. “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology”. en. In: Genet. Med. 17.5 (May 2015), pp. 405–424. [PubMed: 25741868]

[6]. Beaumont Robin N and Wright Caroline F. “Estimating diagnostic noise in panel-based genomic analysis”. en. In: Genet. Med. (Aug. 2022).

[7]. McGuire Amy L et al. “The road ahead in genetics and genomics”. en. In: Nat. Rev. Genet. 21.10 (Oct. 2020), pp. 581–596. [PubMed: 32839576]

[8]. Logsdon Glennis A, Vollger Mitchell R, and Eichler Evan E. “Long-read human genome sequencing and its applications”. en. In: Nat. Rev. Genet. 21.10 (Oct. 2020), pp. 597–614. [PubMed: 32504078]

[9]. 100,000 Genomes Project Pilot Investigators et al. “100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report”. en. In: N. Engl. J. Med. 385.20 (Nov. 2021), pp. 1868–1880. [PubMed: 34758253]

[10]. Schwarz Jana Marie et al. “MutationTaster evaluates disease-causing potential of sequence alterations”. en. In: Nat. Methods 7.8 (Aug. 2010), pp. 575–576. [PubMed: 20676075]

[11]. Adzhubei Ivan A et al. “A method and server for predicting damaging missense mutations”. en. In: Nat. Methods 7.4 (Apr. 2010), pp. 248–249. [PubMed: 20354512]

[12]. Kumar Prateek, Henikoff Steven, and Ng Pauline C. “Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm”. en. In: Nat. Protoc. 4.7 (June 2009), pp. 1073–1081. [PubMed: 19561590]

[13]. Kircher Martin et al. “A general framework for estimating the relative pathogenicity of human genetic variants”. en. In: Nat. Genet. 46.3 (Mar. 2014), pp. 310–315. [PubMed: 24487276]

[14]. Robinson Peter N et al. “The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease”. en. In: Am. J. Hum. Genet. 83.5 (Nov. 2008), pp. 610–615. [PubMed: 18950739]

[15]. da Cunha Leite Ana Julia et al. “Diagnostic yield of patients with undiagnosed intellectual disability, global developmental delay and multiples congenital anomalies using karyotype, microarray analysis, whole exome sequencing from Central Brazil”. en. In: PLoS One 17.4 (Apr. 2022), e0266493. [PubMed: 35390071]

[16]. Clift Kristin et al. “Patients’ views on variants of uncertain significance across indications”. In: J. Community Genet. 11.2 (2020), pp. 139–145. [PubMed: 31432391]

[17]. Makhnoon Sukh et al. “Experiences of patients seeking to participate in variant of uncertain significance reclassification research”. en. In: J. Community Genet. 10.2 (Apr. 2019), pp. 189–196. [PubMed: 30027524]

[18]. van Dijk Sandra et al. “Clinical Characteristics Affect the Impact of an Uninformative DNA Test Result: The Course of Worry and Distress Experienced by Women Who Apply for Genetic Testing for Breast Cancer”. In: J. Clin. Oncol. 24.22 (2006), pp. 3672–3677. [PubMed: 16877736]

[19]. Murray Mitzi L et al. “Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions”. en. In: Genet. Med. 13.12 (Dec. 2011), pp. 998–1005. [PubMed: 21811163]

[20]. Hamburg Margaret A and Collins Francis S. “The path to personalized medicine”. en. In: N. Engl. J. Med. 363.4 (July 2010), pp. 301–304. [PubMed: 20551152]

[21]. Ashley Euan A. “Towards precision medicine”. en. In: Nat. Rev. Genet. 17.9 (Aug. 2016), pp. 507–522. [PubMed: 27528417]

[22]. Helen K Brittain Richard Scott, and Thomas Ellen. “The rise of the genome and personalised medicine”. en. In: Clin. Med. 17.6 (Dec. 2017), pp. 545–551.

[23]. Coudray Nicolas et al. “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning”. en. In: Nat. Med. 24.10 (Oct. 2018), pp. 1559–1567. [PubMed: 30224757]

[24]. Hosny Ahmed et al. “Artificial intelligence in radiology”. en. In: Nat. Rev. Cancer 18.8 (Aug. 2018), pp. 500–510. [PubMed: 29777175]

[25]. Killock David. “AI outperforms radiologists in mammographic screening”. en. In: Nat. Rev. Clin. Oncol. 17.3 (Mar. 2020), p. 134.

[26]. Lu Ming Y et al. “AI-based pathology predicts origins for cancers of unknown primary”. en. In: Nature 594.7861 (June 2021), pp. 106–110. [PubMed: 33953404]

[27]. Poplin Ryan et al. “A universal SNP and small-indel variant caller using deep neural networks”. en. In: Nat. Biotechnol. 36.10 (Nov. 2018), pp. 983–987. [PubMed: 30247488]

[28]. Sundaram Laksshman et al. “Predicting the clinical impact of human mutation with deep neural networks”. en. In: Nat. Genet. 50.8 (Aug. 2018), pp. 1161–1170. [PubMed: 30038395]

[29]. Wick Ryan R, Judd Louise M, and Holt Kathryn E. “Performance of neural network base-calling tools for Oxford Nanopore sequencing”. en. In: Genome Biol. 20.1 (June 2019), p. 129. [PubMed: 31234903]

[30]. Köhler Sebastian et al. “Clinical diagnostics in human genetics with semantic similarity searches in ontologies”. en. In: Am. J. Hum. Genet. 85.4 (Oct. 2009), pp. 457–464. [PubMed: 19800049]

[31]. Robinson Peter N et al. “Improved exome prioritization of disease genes through cross-species phenotype comparison”. en. In: Genome Res. 24.2 (Feb. 2014), pp. 340–348. [PubMed: 24162188]

[32]. Zemojtel Tomasz et al. “Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome”. en. In: Sci. Transl. Med. 6.252 (Sept. 2014), 252ra123.

