PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework

Dingemans, A.J.M.; Hinne, M.; Truijen, K.M.G.; Goltstein, L.; Reeuwijk, J. van; Leeuw, N. de; ... ; Vries, B.B.A. de

nature genetics

Technical Report

https://doi.org/10.1038/s41588-023-01469-w

PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework

Received: 6 September 2022

Accepted: 5 July 2023

Published online: 7 August 2023

Check for updates

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[1,17] & Bert B. A. de Vries[1,17]

Several molecular and phenotypic algorithms exist that establish genotype–phenotype correlations, including facial recognition tools. However, no unified framework that investigates both facial data and other 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. Here we show 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 an 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.

A substantial portion of individuals with clinically and genetically heterogeneous rare diseases, such as neurodevelopmental disorders (NDDs), has been molecularly diagnosed in the last decade using whole-exome sequencing (WES)[1–4]. Clinical WES data interpretation relies on filtering and prioritization for rare genetic variants that are subsequently interpreted in the context of the patient's clinical presentation[5]. Although this strategy is essential to identify the disease-causing variant(s), it is estimated that dozens of variants are

A full list of affiliations appears at the end of the paper. e-mail: lisenka.vissers@radboudumc.nl; bert.devries@radboudumc.nl

https://doi.org/10.1038/s41588-023-01469-w

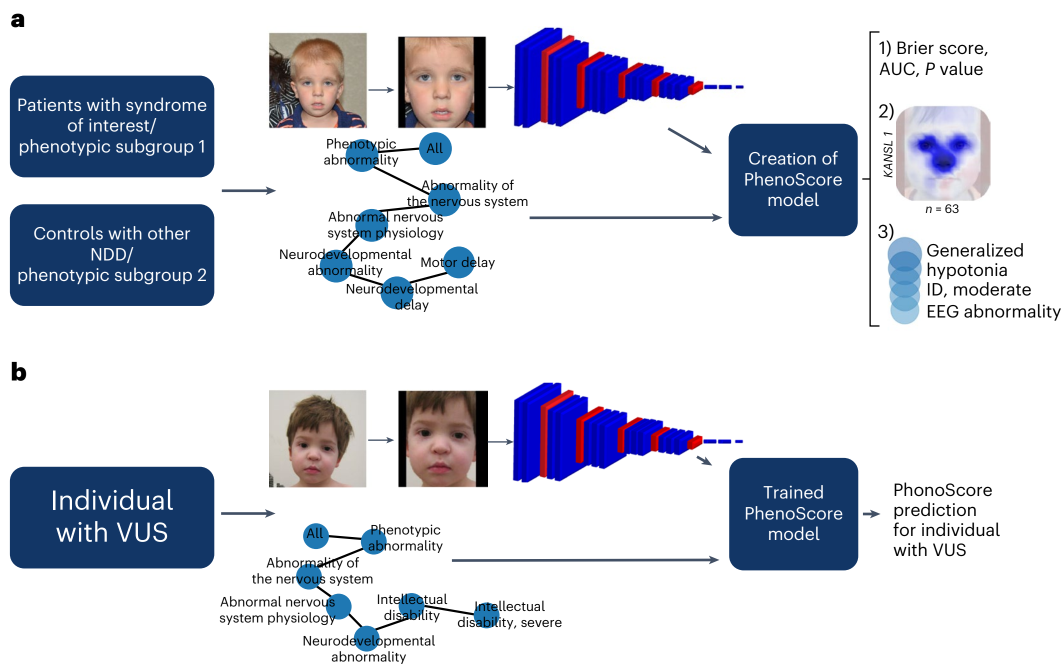


Fig. 1 | Overview of PhenoScore. a, Here the global workflow of this study is displayed, with the training and construction of PhenoScore. *n* individuals and *n* age-, sex- and ethnicity-matched controls are selected for each syndrome. The facial features are extracted using a convolutional neural network, 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 visualizations 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.

prioritized as diagnostic noise[6]—and this number is expected to rise with technological innovations (such as long-read whole-genome sequencing, RNA sequencing and optical genome mapping, enabling the discovery of noncoding variants and complex structural variation) finding their way into the diagnostic arena[7–11].

At the molecular level, several computational methods, such as MutationTaster[12], PolyPhen[13], Sorting Intolerant from Tolerant (SIFT)[14] and Combined Annotation-Dependent Depletion (CADD) score[15], have been designed to effectively prioritize causal variants. 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[16]. However, equivalent to molecular tools, algorithms using these HPO data to quantify phenotypic HPO similarity between individuals with genetic disorders would provide substantial benefits to diagnosing rare diseases. Such a quantitative phenotypic score could, for instance, assist with the interpretation of variants of unknown significance (VUS), which constitute 10–30% of variants assessed[4,17]. Reducing the number of VUSs is of the essence because studies have shown that families usually do not comprehend its meaning[18,19], potentially leading to frustration due to the uncertainty involving a possible diagnosis and course of the disease. Importantly, VUSs have also been shown to inflict inappropriate medical decisions[20,21].

Next to reclassifying VUSs, quantifying phenotypic HPO similarity at the cohort level could also help to provide further steps toward personalized medicine by automatically recognizing distinct phenotypic subtypes leading to more tailored clinical prognosis[22–24].

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[25–28]. In genetics, these new techniques have been used in the assisted interpretation of genomic variants[29–31] and combining molecular and phenotypic evaluations, mainly looking at methods to use phenotypic data to automatically prioritize genetic variants[32–38]. Furthermore, advances in computer vision have led to the application of facial recognition technology in clinical genetics[39–44], with the current state-of-the-art application GestaltMatcher achieving a top-10 accuracy of 64%[44]. Facial recognition can assist in the recognition of (neuro) developmental syndromes because the development of the brain and facial shape are closely linked[45–48]—and therefore, a substantial part of genetic disorders have distinct facial features[49]. However, not all genetic syndromes have a clear, recognizable, facial gestalt, which hinders methods solely looking at facial features. Although tools have previously looked at either combining molecular data with either HPO, or alternatively, with facial features[1,41], 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 HPO from deep phenotyping.

Results

The PhenoScore framework

PhenoScore is a framework that currently consists of the following two modules: a component that extracts the facial features from a 2D-facial photograph and a second module that calculates HPO-based phenotypic similarity. The AI-based framework then provides the following three outputs: a Brier score and corresponding *P* value, defining how well PhenoScore is able to distinguish the investigated syndrome; a facial heatmap, highlighting important facial features

https://doi.org/10.1038/s41588-023-01469-w

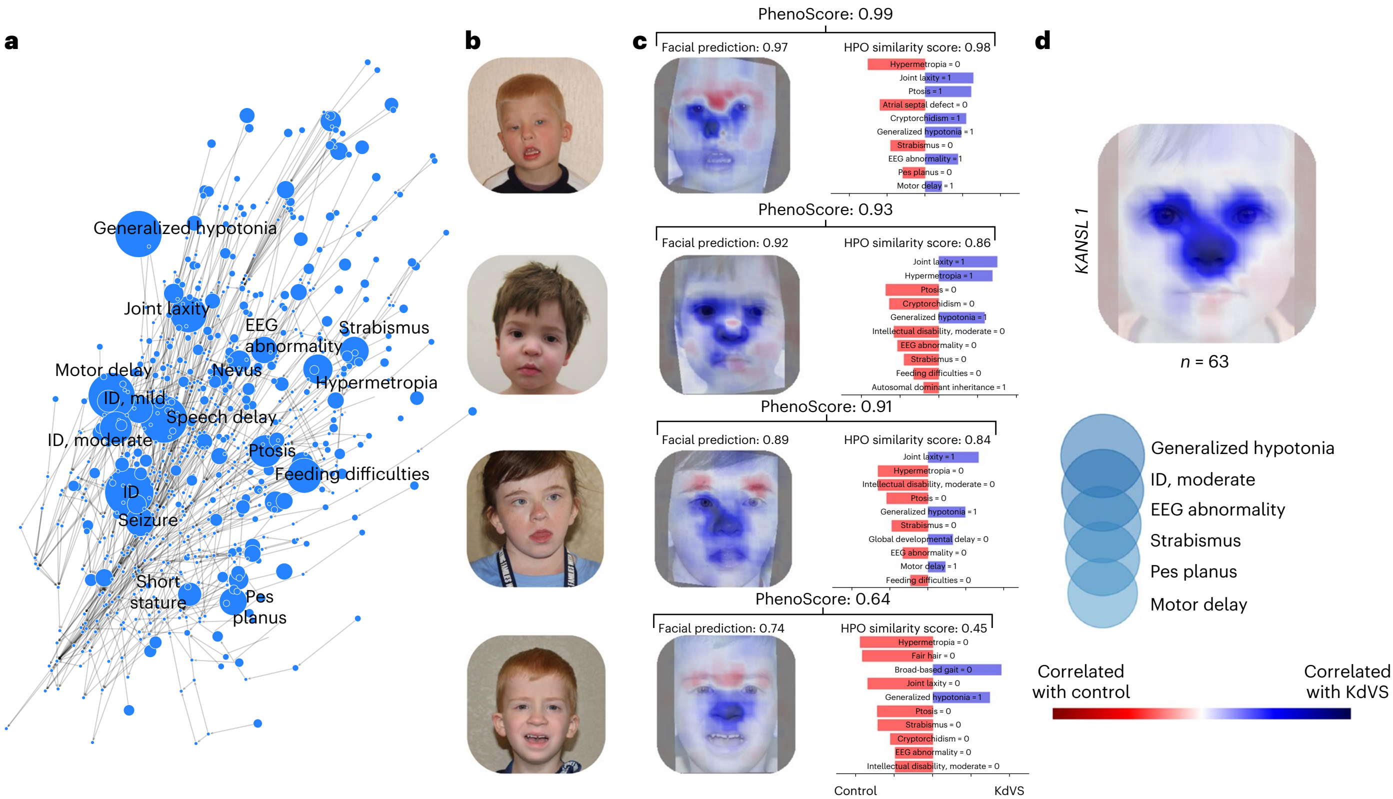


Fig. 2 | PhenoScore for KdVS. a, The HPO terms of all included individuals with KdVS are shown here. HPO terms present in 20% or more of the individuals are annotated with text, and larger nodes correspond to a higher prevalence of that specific clinical feature. The graph structure corresponds to that of the HPO terms. **b**, Four individuals diagnosed with KdVS 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. **c**, For the four randomly selected individuals, the following three predictions are shown: using the facial image, using the phenotypic data and, finally, the PhenoScore, which combines both. Furthermore, heatmaps are generated using 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 KdVS. 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 KdVS 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.

and a visualization of the most important other 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 NDDs (Fig. 1a). Next, the facial features are automatically extracted from the facial photographs and the phenotypic HPO similarity is calculated (with several HPO terms 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 inter-observer 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 (Fig. 1b). 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, enabling PhenoScore to generate facial heatmaps and visualizations on the most important clinical features.

Proof-of-concept: PhenoScore for Koolen-de Vries syndrome (KdVS)

First, we investigated whether using our combined PhenoScore was an improvement on solely using either facial or phenotypic data. The SVM was trained on both separate feature sets alone 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 the predicted outcome and the 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 KdVS (OMIM, 610443), caused by either 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[51–53], for which the interdependence in our modeling is preserved using the graph structure of the HPO terms (Fig. 2a). 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/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 (Fig. 2b) from the training dataset and retrained PhenoScore, evaluating the performance when treating them as if diagnoses of KdVS were unknown. PhenoScore then generated predictions for these four individuals when comparing

Table 1 | Demographics of individuals included in this study

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	1.35×10^{-6}
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	1.35×10^{-6}
ADNP (Helsmoortel-van der Aa syndrome)	615873	33	15/18 (45%/55%)	5.0	0.175	0.118	0.117	0.91	0.84	1.35×10^{-6}
ANKRD11 (KBG syndrome)	148050	22	15/7 (68%/32%)	9.5	0.236	0.216	0.203	0.78	0.70	1.46×10^{-5}
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	1.35×10^{-6}
ATN1 (CHEDDA)	618494	7	2/5 (29%/71%)	5.0	0.233	0.090	0.102	0.99	0.91	1.35×10^{-6}
CHD3 (Snijders Blok–Campeau syndrome)	618205	27	11/16 (41%/59%)	10.0	0.198	0.122	0.118	0.92	0.84	1.35×10^{-6}
CHD8 (IDDAM)	615032	20	15/5 (75%/25%)	11.0	0.247	0.195	0.183	0.80	0.72	7.52×10^{-6}
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	1.35×10^{-6}
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 Silfhout–de Vries syndrome)	615828	13	10/3 (77%/23%)	7.0	0.257	0.091	0.086	0.92	0.91	1.35×10^{-6}
DYRK1A (MRD7)	614104	13	7/6 (54%/46%)	12.0	0.204	0.156	0.133	0.89	0.81	2.40×10^{-6}
EHMT1 (Kleefstra syndrome)	610253	29	12/17 (41%/59%)	6.0	0.206	0.117	0.109	0.93	0.84	1.35×10^{-6}
FBXO11 (IDDFBA)	618089	18	14/4 (78%/22%)	7.0	0.261	0.238	0.220	0.74	0.70	8.25×10^{-5}
IQSEC2 (XLID1)	309530	10	4/6 (40%/60%)	10.5	0.254	0.084	0.086	0.97	0.91	1.35×10^{-6}
KANSL1 (KdVS)	610443	63	28/35 (44%/56%)	6.0	0.128	0.096	0.082	0.94	0.90	1.35×10^{-6}
KDM3B (Diets-Jongmans syndrome)	618846	13	7/6 (54%/46%)	7.0	0.254	0.178	0.176	0.81	0.77	1.84×10^{-5}
MECP2 duplication (MRXSL)	300260	5	5/0 (100%/0%)	8.0	0.184	0.198	0.195	0.83	0.76	6.97×10^{-4}
MED13L (MRFACD)	616789	22	13/9 (59%/41%)	6.0	0.196	0.091	0.075	0.98	0.90	1.35×10^{-6}
NAA10 (Ogden syndrome)	300855	64	14/50 (22%/78%)	7.0	0.181	0.071	0.066	0.95	0.92	1.35×10^{-6}
NAA15 (MRD50)	617787	33	26/7 (79%/21%)	7.0	0.271	0.136	0.131	0.88	0.83	1.35×10^{-6}
PACS1 (Schuurs-Hoeijmakers syndrome)	615009	15	10/5 (67%/33%)	4.0	0.226	0.135	0.125	0.90	0.81	1.35×10^{-6}
PHIP (Chung-Jansen syndrome)	617991	16	9/7 (56%/44%)	12.0	0.224	0.275	0.231	0.72	0.64	4.17×10^{-4}
PPM1D (Jansen-de Vries syndrome)	617450	11	5/6 (45%/55%)	7.0	0.254	0.180	0.142	0.94	0.75	1.05×10^{-5}
PURA (NEDRIHF)	616158	33	18/15 (55%/45%)	9.0	0.211	0.090	0.076	0.96	0.89	1.35×10^{-6}
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	1.35×10^{-6}
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	1.35×10^{-6}
SMARCC2 (Coffin–Siris syndrome 8)	618362	10	8/2 (80%/20%)	9.0	0.252	0.116	0.111	0.96	0.89	4.3×10^{-6}
SON (ZTTK syndrome)	617140	25	13/12 (52%/48%)	6.0	0.237	0.140	0.132	0.89	0.82	1.35×10^{-6}
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	1.96×10^{-5}
TRRAP (DEDDFA)	618454	17	6/11 (35%/65%)	11.0	0.244	0.198	0.167	0.84	0.78	2.40×10^{-6}
WAC (DeSanto-Shinawi syndrome)	616708	9	3/6 (33%/67%)	4.0	0.246	0.133	0.132	0.92	0.82	4.25×10^{-6}
YY1 (Gabriele-de Vries syndrome)	617557	10	5/5 (50%/50%)	8.0	0.255	0.166	0.142	0.90	0.82	1.35×10^{-6}
ZSWIM6 (NEDMAGA)	617865	7	3/4 (43%/57%)	7.0	0.265	0.146	0.138	0.91	0.79	1.46×10^{-5}

The number of individuals per genetic syndrome included in our analysis is shown in this table. For every individual, a facial photograph, phenotypic data, and age-, sex- and ethnicity-matched control with a neurodevelopmental disorder are available (otherwise, the individual was excluded). Per genetic syndrome, the sex distribution, the median age and the results of the 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 the 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, calculated using a one-sided Fisher's combined probability test (Supplementary Data). All but three are significant at the 0.05 level, as expected when inspecting the classification results.

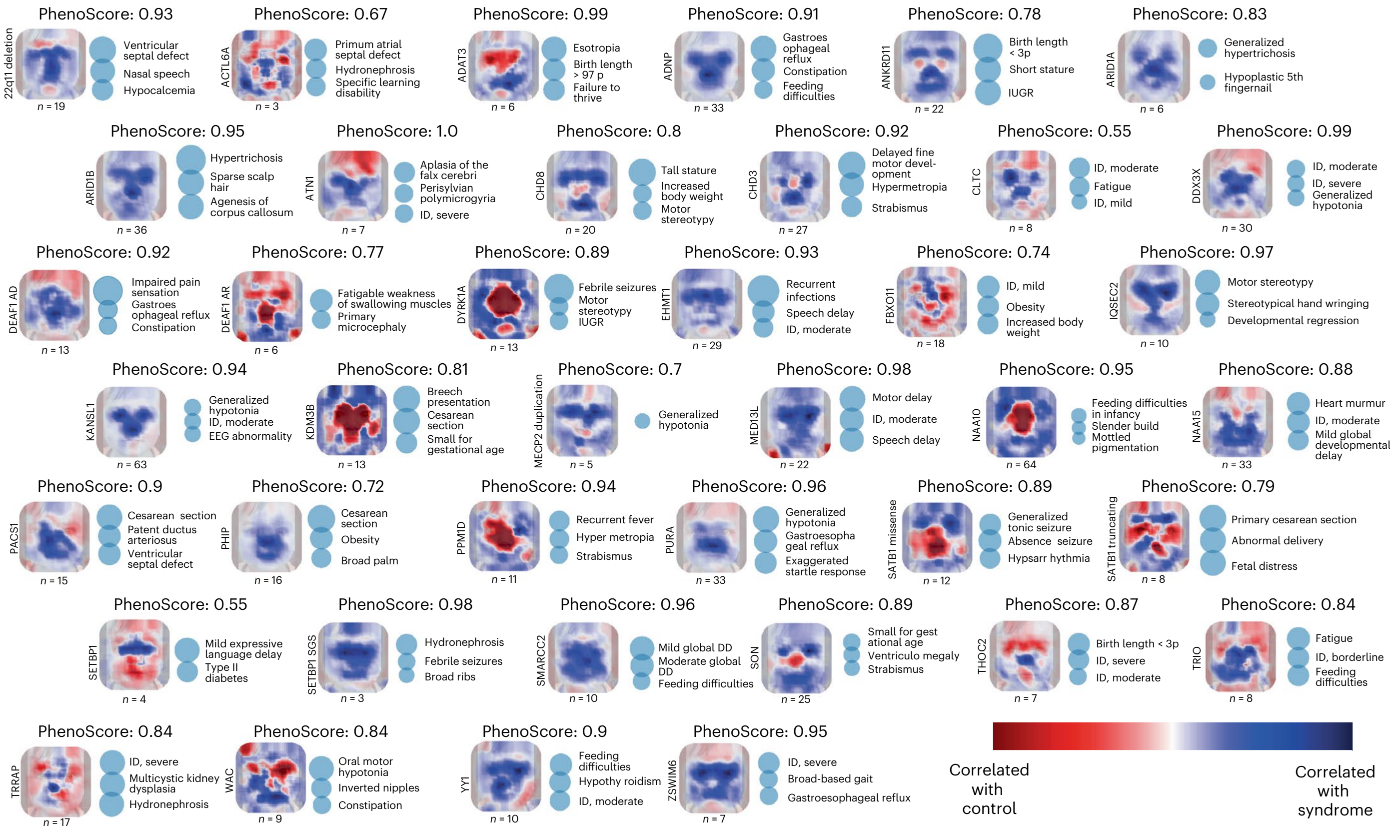


Fig. 3 | Generalization of PhenoScore to 40 syndromes. The heatmaps and most important clinical features of all 40 genetic syndromes included in this study are displayed in this figure. The facial heatmaps and the phenotypic data are the average LIME heatmaps of the five individuals per genetic syndrome with the highest predictive score. For the phenotypic data, in this figure, only the top three features positively correlated with the genetic syndrome of interest are included, and a larger bubble indicates a higher importance of that HPO term for that disorder. The standard face used as background is a nonexistent person generated using StyleGAN[71]. In general, the facial heatmaps correspond well to dysmorphic features known in the 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 be expected to be neutral. The PhenoScore in this figure refers to the AUC of the model for that genetic syndrome. DD, developmental delay; IUGR, intrauterine growth restriction.