[33]. Smedley Damian and Robinson Peter N. “Phenotype-driven strategies for exome prioritization of human Mendelian disease genes”. en. In: Genome Med. 7.1 (July 2015), p. 81. [PubMed: 26229552]

[34]. Smedley Damian et al. “Next-generation diagnostics and disease-gene discovery with the Exomiser”. en. In: Nat. Protoc. 10.12 (Dec. 2015), pp. 2004–2015. [PubMed: 26562621]

[35]. Hsieh Tzung-Chien et al. “PEDIA: prioritization of exome data by image analysis”. en. In: Genet. Med. 21.12 (Dec. 2019), pp. 2807–2814. [PubMed: 31164752]

[36]. Robinson Peter N et al. “Interpretable Clinical Genomics with a Likelihood Ratio Paradigm”. en. In: Am. J. Hum. Genet. 107.3 (Sept. 2020), pp. 403–417. [PubMed: 32755546]

[37]. Ferry Quentin et al. “Diagnostically relevant facial gestalt information from ordinary photos”. en. In: Elife 3 (June 2014), e02020. [PubMed: 24963138]

[38]. Dudding-Byth Tracy 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”. en. In: BMC Biotechnol. 17.1 (Dec. 2017), p. 90. [PubMed: 29258477]

[39]. van der Donk Roos et al. “Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders”. en. In: Genet. Med. (Dec. 2018).

[40]. Gurovich Yaron et al. “Identifying facial phenotypes of genetic disorders using deep learning”. en. In: Nat. Med. 25.1 (Jan. 2019), pp. 60–64. [PubMed: 30617323]

[41]. Dingemans Alexander J M et al. “Quantitative facial phenotyping for Koolen-de Vries and 22q11.2 deletion syndrome”. en. In: Eur. J. Hum. Genet. 29.9 (Feb. 2021), pp. 1418–1423. [PubMed: 33603161]

[42]. Hsieh Tzung-Chien et al. “GestaltMatcher facilitates rare disease matching using facial phenotype descriptors”. en. In: Nat. Genet. (Feb. 2022).

[43]. Claes Peter et al. “Genome-wide mapping of global-to-local genetic effects on human facial shape”. en. In: Nat. Genet. 50.3 (Mar. 2018), pp. 414–423. [PubMed: 29459680]

[44]. White Julie D et al. “Insights into the genetic architecture of the human face”. en. In: Nat. Genet. 53.1 (Jan. 2021), pp. 45–53. [PubMed: 33288918]

[45]. Naqvi Sahin et al. “Shared heritability of human face and brain shape”. en. In: Nat. Genet. 53.6 (June 2021), pp. 830–839. [PubMed: 33821002]

[46]. Zhang Manfei et al. “Genetic variants underlying differences in facial morphology in East Asian and European populations”. en. In: Nat. Genet. 54.4 (Apr. 2022), pp. 403–411. [PubMed: 35393595]

[47]. Silfhout Anneke T Vulto-van et al. “Clinical significance of de novo and inherited copy-number variation”. en. In: Hum. Mutat. 34.12 (Dec. 2013), pp. 1679–1687. [PubMed: 24038936]

[48]. Ribeiro Marco Tulio, Singh Sameer, and Guestrin Carlos. “”Why should i trust you?” Explaining the predictions of any classifier”. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, pp. 1135–1144.

[49]. Ras Gabrielle et al. “Explainable Deep Learning: A Field Guide for the Uninitiated”. en. In: J. Artif. Intell. Res. 73 (Jan. 2022), pp. 329–396.

[50]. Brier GW. “Verification of forecasts expressed in terms of probability”. In: Mon. Weather Rev. (1950).

[51]. Koolen David A et al. “Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome”. en. In: Nat. Genet. 44.6 (Apr. 2012), pp. 639–641. [PubMed: 22544363]

[52]. Zollino Marcella et al. “Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype”. en. In: Nat. Genet. 44.6 (Apr. 2012), pp. 636–638. [PubMed: 22544367]

[53]. Koolen David A et al. “The Koolen-de Vries syndrome: a phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant”. In: Eur. J. Hum. Genet. 24.5 (2016), pp. 652–659. [PubMed: 26306646]

[54]. Köhler Sebastian et al. “Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources”. en. In: Nucleic Acids Res. 47.D1 (Jan. 2019), pp. D1018–D1027. [PubMed: 30476213]

[55]. den Hoed Joery et al. “Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction”. In: Am. J. Hum. Genet. 108.2 (Feb. 2021), pp. 346–356. [PubMed: 33513338]