them with 59 remaining individuals with KdVS in the training set. The output was displayed using local interpretable model-agnostic explanations (LIME), providing heatmaps of prioritized facial information according to PhenoScore (Fig. 2c). In addition, the most important clinical features according to PhenoScore to be predictive for KdVS were summarized. According to PhenoScore, the nose and eyes are the most important facial parts when recognizing KdVS while the presence of hypotonia, moderate intellectual disability, electroencephalography abnormalities, strabismus, pes planus and motor delay are the clinical features of interest (Fig. 2d). This is consistent with expert opinion and the literature[51–53] and shows that harnessing the power of both facial and phenotypic data outperforms the separate predictions.

Expanding PhenoScore to 40 syndromes

After our proof-of-concept using KdVS, we assessed the performance of PhenoScore for the classification of other genetic syndromes. Hereto, we selected 39 further syndromes (Table 1 and Extended Data Table 1) including both clinically well-recognizable syndromes based on facial gestalt, such as Kleefstra syndrome (OMIM, 610253; caused by pathogenic variants in *EHMT1*, which encodes euchromatic histone-lysine N-methyltransferase 1), Helsmoortel-van der Aa syndrome (OMIM, 615873; *ADNP*, encoding Activity Dependent Neuroprotective Protein) and Coffin–Siris syndrome (OMIM, 135900; *ARID1B*, which encodes AT-rich interactive domain-containing protein 1B) but also more recently identified syndromes for which facial gestalt is less prominent, including intellectual developmental disorder with autism and macrocephaly (IDDAM, OMIM, 615032; *CHD8*, which encodes chromodomain-helicase-DNA-binding protein 8) and intellectual developmental disorder with dysmorphic facies and behavioral abnormalities (IDDFBA, OMIM, 618089; *FBXO11*, which encodes F-box only protein 11).

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 (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 (Extended Data Table 2 and Supplementary Data). To compare the performance of PhenoScore to other approaches, we generated predictions for all individuals with a genetic syndrome in the dataset using Phenomizer[32,54] and LIkelihood Ratio Interpretation of Clinical AbnormaLities (LIRICAL)[38]. 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; Extended Data Fig. 1 and Extended Data Table 3).

For 37 (93%) of 40 syndromes, PhenoScore was able to identify predictive features that characterized these syndromes and recognized a distinct phenotypic entity (Table 1 and Extended Data Fig. 2). As

https://doi.org/10.1038/s41588-023-01469-w

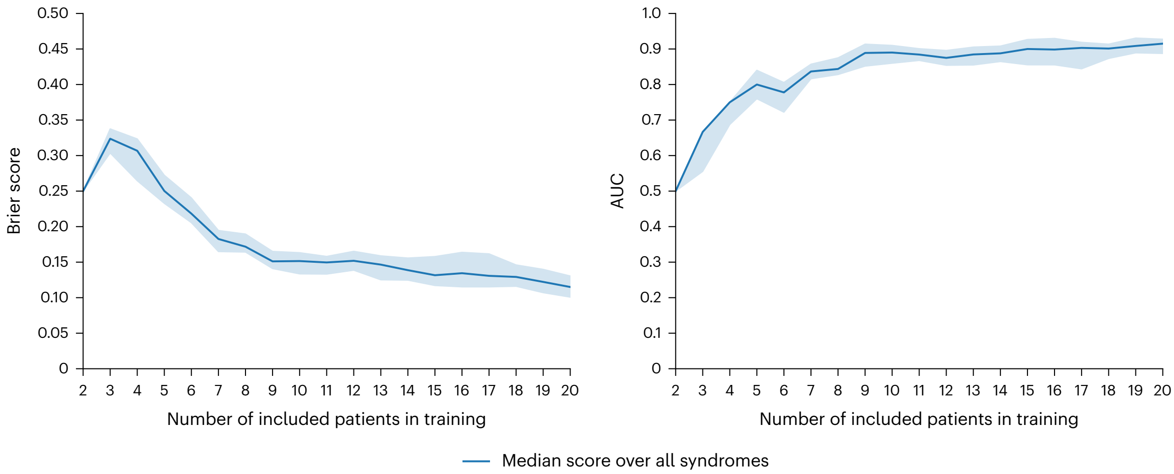


Fig. 4 | Number of individuals needed for training. The performance of the SVM using both facial and HPO features with different sizes of the training set is shown here. Data are presented as median values with confidence intervals. Both the median Brier score and the median AUC improve if the number 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.

expected and visualized in the LIME heatmaps (Fig. 3), these features corresponded remarkably well with those described in the literature.

Moreover, for a genetic syndrome that lacks explicit facial features, like IDDAM, 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* (the gene encoding dual-specificity tyrosine-(Y)-phosphorylation-regulated kinase 1 A) – 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 of genetic syndromes and highlights that the addition of nonfacial phenotypic data in HPO is essential.

Finally, we demonstrate that the performance of PhenoScore is stable over different age and population of origin subgroups (Extended Data 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 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 are of non-Caucasian/non-Western descent, the subgroups for the population of origin analysis do not seem to lead to overt differences in predictive performance between ethnicities.

PhenoScore requires a low number of individuals for training

Most genetic disorders are individually rare, with sometimes only three to five individuals reported worldwide. We therefore next investigated how many individuals PhenoScore is required 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 two 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; Fig. 4). The classification performance can be further improved when the training sets increase in size (median AUC is 0.89 with ten individuals, while with 20 individuals, the median AUC is 0.92).

Use case 1: objective clinical quantification of VUS

To display the power of PhenoScore in the clinical interpretation of variants at an individual level, we reassessed reported VUSs (American College of Medical Genetics and Genomics class 3) in the Radboudumc Department of Human Genetics. These individuals were not included in the training of PhenoScore and can therefore be considered 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 (Extended Data 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 nine 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, which 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 because 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.

Use case 2: 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 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 without being able to determine what aspects clinically distinguish subtypes. We tested

 https://doi.org/10.1038/s41588-023-01469-w

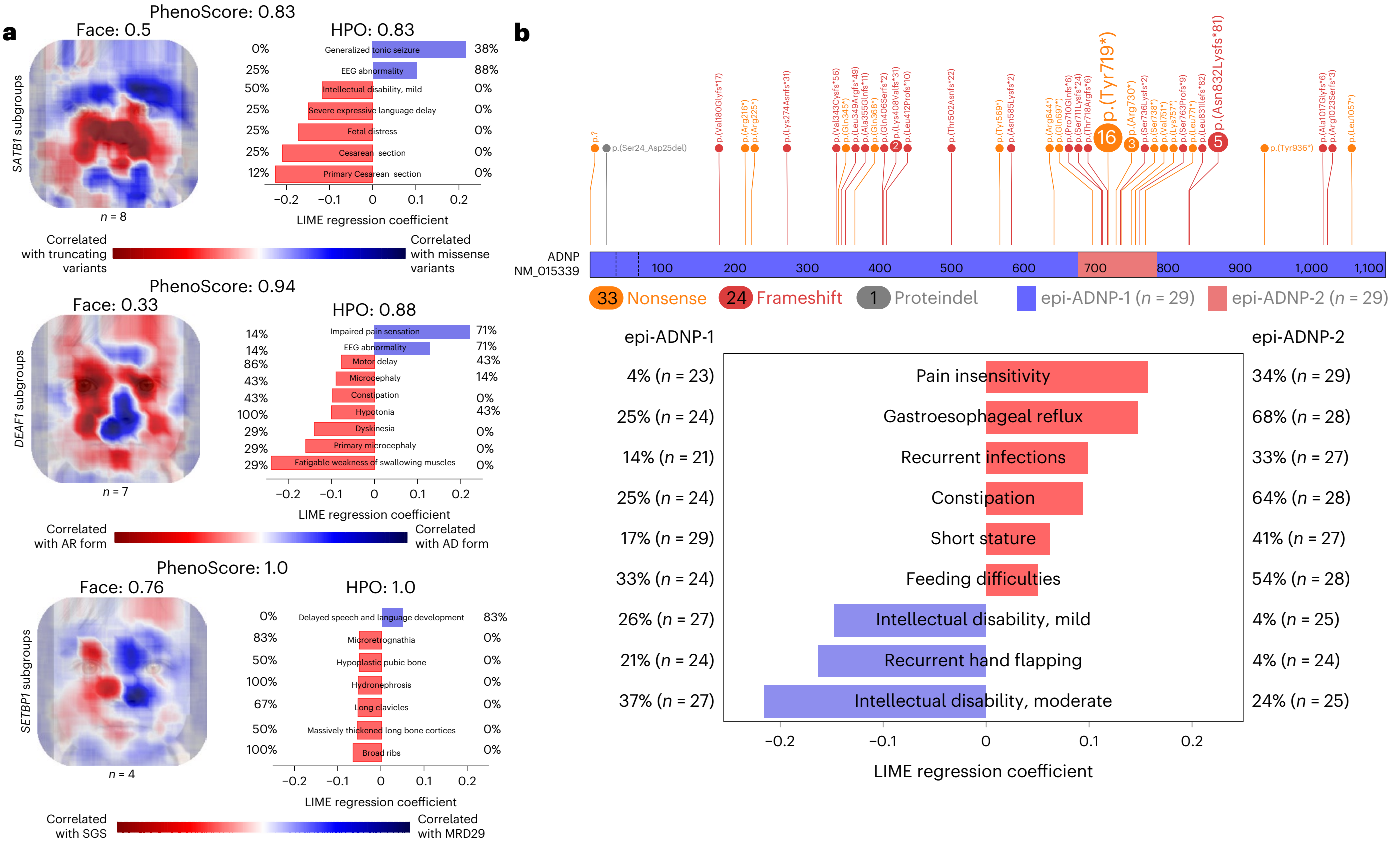


Fig. 5 | Genotype–phenotype correlations and subgroup detection. a, The facial heatmaps and most important clinical features for the three confirmatory subgroup analyses. First, the top panel shows the analysis when comparing the two phenotypic subgroups associated with pathogenic variants in *SATB1*; the middle panel shows the PhenoScore results when analyzing the subgroups for *DEAF1* and, finally, the bottom panel displays the outcome for *SETBP1*. The PhenoScores in this figure correspond to the AUC when training the model. **b**, Top: 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 versus 29). We then trained a classifier on the two groups and found a significant difference (Brier score of 0.24, AUC of 0.71, $P = 0.01$ with one-sided Mann–Whitney U test). Bottom: the most important clinical features according to our model (determined using LIME) and the corresponding prevalence in both groups.

whether PhenoScore can improve these hypothesis-driven approaches to distinguish, or discover, clinical subtypes.

For four genes in our dataset, that is, *ADNP*, *DEAF1* (encodes deformed epidermal autoregulatory factor-1 homolog), *SATB1* (encodes special AT-rich sequence binding protein 1) and *SETBP1* (encodes SET binding protein 1),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 the mode of inheritance, with one 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]. Analyzing both these subgroups shows that PhenoScore distinguishes these groups (for *SETBP1*, Brier score of 0.02 and AUC of 1.0, $P < 0.001$; *DEAF1* leads to a Brier score of 0.13 and AUC of 0.94, $P < 0.001$; Fig. 5a), 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 (p.667) and c.2340 (p.780); 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]. Before 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 score, 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 the HPO of 58 individuals (29 in each group), and on these data, PhenoScore did show evidence for two phenotypically different entities (Brier score of 0.24, AUC of 0.71, $P = 0.01$). Inspecting the generated PhenoScore explanations for clinically relevant differences (Fig. 5b), it seems that recurrent infections and gastrointestinal problems (reflux, constipation and feeding difficulties) are two to three 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. 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*), because these individuals are (much) more phenotypically alike investigated syndrome than the control population (Extended Data Table 6).

Discussion

PhenoScore provides a substantial step in the advancement of AI in clinical genetics—a 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[32,33,38,61]. The utilization of facial recognition technology to assist clinicians in diagnosing individuals has been successful too, with most, unfortunately, relying on proprietary commercial algorithms[37,39–44]. We now show the 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%), which is a substantial 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 level 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 NDDs—but individuals with Schinzel–Giedeon syndrome are phenotypically different from these controls (Fig. 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 (Extended Data Table 6) show 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 NDDs, these assays are scarce because they need to be developed on a gene-per-gene basis, and for these rare disorders, this is usually not cost-effective and is solely done for research purposes. Other methods to assess genetic variants include protein structural analysis[64], which 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 into 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[66–68]. Here the use of a selected number of HPO terms in combination with the use of Resnik scores minimizes the interobserver variability between clinicians. Although these measures should minimize any difference in predictive performance when applying PhenoScore in other institutions, further prospective clinical validation studies, preferably in a multicenter prospective design also including institutions from non-Western countries, are needed to confirm this.

PhenoScore also objectively obtained genotype–phenotype correlations by training on suspected phenotypic subgroups combined with 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 substantially different, but recognizable, phenotype. Although for these genes the associated different phenotypes were also subjectively identifiable 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*[59,60,69], 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 separate the two clinical entities. For instance, neurodevelopmental problems are more common in *ADNP*-type 1, while gastrointestinal symptoms, recurrent infections and short stature are two to three times more common in *ADNP*-type 2. These 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 insights. For *ADNP*, these clinical features have a substantial 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.

Detailed genotype–phenotype analysis could, in theory, be performed for every (genetic) syndrome, suggesting that PhenoScore may be a valuable tool to also foster molecular insights. That is, for many of the 1,600 known genes associated with an NDD phenotype, multiple types of genetic variants (for example, SNVs and CNVs) may cause the disorder. Although the molecular mechanism for CNVs often relates 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 used to find phenotypic outliers, of which the molecular mechanism leading to disease might be different. By quantifying the complete phenotypic similarity and visualizing differences between (sub)groups, PhenoScore empowers detailed genotype–phenotype studies, leading to insights on both the genetic and phenotypic levels.

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 Systematized Nomenclature of Medicine (SNOMED)[70], can readily be integrated. The PhenoScore 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.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01469-w.

References

1. Vissers, L. E. L. M. et al. A de novo paradigm for mental retardation. *Nat. Genet.* **42**, 1109–1112 (2010).
2. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
3. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
4. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
5. Richards, S. 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. *Genet. Med.* **17**, 405–424 (2015).
6. Beaumont, R. N. & Wright, C. F. Estimating diagnostic noise in panel-based genomic analysis. *Genet. Med.* **24**, 2042–2050 (2022).
7. McGuire, A. L. et al. The road ahead in genetics and genomics. *Nat. Rev. Genet.* **21**, 581–596 (2020).
8. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
9. 100,000 Genomes Project Pilot Investigators. et al.100,000 genomes pilot on rare-disease diagnosis in health care—preliminary report. N. Engl. J. Med. 385, 1868–1880 (2021).
10. Neveling, K. et al. Next-generation cytogenetics: comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. *Am. J. Hum. Genet.* **108**, 1423–1435 (2021).
11. Mantere, T. et al. Optical genome mapping enables constitutional chromosomal aberration detection. *Am. J. Hum. Genet.* **108**, 1409–1422 (2021).
12. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
13. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
14. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
15. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
16. 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).
17. Leite, A. J. D. C. 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. *PLoS ONE* **17**, e0266493 (2022).
18. Clift, K. et al. Patients' views on variants of uncertain significance across indications. *J. Community Genet.* **11**, 139–145 (2020).
19. Makhnoon, S., Garrett, L. T., Burke, W., Bowen, D. J. & Shirts, B. H. Experiences of patients seeking to participate in variant of uncertain significance reclassification research. *J. Community Genet.* **10**, 189–196 (2019).
20. Van Dijk, S. 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. *J. Clin. Oncol.* **24**, 3672–3677 (2006).
21. Murray, M. L., Cerrato, F., Bennett, R. L. & Jarvik, G. P. Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions. *Genet. Med.* **13**, 998–1005 (2011).
22. Hamburg, M. A. & Collins, F. S. The path to personalized medicine. *N. Engl. J. Med.* **363**, 301–304 (2010).
23. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).
24. Brittain, H. K., Scott, R. & Thomas, E. The rise of the genome and personalised medicine. *Clin. Med.* **17**, 545–551 (2017).
25. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
26. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
27. Killock, D. AI outperforms radiologists in mammographic screening. *Nat. Rev. Clin. Oncol.* **17**, 134 (2020).
28. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
29. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
30. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
31. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
32. Köhler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
33. Robinson, P. N. et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* **24**, 340–348 (2014).
34. Zemojtel, T. et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* **6**, 252ra123 (2014).
35. Smedley, D. & Robinson, P. N. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* **7**, 81 (2015).
36. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
37. Hsieh, T.-C. et al. PEDIA: prioritization of exome data by image analysis. *Genet. Med.* **21**, 2807–2814 (2019).
38. Robinson, P. N. et al. Interpretable clinical genomics with a likelihood ratio paradigm. *Am. J. Hum. Genet.* **107**, 403–417 (2020).
39. Ferry, Q. et al. Diagnostically relevant facial gestalt information from ordinary photos. *eLife* **3**, e02020 (2014).
40. 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).