[56]. Sá Maria J Nabais et al. “De novo and biallelic DEAF1 variants cause a phenotypic spectrum”. en. In: Genet. Med. 21.9 (Sept. 2019), pp. 2059–2069. [PubMed: 30923367]

[57]. Hoischen Alexander et al. “De novo mutations of SETBP1 cause Schinzel-Giedion syndrome”. en. In: Nat. Genet. 42.6 (June 2010), pp. 483–485. [PubMed: 20436468]

[58]. Filges Isabel et al. “Reduced expression by SETBP1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome”. en. In: J. Med. Genet. 48.2 (Feb. 2011), pp. 117–122. [PubMed: 21037274]

[59]. Bend Eric G et al. “Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome”. en. In: Clin. Epigenetics 11.1 (Apr. 2019), p. 64. [PubMed: 31029150]

[60]. Breen Michael S et al. “Episignatures Stratifying Helsmoortel-Van Der Aa Syndrome Show Modest Correlation with Phenotype”. en. In: Am. J. Hum. Genet. 107.3 (Sept. 2020), pp. 555–563. [PubMed: 32758449]

[61]. Jagadeesh Karthik A et al. “Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization”. en. In: Genet. Med. 21.2 (Feb. 2019), pp. 464–470. [PubMed: 29997393]

[62]. Lyra Paulo C M Jr et al. “Integration of functional assay data results provides strong evidence for classification of hundreds of BRCA1 variants of uncertain significance”. en. In: Genet. Med. 23.2 (Feb. 2021), pp. 306–315. [PubMed: 33087888]

[63]. Frederiksen Jane H et al. “Classification of MSH6 Variants of Uncertain Significance Using Functional Assays”. en. In: Int. J. Mol. Sci. 22.16 (Aug. 2021).

[64]. Caswell Richard C et al. “Assessing the clinical utility of protein structural analysis in genomic variant classification: experiences from a diagnostic laboratory”. en. In: Genome Med. 14.1 (July 2022), p. 77. [PubMed: 35869530]

[65]. Dingemans Alexander J M et al. “Human disease genes website series: An international, open and dynamic library for up-to-date clinical information”. en. In: Am. J. Med. Genet. A (Jan. 2021).

[66]. McKusick Victor A. “Mendelian Inheritance in Man and its online version, OMIM”. en. In: Am. J. Hum. Genet. 80.4 (Apr. 2007), pp. 588–604. [PubMed: 17357067]

[67]. Firth Helen V et al. “DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources”. en. In: Am. J. Hum. Genet. 84.4 (Apr. 2009), pp. 524–533. [PubMed: 19344873]

[68]. Adam Margaret P et al. GeneReviews®. Ed. by Adam Margaret P et al. Seattle (WA): University of Washington, Seattle, Mar. 2010.

[69]. Helsmoortel C et al. “A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP”. In: Nat. Genet. 46.4 (Apr. 2014), pp. 380–384. [PubMed: 24531329]

[70]. Côté RA and Robboy S. “Progress in medical information management. Systematized nomenclature of medicine (SNOMED)”. en. In: JAMA 243.8 (1980), pp. 756–762. [PubMed: 6986000]

[71]. Parkhi Omkar M, Vedaldi Andrea, and Zisserman Andrew. “Deep Face Recognition”. In: Proceedings of the British Machine Vision Conference 2015 (2015), pp. 41.1–41.12.

[72]. Cao et al. “VGGFace2: A Dataset for Recognising Faces across Pose and Age”. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Vol. 0. May 2018, pp. 67–74.

[73]. Dingemans Alexander J M et al. “Comparing facial feature extraction methods in the diagnosis of rare genetic syndromes”. en. In: medRxiv (Aug. 2022), p. 2022.08.26.22279217.

[74]. Resnik P. “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language”. en. In: J. Artif. Intell. Res. 11 (July 1999), pp. 95–130.

[75]. Pesquita Catia et al. “Metrics for GO based protein semantic similarity: a systematic evaluation”. en. In: BMC Bioinformatics 9 Suppl 5 (Apr. 2008), S4.

[76]. Arvai Kevin and Borroto Carlos and Gainullin Vladimir and Retterer Kyle. Phenopy. Nov. 2019.

[77]. Köhler Sebastian et al. “The Human Phenotype Ontology in 2017”. en. In: Nucleic Acids Res. 45.D1 (Jan. 2017), pp. D865–D876. [PubMed: 27899602]

[78]. Yuan Xiao et al. “Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases”. en. In: Brief. Bioinform. 23.2 (Mar. 2022).

[79]. Manders Peggy et al. “The Radboud biobank: A central facility for disease-based biobanks to optimise use and distribution of biomaterial for scientific research in the Radboud university medical center, Nijmegen”. en. In: Open J. Bioresour. 5 (Feb. 2018).