41. Van der Donk, R. et al. Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. *Genet. Med.* **21**, 1719–1725 (2019).
42. Gurovich, Y. et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **25**, 60–64 (2019).
43. Dingemans, A. J. M. et al. Quantitative facial phenotyping for Koolen-de Vries and 22q11.2 deletion syndrome. *Eur. J. Hum. Genet.* **29**, 1418–1423 (2021).
44. Hsieh, T.-C. et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat. Genet.* **54**, 349–357 (2022).
45. Claes, P. et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat. Genet.* **50**, 414–423 (2018).
46. White, J. D. et al. Insights into the genetic architecture of the human face. *Nat. Genet.* **53**, 45–53 (2021).
47. Naqvi, S. et al. Shared heritability of human face and brain shape. *Nat. Genet.* **53**, 830–839 (2021).
48. Zhang, M. et al. Genetic variants underlying differences in facial morphology in East Asian and European populations. *Nat. Genet.* **54**, 403–411 (2022).
49. Vulto-van Silfhout, A. T. et al. Clinical significance of de novo and inherited copy-number variation. *Hum. Mutat.* **34**, 1679–1687 (2013).
50. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
51. Koolen, D. A. et al. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).
52. Zollino, M. et al. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nat. Genet.* **44**, 636–638 (2012).
53. Koolen, D. A. et al. The Koolen-de Vries syndrome: a phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant. *Eur. J. Hum. Genet.* **24**, 652–659 (2016).
54. Köhler, S. et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
55. den Hoed, J. et al. Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction. *Am. J. Hum. Genet.* **108**, 346–356 (2021).
56. Nabais Sá, M. J. et al. De novo and biallelic DEAF1 variants cause a phenotypic spectrum. *Genet. Med.* **21**, 2059–2069 (2019).
57. Hoischen, A. et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
58. Filges, I. et al. Reduced expression by SETBP1 haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome. *J. Med. Genet.* **48**, 117–122 (2011).
59. Bend, E. G. et al. Gene domain-specific DNA methylation episignatures highlight distinct molecular entities of ADNP syndrome. *Clin. Epigenetics* **11**, 64 (2019).
60. Breen, M. S. et al. Episignatures stratifying Helsmoortel-Van Der Aa syndrome Show modest correlation with phenotype. *Am. J. Hum. Genet.* **107**, 555–563 (2020).
61. Jagadeesh, K. A. et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet. Med.* **21**, 464–470 (2019).
62. Lyra Jr, P. C. M. et al. Integration of functional assay data results provides strong evidence for classification of hundreds of BRCA1 variants of uncertain significance. *Genet. Med.* **23**, 306–315 (2021).
63. Frederiksen, J. H., Jensen, S. B., Tümer, Z. & Hansen, T. V. O. Classification of MSH6 variants of uncertain significance using functional assays. *Int. J. Mol. Sci.* **22**, 8627 (2021).
64. Caswell, R. C., Gunning, A. C., Owens, M. M., Ellard, S. & Wright, C. F. Assessing the clinical utility of protein structural analysis in genomic variant classification: experiences from a diagnostic laboratory. *Genome Med.* **14**, 77 (2022).
65. Dingemans, A. J. M. et al. Human disease genes website series: an international, open and dynamic library for up-to-date clinical information. *Am. J. Med. Genet. A* **185**, 1039–1046 (2021).
66. McKusick, V. A. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
67. Firth, H. V. et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
68. Adam, M. P. et al. *GeneReviews* (Univ. Washington, 2010).
69. Helsmoortel, C. et al. A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nat. Genet.* **46**, 380–384 (2014).
70. Côté, R. A. & Robboy, S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA* **243**, 756–762 (1980).
71. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4217–4228 (2021).

[1]Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands. [2]Department of Artificial Intelligence, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands. [3]Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. [4]Department of Human Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands. [5]Department of Pediatrics, University of Montreal, Montreal, Quebec, Canada. [6]Faculty of Medicine and Health, UNSW Sydney, Sydney, New South Wales, Australia. [7]Sydney Children's Hospitals Network, Sydney, New South Wales, 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, NY, USA. [10]Biology PhD Program, The Graduate Center, The City University of New York, New York City, NY, USA. [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 Zürich, Zürich, Switzerland. [13]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [14]Faculty of Medicine, University of Southampton, Southampton, UK. [15]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. [16]Department of Medical Genetics, University of Antwerp, Antwerp, Belgium. [17]These authors contributed equally: Lisenka E. L. M. Vissers, Bert B. A. de Vries. ✉e-mail: lisenka.vissers@radboudumc.nl; bert.devries@radboudumc.nl

 https://doi.org/10.1038/s41588-023-01469-w

Methods

Inclusion of individuals

The literature was searched for clinical studies that included facial photographs of 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 nonfamilial 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 Extended Data Table 1 includes 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.

Ethics declaration

In this study, data from the Biobank Intellectual Disability, which is part of the Radboud Biobank initiative (for more information, see[72] 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. This research complies with all relevant ethical regulations, and the use of this dataset was approved by the ethical committee of the Radboud university medical center (2020-6151 and 2020-7142). Written informed consent was obtained for the publication of the facial images included in this study.

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 NDD seen at our outpatient clinic at the Radboud University Medical Center were selected from our internal control database with over 1,200 individuals with both facial image and quantitative phenotypic data available (for a complete overview of the workflow of this study, see Fig. 1a). 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, fivefold cross-validation was used, otherwise, leave-one-out cross-validation was chosen. The hyperparameters of the model 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 substantially 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 substantially 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 (Supplementary Data).

Extraction of facial features

The facial features were extracted using VGGFace2 (refs. 73,74), as it was previously shown to be the best-performing open-source solution for this task[75]. VGGFace2 is a state-of-the-art facial recognition method that uses a deep neural network. To avoid overfitting, we did not retrain VGGFace2 but used its pretrained weights instead on the database of 3.1 million images. The facial images of the individuals in our study 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.

Phenotypic similarity

To create a homogeneous dataset, the phenotype of every individual in this study was manually converted into HPO terms[16]. A selection of HPO terms and all their child nodes were removed to eliminate any subjectivity in assessing an individual. These were as follows: 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, after we investigated what the consequences of including all HPO terms were and concluding that the inclusion of facial data to HPO data improves the performance of models significantly in each scenario ($P < 0.001$ using a two-sided Wilcoxon signed-rank test). To further reduce possible interobserver variability, the phenotypic similarity between individuals was calculated using the Resnik score[76], because it takes the semantic similarity between symptoms into account. The Resnik score uses 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), substantially further down the HPO tree, has an IC of 8.97. This corresponds to our intuition—rare features provide more information than common features, because 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. Because 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[77]. 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 $n \times 2$ 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 (ref. 78).

Construction of machine-learning model

Finally, the data were used to train a binary classifier. We selected an 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

 https://doi.org/10.1038/s41588-023-01469-w

values investigated being 1×10^{-5}, 1×10^{-3}, 1, 1×10^{3} and 1×10^{5}. 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 much 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 used 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 (Fig. 4).

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

Explainability of predictions

To see which features contained important information for our model, we generated LIME[79,80]. 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 are 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 the 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 substantial 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 were generated for the individuals with the investigated genetic syndrome or phenotypic subgroup 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.

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, see Supplementary Data.

Benchmarking PhenoScore

To determine whether our approach is an improvement over existing methods, we used the Phenomizer algorithm[32,81] and LIRICAL[38] (considered as state-of-the-art[82]) to generate predictions for all individuals with a genetic syndrome in our dataset (except for the genetic syndrome associated with *ACTL6A*, as the absence of an OMIM number prohibits Phenomizer and LIRICAL to generate predictions). Because 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, because 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) and abnormal eye morphology (HP:0012372)) and 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 statistically 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 (Extended Data Table 3).

Statistics and reproducibility

No statistical method was used to predetermine sample size—because data were collected from the literature, the number of cases available with both phenotypic data and facial photographs was the limiting factor. Data were only excluded if individuals were related to each other, to avoid the introduction of bias, because family members are facially similar. Therefore, including family members could unjustly overinflate the results of our analysis. The investigators were not blinded to allocation during experiments and outcome assessment, although cross-validation was used during all analyses, which is equivalent to blinding for algorithms and models.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The used dataset in this study is not publicly available due to both IRB and General Data Protection Regulation (EU GDPR) restrictions because the data might be (partially) traceable. However, access to the data may be requested from the data availability committee by contacting the corresponding authors via e-mail with a research proposal, who will respond within 14 d.

Code availability

The code of PhenoScore version 1.0.0 created during this study is freely available at https://github.com/ldingemans/PhenoScore ref. 83, to enable anyone to apply PhenoScore to their own dataset. Included in PhenoScore are the following two examples: the data for the *SATB1* subgroups (positive example) and random data (negative example).

References

72. Manders, P., Lutomski, J. E., Smit, C., Swinkels, D. W. & Zielhuis, G. A. 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. *Open J. Bioresour.* **5**, 2 (2018).
73. Parkhi, O. M., Vedaldi, A. & Zisserman, A. Deep face recognition. *Proceedings of the British Machine Vision Conference* (eds Xianghua X. et al.) 41.1–41.12 (BMVA Press, 2015).
74. Cao, Q. Shen, L., Xie, W. Parkhi, O. M. & Zisserman, A. VGGFace2: a dataset for recognising faces across pose and age. *Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (F&G)* pp. 67–74 (IEEE, 2018).
75. Dingemans, A. J. M., de Vries, B. B. A., Vissers, L. E. L., van Gerven, M. A. J. & Hinne, M. Comparing facial feature extraction methods in the diagnosis of rare genetic syndromes. Preprint at *medRxiv* https://doi.org/10.1101/2022.08.26.22279217 (2022).
76. Resnik, P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999).
77. Pesquita, C. et al. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9**, S4 (2008).
78. Arvai, K., Gainullin, V. & Borroto, C. GeneDx/phenopy. *Zenodo* https://doi.org/10.5281/zenodo.4587231 (2019).
79. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?' Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and DATA MINIng* 1135–1144 (Association for Computing Machinery, 2016).
80. Ras, G., Xie, N., van Gerven, M. & Doran, D. Explainable deep learning: a field guide for the uninitiated. *J. Artif. Intell. Res.* **73**, 329–396 (2022).
81. Köhler, S. et al. The human phenotype ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876 (2017).
82. Yuan, X. et al. Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief. Bioinform.* **23**, bbac019 (2022).
83. Dingemans, L. ldingemans/PhenoScore: v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.7892317 (2023).