[80]. Karras Tero, Laine Samuli, and Aila Timo. “A Style-Based Generator Architecture for Generative Adversarial Networks”. en. In: IEEE Trans. Pattern Anal. Mach. Intell. 43.12 (Dec. 2021), pp. 4217–4228. [PubMed: 32012000]

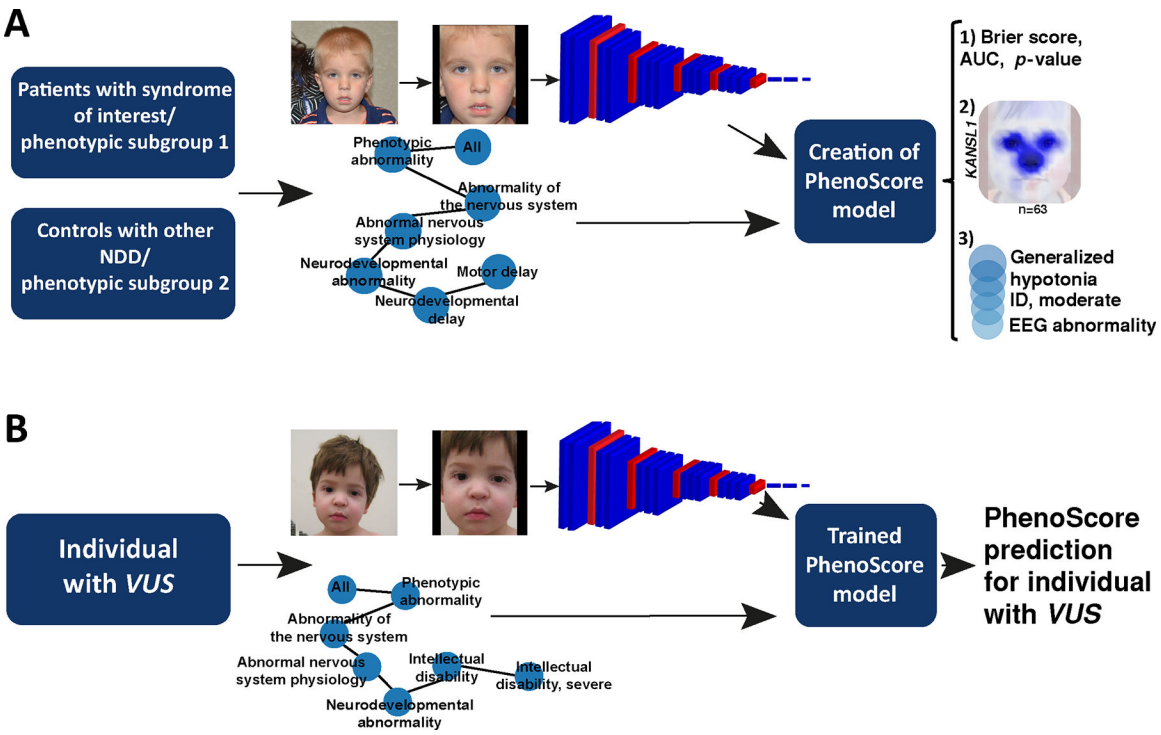

Figure 1:

A) Here, the global workflow of this study is displayed, with the training and construction of PhenoScore. *n* individuals and *n* age-, sex- and ethnicity matched controls are selected for each syndrome. The facial features are extracted using a convolutional neural network, VGGFace2, and in parallel, the phenotypic similarity of individuals and controls is calculated. PhenoScore is then trained on both the facial features and the HPO similarity combined. PhenoScore outputs the classification metrics (the Brier score, AUC and corresponding *p*-value) to report how well it is able to distinguish the investigated phenotypic groups. Furthermore, facial heatmaps and visualisations for the most important phenotypic features are generated as well. B) The trained PhenoScore model for a specific syndrome is used for a new individual with a VUS. Again, the phenotypic similarity and facial distances are calculated, and these are used as input for PhenoScore after training. The output is a score and assesses whether the individual of interest has that specific syndrome, thus the VUS being (likely) pathogenic.

VUS: variant of unknown significance; NDD: neurodevelopmental disorders

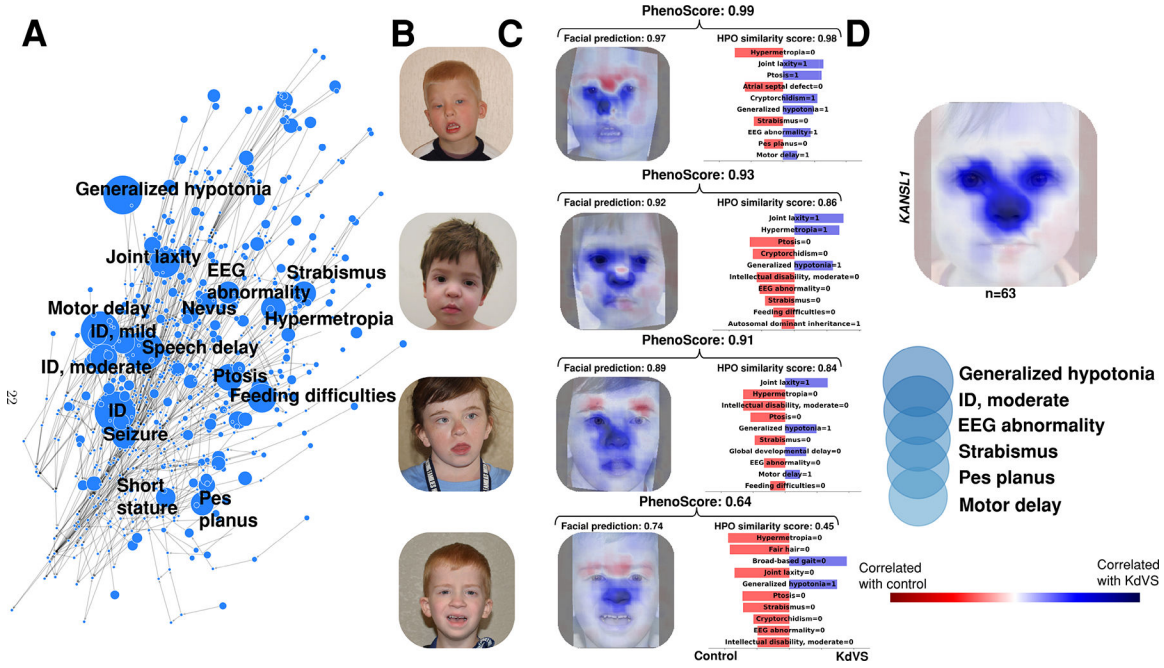

Figure 2:

A) The HPO terms of all included individuals with Koolen-de Vries (KdVS) are shown here. HPO terms present in 20% or more of the individuals are annotated with text, and larger nodes correspond to a higher prevalence of that specific clinical feature. The graph structure corresponds to that of the HPO terms. ID = intellectual disability. B) Four individuals diagnosed with Koolen-de Vries syndrome are presented here (written informed consent for the publication of these facial images was obtained). These were randomly selected from the included dataset without any selection criterion.

Figure 2: C) For the four randomly selected individuals, three predictions are shown: using the facial image, using the phenotypic data, and finally, the PhenoScore, which combines both. Furthermore, heatmaps are generated using local interpretable model-agnostic explanations (LIME) to see which facial areas are most important according to our model, where blue correlates with KdVS and red areas correlate with controls. The nose and eyes are clearly prioritized, corresponding to the known dysmorphic features in Koolen-de Vries. Furthermore, the most important clinical features are shown for each individual and the contribution (corresponding to the LIME regression coefficient) of that feature to the prediction. D) Finally, a summarized heatmap was generated to investigate the overall most important facial and phenotypic features. We averaged the heatmaps of the five individuals with Koolen-de Vries with the highest prediction. Next to that, to obtain the most important clinical features, too, we averaged the LIME regression coefficient for the different symptoms of the five highest-scoring individuals based on HPO. Shown clinical features are ordered based on importance, and the size of the circle indicates the relative importance of the feature. ID=intellectual disability