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 from 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 2021/GNT2008166). Furthermore, we are grateful to the Dutch Organization 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 (to 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 Unions Horizon 2020 research and innovation program under grant agreement 779257. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.J.M.D., M.H., L.E.L.M.V., B.B.A.d.V. and M.A.J.v.G. designed the study. 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. and B.B.A.d.V. collected and curated the data. A.J.M.D. and M.H. performed the formal analyses. L.E.L.M.V. and B.B.A.d.V. acquired the funding. A.J.M.D. and M.H. completed the modeling and investigations. A.J.M.D. developed the software. A.J.M.D., M.H., L.E.L.M.V., B.B.A.d.V. and M.A.J.v.G. wrote the original draft. All authors reviewed and edited the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

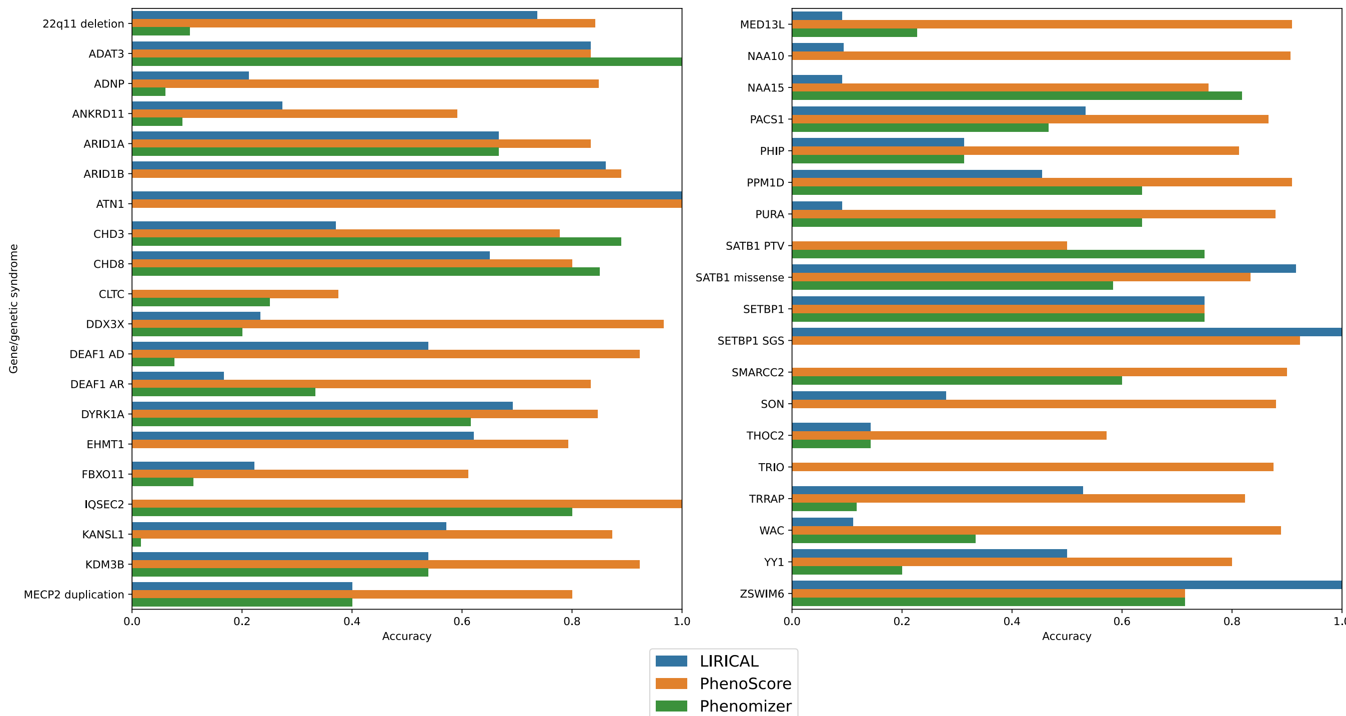
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01469-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01469-w.

Correspondence and requests for materials should be addressed to Lisenka E. L. M. Vissers or Bert B. A. de Vries.

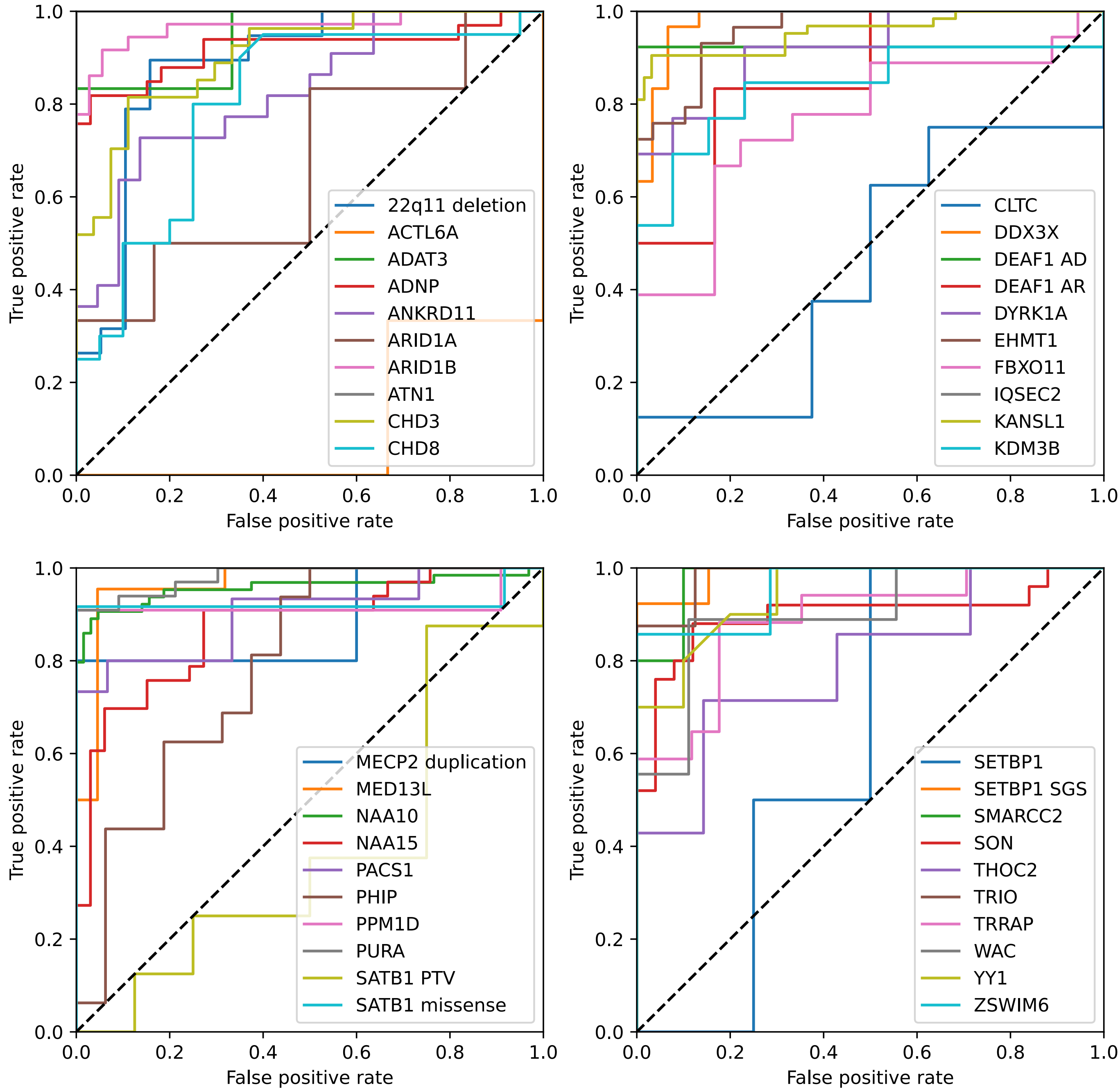
Peer review information *Nature Genetics* thanks Xinran Dong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-023-01469-w



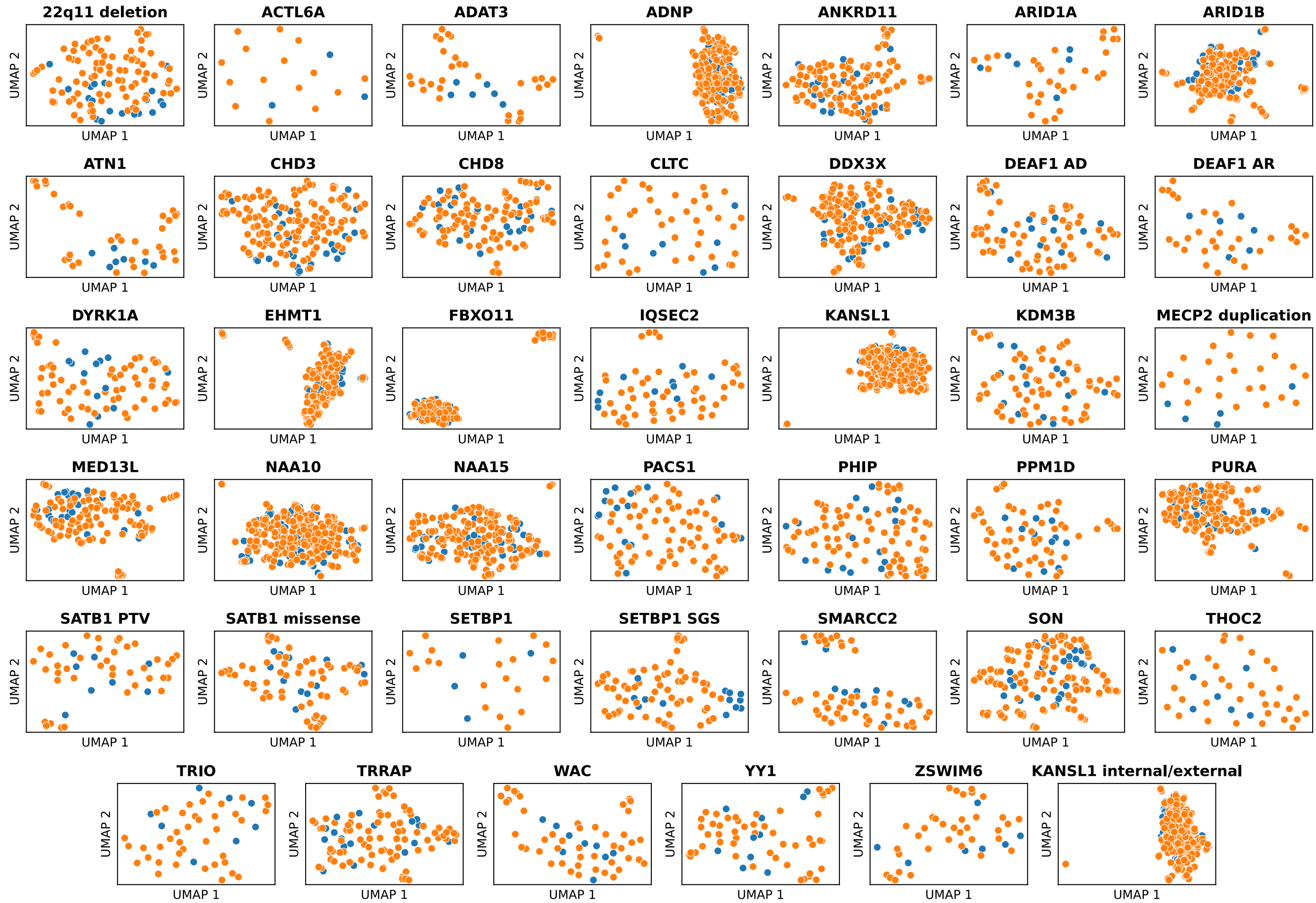
Extended Data Fig. 1 | Benchmarking PhenoScore. The predictive accuracies of LIRICAL, Phenomizer and PhenoScore [118-120] for every included genetic syndrome are displayed here, except for *ACTL6A*, since the associated phenotype has no OMIM number and therefore Phenomizer and LIRICAL do not include it in its predictions. For PhenoScore and LIRICAL, to calculate the accuracy, a cut-off value of 0.5 for the predictions was used, while for Phenomizer in this case, 0.05 was chosen. For almost every investigated syndrome, PhenoScore outperforms Phenomizer and LIRICAL.