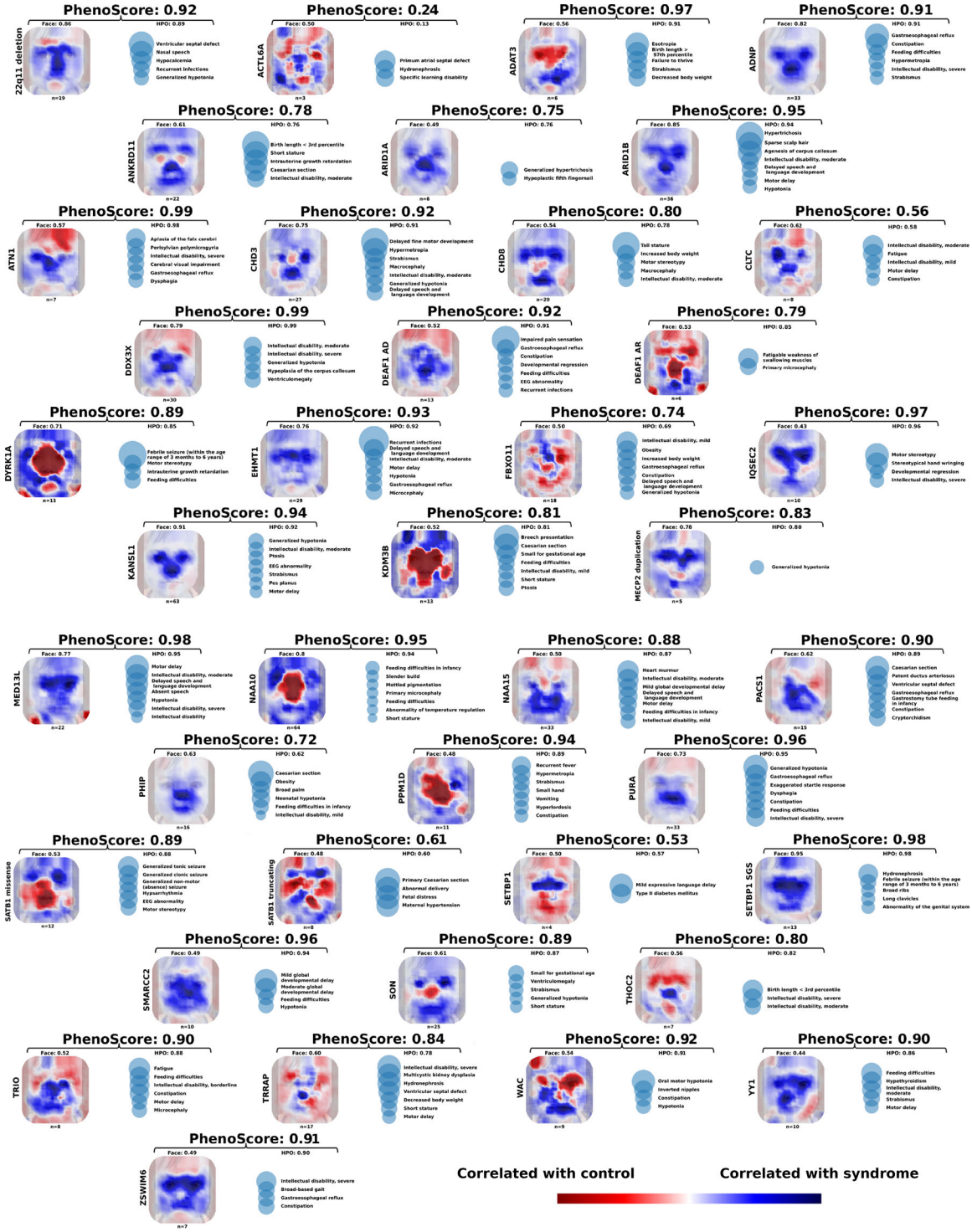

Figure 3:

The heatmaps and most important clinical features of all 40 genetic syndromes included in this study are displayed in this figure. The facial heatmaps and the phenotypic data are the average LIME heatmaps of the five individuals per genetic syndrome with the highest predictive score. For the phenotypic data, in this figure, only features positively correlated

with the genetic syndrome of interest are included. The standard face used as background is a non-existent person generated using StyleGAN [80]. In general, the facial heatmaps correspond well to dysmorphic features known in literature of the investigated syndromes. In specific regions, however, faces from cases are more similar to controls than to other cases (in red), signifying that random facial variance also contributes to the predictions whereas these would expected to be neutral. The PhenoScore in this figure refers to the AUC of the model for that genetic syndrome.

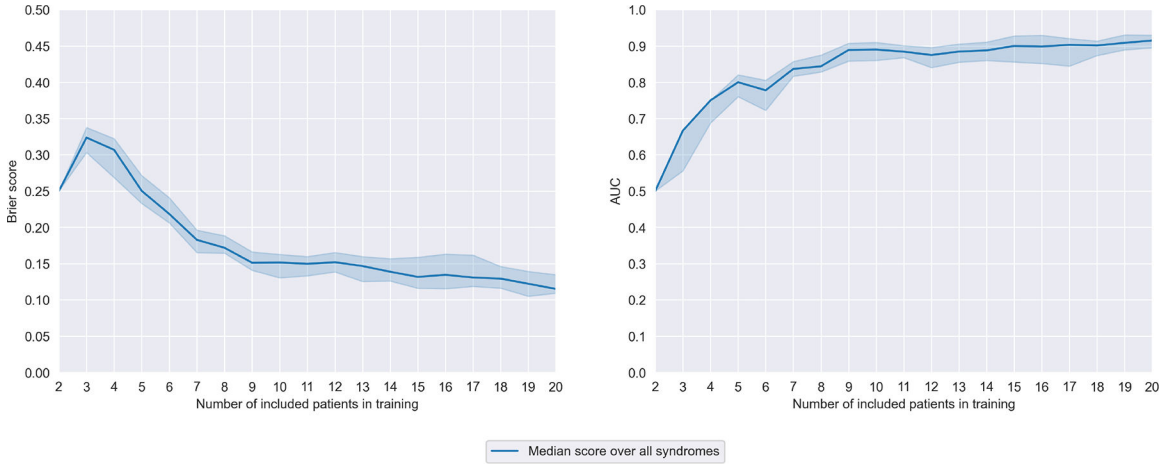

Figure 4:

The performance of the SVM using both facial- and HPO features with different sizes of the training set is shown here. Both the median Brier score and the median AUC improve if the number of individuals to train on is larger — as would be expected. Interestingly, only five individuals are needed for an already acceptable classification performance, with performance increasing with a larger training set, as is expected.

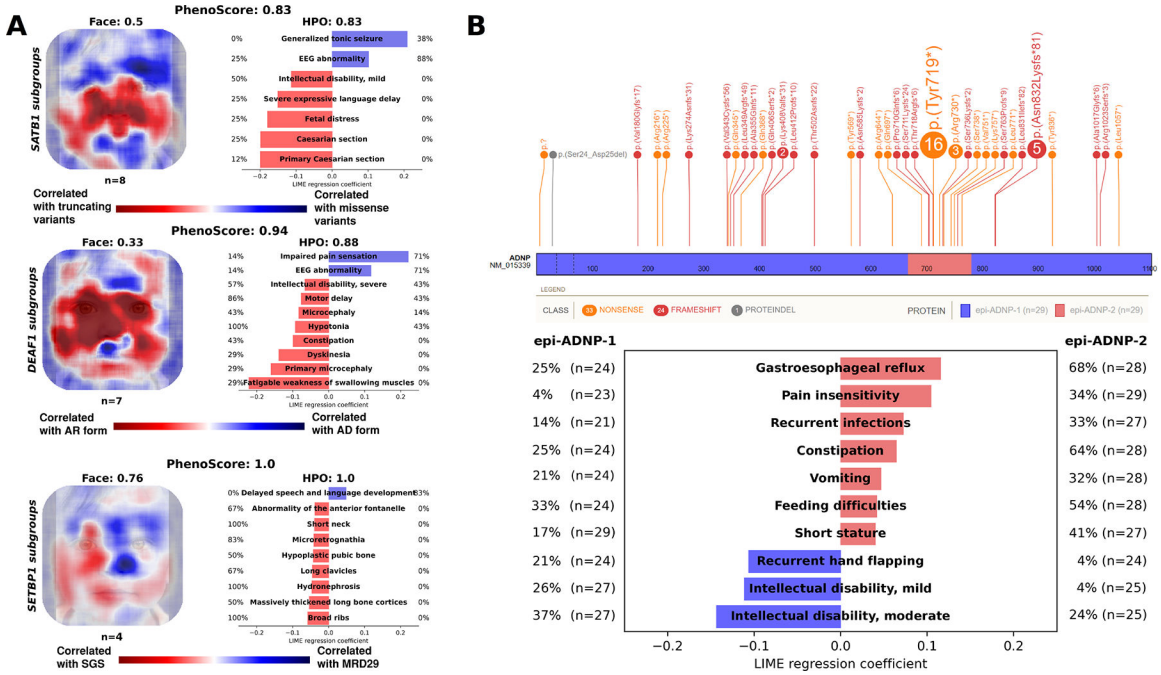

Figure 5:

A) The facial heatmaps and most important clinical features for the three confirmatory subgroup analyses. First (top-left), the analysis when comparing the two phenotypic subgroups associated with pathogenic variants in *DEAF1*; top-right shows the PhenoScore results when analysing the subgroups for *SATB1* and finally, in the bottom panel the outcome for *SETBP1* is displayed. The PhenoScores in this figure correspond to the AUC when training the model.

B) Above: a lollipop plot (generated using St. Jude's ProteinPaint) of the genetic variants currently collected using the *ADNP* HDG website [65]. Of the 58 included individuals, 29 had a variant in the c.2000–2340 region, indicated by others as having a different methylation signature than variants outside this region [59]. Using only the HPO module of our PhenoScore framework, we first matched the groups on sex-, ethnicity- and age when possible to create two groups of the same size (29 vs. 29). We then trained a classifier on the two groups and found a significant difference (Brier score of 0.24, AUC of 0.71, $p = 0.01$). Below: the most important clinical features according to our model (determined using LIME) and the corresponding prevalence in both groups.

Table 1:

The number of individuals per genetic syndrome included in our analysis are shown here. For every individual, a facial photograph, phenotypic data, and an age-, sex- and ethnicity control with a neurodevelopmental disorder is available (otherwise, the individual was excluded). Per genetic syndrome, the sex distribution, the median age and the results of the support vector machine (SVM) classifier are displayed here. The Brier score, for which lower is better, per syndrome is shown — with the numbers shown corresponding to the mean of the scores during the five iterations in which matched controls were sampled. The AUC (higher is better) and accuracy (with 0.5 as cut-off) are included as well. For almost all syndromes, the combination of facial- and phenotypic data is an improvement over using either dataset alone. Furthermore, the last column of this table displays the calculated *p*-values for the investigated syndromes using the random permutation test. All but three are significant at the 0.05 level, as expected when inspecting the classification results.

Gene/genetic syndrome	OMIM number	Number of individuals	Sex (male/female)	Age (median in years)	Brier Facial data only	Brier HPO data only	PhenoScore (Brier)	PhenoScore (AUC)	PhenoScore (accuracy)	*p*-value
22q11 deletion syndrome	188400	19	10/9 (53%/47%)	5.0	0.147	0.138	0.108	0.92	0.85	<0.001
ACTL6A	NA	3	2/1 (67%/33%)	6.0	0.250	0.709	0.575	0.24	0.33	0.90
ADAT3 (NEDBGF)	615286	6	3/3 (50%/50%)	7.5	0.256	0.112	0.087	0.97	0.88	<0.001
ADNP (Helsmoortel-van der Aa syndrome)	615873	33	15/18 (45%/55%)	5.0	0.175	0.118	0.117	0.91	0.84	<0.001
ANKRD11 (KBG syndrome)	148050	22	15/7 (68%/32%)	9.5	0.236	0.216	0.203	0.78	0.70	<0.001
ARID1A (Coffin-Siris syndrome 2)	614607	6	3/3 (50%/50%)	9.5	0.261	0.244	0.262	0.75	0.63	0.02
ARID1B (Coffin-Siris syndrome)	135900	36	16/20 (44%/56%)	5.5	0.162	0.096	0.075	0.95	0.91	<0.001
ATN1 (CHEDDA)	618494	7	2/5 (29%/71%)	5.0	0.233	0.090	0.102	0.99	0.91	<0.001
CHD3 (Snijders Blok-Campeau syndrome)	618205	27	11/16 (41%/59%)	10.0	0.198	0.122	0.118	0.92	0.84	<0.001
CHD8 (IDDAM)	615032	20	15/5 (75%/25%)	11.0	0.247	0.195	0.183	0.80	0.72	<0.001
CLTC (MRD56)	617854	8	4/4 (50%/50%)	14.5	0.240	0.278	0.275	0.56	0.56	0.13
DDX3X (MRXSSB)	300958	30	0/30 (0%/100%)	8.5	0.189	0.035	0.034	0.99	0.96	<0.001
DEAF1 (NEDHELS)	617171	6	3/3 (50%/50%)	8.0	0.256	0.224	0.239	0.79	0.67	0.01
DEAF1 (Vulto-van Silfout-de Vries syndrome)	615828	13	10/3 (77%/23%)	7.0	0.257	0.091	0.086	0.92	0.91	<0.001
DYRK1A (MRD7)	614104	13	7/6 (54%/46%)	12.0	0.204	0.156	0.133	0.89	0.81	<0.001
EHMT1 (Kleefstra syndrome)	610253	29	12/17 (41%/59%)	6.0	0.206	0.117	0.109	0.93	0.84	<0.001