https://doi.org/10.1038/s41588-023-01469-w



Extended Data Fig. 2 | AUC curves of PhenoScore per genetic syndrome. The receiver operating characteristic curve of all 40 genetic syndromes included in this study.

 https://doi.org/10.1038/s41588-023-01469-w



Extended Data Fig. 3 | UMAP plots of facial feature vectors. The Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP[3]) plot for the VGGFace2 vectors of all included genetic syndromes, and for the extra systematic confounder analysis for which the individuals with Koolen-de Vries syndrome seen at other centers were compared to individuals seen at our outpatient clinic. For all plots (except the *KANSL1* internal/external plot), the feature vectors of all sampled controls during five iterations and the feature vectors of the included patients were provided as input to UMAP. The classes are not separable in this projected space, which provides evidence that the classification is not based on a systematic confounder.

https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 1 | List of publications for data collection

Genetic syndrome	PMID of used publications
22q11 deletion syndrome	15831592,17041934,18636631,1956057,21200182,25317860,3816857,12548732 [4–11]
ACTL6A	28649782 [12]
ADAT3	23620220,26842963 [13, 14]
ADNP	24531329,25217958,27031564,29724491 [15–17, 69]
ANKRD11	19920853,21527850,21654729,22307766,23494856,26269249 [18–23]
ARID1A	23815551,23929686,32888375,35579625 [24–27]
ARID1B	19034313,22405089,22426309,22585544,23906836,24569609,26395437,26754677,27112773,28323383,30055038,30349098,31981384,32339967 [28–41]
ATN1	30827498 [42]
CHD3	30397230,32483341 [43, 44]
CHD8	23160955,24998929,31001818,31721432,36182950 [45–49]
CLTC	31776469 [50]
DDX3X	26235985 [51]
DEAF1	30451703,30923367,31688097,31929336 [52–54, 56]
DYRK1A	18405873,21294719,23099646,23160955,25707398,Not published [45, 55–58]
EHMT1	22670141,Not published [59]
FBXO11	27479843,28343630 [60, 61]
IQSEC2	23674175,30666632 [62, 63]
KANSL1	16906164,17601928,18628315,21094706,22544363,26306646,Not published [51, 53, 64–67]
KDM3B	30929739 [68]
MECP2 duplication	18854860,18985075 [69, 70]
MED13L	23403903,24781760,25712080,28645799,29511999,29959045 [71–76]
NAA10	27094817,31093388,31127942,32698785,34075687,34200686,35039925 [77–84]
NAA15	26785492,29656860,28191889,31127942 [79, 85–87]
PACS1	26842493 [88]
PHIP	29209020 [89]
PPM1D	28343630 [61]
PURA	27148565,29097605,29150892 [90–92]
SATB1	33513338 [55]
SETBP1	18461363,20436468,28346496,21037274,33867525 [57, 58, 93–95]
SMARCC2	30580808 [96]
SON	24896178,27256762,27545680 [4, 97–99]
THOC2	29851191,32116545 [100, 101]
TRIO	26721934,27418539 [102, 103]
TRRAP	308274965 [104]
WAC	23033978,26264232,26757981 [2, 105, 106]
YY1	21076407,28575647 [1, 107]
ZSWIM6	29198722 [108]

A list is shown of the used publications per syndrome to create the dataset by extracting the phenotypic data and photographs of individuals in these papers. For several syndromes, not (yet) published individuals were added to the dataset, as indicated by Not published.

 https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 2 | Performance of PhenoScore with other syndromes as control

Genetic syndrome	Facial data only	HPO data only	PhenoScore (Brier)
22q11 deletion syndrome	0.221	0.106	0.093
ACTL6A	0.250	0.606	0.549
ADAT3	0.244	0.083	0.074
ADNP	0.233	0.178	0.170
ANKRD11	0.242	0.158	0.143
ARID1A	0.337	0.224	0.235
ARID1B	0.175	0.117	0.105
ATN1	0.233	0.116	0.107
CHD3	0.232	0.134	0.134
CHD8	0.235	0.169	0.132
CLTC	0.259	0.316	0.280
DDX3X	0.213	0.039	0.037
DEAF1 (AR)	0.272	0.164	0.164
DEAF1 (AD)	0.266	0.113	0.107
DYRK1A	0.226	0.165	0.143
EHMT1	0.173	0.119	0.092
FBXO11	0.280	0.228	0.211
IQSEC2	0.257	0.115	0.104
KANSL1	0.112	0.113	0.097
KDM3B	0.252	0.218	0.202
MECP2 duplication	0.247	0.395	0.387
MED13L	0.238	0.187	0.165
NAA10	0.183	0.086	0.086
NAA15	0.257	0.153	0.143
PACS1	0.244	0.150	0.154
PHIP	0.230	0.161	0.137
PPM1D	0.259	0.148	0.132
PURA	0.247	0.115	0.113
SATB1 (truncating)	0.311	0.197	0.191
SATB1 (missense)	0.260	0.206	0.187
SETBP1 (SGS)	0.139	0.053	0.051
SETBP1	0.255	0.216	0.265
SMARCC2	0.265	0.141	0.137
SON	0.249	0.135	0.132
THOC2	0.266	0.183	0.171
TRIO	0.259	0.133	0.133
TRRAP	0.264	0.185	0.165
WAC	0.244	0.198	0.171
YY1	0.253	0.287	0.267
ZSWIM6	0.271	0.194	0.164

The Brier scores of the support vector machine (SVM) classifier are displayed here, now with the other individuals included in this study as the control dataset, instead of the controls from the Radboud university medical center. A lower Brier score indicates a better result. The results are slightly worse than on the RUMC control dataset, as expected, since not for not every individual, a control is available because the RUMC control dataset is significantly larger than the number of individuals included in this study. AD=autosomal dominant, AR=autosomal recessive, SGS=Schinzel-Giedion-syndrome.

 https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 3 | Benchmarking PhenoScore

	Phenomizer threshold 0.001	Phenomizer threshold 0.01	Phenomizer threshold 0.05	Phenomizer threshold 0.10	LIRICAL threshold 0.3	LIRICAL threshold 0.5	LIRICAL threshold 0.7	LIRICAL threshold 0.9
PhenoScore (HPO-only) threshold 0.3	6%/91%***	16%/91%***	29%/91%***	35%/91%***	41%/91%***	39%/91%***	37%/91%***	33%/91%***
PhenoScore (HPO-only) threshold 0.5	6%/82%***	16%/82%***	29%/82%***	35%/82%***	41%/82%***	39%/82%***	37%/82%***	33%/82%***
PhenoScore (HPO-only) threshold 0.7	6%/70%***	16%/70%***	29%/70%***	35%/70%***	41%/70%***	39%/70%***	37%/70%***	33%/70%***
PhenoScore (HPO-only) threshold 0.9	6%/46%***	16%/46%***	29%/46%***	35%/46%***	41%/46%	39%/46%	37%/46%	33%/46%***
PhenoScore threshold 0.3	6%/92%***	16%/92%***	29%/92%***	35%/92%***	41%/92%***	39%/92%***	37%/92%***	33%/92%***
PhenoScore threshold 0.5	6%/84%***	16%/84%***	29%/84%***	35%/84%***	41%/84%***	39%/84%***	37%/84%***	33%/84%***
PhenoScore threshold 0.7	6%/72%***	16%/72%***	29%/72%***	35%/72%***	41%/72%***	39%/72%***	37%/72%***	33%/72%***
PhenoScore threshold 0.9	6%/46%***	16%/46%***	29%/46%***	35%/46%***	41%/46%	39%/46%	37%/46%	33%/46%***

The predictive accuracy of LIRICAL, Phenomizer and PhenoScore [118-120] for all individuals (except *ACLT6A*) included in this study, with different cut-off values for each algorithm. The Phenomizer or LIRICAL accuracy is shown first in every cell, and then the PhenoScore accuracy with those specific thresholds. Even with the least strict threshold for the p-values of Phenomizer (0.1) or LIRICAL (0.3) and the most stringent for the output of PhenoScore (0.9), PhenoScore still outperforms both. ***significant at the 0.001 level using a two-sided chi-squared test.

https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 4 | Subgroup analyses

Age group	Brier score	Accuracy	n in group
<2 years	0.12	0.84	61
2 — 5 years	0.10	0.85	222
6 — 11 years	0.12	0.84	213
12 — 17 years	0.09	0.86	115
18 years and over	0.13	0.77	100
Population of origin	**Brier score**	**Accuracy**	**n in group**
Caucasian/Western	0.11	0.84	643
Ottoman/Middle Eastern	0.07	0.91	45
Asian	0.16	0.88	17
African	0.08	1.00	6

The performance of PhenoScore when ignoring the specific genetic syndrome diagnoses, but focusing on different subgroups of the study population: based on age and population of origin. Here, we calculated both the Brier score (for which lower is better) and accuracy (with 0.5 as cut-off, higher is better) using the predictions of all included individuals in this study when not in the training set. The predictions were calculated for the subgroups, demonstrating that the predictive performance is relatively stable.

https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 5 | Classifying variants of uncertain significance

Gene	Variant (genomic)	Variant (RNA)	Variant (protein)	Classification at first diagnostic evaluation	Current status/comments	AUC this gene	PhenoScore (prediction probability)	PhenoScore Classification
ANKRD11	Chr16(GRCh37):g.89349673C>T	NM_001256182.1:c.3277G>A	p.(Gly1093Arg)	class 3	VUS, clinician of opinion that phenotype fits	0.78	0.21	No phenotypic match
ANKRD11	Chr16(GRCh37):g.89349812_89349813delinsCA	NM_013275.5:c.3137_3138delinsTG	p.(Cys1046Leu)	class 3	VUS	0.78	0.52	Inconclusive
ANKRD11	Chr16(GRCh37):g.89347788G>A	NM_013275.5:c.5162C>T	p.(Thr1721Met)	class 3	VUS, clinician of opinion that phenotype fits	0.78	0.82	**Phenotypic match**
ARID1B	Chr6(GRCh37):g.157406024C>G	NM_020732.3:c.2266C>G	p.(Pro756Ala)	class 3	VUS, clinician of opinion that phenotype fits	0.95	0.03	No phenotypic match
ARID1B	Chr6(GRCh37):g.157488314C>T	NM_001346813.1:c.2981C>T	p.(Ser994Leu)	class 3	Other genetic diagnosis confirmed and only HPO data available	0.95	0.03	No phenotypic match
CHD8	Chr14(GRCh37):g.21870111C>G	NM_001170629.1:c.4062+5G>C	p.(?)	class 3	Pathogenic after RNA analysis	0.80	0.93	**Phenotypic match**
DDX3X	ChrX(GRCh37):g.41203603C>T	NM_001356.4:c.976C>T	p.(Arg326Cys)	class 3	VUS, clinician of opinion that phenotype fits	0.99	0.02	No phenotypic match
DDX3X	ChrX(GRCh37):g.41205825T>C	NM_001356.4:c.1565T>C	p.(Ile522Thr)	class 3	VUS	0.99	0.02	No phenotypic match
DEAF1	Chr11(GRCh37):g.654020A>G	NM_021008.3:c.1535T>C	p.(Met512Thr)	class 3	VUS	0.77	0.41	Inconclusive
EHMT1	Chr9(GRCh37):g.140674167T>C	NM_024757.4:c.2273T>C	p.(Leu758Pro)	class 3	Pathogenic after EpiSign analysis	0.93	0.04	No phenotypic match
IQSEC2	ChrX(GRCh37):g.53277979G>A	NM_001111125.2:c.2383C>T	p.(Arg795Trp)	class 3	VUS	0.97	0.35	Inconclusive
IQSEC2	ChrX(GRCh37):g.53277329G>T	NM_001111125.2:c.2549C>A	p.(Ala850Asp)	class 3	VUS, clinician of opinion that phenotype fits	0.97	0.52	Inconclusive
KDM3B	Chr5(GRCh37):g.137727707A>T	NM_016604.3:c.2386A>T	p.(Arg796Trp)	class 3	VUS	0.81	0.43	Inconclusive
NAA10	ChrX(GRCh37):g.153196242G>A	NM_003491.3:c.445C>T	p.(Arg149Trp)	class 3	VUS, clinician of opinion that phenotype fits	0.95	0.04	No phenotypic match
PHIP	Chr6(GRCh37):g.79707136T>C	NM_017934.6:c.2196A>G	p.(Val732=)	class 3	VUS	0.72	0.88	**Phenotypic match**
PPM1D	Chr17(GRCh37):g.58711260C>T	NM_003620.3:c.748C>T	p.(Arg250*)	class 3	VUS	0.94	0.61	Inconclusive
PURA	Chr5(GRCh37):g.139494093G>T	NM_005859.4:c.327G>T	p.(Glu109Asp)	class 3	VUS	0.96	0.05	No phenotypic match
SMARCC2	Chr12(GRCh37):g.56577703G>A	NM_003075.3:c.574C>T	p.(Arg192*)	class 3	Benign after EpiSign analysis	0.96	0.48	Inconclusive
SMARCC2	Chr12(GRCh37):g.56579937A>T	NM_003075.3:c.317+2T>A	p.(?)	class 3	Benign after EpiSign analysis	0.96	0.37	Inconclusive
TRIO	Chr5(GRCh37):g.14374354G>A	NM_007118.3:c.3233G>A	p.(Arg1078Gln)	class 3	VUS, clinician of opinion that phenotype fits	0.84	0.27	No phenotypic match
TRRAP	Chr7(GRCh37):g.98592220A>G	NM_001244580.1:c.10016A>G	p.(Gln3339Arg)	class 3	VUS	0.84	0.19	No phenotypic match
WAC	Chr10(GRCh37):g.28906631A>G	NM_016628.4:c.1792A>G	p.(Met598Val)	class 3	VUS	0.84	0.34	Inconclusive

The 22 individuals with a VUS in one of the 40 included syndromes are displayed here, including the genetic information and the PhenoScore — both the score between 0 and 1 in which higher score indicates increased phenotypic similarity with the syndrome of interest and a PhenoScore classification using cut-offs of 0.3 and 0.7. Next to that, the area under the curve (AUC) of that gene is displayed, in which a higher score indicates that PhenoScore is better to distinguish that genetic syndrome in general.

https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 6 | PhenoScore with phenotypically similar individuals

Gene/genetic syndrome	Input as VUS	Phenotypic Match	No Match	VUS	Mean score of all
DEAF1 AD	**DEAF1 AR**	33%	33%	33%	0.52
DEAF1 AR	**DEAF1 AD**	77%	0%	23%	0.77
SATB1 PTV	**SATB1 missense**	25%	0%	75%	0.52
SATB1 missense	**SATB1 PTV**	12%	25%	62%	0.49
SETBP1	**SETBP1 SGS**	0%	92%	8%	0.23
SETBP1 SGS	**SETBP1**	0%	100%	0%	0.13
ADNP methylation 1	**ADNP methylation group 2**	97%	3%	0%	0.89
ADNP methylation 2	**ADNP methylation group 1**	41%	34%	24%	0.51

For these analyses, a model was trained on a specific subgroup for a gene and that model was then used to classify individuals diagnosed with the other subgroup of that gene. For instance, a model was trained for individuals with the syndrome associated with the autosomal dominant form of *DEAF1*. Individuals with the recessive genetic syndrome associated with *DEAF1* were then classified using this model. These analyses show that clinicians and researchers should be careful when interpreting the results of PhenoScore when investigating phenotypically similar syndromes, as the number of false-positives could be elevated in that case.

https://doi.org/10.1038/s41588-023-01469-w

Extended Data Table 7 | Systematic confounder analysis using Koolen-de Vries syndrome

Individual	PhenoScore (prediction probability)	Score HPO only	Score facial data only	PhenoScore classification
KANSL1 Nijmegen 1	0.13	0.06	0.86	No phenotypic match
KANSL1 Nijmegen 2	0.49	0.39	0.61	Inconclusive
KANSL1 Nijmegen 3	0.96	0.97	0.64	**Phenotypic match**
KANSL1 Nijmegen 4	0.97	0.98	0.72	**Phenotypic match**
KANSL1 Nijmegen 5	0.61	0.66	0.68	Inconclusive
KANSL1 Nijmegen 6	0.97	0.98	0.64	**Phenotypic match**
KANSL1 Nijmegen 7	0.98	0.97	0.93	**Phenotypic match**
KANSL1 Nijmegen 8	0.98	0.99	0.48	**Phenotypic match**
KANSL1 Nijmegen 9	0.91	0.94	0.35	**Phenotypic match**
KANSL1 Nijmegen 10	0.94	0.95	0.76	**Phenotypic match**
KANSL1 Nijmegen 11	0.99	0.98	0.93	**Phenotypic match**
KANSL1 Nijmegen 12	0.97	0.96	0.68	**Phenotypic match**
KANSL1 Nijmegen 13	0.95	0.97	0.39	**Phenotypic match**
KANSL1 Nijmegen 14	0.88	0.87	0.60	**Phenotypic match**
KANSL1 Nijmegen 15	0.97	0.97	0.82	**Phenotypic match**
KANSL1 Nijmegen 16	0.51	0.21	0.94	Inconclusive
KANSL1 Nijmegen 17	0.98	0.98	0.36	**Phenotypic match**
KANSL1 Nijmegen 18	0.54	0.67	0.52	Inconclusive

The classification of 18 individuals that were seen at our outpatient clinic in the Radboud university medical center, the same clinic as all control individuals. For this analysis, these 18 individuals were left out of the training data, so that only individuals with Koolen-de Vries syndrome seen outside our outpatient clinic were included when training PhenoScore. The scores displayed here are then the predictions when using this model to generate predictions for the individuals seen at our institution. The median PhenoScore (here a posterior probability) is 0.95, and 13 out of 18 are correctly identified as having KdVS, while only one is incorrectly labeled as negative (mainly because of the phenotype that is not matching: facial predictions are high in this case). Furthermore, when calculating the Brier score using these predictions, it is 0.0917—strikingly close to the Brier score of the regular model. This is all indication of the absence of a systematic confounder related to the origin of the data.