Gene/genetic syndrome	OMIM number	Number of individuals	Sex (male/female)	Age (median in years)	Brier Facial data only	Brier HPO data only	PhenoScore (Brier)	PhenoScore (AUC)	PhenoScore (accuracy)	*p*-value
FBXO11 (IDDFBA)	618089	18	14/4 (78%/22%)	7.0	0.261	0.238	0.220	0.74	0.70	<0.001
IQSEC2 (XLID1)	309530	10	4/6 (40%/60%)	10.5	0.254	0.084	0.086	0.97	0.91	<0.001
KANSL1 (Koolen-De Vries syndrome)	610443	63	28/35 (44%/56%)	6.0	0.128	0.096	0.082	0.94	0.90	<0.001
KDM3B (Diets-Jongmans syndrome)	618846	13	7/6 (54%/46%)	7.0	0.254	0.178	0.176	0.81	0.77	<0.001
MECP2 duplication (MRXSL)	300260	5	5/0 (100%/0%)	8.0	0.184	0.198	0.195	0.83	0.76	<0.001
MED13L (MRFACD)	616789	22	13/9 (59%/41%)	6.0	0.196	0.091	0.075	0.98	0.90	<0.001
NAA10 (Ogden syndrome)	300855	64	14/50 (22%/78%)	7.0	0.181	0.071	0.066	0.95	0.92	<0.001
NAA15 (MRD50)	617787	33	26/7 (79%/21%)	7.0	0.271	0.136	0.131	0.88	0.83	<0.001
PACS1 (Schuurs-Hoeijmakers syndrome)	615009	15	10/5 (67%/33%)	4.0	0.226	0.135	0.125	0.90	0.81	<0.001
PHIP (Chung-Jansen syndrome)	617991	16	9/7 (56%/44%)	12.0	0.224	0.275	0.231	0.72	0.64	<0.001
PPM1D (Jansen-de Vries syndrome)	617450	11	5/6 (45%/55%)	7.0	0.254	0.180	0.142	0.94	0.75	<0.001
PURA (NEDRIHF)	616158	33	18/15 (55%/45%)	9.0	0.211	0.090	0.076	0.96	0.89	<0.001
SATB1 (DEFDA)	619228	8	3/5 (38%/62%)	6.5	0.282	0.262	0.261	0.61	0.56	0.03
SATB1 (Kohlschutter-Tonz syndrome-like)	619229	12	5/7 (42%/58%)	11.5	0.270	0.123	0.123	0.89	0.85	<0.001
SETBP1 (MRD29)	616078	4	1/3 (25%/75%)	13.5	0.250	0.287	0.385	0.53	0.55	0.21
SETBP1 (Schinzel-Giedion syndrome)	269150	13	7/6 (54%/46%)	1.0	0.091	0.065	0.061	0.98	0.91	<0.001
SMARCC2 (Coffin-Siris syndrome 8)	618362	10	8/2 (80%/20%)	9.0	0.252	0.116	0.111	0.96	0.89	<0.001
SON (ZTTK syndrome)	617140	25	13/12 (52%/48%)	6.0	0.237	0.140	0.132	0.89	0.82	<0.001
THOC2 (XLID12)	300957	7	7/0 (100%/0%)	6.0	0.256	0.201	0.192	0.80	0.69	0.001
TRIO (MRD63)	618825	8	3/5 (38%/62%)	10.5	0.264	0.144	0.137	0.90	0.86	<0.001
TRRAP (DEDDFA)	618454	17	6/11 (35%/65%)	11.0	0.244	0.198	0.167	0.84	0.78	<0.001
WAC (DeSanto-Shinawi syndrome)	616708	9	3/6 (33%/67%)	4.0	0.246	0.133	0.132	0.92	0.82	<0.001
YY1 (Gabriele-de Vries syndrome)	617557	10	5/5 (50%/50%)	8.0	0.255	0.166	0.142	0.90	0.82	<0.001

Gene/genetic syndrome	OMIM number	Number of individuals	Sex (male/female)	Age (median in years)	Brier Facial data only	Brier HPO data only	PhenoScore (Brier)	PhenoScore (AUC)	PhenoScore (accuracy)	*p*-value
ZSWIM6 (NEDMAGA)	617865	7	3/4 (43%/57%)	7.0	0.265	0.146	0.138	0.91	0.79	<0.001